Preemptible Queues with Advance Reservations: Strategic Behavior and Revenue Management

Jonathan Chamberlain*, Eran Simhon, David Starobinski

Division of Systems Engineering, Boston University, 15 St Mary's Street, Boston MA, 02215, USA

Abstract

Consider an $M/G/1$ queuing system that supports advance reservations. In this system, strategic customers must decide whether to reserve a server in advance (thereby gaining higher priority) or forgo reservations. Reserving a server in advance bears a cost. The provider can further impact the customers' reservation decisions via implementation of one of several priority-based preemption policies: (i) one in which any customer is subject to service preemption by a higher priority customer (PR); (ii) one in which service preemption does not occur (NP); and (iii) a hybrid policy in which only customers without a priority reservation are subject to service preemption (HPR). In this work, we characterize the strategic behavior of customers, equilibrium outcomes, and provider's revenue maximization under each of these policies. In all the cases, we prove that (i) the only possible type of Nash equilibria is a threshold one based on the customers' priorities; and (ii) the system load impacts both the structure and number of Nash equilibria. We also prove that HPR is the only policy in which (i) an equilibrium where all customers make reservations may exist; and (ii) the second moment of service impacts the equilibria. Finally, we prove that for any system load and any service distribution, the HPR policy yields the highest maximum revenue, followed in turn by the PR policy and the NP policy. We further show that the relative difference in the performance of the HPR and PR policies is greatest at low system load and under low service variance.

Keywords: Game theory, Queuing, Revenue management

*Corresponding author
Email address: jdchambo@bu.edu (Jonathan Chamberlain)

1. Introduction

Many services, such as health care, transportation, and cloud computing combine both a first-come-first-served policy and advance reservations. Advance reservations benefit a service provider since knowledge about future demand can improve resource management and quality-of-service (e.g., (Virtamo & Aalto, 1991; Charbonneau & Vokkarane, 2011; Du & Larsen, 2017; Simhon & Starobinski, 2018; Izady, 2019)). Customers are also motivated to reserve in advance, since it decreases their expected waiting time. However, typically, reservations bear an additional cost for customers. This cost can be a reservation fee, the time or resources required for making the reservation, the cost of financing advance payment, or the cost of cancellation if needed.

Customers making reservations are typically guaranteed usage of resources while in service. On the other hand, customers that forgo reservations avoid bearing the reservation cost, but may be subject to service preemption if the resources are needed by a customer with a reservation. The ability to preempt customers is particularly relevant in computing systems. For instance, while Microsoft Azure allows customers to purchase reserved instances (Microsoft, 2019), there is also an option to utilize low priority instances (Microsoft, 2018). These low priority Azure instances utilize spare capacity but are subject to preemption if a reserved Azure instance requires the resources.

A customer's choice in such systems is not made in isolation, as the decision is not only based on the customer's cost to make the reservation but also by the choice made by other customers. The customer's wait time is also influenced by the provider's decision of whether to preempt customers or not based on their priorities. The above Azure example is one possible implementation, based on a policy decision as opposed to technical limitations. In this paper, we study scenarios where a provider implements one of the following policies:

- Non-preemptive (NP): no customer is preempted once in service, regardless of its assigned priority.
- Preemptive resume (PR): all customers are subject to preemption while in service by an arriving customer with higher priority. When re-entering service, preempted customers resume from the point of interruption.
- Hybrid preemptive/non-preemptive (HPR): customers making a priority reservation are not preempted once in service. Customers who do not make priority reservations are subject to PR behavior, and are preempted if a customer with priority reservation arrives.

Under the PR policy, if the service time of each customer is deterministic or known in advance, then at the time of the reservation, the provider is able to provide precise start and end times of service. Under the NP and HPR policies, if service times are deterministic, an estimate on the start and end of the service time can be provided based upon knowledge of the customers already reserving and the system load.

Our objective is to compare the effects of the NP, PR, and HPR policies on the customers' decisions and the resulting impact on the revenue that the provider collects from reservation payments. Rather than modeling a specific provider, we introduce a model that is analytically tractable and captures key insights of strategic customer behavior in the face of priority reservations and preemption. As a result, rather than purchasing access to a particular instance or service period slot, here customers have the opportunity to purchase priority to secure service as close to their desired start time as possible. Since the decisions of the providers and customers affect each others' payoffs from service, game theory is the solution of choice for studying such systems.

In particular, we consider a game based around advance reservations. We assume that the time axis is divided into two distinct time-periods: a *reservation period* in which customers desiring service make reservation requests and a *service period* in which the customers' jobs are processed. Customers making reservations early get higher priority over customers making reservations later. Similar assumptions are commonly made in the literature of advance reservations (Virtamo, 1992), (Yessad et al., 2007), (Syed et al., 2008). A key difference between this advance reservation model and standard priority queue models is how the priority of customers is determined. In the advance reservation model, each customer realizes during the reservation period that it desires service at some future time point. Upon such a realization, the customer is offered a potential priority for a cost. This potential priority is a *relative priority* based on how many customers have requested reservations thus far. The customer then has to decide between bearing the reservation cost to exercise this potential relative priority, and forgoing the reservation cost but defaulting to the lowest priority. An important aspect of this model is that customers cannot improve on their offered priority by declining the initial offer and attempting to request a reservation later. We assume that the arrival order of jobs during the service period follows a Poisson process, which is independent from the arrival order of requests during the reservation period. Our analysis of the waiting time during the service period assumes that the system has reached steady-state.

Given this model, the objective is to determine: (i) whether an equilibrium state exists, and if so whether it unique; (ii) the impact of the reservation cost C and system parameters on the equilibrium outcomes; (iii) assuming that the cost C is a fee set by the provider, the revenue maximizing cost for each policy; (iv) the policy under which the provider optimizes revenue from reservations. The primary focus of our analysis is in the deterministic service case, wherein the service time is fixed and identical for all customers. Nevertheless, we show that our results are extensible to general service distributions. Our main contributions under this analysis are as follows:

- The only type of equilibrium that is possible, regardless of the provider's choice of preemption policy, is a threshold equilibrium based on customer priorities. All customers with potential priority greater than the threshold will choose to make a reservation, and all others forgo reservation.
- Equilibrium states where all customers opt to make a reservation are only

possible under a HPR policy.

- If the provider implements a NP or a PR policy, the possible equilibrium states depend only on the system load (i.e., it only depends on the first moments of the inter-arrival distribution and the service distribution).
- Conversely, in the HPR policy, the possible equilibrium states also depend on the value of the second moment of the service distribution.
- A rational provider will opt to implement the HPR policy, as the maximum possible revenue obtainable under this policy will be greater than that under the NP or PR policies, *for any system load and any service distribution*. Likewise, among the three policies, the HPR policy always maximizes the fraction of customers making reservations.
- As the system load tends to 0, the maximum revenue under PR is infinitely better than under NP, while the maximum revenue under HPR is at most 8 times greater than under PR. This is shown theoretically for deterministic service and numerically for other service distributions.
- As the system load tends to 1, the relative difference between the maximum revenues of each of these policies vanishes.
- The relative difference between the maximum revenues of the HPR and PR policies decreases with the service variance.

Note that the provider controlling the reservations need not be the same entity that controls the server. For instance, services such as OpenTable allow customers to make reservations at partner restaurants (Table, 2020). Thus, the equilibrium analysis of the advance reservation model should be of interest independently of the profit maximization aspect.

The remaining sections are organized as follows. In Section 2 we review related work and discuss how our model and analysis differ from prior work in the literature. In Section 3 we introduce our model, and establish the possible types of equilibrium for each preemption policy. In Section 4 we analyze the model, derive the Nash equilibria, and compare the revenue outcomes for each of the preemption policies, under the assumption of deterministic service. Section 5 extends the analysis to general service distributions. We conclude and suggest future research directions in Section 6.

2. Related Work

Our work here relates to strategic behavior in queues, commonly referred to as *queuing games*. The scenarios that we study also relate to the cloud computing marketplace, thus we also contrast our model with those used in prior studies of cloud computing economics.

Analysis of queuing games was pioneered by Naor (Naor, 1969) and has been studied extensively since. In that seminal paper, the author studies an $M|M|1$

queue where customers decide whether to join or balk after observing the queue length. The model developed in that work has been used subsequently, with the books (Hassin & Haviv, 2003) and (Hassin, 2016) providing an extensive review of the field. In particular, the model that we study here relates to *queuing games* with priority, wherein customers can purchase their priority level, a consideration featured in several papers in the literature (Balachandran, 1972; Mendelson & Whang, 1990; Hassin & Haviv, 1997). These works consider discrete levels of priority, while our work considers a continuum of priorities. A system featuring such a continuum of priorities is studied in (Master et al., 2017), however this work focuses purely on the performance aspects of an $M|M|1$ queue with such priorities and is not analyzed in the context of a game setting as is the case here.

Many previous works on queuing games assume that customers are able to choose their priority freely, regardless of the order in which the customers make their decisions on which priority to take (Haviv, 2011; Gavirneni & Kulkarni, 2016; Gurvich et al., 2019). A popular model is one in which priority is tied to payment, thus customers willing to pay additional fees are able to achieve higher priority (Hassin, 1995). Other possible schemes include the use of contracts whereby customers agree to pay some $h(X)$ amount to join, based on some random variable X whose realization is unknown to the customer while deciding when to join, assigning preemptive priorities in a randomized fashion, or randomized entry fees (Haviv & Oz, 2018). Or, similarly to the model we consider here, the base price for service may be zero but customers may pay additional fees for higher priorities; customers then choose between joining or balking from the queue, and if joining how much to pay for priority (Wang et al., 2019). Alternatively, a scheme may exist which grants priority to customers willing to refer new customers to a wait list (Yang & Debo, 2019).

By contrast, in our model, the priority that a customer may exercise is related to the time at which it makes the priority reservation - making a reservation only enables a customer to have priority over those customers who make reservations later and those who do not make reservations at all. Thus, the time at which a customer requests a priority reservation plays a crucial role in the priority level it ultimately acquires, and thus there is no incentive to delay making the reservation decision, as better priority cannot be gained by waiting. From a provider perspective, this priority scheme does have the advantage of pre-planning, (e.g., allocating service resources), as customers are declaring their desired service in advance. This additionally allows for quality of service guarantees (Charbonneau & Vokkarane, 2011).

Other methods of exploiting customer delay sensitivity within the context of providers with finite capacity include the use of priority auctions, whether for preemptive or non-preemptive priorities (Afèche & Mendelson, 2004), and ones in which customers are segmented by lead time and subject to pooling types with different delay costs into a single class or focusing on serving the most impatient and patient customers, while pricing out those who fall in the middle (Afeche & Pavlin, 2016). The queue may also have mechanism to accept bribes, where customers may jump ahead in line upon entry, based on the amount

paid (Lui, 1985). Alternatively, customers may have the ability to rearrange themselves in the queue based on their wait tolerance, with customers bumped back in the queue paid monetary compensation in exchange (Yang et al., 2017).

By contrast, here our focus is on controlling behavior via various preemption policies based upon priority. In particular, we consider the HPR policy where a customer's priority reservation also implies that its job cannot be preempted by higher priority jobs. To the best of our knowledge, this hybrid form of preemption is novel and has not been considered in prior work. We note that preemption is an important feature in both private and public computing services. For instance, Two Sigma's *Cook* scheduler divides jobs into batch and interactive, prioritizing interactive jobs and preempting a batch job if necessary (Jin, 2016). Per Two Sigma's benchmarks, the purported benefit of *Cook* is low latency and high throughput compared to a non-optimized system, thanks to the prioritization and preemption of jobs (Jin, 2016). This sort of preemption is the inspiration for our HPR model, which we combine with the AR mechanism in order to incentive the customer to purchase the offered priority. Thus, the mechanism is designed to encourage customers to reserve as early as possible as opposed to mechanics based upon exploiting the delay tolerance of customers attempting to be served right in that moment.

The works in (Juneja & Jain, 2009), (Jain et al., 2011), and (Honnappa & Jain, 2015) analyze *concert queuing games*. In such games, customers strategically choose their arrival times to minimize their wait times. The similar meeting game features members timing their arrivals in order to minimize time spent waiting before a quorum is achieved for the meeting to begin (Guéant et al., 2011). Alternatively, a game can be formulated in which customers trade off the cost of waiting for a scheduled airplane flight to take off with the benefit from gaining a better seat (Talak et al., 2019). In contrast, in our model, the point of time during the reservation period at which a customer realizes that it will need service in the future is a random variable. Furthermore, there is no possible benefit to waiting to make a reservation under our system. Unlike in a concert queue and meeting games where a later arrival may minimize the wait time, a later reservation under our model cannot result in a higher priority offer. The most likely outcome from waiting to make the reservation is a lower priority offer and thus a longer expected wait for the customer.

Advance reservations have been researched from various other perspectives in the literature, including scheduling and routing algorithms for communication networks, methods for revenue maximization, and performance analysis of queuing systems. Thus, the work in (Wang et al., 2013) describes a distributed architecture for advance reservation, while (Smith et al., 2000) proposes a scheduling model that supports AR and evaluates several performance metrics. Queuing analysis of advance reservation systems is notoriously difficult (Chen et al., 2017a). The work in (Virtamo, 1992) analyzes the impact of advance reservations on server utilization under a stochastic arrival model, and (Reiman & Wang, 2008) considers admission control strategies in reservation systems with different classes of customers. The use of AR is particularly beneficial when dealing with renewable resources which can be reserved repeat-

edly, as highlighted in the work in (Chen et al., 2017b); however, this work and a similar non-AR version in (Levi & Radovanović, 2010) consider loss systems where if the server is unavailable the request for the resource is dropped. Here we consider a system where if the server is not available at the desired time, the reservation is rescheduled for the first available point after the desired start time.

While there exists a rich literature on advance reservations, few works study advance reservation systems as a game. The strategic behavior of customers in a system that support AR is studied in (Simhon & Starobinski, 2014) and (Simhon et al., 2015). These two papers study a *loss system*, i.e., a system with no queue. In our paper, instead, we focus on a queuing system (i.e., customers that encounter a busy server wait for service). This leads to a different model and, interestingly, more explicit results than in those prior works. We show that the server utilization (system load) plays a key role in the behavior of the system and, specifically, in the number of equilibria. Furthermore, the impact of preemption (or lack thereof) is ignored in those prior works. In addition, while some works on reservation consider the issue customers reneging on their reservations, such as (Oh & Su, 2018), here we assume that the customers who make a reservation within this queue do so because they require service and are not attempting to pay as a form of insurance, particularly as the full value of the reservation fee is payed in advance as opposed to it being a partial deposit.

One of the motivations of our model is cloud computing. While we do not claim to model the practices of a specific provider, our model is inspired by the use case of cloud computing and our goal is to provide insight into customer behavior given competition for a cloud-based resource, and in particular, the customer decision of whether to make a priority reservation or not. In contrast, many previous works in the literature assume that only one type of instance exists, typically a non-preemptive type. This applies to the context of determining pricing levels under demand uncertainty (Niu et al., 2012), pricing under a duopoly competition situation (Li et al., 2014), or evaluating the socially optimal pricing model with a single instance type (Menache et al., 2011). The actual scheduling of jobs and allocation of computing time are inherently tied to the price customers are willing to pay for priority, and this issue has been studied under multiple contexts. These include scheduling based on sensitivity to deadlines (Lucier et al., 2013), maximizing the social welfare of scheduling based upon user supplied job parameters (Chawla et al., 2017a; Azar et al., 2015), or scheduling where customers' willingness to pay is tied to how rapidly a job is completed (Jain et al., 2014). However, these works are ultimately concerned with solving other aspects of the pricing problem, such as scheduling jobs in an on demand fashion and scheduling in a system where customers are misreporting the properties of their job to improve their own allocation, rather than studying how preemption policies impact willingness to purchase priority reservations as we do here.

Markets mixing multiple instance types have also been studied in the literature. While (Javadi et al., 2011; Song & Guerin, 2017) investigate customer behavior when bidding on spot instances, the model employed in these works

assume customers have already made the decision to enter the sport market. In contrast, the authors in (Dierks & Seuken, 2019) consider a situation in which customers are choosing between fixed price reserved service and variable priced preemptive spot service. The work is concerned with the problem of how to price the spot market in such a manner so that the fixed price market is not ignored in the process. Our model operates on a similar principle of having two instance types, but in our work we are concerned with how the customers' priority reservation decisions are impacted by changes in the provider decisions, rather than viewing the reserved and spot markets as two services essentially competing one against the other, which is the heart of the work in (Dierks & Seuken, 2019).

Alternate pricing mechanisms based upon flexible reservation and time-of-use pricing are studied in (Wang et al., 2015; Chawla et al., 2017b). The former paper focuses on the issue of fixed term reservations, in which customers pay for the right to claim instance availability for some number of weeks, months, or years. The latter is concerned with the issue of load shifting to account for variations in demand for resources. Both of these are separate from the questions which we consider in this paper, such as whether customers are willing to be charged a certain cost to receive relative priority over other customers, and how different preemption policies implemented by a provider impact strategic behavior of said customers and the provider's revenue.

3. Preliminaries

In this section, we first introduce our model, including notation and assumptions. Next, we characterize the possible types of equilibria, and show that any equilibrium must be of threshold type.

3.1. Model

We consider a queuing system that utilizes advance reservation mechanics and in which customers do not cooperate, thus leading to a non-cooperative game situation. Generally, customers will approach a provider without knowledge about other customers attempting to enter service and the system state. We assume that the provider does not disclose this information, nor is there a trivial means to deduce the information in order to communicate directly with others seeking service. Thus, customers do not consult each other, and their decision is based solely on the information regarding their expected wait times and the reservation cost C.

Within the game, we consider the following parameters related to the $M|G|1$ queue and customer priorities:

- $\hat{p}$ - The priority made available to a customer, related to the time of its reservation request.
- p - The priority assigned to the customer following the decision to make a reservation or not.

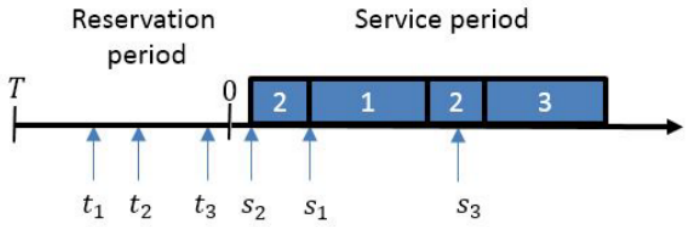


Figure 1: An Advance Reservation scheme with Preemptive Resume (PR) policy in effect. Customers 1,2, and 3 make a reservation during the reservation period at times t_1, t_2, t_3, respectively. As such, customer 1 has the highest priority, and customer 3 has the lowest. However, while customer 1 has the highest priority, customer 2 has the earliest requested start. Thus, customer 2 is permitted to enter service prior to customer 1, but is preempted at customer 1's desired service start time.

- C - The cost set to make a priority reservation.
- λ - The arrival rate of customers to the service queue.
- X - The random variable describing the length of service.
- μ - The service rate, equal to $1/E[X]$ by definition.
- ρ - The system load, which is the ratio of arrival rate to service rate; i.e., $\rho = \lambda/\mu$.

As in other work on advance reservations (e.g. (Virtamo, 1992)), we assume that the time axis is divided into two distinct periods, the *reservation period* and the *service period*, to make the analysis of the game tractable. The reservation period lasts from the time the first customer makes its priority reservation decision until the scheduled start of the service period, at which time the queue begins processing jobs. We describe what occurs during each period below.

Reservation Period: During this period, a customer determines that a job needs to be processed during the upcoming service period. At this point, the customer approaches the provider with the service request. We assume that once the customer determines that it wants service, it will not balk or renege from the queue. The customer is offered the option to make a priority reservation for a cost C.

In advance reservation systems, customers making reservations early typically receive higher priority over customers making reservations later. Therefore, each individual customer receives an individualized priority offer based on the time at which the reservation is made. The customer's potential priority $\hat{p}$ is relative to how long prior to the start of the service period the customer is making its decision.

The timing of the reservation decisions during this period follows an arbitrary continuous probability distribution. Using the probability integral trans-

formation theorem (Dodge, 2003), we can map such distribution to a uniform distribution between 0 and 1. As such, we let the offered potential priority $\hat{p}$ be drawn from $U[0,1]$, with 1 as the highest priority. Based on this transformation, we can view $\hat{p}$ as the percentage of other customers that the customer is expected to have priority over. For example, a tagged customer with potential priority $\hat{p} = 0.75$ has the opportunity to make a reservation request yielding priority over 75 % of the other customers in the system which are expected to make their decisions after the tagged customer. This priority mechanism is a key feature of our system. As proven in (Simhon & Starobinski, 2017), there is no advantage to waiting to make this request as the offered priority will decrease over time as other customers make their requests. Thus, waiting cannot confer a greater benefit.

If the customer exercises it potential priority, the customer's priority is $p = \hat{p}$. Otherwise, the priority is $p = 0$. Thus, the customer's cost if exercising the potential priority is the cost of waiting with priority $\hat{p}$, plus the cost C to make the priority reservation. If electing to forgo reservations, the customer is defaulted to the lowest priority. Thus, the customer's total cost of waiting is the cost of waiting with priority 0. The priority and requested start time of the job are noted for the service period.

Service Period: During this period, the jobs are queued up and processed by the provider. Jobs are served in priority order, jobs with equal priority are served in First Come First Serve order. A job's arrival time to the queue is its requested start time. The arrival times of the jobs follow a Poisson process with parameter λ, while the service times follow a general distribution with mean service time $1/\mu$.

We assume that the order in which jobs arrive to the queue is independent of their priority. Jobs enter service as they reach the head of the service queue. Whether a newly arriving job may preempt a lower priority job currently in service depends on the preemption policy that the provider chooses to implement. However, all jobs are permitted to enter service regardless of whether they will be preempted during service, as the preemption policies being considered are work conserving. Thus, there is no concern for service being wasted.

An example of the process is visualized in Figure 1 for a system with deterministic service time. In this example, there are three customers. In the reservation period, they make their reservation requests at times t_1, t_2, and t_3. Therefore, customer 1 has the highest priority, and customer 3 has the lowest priority. During the service period, customers are served according to a Preemptive Resume $M/D/1$ queue. Customer 2 has the earliest start time s_2; as no one is in service at that point customer 2 may enter service immediately. However, as customer 1 has higher priority, customer 2 is preempted at the desired service start time s_1 of customer 1. Once the service of customer 1 is completed, the service of customer 2 resumes from the point of interruption. Customer 3 has a requested start during the service period of customer 2, but since customer 2 has higher priority than customer 3, customer 3 must wait until the service of customer 2 is completed.

In summary the system during the service period can be modeled as an

unobservable $M/G/1$ queue based upon the model introduced in (Edelson & Hilderbrand, 1975), where each incoming job has a random priority $p \in [0, 1]$. If the customer associated with the job made a priority reservation then $p = \hat{p}$ otherwise $p = 0$. The customer pays C if opting to exercise the offered $\hat{p}$, and 0 otherwise. A similar payment model is utilized in Boston University's Shared Computing Cluster, wherein *shared nodes* may be freely used, while *buy-in* nodes may be purchased for dedicated access (Liao et al., 2018). We can imagine that our AR system is operating under a similar philosophy, whereby it is being made available to a wider community in order to conduct work, while offering a higher tier of service in exchange for a fee.

3.2. Equilibrium Types

With the model established, we wish to determine the structure of equilibria which may arise in the game. In considering the potential equilibria, we assume that the system has reached steady-state. We first introduce the action space, which consists of the possible actions that customers may take.

Definition 1. *Let σ denote the* action *taken by the customer. There are two actions available: the Reserved Instance (RI) action to make a reservation, and the Non-Reserved Instance (NRI) action to forgo a reservation.*

We will show in the sequel that the equilibrium strategies are pure, namely customers will always choose one of the above two actions and not mix (randomize) between them.

While the provider can choose between implementation of the NP, PR, or HPR policies, in order to determine the relationship between an equilibrium state and the cost leading to that state, we begin by performing a general analysis of the game without respect to the specific preemption policies.

Per general game theoretic assumptions, customers are rational. Thus, the reservation decision is based solely on which option will incur the least cost. Let $W(\sigma, p)$ be the expected wait time in the queue for a customer taking action σ, with priority p. Then $W(RI, \hat{p})$ is the expected wait time for a customer who decides to exercise priority $\hat{p}$. Similarly, $W(NRI, 0)$ is the expected wait time of a customer who does not exercise its potential priority, and accepts the lowest priority of 0.

The decision to make a priority reservation or not will then be based upon the relation between the costs of the expected wait times and the priority reservation cost C. If α is the customer's cost of waiting per time unit, the following must hold for a customer to choose to make the reservation:

$$\alpha W(RI, \hat{p}) + C \leq \alpha W(NRI, 0). \tag{1}$$

As customers are statistically identical in our model, α is identical for all customers, and WLOG we let $\alpha = 1$ to consider the normalized cost. We are particularly interested in establishing the equilibrium strategy followed by customers. An equilibrium state is reached if no customer can unilaterally change its strategy and improve its payoff. As Equation (1) is defined in terms of wait times with priority, we claim that a threshold strategy will be followed:

Cost C	**Possible Equilibria**
$\underline{C} < C < \overline{C}$	At least one $some - reserve$ possible.
$C < C_0$	$All - reserve$ possible.
$C < \underline{C}$	$All - reserve$ only equilibrium.
$C > C_1$	$None - reserve$ possible.
$C > \overline{C}$	$None - reserve$ only equilibrium.

Table 1: Possible equilibria associated to a cost C based on the relationship of C to the values defined in Equation (3).

Definition 2. *A threshold strategy is a strategy where for some $\phi \in [0,1]$, customers make the following decision based on their offered potential priority $\hat{p}$:*

- *RI if $\hat{p} > \phi$;*
- *NRI if $\hat{p} \leq \phi$.*

Lemma 1. *Under the game model, at equilibrium, all customers follow a threshold strategy.*

The proof of this lemma is provided in Appendix A. Based on Lemma 1, any equilibrium must be of the following type:

1. $All - reserve$: all customers choose RI.
2. $None - reserve$: all customers choose NRI.
3. $Some - reserve$: only some of the customers choose RI, the rest choose NRI.

Next, we define a function that relates a threshold priority $\phi \in [0,1]$ to the cost leading to an equilibrium state at that threshold:

$$C(\phi) \triangleq W(NRI, 0) - W(RI, \phi). \tag{2}$$

As shown in the sequel, the wait time formulas are derived from the known formulas in (Conway et al., 1967, Ch. 8) which are continuous.Therefore, $C(\phi)$ is itself continuous. As a result, we determine which equilibria types are associated with a given reservation cost C based on the relationship between C and certain special values of $C(\phi)$:

$$\begin{aligned} C_0 &\triangleq C(0); \\ C_1 &\triangleq C(1); \\ \underline{C} &\triangleq \min_{\phi \in [0,1]} C(\phi); \\ \overline{C} &\triangleq \max_{\phi \in [0,1]} C(\phi). \end{aligned} \tag{3}$$

Table 1 contains the association between C and the values of $C(\phi)$ that lead to the equilibrium types. In general, at least one $some - reserve$ equilibrium

is possible if a solution to $C = C(\phi)$ exists. An $all - reserve$ equilibrium is possible if C is less than the cost leading to a threshold of 0, where all customers have incentive to make a reservation. Similarly, a $none - reserve$ equilibrium is possible if C is greater than the cost leading to a threshold of 1, where no customers have incentive to make a reservation.

As shown in the sequel, we note that $C(\phi)$ under each policy may be monotone increasing, monotone decreasing, or unimodal based upon the values of the second moment and service load as well as the preemption policy in place. Therefore, it is not universally the case that $\underline{C} = C_0$ and $\overline{C} = C_1$. As such, it is necessary to derive the specific cost function for the policy and service distribution in place to determine the behavior of $C(\phi)$.

4. Equilibrium and Revenue Analysis under Deterministic Service

In this section, we derive the equilibrium outcomes for each preemption policy and establish the regions for which each equilibrium type is possible. We use the notation from Equation (3) for each preemption policy, with the understanding that they are being applied to that policy's cost function. Once the equilibrium outcomes for all three policies are derived, we compare the maximal revenue achievable under each policy and determine the preemption policy that a rational provider would implement.

To streamline the presentation of the analysis, we assume an $M|D|1$ queue is in effect (i.e., the service time is deterministic). Deterministic service time is an important special case. Indeed, under the PR policy, the provider can inform all the customers making reservation about the precise beginning and end times of their services, since they will have preemptive priority over all the customers making reservation later during the reservation period (and those not making reservation at all).

Under the NP or HPR policies, it is not possible to preempt lower priority customers. However, because of how reservations work, a future customer in the reservation period cannot get higher priority. Hence, it is still possible to generate a window for the service time. An initial slot can be computed based upon the reservations to that point. Because service times are deterministic, there is a limited window for a lower priority customer to impact higher priority ones. The worst-case added delay to the service time in that case is $1/\mu$. Thus while exact start and finish times cannot be determined in advance in contrast to the PR policy, the reservation system allows the provider to make a service guarantee by supplying a window for the completion of service based on the information provided by the customer.

Note that if the time quotation were given to a customer before he/she makes a reservation, it would reveal to the customer how many prior reservations were made and affect his/her decision. Thus, to preserve the notion of an unobservable queue, we consider here the situation where the time quotation is not given until after the customer decides to make a reservation. We leave the analysis of the case where the quotation is made prior to reservation for future work.

While the equilibrium outcomes of the NP and PR policies are insensitive to the service distribution except for the mean, the analysis of the HPR policy results in multiple sub-cases depending on the second moment of service. The presentation of results pertaining to the $M|G|1$ queue is hence deferred to Section 5.

4.1. NP Policy

We first consider the analysis of the system under the Non Preemptive (NP) policy. That is, once in service a customer cannot be preempted. Any higher priority customer that arrives must wait in the queue for the current customer to complete service before entering the server.

Let ρ_p be the system load of customers of priority p, and ρ_a be the system load of customers with priority higher than p. Then the wait time for a class p customer choosing action σ is derived from the formula in (Conway et al., 1967, p.164) for wait time in an NP queue as follows:

$$W(\sigma, p) = \frac{\rho}{2\mu(1-\rho_a-\rho_p)(1-\rho_a)}. \tag{4}$$

Given a class p customer makes a reservation, the class of customers with higher priority is simply the fraction of customers who made priority reservations prior to the current customer, thus $\rho_a = (1-p)\rho$. Further, by definition only one customer can have the specific priority p thus the effective arrival rate is 0 for customers of this class, so $\rho_p = 0$. The resulting wait time for a (virtual) customer with priority $p = \phi$ is given by:

$$W(RI, \phi) = \frac{\rho}{2\mu(1-\rho(1-\phi))^2}. \tag{5}$$

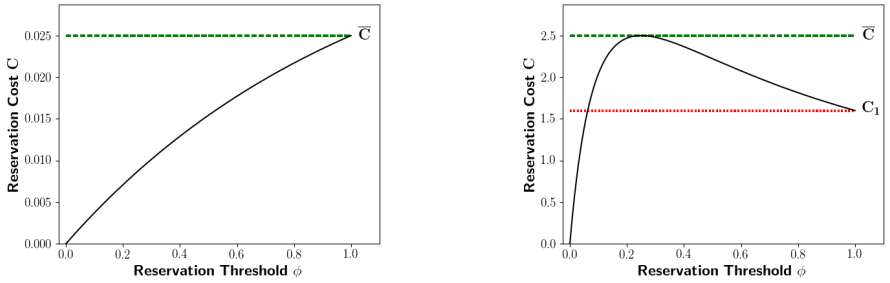
For a class 0 customer who did not make a reservation, the class of customers of higher priority are all of the customers who did make the reservation, and so $\rho_a = (1-\phi)\rho$. The class of customers with the same priority class will be all customers who did not make a reservation and $\rho_p = \phi\rho$. Therefore the resulting wait time for a customer with no reservation is given as:

$$W(NRI, 0) = \frac{\rho}{2\mu(1-\rho)(1-\rho(1-\phi))}. \tag{6}$$

Using Equations (5) and (6), we can then derive the cost function under the NP policy from Equation (2):

$$C_{NP}(\phi) \triangleq \frac{\rho^2\phi}{2\mu(1-\rho)(1-\rho(1-\phi))^2}. \tag{7}$$

Computing the derivative of this function with respect to ϕ, we determine that it is monotone increasing if $\rho \leq 1/2$ and unimodal with a unique maximum at $\phi = (1-\rho)/\rho$ otherwise. Based upon this information, we apply the definitions in Equation (3) to C_{NP} that yield the following quantities:



(a) $\rho = 0.2$ (b) $\rho = 0.8$

Figure 2: Plots of the cost $C_{NP}(\phi)$ leading to equilibrium ϕ for $\rho \in \{0.2, 0.8\}$ and $\mu = 1$. At low load, the function is monotone increasing. However, at high load, the cost function is unimodal, reflecting how incentives to reserve vary as the system load changes.

$$\underline{C} = C_0 = 0;$$
$$C_1 = \frac{\rho^2}{2\mu(1-\rho)};$$
$$\overline{C} = \begin{cases} C_1 & \text{for} \quad \rho \leq \frac{1}{2}; \\ \frac{\rho}{8\mu(1-\rho)^2} & \text{for} \quad \rho > \frac{1}{2}. \end{cases}$$

We note that as $C_0 = 0$, we cannot have an $all - reserve$ equilibrium here. Further, the value of ρ determines what equilibrium types are possible for a given cost C. We formalize our observations with the following theorem, which applies the results from Table 1 to $C_{NP}(\phi)$:

Theorem 1. *Under the NP policy, the equilibrium outcomes are as follows: If* $\rho \leq 1/2$*:*

- *If* $C < \overline{C}$*, there is a unique* $some - reserve$ *equilibrium.*
- *If* $C > \overline{C}$*, there is a unique* $none - reserve$ *equilibrium.*

Else, if $1/2 < \rho < 1$*:*

- *If* $C < C_1$*, there is a unique* $some - reserve$ *equilibrium.*
- *If* $C_1 < C < \overline{C}$*, there are two* $some - reserve$ *equilibria and one* $none - reserve$ *equilibrium.*
- *If* $C > \overline{C}$*, there is a unique* $none - reserve$ *equilibrium.*

Example 1. Figure 2 illustrates Theorem 1 for different loads ρ, and $\mu = 1$. Fig. 2(a) shows an example of the cost function behavior at a load less than $1/2$, namely $\rho = 0.2$. We notice that the function is monotone increasing with

respect to the threshold ϕ. Thus, for a cost lower than 0.025, there is exactly one solution to $C = C(\phi)$ yielding a unique $some-reserve$ equilibrium for that cost. If the cost is larger than 0.025, there are no solutions to $C = C(\phi)$ and thus a $none-reserve$ equilibrium prevails.

This example reflects the intuitive incentive to make a priority reservation. A customer with a higher priority will bypass waiting for a greater fraction of customers, thus such customers have a higher incentive to make a reservation, and so a larger threshold priority corresponds to a higher cost. However, as the priority decreases, customers gain advantage over fewer customers. At priority 0, there is no incentive to make a reservation, as the customer is locked into the lowest priority regardless. Thus, while a lower value of ϕ corresponds to a larger proportion of customers making reservations, customers with lower priority would do so only at a lower cost.

Fig. 2(b) shows an example for a load greater than 1/2, namely $\rho = 0.8$. The function is unimodal, thus there is a range of costs leading to two solutions to $C(\phi) = C$. For example, if $C = 2$, there are two solutions to $C(\phi) = C$ and thus there are two $some - reserve$ equilibria associated to the cost. In addition, $C > C_1 = 1.6$, so the cost is greater than the cost that leads to the $none-reserve$ equilibrium state. Therefore, there are a total of three equilibria associated with this cost.

We observe that for the range of cost $C < C_1$, the behavior is similar to the low load case (i.e., a single $some - reserve$ equilibrium prevails). However, in the range $C_1 < C < \overline{C}$, three equilibria are possible. In that case, due to the high reservation cost and the high arrival rate of jobs and associated wait time, customers tend to follow the same behavior as those of other customers. For instance, a customer has no incentive to make a reservation if all other customers decline making a reservation. However, if a large fraction of customers make a reservation, only customers with the lowest priorities will forgo reservations.

4.2. PR Policy

We next consider the analysis of the cost function that results under a Preemptive Resume (PR) policy. That is, a policy under which all customers in service are eligible to be preempted on the arrival of a customer with higher priority. Any preempted customers reenter the queue at a later time, resuming service from the point of interruption.

To analyze the system behavior under this policy, we begin by noting that the wait time for a class p customer taking action σ can be derived from the formula for system wait time under a PR policy, given in (Conway et al., 1967, p.175) as follows:

$$W(\sigma, p) = \frac{1}{\mu(1-\rho_a)} + \frac{\rho_a + \rho_p}{2\mu(1-\rho_a)(1-\rho_a-\rho_p)} - \frac{1}{\mu}, \tag{8}$$

where ρ_a and ρ_p are defined as in Subsection 4.1. From the perspective of a class p customer making a reservation, the arrival rate of the higher class customers is $\rho_a = (1-p)\rho$ and the arrival rate of the current class is $\rho_p = 0$ as before.

Similarly, from the perspective of a class 0 customer not making a reservation, $\rho_a = (1-\phi)\rho$ and $\rho_p = \phi\rho$ as in the NP policy. We can derive $W(RI, \phi)$ and $W(NRI, 0)$ as in Subsection 4.1 using Equation (8).

As a result, from $W(RI, \phi)$, $W(NRI, 0)$, and Equation (2) the cost function with respect to threshold ϕ under the PR policy is:

$$C_{PR}(\phi) \triangleq \frac{\rho\phi}{2\mu(1-\rho)(1-\rho(1-\phi))^2}. \tag{9}$$

Comparing Equations (7) and (9) we find $C_{PR}(\phi) = (1/\rho)C_{NP}(\phi)$. As the two functions differ by a multiplicative factor, $C_{PR}(\phi)$ *follows the same behavior as* $C_{NP}(\phi)$, which is confirmed by computing the derivative of $C_{PR}(\phi)$ with respect to ϕ. Accordingly, the values derived from Equation (3) and resulting equilibrium structure are:

$$\begin{aligned}
\underline{C} &= C_0 = 0; \\
C_1 &= \frac{\rho}{2\mu(1-\rho)}; \\
\overline{C} &= \begin{cases} C_1 & \text{for} \quad \rho \leq \frac{1}{2}; \\ \frac{1}{8\mu(1-\rho)^2} & \text{for} \quad \rho > \frac{1}{2}. \end{cases}
\end{aligned}$$

Theorem 2. *Under the PR policy, the equilibrium outcomes are as follows:*
If $\rho \leq 1/2$*:*

- *If* $C < \overline{C}$*, there is a unique some − reserve equilibrium.*
- *If* $C > \overline{C}$*, there is a unique none − reserve equilibrium.*

If $1/2 < \rho < 1$*:*

- *If* $C < C_1$*, there is a unique some − reserve equilibrium.*
- *If* $C_1 < C < \overline{C}$*, there are two some − reserve equilibria, and one none − reserve equilibrium.*
- *If* $C > \overline{C}$*, there is a unique none − reserve equilibrium.*

That the same equilibrium structure exists in the NP and PR cases is related to the nature of the reservation decision. Whether preemption is in effect or not, in both cases the decision is between having priority $\hat{p}$ and priority 0. The reservation decision does not impact whether a customer is eligible for preemption in service by higher priority customers. Thus, while there is incentive to make a reservation under the PR case in order to save wait time by preempting lower priority customers, the overall advantages and disadvantages of reserving with a given priority $\hat{p}$ are the same as in the NP case. In contrast, as seen below, the HPR policy changes the nature of the incentive to make a reservation, resulting in a different structure.

4.3. HPR Policy

We next consider the system under the Hybrid Preemptive/ Non-Preemptive (HPR) policy. In the HPR policy, a customer who has made a priority reservation cannot have its service preempted by a customer arriving with higher priority. However, a customer who does not make a reservation is preempted if a higher priority customer (i.e., a customer who made a reservation) arrives. As with the PR policy, any such customer preempted will resume from the point of interruption when reentering service.

Thus, under the HPR policy, we can consider customers making a reservation to be acting as if they are in an NP queue where the only customers present are those who also made reservations. Thus, Equation (4) is modified to reflect the wait time amongst this group of customers. Letting $p = \phi$, the wait time at the threshold becomes

$$W(RI, \phi) = \frac{\rho(1-\phi)}{2\mu(1-\rho(1-\phi))^2}.$$

Meanwhile, the customers who do not make reservations act as if they are the lowest priority customers in a queue operating under a PR policy. As a result, $W(NRI, 0)$ is the same under the HPR and PR policies. We then derive the cost function by taking the difference between $W(RI, \phi)$ and $W(NRI, 0)$:

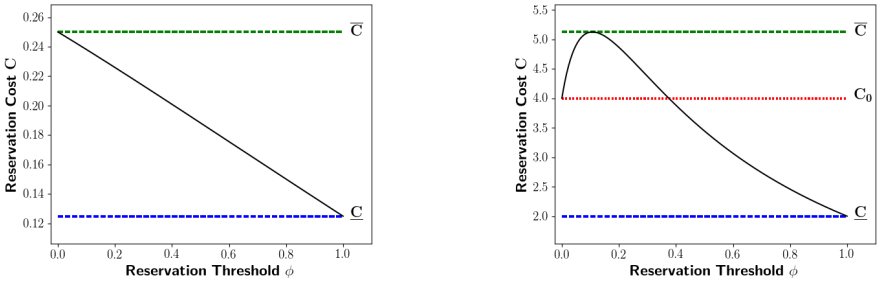
$$C_{HPR}(\phi) = \frac{1}{\mu}\left(\frac{2(1-\rho)(1-\rho(1-\phi)) + \rho\phi}{2(1-\rho)(1-\rho(1-\phi))^2} - 1\right). \tag{10}$$

Analyzing the derivative with respect to ϕ, we find that if $\rho < 1/2$, the function is monotone decreasing, otherwise it is unimodal with a unique maximum at $\phi = (2\rho^2 - 3\rho + 1)/(2\rho^2 - 3\rho)$. We then obtain the following by applying the definitions in Equation (3):

$$\begin{aligned}
C_0 &= \frac{\rho}{\mu(1-\rho)};\\
C_1 &= \frac{\rho}{2\mu(1-\rho)};\\
\underline{C} &= C_1;\\
\overline{C} &= \begin{cases} C_0 & \text{for} \rho \in \left(0, \frac{1}{2}\right);\\ \frac{-4\rho^2+4\rho+1}{8\mu(1-\rho)^2} & \text{otherwise.} \end{cases}
\end{aligned}$$

Theorem 3. *Under the HPR policy, the equilibrium outcomes are as follows: When $\rho < 1/2$:*

- *If $C < \underline{C}$, there is a unique $all - reserve$ equilibrium.*
- *If $\underline{C} < C < \overline{C}$, there are multiple equilibria: one each of $all - reserve$, $some - reserve$, and $none - reserve$.*
- *If $C > \overline{C}$, there is a unique $none - reserve$ equilibrium.*



(a) $\rho = 0.2$ (b) $\rho = 0.8$

Figure 3: Plots of $C_{HPR}(\phi)$ for $\rho \in \{0.2, 0.8\}$ and $\mu = 1$. In contrast to the NP and PR cases, the minimum values of the cost functions are not 0, thus $all - reserve$ equilibria are possible. In fact, under low loads the $all - reserve$ equilibrium is associated with the highest cost customers are willing to pay for priority.

When $1/2 < \rho < 1$*:*

- *If* $C < \underline{C}$*, there is a unique* $all - reserve$ *equilibrium.*
- *If* $\underline{C} < C < C_0$*, there are multiple equilibria: one each of* $all - reserve$*,* $some - reserve$*, and* $none - reserve$*.*
- *If* $C_0 < C < \overline{C}$*, there are multiple equilibria: two* $some - reserve$ *and one* $none - reserve$*.*
- *If* $C > \overline{C}$*, there is a unique* $none - reserve$ *equilibrium.*

A significant difference with the NP or PR policies is that the $all - reserve$ equilibrium is possible under HPR since $C_0 > 0$. This is due to the difference in the consequences of the reservation decision compared to the previous cases. Under NP or PR, making a reservation only grants a priority, but does not alter whether a customer is eligible for preemption while in service. Under the HPR policy, a customer making a reservation is granted both priority *and* protection against preemption by customers with higher priority.

Example 2. Considering Figure 3, we find that the difference in priority advantages are best exemplified for thresholds near 0. Because making a reservation under the HPR policy prevents preemption, customers offered low priorities receive a tangible advantage when reserving, and thus have incentive to do so unlike in the NP or PR policies.

As seen in Figure 3(a), under low load (i.e., $\rho = 0.2$), the cost leading to a particular threshold decreases as ϕ approaches 1. If no customers make a reservation, the queue reverts to a simple FCFS queue as all customers have equal priority. However, as customers begin to make reservations, those who do not are negatively impacted by preemption. This in turn results in greater incentive to make a reservation, which results in customers with lower priorities

being willing to bear a higher cost to make a reservation. In fact, here the highest cost is associated with the threshold $\phi = 0$, corresponding to the $all - reserve$ equilibrium.

For $\underline{C} = 0.13 < C < \overline{C} = 0.25$, there is exactly one solution to $C(\phi) = C$, i.e., a $some-reserve$ equilibrium, but there is also an $all-reserve$ equilibrium and a $none-reserve$ equilibrium associated to that cost. We find a similar situation at higher load (see Figure 3(b) for $\rho = 0.8$), except that if $C_0 = 4 < C < \overline{C} = 5.13$, we instead find two $some-reserve$ and one $none-reserve$ equilibria associated with the cost, as there is a limit to the reservation costs customers are willing to bear. Beyond a certain cost level, the lowest priority customers are better off accepting a longer wait.

In the next section, we consider the resulting revenues obtained by the providers under each policy, and the resulting impact on the maximum revenue obtainable when implementing one policy over another.

4.4. Provider Revenue Maximization

In this subsection, we consider the question of which policy a provider should implement to realize the maximum revenue intake from priority purchases. We assume that the cost C is a fee that a customer needs to pay in order to make a reservation and exercise its potential priority. For a given system load ρ, we demonstrate that the HPR policy always leads to the maximum revenue. We also compare the maximum revenue achieved under the different policies and show that while the performance ratio between HPR and PR is bounded, the performance ratio between HPR (or PR) and NP is unbounded.

We first define the revenue function. As customers arrive via a stochastic process, we evaluate the expected revenue per time unit. With ϕ as the threshold, the arrival rate of customers purchasing priority is $\lambda(1-\phi)$. As revenue is cost multiplied by the number of paying customers, the resulting function is:

$$R(\phi) \triangleq \lambda(1-\phi)C(\phi). \tag{11}$$

We note that $R(\phi)$ is continuous for $\phi \in [0,1]$, as it is the product of three continuous functions. As such, we are guaranteed to have a maximum value of $R(\phi)$ for $\phi \in [0,1]$. To show that the HPR policy leads to the greatest maximum revenue for any load ρ, we first need the following Lemma:

Lemma 2. *For fixed λ and μ, and given $\phi \in [0,1]$, $C_{HPR}(\phi) \geq C_{PR}(\phi) \geq C_{NP}(\phi)$.*

Proof. As λ and μ are fixed, ρ is also fixed by definition. As $(1/\rho)C_{PR}(\phi) = C_{NP}(\phi)$ and $\rho < 1$, we immediately conclude that $C_{PR}(\phi) \geq C_{NP}(\phi)$. To show that $C_{HPR}(\phi) \geq C_{PR}(\phi)$ also holds, we show that $C_{HPR}(\phi) - C_{PR}(\phi) \geq 0$ must hold. Referring to Equations (9) and (10), we note the common denominator $2\mu(1-\rho)(1-\rho(1-\phi))^2$. Multiplying both equations by this quantity, $C_{HPR}(\phi) - C_{PR}(\phi) \geq 0$ holds if the following also holds:

$$2(1-\rho)(1-\rho(1-\phi))(1-(1-\rho(1-\phi))) + \rho\phi(1-\rho) \geq 0.$$

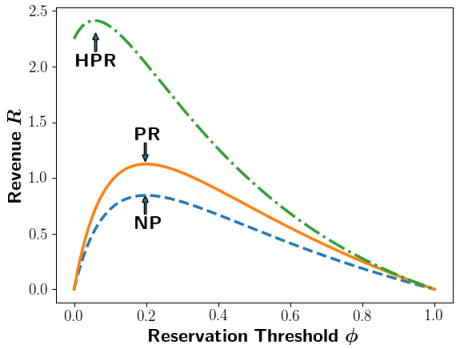


Figure 4: Plots of the revenue functions $R(\phi)$ corresponding to each preemption policy for $\rho = 0.75$. The highest revenue is obtained under the HPR policy. As expected, the NP and PR policies achieve their highest revenues at the same threshold point.

As $\rho \in (0,1)$ and $\phi \in [0,1]$, all terms on the left hand side are non-negative. Thus the inequality holds, and $C_{HPR}(\phi) \geq C_{PR}(\phi)$. □

Definition 3. *The maximum possible revenues of the NP, PR, and HPR policies are denoted by R^*_{NP}, R^*_{PR}, and R^*_{HPR}, respectively.*

Corollary 1. *Applying Lemma 2 to the definition of the revenue function given in Equation* (11)*, we deduce that $R^*_{HPR} \geq R^*_{PR} \geq R^*_{NP}$ for fixed λ and μ.*

Example 3. In Figure 4 we plot the revenue functions under each policy with $\rho = 0.75$. In particular, the maximum revenue obtainable under HPR is 2.414, with an associated threshold of $\phi = 0.057$. The PR and NP policies result in smaller maximum revenues of 1.125 and 0.844 respectively, at a threshold of $\phi = 0.2$. Thus, the HPR policy leads to the greatest maximum revenue. Further, the maximum revenue under HPR corresponds to a smaller threshold value of ϕ than under PR or NP, and thus a larger fraction of customers will choose to make a reservation under this policy.

In addition, referring back to Section 4.3, under the HPR policy an $all-reserve$ equilibrium is possible, and in fact for low loads the highest cost customers are willing to pay occurs at the threshold $\phi = 0$ corresponding to the $all-reserve$ equilibrium. However, as seen in Figure 4, it is not always the case that the maximum revenue under HPR is obtained at $\phi = 0$. We will see below in Theorem 4 that for low loads R^*_{HPR} is equal to $R_{HPR}(0)$ and the $all-reserve$ is the provider's optimal equilibrium state. However, for sufficiently high loads there is a $some-reserve$ equilibrium that leads to the maximum revenue, as the higher fee C outweighs the decrease in the fraction of customers willing to pay that cost.

We make an assumption that the provider is able to steer customers toward the equilibrium achieving maximum revenue. While we proceed with this assumption for the remainder of this section, it may be the case that the cost (fee)

leading to the maximum revenue has multiple equilibria associated with it, resulting in the possibility of the system converging on an equilibrium state which does not result in the maximum possible revenue to the provider. We show in Appendix E that Corollary 1 is extendable to the case where a provider operates under the constraint that the reservation cost must be associated to a unique $some-reserve$ or $all-reserve$ equilibrium.

Next, using Equations (7), (9) , (10), and (11) we explicitly define $R(\phi)$ under each policy:

$$\begin{aligned} R_{NP}(\phi) &= \frac{\rho^3\phi(1-\phi)}{2(1-\rho)(1-\rho(1-\phi))^2}; \\ R_{PR}(\phi) &= \frac{\rho^2\phi(1-\phi)}{2(1-\rho)(1-\rho(1-\phi))^2}; \\ R_{HPR}(\phi) &= \frac{2\rho(1-\rho)(1-\phi)(1-\rho(1-\phi))+\rho^2\phi(1-\phi)}{2(1-\rho)(1-\rho(1-\phi))^2} \\ &\quad -\rho(1-\phi). \end{aligned} \tag{12}$$

The next theorem explicitly characterizes the maximum revenues for the NP and PR policies. For the HPR policy, an expression is provided for sufficient low load (i.e., $\rho < 0.633$)).

Theorem 4. *The maximum revenue under each policy is equal to*

$$\begin{aligned} R^*_{NP} &= \frac{\rho^3}{8(1-\rho)^2}; \\ R^*_{PR} &= \frac{\rho^2}{8(1-\rho)^2}; \\ R^*_{HPR} &= \begin{cases} \frac{\rho^2}{1-\rho} & \text{for } \rho \in \left(0, \frac{3-\sqrt{3}}{2}\right]; \\ R_{HPR}(\phi^*) & \text{for } \rho \in \left(\frac{3-\sqrt{3}}{2}, 1\right), \text{ where } \phi^* \in \left[0, \frac{1-\rho}{2-\rho}\right). \end{cases} \end{aligned}$$

The proof involves taking the derivative of each revenue function. In doing so for $R_{NP}(\phi)$ and $R_{PR}(\phi)$, we find that both functions are unimodal with a unique maximum at $\phi^* = (1-\rho)/(2-\rho)$. As noted in Subsection 4.2, the cost functions have the same behavior as they differ by a factor of ρ, thus leading to the optimal revenues under each policy occurring at the same threshold value. This is a consequence of the incentives to purchase a reservation being largely the same under both policies - the priority offer only secures priority over customers who are making reservations at a later point in time, and does not include protection from preemption.

However, taking the derivative of $R_{HPR}(\phi)$ involves evaluating a cubic equation in ϕ to determine the increasing/decreasing behavior of the original function. In Appendix C we show that if $\rho \in (0, (3-\sqrt{3})/2]$, $R_{HPR}(\phi)$ is monotone decreasing with maximum at $\phi^* = 0$, and thus for such ρ the $all-$

reserve equilibria does indeed lead to the maximum possible revenue for the provider. For any other ρ, $R_{HPR}(\phi)$ is unimodal with a unique maximum $\phi^* \in [0, (1-\rho)/(2-\rho))$. A simplified closed form expression for R^*_{HPR} does not exist for such ρ as ϕ^* is the solution to a cubic equation, thus there are three possible forms ϕ^* could take. However, as seen in Appendix C, it is possible to guarantee that there must be a $some - reserve$ in the interval which leads to the maximum revenue.

In comparing the value of ϕ^* under HPR to PR or NP, we note that for loads smaller than $(3-\sqrt{3})/2$, $\phi^* = 0$ and so the optimum result is for all customers to make a reservation, something which is never the case under PR or NP. In the high load case, we establish in Appendix C that there is a strict upper bound on ϕ^* equal to $(1-\rho)/(2-\rho)$. As this is the threshold leading to optimal revenues under the PR and NP polices,we can conclude that ϕ^* will always be smaller under HPR than under PR or NP. As a smaller ϕ^* corresponds to a larger fraction of customers reserving, we conclude that when a provider is acting in a revenue maximizing fashion, a greater fraction of customers will do so as compared to PR or NP. This follows from the fact that under HPR, purchasing a reservation grants both priority and protection from preemption thus resulting in customers having greater incentive to purchase a reservation in general.

4.4.1. Comparison of Maximum Revenues

We next turn our attention to the question of determining by what factor the maximum revenue of the HPR policy is greater than the maximum revenues of the other policies. Figure 5 plots the ratios R^*_{PR}/R^*_{NP} and R^*_{HPR}/R^*_{PR} as a function of the load ρ. From this plot, we observe two results of interest: The first is that as ρ approaches 1, both ratios approach 1 implying that the maximum revenues are asymptotically equivalent under high system loads. The second is that under low loads, the ratio of maximum revenues in choosing PR over NP becomes infinitely large while the ratio when choosing HPR over PR has a finite limit of 8. The reminder of this section formalizes these observations.

Lemma 3. $\lim_{\rho\to 1} R^*_{PR}/R^*_{NP} = 1$ *and* $\lim_{\rho\to 1} R^*_{HPR}/R^*_{PR} = 1$.

Proof. To show this for R^*_{PR}/R^*_{NP} is straightforward due to the explicit relationship $(1/\rho)R^*_{NP} = R^*_{PR}$. Thus $R^*_{NP}/R^*_{PR} = 1/\rho$, and as ρ approaches 1 the ratio $1/\rho$ also approaches 1.

To show this for R^*_{HPR}/R^*_{PR} is rendered complicated by the fact that as ρ approaches 1, there is no closed form expression for R^*_{HPR}. We know that R^*_{HPR} occurs at some $\phi^* \in [0, (1-\rho)/(2-\rho))$. Thus, we consider an upper bound on R^*_{HPR}. To do so, we consider each ϕ term of $R_{HPR}(\phi)$ individually, and select the value of ϕ in the interval $[0, (1-\rho)/(2-\rho))$ which maximizes that term. If ϕ is negative, then letting $\phi = 0$ maximizes the term. If ϕ is positive, then $\phi = (1-\rho)/(2-\rho)$ maximizes the term. This results in the quantity:

$$R_{UB} \triangleq \frac{\rho(8 - 12\rho + 8\rho^2 - 3\rho^3)}{8(2-\rho)(1-\rho)^2}.$$

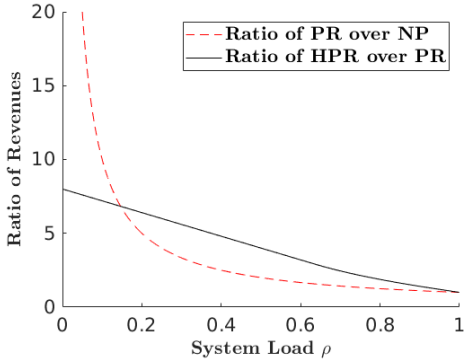


Figure 5: A plot of the ratios of the maximum revenues of PR and NP, and HPR and PR. As the system load ρ approaches 1, both ratios approach 1, meaning that the maximum revenues under all policies are asymptotically equivalent. Meanwhile, as ρ approaches zero, we find that the ratio of the maximum revenues of PR and NP is unbounded while the ratio of the maximum revenues of HPR and PR is bounded by 8.

To confirm this is indeed an upper bound, we evaluate $R_{UB} - R_{HPR}$. Showing that this expression is non negative for ρ approaching 1 and $\phi \in [0, (1-\rho)/(2-\rho))$ ultimately reduces to showing that the denominator is always positive and that the numerator reduces to ϕ^2 as ρ approaches 1. Therefore, $R_{UB} - R_{HPR} \geq 0$ follows.

Thus, we have $R^*_{PR} \leq R^*_{HPR} \leq R_{UB}$, which in turn implies that $1 \leq R^*_{HPR}/R^*_{PR} \leq R_{UB}/R^*_{PR}$. Calculating R_{UB}/R^*_{PR} yields

$$\frac{8 - 12\rho + 8\rho^2 - 3\rho^3}{\rho(2-\rho)}.$$

Therefore as ρ approaches 1, R_{UB}/R^*_{PR} approaches 1 and therefore by the Squeeze Theorem so must R^*_{HPR}/R^*_{PR}. □

To establish intuition for the behavior of R^*_{PR}/R^*_{NP} under high loads, we note that the maximum revenues under both policies occur at the same threshold $\phi^* = (1-\rho)/(2-\rho)$. As ρ approaches 1, this results in a threshold of $\phi^* = 0$, meaning that all customers will purchase a reservation. This follows as a consequence of the system load definition - as the arrival rate of customers approaches the service rate, wait times for the lowest priority customers approach infinity. As a result, customers have greater incentive to purchase priority to shorten waiting time. In addition, we have the explicit relationship between the corresponding cost functions $C_{PR}(\phi) = (1/\rho)C_{NP}(\phi)$. As ρ approaches 1, this results in $C_{PR}(\phi) = C_{NP}(\phi)$, and the revenues converge due all customers being willing to pay for priority reservation under both policies, at the same reservation cost.

Similarly, in examining the behavior of R^*_{HPR}/R^*_{PR} under high loads we note that the the threshold leading to R^*_{HPR} must lie in the interval $\phi^* \in$

$[0, (1-\rho)/(2-\rho))$. This interval shrinks to a single point 0 as ρ approaches 1. This results in the maximum revenue under HPR occurring at a threshold where all customers purchase reservations. It follows from the definition of the revenue functions that customers must also be willing to pay the same price for a reservation under either policy as ρ approaches 1. This occurs for the same reasons as noted above - customers have increased willingness to purchase the priority reservation to avoid infinitely long wait times associated with the lowest priority as the system load ρ approaches the stability threshold of 1.

We now consider the ratios of the maximum revenues under low system loads.

Lemma 4. $\lim_{\rho\to 0} R^*_{PR}/R^*_{NP} = \infty$ *and* $\lim_{\rho\to 0} R^*_{HPR}/R^*_{PR} = 8$.

Proof. The ratio R^*_{PR}/R^*_{NP} is equal to $1/\rho$ as established in the proof of Lemma 3. Thus as ρ approaches 0, the ratio grows in an unbounded fashion.

To show the result for R^*_{HPR}/R^*_{PR}, we note that for low loads ρ an explicit expression for R^*_{HPR} does exist by Theorem 4 and we can compute the ratio as equal to $8(1-\rho)$. Thus, the ratio approaches 8 as ρ approaches 0. □

In considering this behavior under low loads, we consider that as ρ approaches 0, it is unlikely that a newly arriving customer will observe a customer already in service. In addition, it is vanishingly unlikely that a newly arriving customer will observe any customers in the queue. Thus, under the NP policy there is no incentive to purchase priority as there is almost zero probability of arriving to find a lower priority customer in the queue, and there is no ability to preempt a lower priority customer already in service. In contrast, because a higher priority customer is able to preempt a lower priority customer under the PR policy, there exists a larger incentive to purchase priority as insurance against waiting for a customer in service. As a result, the ratio of the maximum revenues under PR to NP will approach infinity as the preemption mechanic determines whether any incentive to reserve priority exists at all under low loads.

In contrast, we note that as ρ goes to 0 R^*_{HPR} is at most 8 times greater than R^*_{PR}. The intuition for this result comes from the fact that both policies implement preemption mechanics. Therefore, the same incentive exists under both policies to purchase priority as an insurance against having to wait for no priority jobs to complete service. Under HPR however, customers have the additional incentive to reserve to prevent preemption by a higher priority customer which does not exist under PR. Thus, customers under HPR have slightly higher incentive to purchase priority reservations, but this incentive is mitigated by the low number of customers entering the system.

The incentives to purchase a reservation are ultimately similar enough under HPR and PR to result in the maximum revenues being bounded by a constant. Indeed, we note that the threshold leading to the optimal revenue is 0 in the HPR policy (i.e., all customers purchase reservation), and the threshold leading to optimal revenue in the PR policy approaches 1/2 for ρ near 0. Thus for low loads, we find that twice as many customers are willing to purchase a reservation

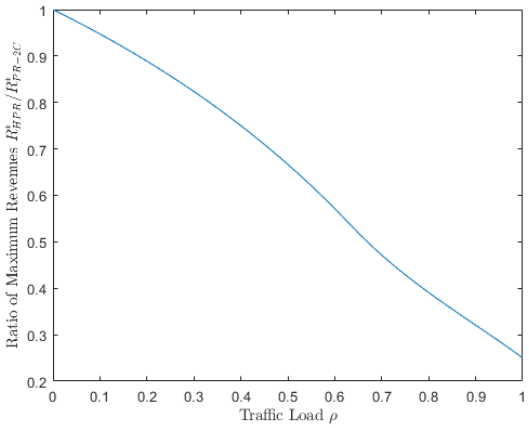


Figure 6: Plot of the ratio of the maximum revenue under HPR to the maximum revenue under a two-class Preemptive Resume (PR-2C) as ρ varies in $(0, 1)$. We find that as the traffic load increases (and thus, customers arrive at a greater rate), implementing an AR based queue does result in revenue shortfalls compared to a system with reservation. However, while this system does result in lower potential revenues, there are other considerations for the use of AR (such as planning and provisioning of service resources).

under HPR than under PR when the provider charges a revenue maximizing amount. However, customers are only willing to pay 4 times more under the HPR policy than what they would be charged under PR, in contrast to being willing to pay some arbitrarily larger amount under PR compared to NP.

Regardless, a provider is always best off under the HPR policy no matter the value of the system load ρ. However, the factor by which a provider is best off under HPR compared to the other policies varies dramatically as ρ varies, due to the arrival rate of customers impacting the incentives to purchase priority under each policy.

4.5. Comparison to non-AR System

Having established that under our model, HPR is the optimal policy from a provider perspective, we are interested in how our Advance Reservation model compares to other setups. In particular, we consider a comparison to a two-class model without reservations wherein customers have the option to purchase priority access, as in (Chamberlain & Starobinski, 2020). Specifically, this is a Preemptive Resume queue wherein customers are given the option of paying to join the premium class on entry; otherwise, customers will remain in the ordinary class. Assuming strategic customer behavior, the resulting maximum revenue under this model is (Chamberlain & Starobinski, 2020)

$$\mathcal{R}^*_{PR-2C} = \frac{\rho^2(2-\rho)}{2(1-\rho)^2}. \tag{13}$$

Computing the ratios of R^*_{HPR} and R^*_{PR-2C}, we find the following relations hold:

$$\begin{aligned} \frac{R^*_{HPR}}{R^*_{PR-2C}} &= \frac{2(1-\rho)}{2-\rho}, \quad \text{if } \rho \leq \frac{3-\sqrt{3}}{2}; \\ \frac{2(1-\rho)}{2-\rho} \leq \frac{R^*_{HPR}}{R^*_{PR-2C}} &\leq \frac{6-5\rho}{4(2-\rho)}, \quad \text{if } \rho > \frac{3-\sqrt{3}}{2}. \end{aligned} \tag{14}$$

As shown in Figure 6 we find that as ρ increases towards 1 the ratio decreases significantly; using exact solutions for R^*_{HPR} for heavy traffic loads, we find that the ratio decreases to approximately 0.25 for ρ near 1. Thus from a purely revenue generating perspective the AR model is not an optimal implementation, particularly under heavy traffic situations. However, there are additional trade offs between the models that result in the AR model still worth considering. In particular, as we are considering deterministic service times, it is possible to provide actual start and end times to customers upon making reservations under PR, and bounded estimates in advance under the other policies, as explained at the beginning of this section. This can provide multiple benefits to both customer and provider. Indeed, customer sensitivity to deadlines is cited as a reason to offer AR in (Simhon & Starobinski, 2017). Thus, customers have incentive to make reservations not only to reduce total wait time in the system, but to also have greater certainty that their job will be completed within a particular time frame.

The converse is that under the two class model, the provider is ultimately able to charge a larger premium class entry fee as compared to the AR model, particularly under higher traffic loads. Yet, advance knowledge of the system state enables providers to manage quality control and improve planning and provisioning (Charbonneau & Vokkarane, 2011). This is particularly relevant to systems where the server is a renewable resource, which a computing cluster falls under (Chen et al., 2017b). Thus, AR can become a means to avoid customers from switching to competitors, via offering greater certainty in terms of job completion times, and thus increasing revenues via retaining and growing the customer base.

5. Extension to General Service

In this section we outline how the analysis of the previous section extends to general service distributions. The system model is the same as before, except that we do need to make computations with respect to general second moments of service. We will show that the provider is still best off choosing the HPR policy regardless of the distribution of service in effect.

The model is as before, except that there is now an additional parameter to consider as service is no longer assumed to be deterministic. Thus, the second moment of service must be taken into account. We introduce a new parameter K, such that the second moment of service $E[X^2] = K/\mu^2$. Since by definition $E[X] = 1/\mu$, we must have $K \geq 1$, as otherwise the variance would

be negative. We note that the case $K = 1$ corresponds to the $M|D|1$ system analyzed in the previous section, and $K = 2$ correspond to an $M|M|1$ system. Having established this parameter, we derive the cost functions for each policy as before.

5.1. Cost Functions

In order to re-derive the cost functions, we use the same wait time formulas from (Conway et al., 1967, Ch.8) as we did in the previous section. The difference is that now the second moment is defined as K/μ^2 rather than using a fixed service distribution with second moment $1/\mu^2$. For a customer with priority p and action σ in the NP queue, the wait time formula updates to

$$W(\sigma, p) = \frac{K\rho}{2\mu(1-\rho_a-\rho_p)(1-\rho_a)}.$$

Similarly, the formula for wait time for a customer with priority p and action σ in the PR queue updates to

$$W(\sigma, p) = \frac{1}{\mu(1-\rho_a)} + \frac{K(\rho_a+\rho_p)}{2\mu(1-\rho_a)(1-\rho_a-\rho_p)} - \frac{1}{\mu},$$

Using these as a basis, the cost functions under the respective policies in the $M|G|1$ system are as follows:

$$\begin{aligned} C_{NP}(\phi) &= \frac{K\rho^2\phi}{2\mu(1-\rho)(1-\rho(1-\phi))^2}; \\ C_{PR}(\phi) &= \frac{K\rho\phi}{2\mu(1-\rho)(1-\rho(1-\phi))^2}; \\ C_{HPR}(\phi) &= \frac{1}{\mu}\left(\frac{2(1-\rho)(1-\rho(1-\phi)) + K\rho\phi}{2(1-\rho)(1-\rho(1-\phi))^2} - 1\right). \end{aligned} \tag{15}$$

Computing derivatives we find that $C_{NP}(\phi)$ and $C_{PR}(\phi)$ have the same behavior as under $M|D|1$, regardless of the value of K. This is because K is a multiplicative constant with respect to ϕ, and so does not affect the regions where the derivative is positive or negative.

Conversely, K does impact the behavior of $C_{HPR}(\phi)$, and it is possible to show that depending on the values of K and ρ, $C_{HPR}(\phi)$ can be monotone increasing, monotone decreasing, or unimodal. Figure 7 shows an example of how behavior changes for fixed ρ as K increases. Thus, the possible equilibrium regions that can arise have multiple sub cases to consider to take into account how K impacts the function behavior. In Appendix B, we extend Theorem 3 to the $M|G|1$ case to show how the value of K impacts the possible equilibria.

Despite this additional complexity, we can show that for fixed K, the result from Lemma 2 extends to the $M|G|1$ case. The proof of this result is contained in Appendix D, and it follows the same logic as the proof under the $M|D|1$ case.

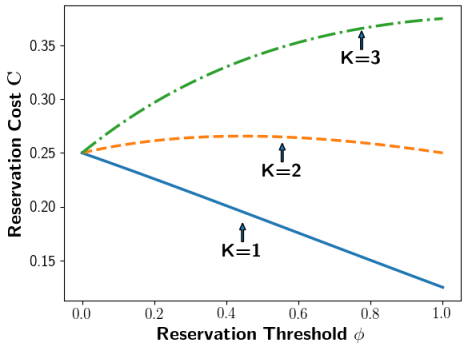


Figure 7: Plots of $C_{HPR}(\phi)$ for $\rho = 0.2$, $\mu = 1$, and $K \in \{1, 2, 3\}$. We find that as K increases, the cost function changes from monotone decreasing behavior, to unimodal behavior, to monotone increasing behavior. Thus, under the HPR policy, the second moment of service must be taken into account when deriving the equilibrium outcomes.

5.2. Comparison of Revenues

Using Equation (15) for the cost functions, the revenue functions for each policy under general service distribution are

$$\begin{aligned} R_{NP}(\phi) &= \frac{K\rho^3\phi(1-\phi)}{2(1-\rho)(1-\rho(1-\phi))^2}; \\ R_{PR}(\phi) &= \frac{K\rho^2\phi(1-\phi)}{2(1-\rho)(1-\rho(1-\phi))^2}; \\ R_{HPR}(\phi) &= \frac{2\rho(1-\rho)(1-\phi)(1-\rho(1-\phi)) + K\rho^2\phi(1-\phi)}{2(1-\rho)(1-\rho(1-\phi))^2} \\ &\quad - \rho(1-\phi). \end{aligned}$$

Corollary 2. *Because Lemma 2 still holds under the $M|G|1$ case and the revenue functions are continuous in $\phi \in [0, 1]$, it follows that $R^*_{HPR} \geq R^*_{PR} \geq R^*_{NP}$, for fixed λ, μ, and K.*

Because $(1/\rho)R_{NP}(\phi) = R_{PR}(\phi)$, the ratio of the maximum revenues under PR to that under NP follows the same behavior in the $M|G|1$ queue as in the $M|D|1$ queue. This relationship holds regardless of the value of K.

Meanwhile, Figure 8 shows that the ratio of the maximum revenues of HPR to PR depends on both ρ and K. The ratio converges to 1 as ρ approaches 1 as expected from the $M|D|1$ result. However, the ratio also converges to 1 as K increases, regardless of the value of ρ. Inspecting $R_{HPR}(\phi)$, we note that there is a single term K present:

$$\frac{K\rho^2\phi(1-\phi)}{2(1-\rho)(1-\rho(1-\phi))^2}.$$

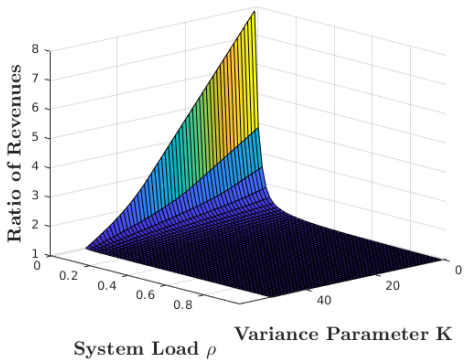


Figure 8: Surface plot of the ratio of maximum revenues under HPR to PR, for $K \in [1, 50]$. We find that while the ratio converges to 1 under high loads as expected from the $M|D|1$ analysis, the ratio also converges to 1 for high values of K regardless of load.

This term happens to be equal to $R_{PR}(\phi)$. As K is unbounded from above while $\rho \in (0, 1)$ and $\phi \in [0, 1]$, this term dominates as K increases. As a result, for large K, $R_{HPR}(\phi)$ and $R_{PR}(\phi)$ differ by a comparatively small constant, and therefore the maximum revenues will converge on each other as K increases for any system load ρ. These numerical results indicate that the upper bound of 8 on the ratio between the maximum revenues of the HPR and PR policies holds across all system loads and service distributions.

6. Conclusions

In this work, we evaluated the impact of implementing different preemption policies on the equilibrium and revenue outcomes in a reservation system. In doing so, we developed a game model where customers choose between making a reservation that grants priority over customers making a reservation later and those accepting the lowest possible priority. From the model, we proved that the equilibrium strategy is a threshold strategy, parameterized as a variable $\phi \in [0, 1]$. This enables derivation of closed form expressions for the costs as a function of ϕ leading to one or more equilibria under each policy.

We determined that in applying the model to an $M|D|1$ queue, the system load ρ determines the equilibrium regions (in terms of the cost C) that prevail under each policy. Under this model, the NP and PR policies feature similar equilibrium structures and in both policies there are no equilibria where all customers will elect to make a reservation. However, such equilibria do exist if the HPR policy is in effect. Indeed, we found that for sufficiently low loads, the greatest costs customers are willing to bear coincides with the threshold $\phi = 0$ which corresponds to the $all - reserve$ equilibrium. This incentive is driven by the fact that under HPR making a reservation yields both higher priority and protection from preemption.

In comparing the relative differences in the maximum revenues under each policy, we found that as ρ tends to 1, the revenues under all three policies coincide. However, as ρ tends to 0, the relative difference between the maximum revenues of PR and NP becomes arbitrarily large, but the relative differences of the maximum revenues between HPR and PR approaches 8. In this regime, while twice as many customers make a reservation in HPR, they are only willing to pay four times as much than in PR.

We addressed the comparison of the reservation model to a two-class priority system without reservation. While a system without reservation may yield larger revenue under the revenue model considered in this paper, the provider may choose to trade this off for greater control over the system given the ability to pre-plan utilization, and issue service level guarantees which may provide incentive for customers to avoid switching to competitors.

We outlined how these results extend to the $M|G|1$ case. We showed that the second moment of service impacts the equilibria structure of the HPR policy, but not that of NP or PR. In addition, we showed that among the three policies, HPR *always* achieves the greatest possible revenues for any given combination of system load and second moment of service. Based on our analytical and numerical results, we conjecture that $R^*_{HPR}/R^*_{PR} \leq 8$ for any service distribution and any load ρ. In fact, the numerical results indicate that the advantage to the provider in choosing HPR over PR decreases as the variance in service increases.

A formal proof of the above conjecture as well as the analysis of other possible preemption policies represent interesting directions for future work. Other avenues for future work include considering the situation where customers vary as to their willingness to make a reservation as opposed to considering homogeneous customers. Alternatively, one may consider a provider that varies the service rate, given the ability to pre-plan available service capacity on the basis of the known information from the customer reservation requests. This information for example would enable the provider to perform maintenance during anticipated low traffic periods, which in turn impacts the service rate and thus the cost a customer is willing to pay. In addition, one could consider the impact of the provider revealing the time quotation prior to the reservation being made on the customers' willingness to make a reservation. Such a mechanism essentially changes the complexion of the queue and thus would require an in-depth analysis similar to the one provided for our base model.

Acknowledgment

This research was supported in part by the NSF under grant CNS-1717858 and CNS-1908087.

References

Afèche, P., & Mendelson, H. (2004). Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Science*, *50*, 869–882.

Afeche, P., & Pavlin, J. M. (2016). Optimal price/lead-time menus for queues with customer choice: Segmentation, pooling, and strategic delay. *Management Science*, *62*, 2412–2436.

Azar, Y., Kalp-Shaltiel, I., Lucier, B., Menache, I., Naor, J. S., & Yaniv, J. (2015). Truthful online scheduling with commitments. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation* EC '15 (pp. 715–732). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/2764468.2764535.

Balachandran, K. R. (1972). Purchasing priorities in queues. *Management Science*, *18*, 319–326. URL: https://doi.org/10.1287/mnsc.18.5.319. arXiv:https://doi.org/10.1287/mnsc.18.5.319.

Chamberlain, J., & Starobinski, D. (2020). Strategic revenue management of preemptive versus non-preemptive queues. arXiv:2007.06764.

Charbonneau, N., & Vokkarane, V. M. (2011). A survey of advance reservation routing and wavelength assignment in wavelength-routed wdm networks. *IEEE Communications Surveys & Tutorials*, *14*, 1037–1064.

Chawla, S., Devanur, N., Kulkarni, J., & Niazadeh, R. (2017a). Truth and regret in online scheduling. In *Proceedings of the 2017 ACM Conference on Economics and Computation* EC '17 (pp. 423–440). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/3033274.3085119.

Chawla, S., Devanur, N. R., Holroyd, A. E., Karlin, A. R., Martin, J. B., & Sivan, B. (2017b). Stability of service under time-of-use pricing. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing* STOC 2017 (pp. 184–197). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/3055399.3055455.

Chen, G., Zhao, Y., Shen, X., & Zhou, H. (2017a). Effisha: A software framework for enabling efficient preemptive scheduling of GPU. *Proceedings of the ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPOPP*, (pp. 3–16).

Chen, Y., Levi, R., & Shi, C. (2017b). Revenue management of reusable resources with advanced reservations. *Production and Operations Management*, *26*, 836–859.

Conway, R., Maxwell, W., & Miller, L. (1967). *Theory of Scheduling*. Addison-Wesley.

Dierks, L., & Seuken, S. (2019). Cloud pricing: The spot market strikes back. *SSRN Electronic Journal*, . URL: https://ssrn.com/abstract=3383420.

Dodge, Y. (Ed.) (2003). *The Oxford Dictionary of Statistical Terms*. Oxford University Press.

Du, B., & Larsen, C. (2017). Reservation policies of advance orders in the presence of multiple demand classes. *European Journal of Operational Research*, *256*, 430 – 438. URL: http://www.sciencedirect.com/science/article/pii/S0377221716304489. doi:https://doi.org/10.1016/j.ejor.2016.06.028.

Edelson, N. M., & Hilderbrand, D. K. (1975). Congestion tolls for poisson queuing processes. *Econometrica: Journal of the Econometric Society*, (pp. 81–92).

Gavirneni, S., & Kulkarni, V. G. (2016). Self-selecting priority queues with burr distributed waiting costs. *Production and Operations Management*, *25*, 979–992.

Guéant, O., Lasry, J.-M., & Lions, P.-L. (2011). Mean field games and applications. In *Paris-Princeton lectures on mathematical finance 2010* (pp. 205–266). Springer.

Gurvich, I., Lariviere, M. A., & Ozkan, C. (2019). Coverage, coarseness, and classification: Determinants of social efficiency in priority queues. *Management Science*, *65*, 1061–1075.

Hassin, R. (1995). Decentralized regulation of a queue. *Management Science*, *41*, 163–173.

Hassin, R. (2016). *Rational Queueing*. CRC Press. URL: https://www.taylorfrancis.com/books/9780429153884.

Hassin, R., & Haviv, M. (1997). Equilibrium threshold strategies: The case of queues with priorities. *Operations Research*, *45*, 966–973. URL: https://doi.org/10.1287/opre.45.6.966. arXiv:https://doi.org/10.1287/opre.45.6.966.

Hassin, R., & Haviv, M. (2003). *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kulwer Academic.

Haviv, M. (2011). Strategic Customer Behavior in a Single Server Queue. In *Wiley Encyclopedia of Operations Research and Management Science*. American Cancer Society. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470400531.eorms0989.

Haviv, M., & Oz, B. (2018). Self-regulation of an unobservable queue. *Management Science*, *64*, 2380–2389.

Honnappa, H., & Jain, R. (2015). Strategic Arrivals into Queueing Networks: The Network Concert Queueing Game. *Operations Research*, *63*, 247–259. URL: http://pubsonline.informs.org/doi/10.1287/opre.2014.1338.

Izady, N. (2019). An integrated approach to demand and capacity planning in outpatient clinics. *European Journal of Operational Research*, *279*, 645 – 656. URL: http://www.sciencedirect.com/science/article/pii/S0377221 71930476X. doi:https://doi.org/10.1016/j.ejor.2019.06.001.

Jain, N., Menache, I., Naor, J. S., & Yaniv, J. (2014). A truthful mechanism for value-based scheduling in cloud computing. *Theory of Computing Systems*, *54*, 388–406. URL: https://doi.org/10.1007/s00224-013-9449-0.

Jain, R., Juneja, S., & Shimkin, N. (2011). The concert queueing game: to wait or to be late. *Discrete Event Dynamic Systems*, *21*, 103–138. URL: http://link.springer.com/10.1007/s10626-010-0097-0.

Javadi, B., Thulasiramy, R. K., & Buyya, R. (2011). Statistical modeling of spot instance prices in public cloud environments. In *2011 Fourth IEEE International Conference on Utility and Cloud Computing* (pp. 219–228).

Jin, L. (2016). Cook: A fair preemptive resource scheduler for compute clusters. URL: https://www.twosigma.com/insights/article/cook-a-fair-pree mptive-resource-scheduler-for-compute-clusters/.

Juneja, S., & Jain, R. (2009). The concert/cafeteria queueing problem: a game of arrivals. In *Proceedings of the 4th International ICST Conference on Performance Evaluation Methodologies and Tools* (pp. 59:1–59:6). ICST. URL: http://eudl.eu/doi/10.4108/ICST.VALUETOOLS2009.7624.

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

Li, X., Gu, B., Zhang, C., Yamori, K., & Tanaka, Y. (2014). Price competition in a duopoly iaas cloud market. In *The 16th Asia-Pacific Network Operations and Management Symposium* (pp. 1–4).

Liao, C., Klausner, Y., Starobinski, D., Simhon, E., & Bestavros, A. (2018). A case study of a shared/buy-in computing ecosystem. *Cluster Computing*, *21*, 1595–1606.

Lucier, B., Menache, I., Naor, J. S., & Yaniv, J. (2013). Efficient online scheduling for deadline-sensitive jobs: Extended abstract. In *Proceedings of the Twenty-fifth Annual ACM Symposium on Parallelism in Algorithms and Architectures* SPAA '13 (pp. 305–314). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/2486159.2486187.

Lui, F. T. (1985). An equilibrium queuing model of bribery. *The Journal of political economy*, *93*, 760–781.

Master, N., Zhou, Z., & Bambos, N. (2017). An infinite dimensional model for a single server priority queue. In *2017 American Control Conference (ACC)* (pp. 1753–1758).

Menache, I., Ozdaglar, A., & Shimkin, N. (2011). Socially optimal pricing of cloud computing resources. In *Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools* VALUETOOLS '11 (pp. 322–331).

Mendelson, H., & Whang, S. (1990). Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations Research*, *38*, 870–883. URL: https://doi.org/10.1287/opre.38.5.870. arXiv:https://doi.org/10.1287/opre.38.5.870.

Microsoft (2018). Use low-priority VMs with batch. URL: https://docs.microsoft.com/en-us/azure/batch/batch-low-pri-vms.

Microsoft (2019). Azure reserved VM instances. URL: https://azure.microsoft.com/en-us/pricing/reserved-vm-instances/.

Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica*, *37*, 15–24. URL: http://www.jstor.org/stable/1909200.

Niu, D., Feng, C., & Li, B. (2012). Pricing cloud bandwidth reservations under demand uncertainty. *SIGMETRICS Perform. Eval. Rev.*, *40*, 151–162.

Oh, J., & Su, X. (2018). Reservation policies in queues: Advance deposits, spot prices, and capacity allocation. *Production and Operations Management*, *27*, 680–695.

Reiman, M. I., & Wang, Q. (2008). An asymptotically optimal policy for a quantity-based network revenue management problem. *Mathematics of Operations Research*, *33*, 257–282.

Simhon, E., Cramer, C., Lister, Z., & Starobinski, D. (2015). Pricing in dynamic advance reservation games. In *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 546–551).

Simhon, E., & Starobinski, D. (2014). Game-theoretic analysis of advance reservation services. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)* (pp. 1–6).

Simhon, E., & Starobinski, D. (2017). Advance reservation games. *ACM Trans. Model. Perform. Eval. Comput. Syst.*, *2*, 10:1–10:21.

Simhon, E., & Starobinski, D. (2018). On the impact of information disclosure on advance reservations: A game-theoretic view. *European Journal of Operational Research*, *267*, 1075–1088.

Smith, W., Foster, I., & Taylor, V. (2000). Scheduling with advanced reservations. In *Proceedings 14th International Parallel and Distributed Processing Symposium. IPDPS 2000* (pp. 127–132).

Song, J., & Guerin, R. (2017). Pricing and bidding strategies for cloud computing spot instances. In *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 647–653).

Syed, A. A., Ye, W., & Heidemann, J. (2008). T-lohi: A new class of mac protocols for underwater acoustic sensor networks. In *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE* (pp. 231–235). IEEE.

Table, O. (2020). About us. URL: `https://www.opentable.com/about/`.

Talak, R., Manjunath, D., & Proutiere, A. (2019). Strategic arrivals to queues offering priority service. *Queueing Systems*, *92*, 103–130.

Virtamo, J. T. (1992). A model of reservation systems. *IEEE Transactions on Communications*, *40*, 109–118.

Virtamo, J. T., & Aalto, S. (1991). Stochastic optimization of reservation systems. *European journal of operational research*, *51*, 327–337.

Wang, C., Ma, W., Qin, T., Chen, X., Hu, X., & Liu, T.-Y. (2015). Selling reserved instances in cloud computing. *International Joint Conference on Artificial Intelligence*, . URL: `https://www.aaai.org/ocs/index.php/IJCAI/IJCAI15/paper/view/11337/10695`.

Wang, J., Cui, S., & Wang, Z. (2019). Equilibrium strategies in M/M/1 priority queues with balking. *Production and operations management*, *28*, 43–62.

Wang, W., Niu, D., Li, B., & Liang, B. (2013). Dynamic cloud resource reservation via cloud brokerage. In *2013 IEEE 33rd International Conference on Distributed Computing Systems* (pp. 400–409).

Yang, L., & Debo, L. (2019). Referral priority program: Leveraging social ties via operational incentives. *Management Science*, *65*, 2231–2248.

Yang, L., Debo, L., & Gupta, V. (2017). Trading time in a congested environment. *Management Science*, *63*, 2377–2395.

Yessad, S., Nait-Abdesselam, F., Taleb, T., & Bensaou, B. (2007). R-mac: Reservation medium access control protocol for wireless sensor networks. In *Local Computer Networks, 2007. LCN 2007. 32nd IEEE Conference on* (pp. 719–724). IEEE.

Appendix A. Proof of Lemma 1

We next provide the proof of Lemma 1. Recall that Lemma 1 states that *Under the game model, all customers follow a threshold strategy.*

Proof. The proof of the Lemma adapts the technique of Lemma 4.4 in (Simhon & Starobinski, 2017). Consider a customer making the decision of which action to take at $\hat{p} \in [0, 1]$. Customers in our model are statistically identical and therefore have the same benefit from service. Thus, the total utility for a customer depends on the cost of waiting. For a RI customer, this cost is $W(RI, \hat{p}) + C$. For a NRI customer the cost is $W(NRI, 0)$, as choosing NRI results in automatically being assigned priority 0. As the expected wait time for the NRI customers does not depend on $\hat{p}$, we can consider $W(NRI, 0)$ to be constant with respect to the potential priority. As a result, we call this quantity T.

For the RI customers, wait time is dependent on $\hat{p}$. With 1 defined as the highest priority, $W(RI, \hat{p})$ is a non-increasing function with respect to $\hat{p}$. Therefore, because C is a constant parameter, $W(RI, \hat{p}) + C$ is also a non-increasing function.

As a result, the quantities $W(RI, \hat{p}) + C$ and T intersect at most once. If they do intersect, then they do so either at a single point ϕ or along some interval $\hat{p} \in [\phi_1, \phi_2]$. If intersecting along a single point, then $W(RI, \hat{p}) + C < T$ for $\hat{p} > \phi$ and $W(RI, \hat{p}) + C > T$ for $\hat{p} < \phi$. So, for $\hat{p} > \phi$ the customer has incentive to make a reservation and will do so. For $\hat{p} \leq \phi$ the customer will not make a reservation as it lacks incentive. Thus, ϕ is the threshold.

If the quantities intersect along an interval, then $W(RI, \hat{p}) + C < T$ for $\hat{p} > \phi_2$, and $W(RI, \hat{p}) + C > T$ for $\hat{p} < \phi_1$. Thus, if $\hat{p} > \phi_2$ the customer makes a reservation, and if $\hat{p} < \phi_1$ the customer does not due to the incentives involved. For $\hat{p} \in [\phi_1, \phi_2]$, $W(RI, \hat{p}) + C = T$, thus the customer is indifferent between the two options. Because there is no incentive to make a reservation, the customer does not do so. Therefore, in this case the threshold is ϕ_2.

If $W(RI, \hat{p}) + C$ and T do not intersect at all, then either $W(RI, \hat{p}) + C < T$ or $W(RI, \hat{p}) + C > T$ for all $\hat{p}$. In the former case, all customers have incentive to make a reservation and so the resulting threshold is $\phi = 0$. In the latter case, no customers have incentive to make a reservation and thus the resulting threshold is $\phi = 1$. □

Appendix B. Extension of Theorem 3 to M/G/1

Similar to under $M|D|1$, in order to determine the behavior of each cost function and derive the resulting equilibria regions for each policy, we compute the derivative of the corresponding $C(\phi)$. As noted in section 5, K is a multiplicative constant in the $C_{NP}(\phi)$ and $C_{PR}(\phi)$ definitions. Thus, we simply factor K out in these cases, and take the derivative as under $M|D|1$.

However, in the definition of $C_{HPR}(\phi)$, K is not a multiplicative constant:

$$C_{HPR}(\phi) = \frac{1}{\mu}\left(\frac{2(1-\rho)(1-\rho(1-\phi)) + K\rho\phi}{2(1-\rho)(1-\rho(1-\phi))^2} - 1\right).$$

We assert that as a result, the value of K impacts how $C_{HPR}(\phi)$ behaves:

Theorem 5. *If $K \in [1,2)$ and $\rho \in (0,(2-K)/2)$, $C_{HPR}(\phi)$ is a monotone decreasing function with respect to ϕ; if $K > 2$ and $\rho \in (0,(K-2)/(2K-2))$, C_{HPR} is a monotone increasing function with respect to ϕ; otherwise, C_{HPR} is unimodal with a unique maximum at $\phi = (2\rho^2 + (K-4)\rho + (2-K))/(2\rho^2 - (K+2)\rho)$.*

Proof. As we are evaluating the increasing/decreasing behavior of a function, we prove this via determining the intervals over which the derivative is positive and negative. Computing the derivative with respect to ϕ results in the following:

$$\frac{dC_{HPR}}{d\phi} = \frac{1}{\mu}\left(\frac{K\rho(1-\rho(1+\phi)) - 2\rho(1-\rho)(1-\rho(1-\phi))}{2(1-\rho)(1-\rho(1-\phi))^3}\right).$$

As $\mu > 0$, the sign of the derivative will be determined by the sign of the expression in the parentheses. As $\rho \in (0,1)$ and $\phi \in [0,1]$ it follows that $(1-\rho)$ and $(1-\rho(1-\phi))^3$ will always be positive. Thus, the sign of the derivative depends on the sign of the numerator.

We note that $-2\rho(1-\rho)(1-\rho(1-\phi))$ will always be negative for the values of ρ and ϕ that we consider, however the sign of $K\rho(1-\rho(1+\phi))$ varies depending on the value of ρ and ϕ. Therefore, the sign of the numerator depends on the relative values of ρ, ϕ, and K.

To determine exact conditions for where the numerator is positive, we solve for where the numerator is positive in terms of ϕ:

$$\phi < \frac{2\rho^2 + (K-4)\rho + (2-K)}{2\rho^2 - (K+2)\rho}.$$

The inequality changes from true to false at $\phi = (2\rho^2 + (K-4)\rho + (2-K))/(2\rho^2 - (K+2)\rho)$. Denote this by ϕ^{opt}. Given that $\phi \in [0,1]$ and that the above condition is linear in ϕ, there are thus three possible cases to consider:

1. $\phi^{opt} < 0$, in which case the derivative is negative for all ϕ being positive and thus C_{HPR} is monotonically decreasing.
2. $0 < \phi^{opt} < 1$, in which case the derivative goes from being negative to positive at ϕ^{opt}, therefore C_{HPR} is unimodal with a unique maximum at ϕ^{opt}.
3. $\phi^{opt} > 1$, in which case the derivative is positive for all ϕ and thus C_{HPR} is monotonically increasing.

Solving each of these cases yields the following.

1. If $K \in [1,2)$ and $\rho \in (0,(2-K)/2)$, the derivative is negative for all $\phi \in [0,1]$ therefore $C_{HPR}(\phi)$ is monotone decreasing.

2. If $K > 2$ and $\rho \in (0, (K-2)/(2K-2))$, the derivative is positive for all $\phi \in [0,1]$ therefore $C_{HPR}(\phi)$ is monotone increasing.
3. Otherwise, the derivative changes from positive to negative at ϕ^{opt}, therefore $C_{HPR}(\phi)$ is unimodal with a unique maximum at ϕ^{opt}.

□

Thus, the behavior of $C_{HPR}(\phi)$ depends on both K and ρ. Indeed, we have three distinct cases to consider based on whether K is greater than, equal to, or less than 2.

Appendix B.1. Case I: Low Service Variance

Applying Equation (3) to the $K < 2$ case specifically results in the following:

$$C_0 = \frac{\rho}{\mu(1-\rho)};$$
$$C_1 = \frac{K\rho}{2\mu(1-\rho)};$$
$$\underline{C} = C_1 = \frac{\rho}{2\mu(1-\rho)};$$
$$\overline{C} = \begin{cases} C_0 & \text{for } \rho \in \left(0, \frac{2-K}{2}\right); \\ \frac{(4-8K)\rho^2+(12K-8)\rho+(K-2)^2}{8K\mu(1-\rho)^2} & \text{otherwise.} \end{cases}$$

And based on these values, we have the following theorem with respect to the equilibrium behavior:

Theorem 6. *Under the HPR policy with low service variance the equilibrium outcomes are as follows:*
Case I; $\rho < (2-K)/2$:

- *If $C < \underline{C}$, there is a unique $all-reserve$ equilibrium.*
- *If $\underline{C} < C < \overline{C}$, there are multiple equilibria: one each of $all-reserve$, $some-reserve$, and $none-reserve$.*
- *If $C > \overline{C}$, there is a unique $none-reserve$ equilibrium.*

Case II; $\rho > (2-K)/2$:

- *If $C < \underline{C}$, there is a unique $all-reserve$ equilibrium.*
- *If $\underline{C} < C < C_0$, there are multiple equilibria: one each of $all-reserve$, $some-reserve$, and $none-reserve$.*
- *If $C_0 < C < \overline{C}$, there are multiple equilibria: two $some-reserve$ and one $none-reserve$*
- *If $C > \overline{C}$, there is a unique $none-reserve$ equilibrium.*

Letting $K = 1$ reduces this to the deterministic service case examined in Section 4.3, as expected.

Appendix B.2. Case II: Exponentially Distributed Service

Applying the definitions from Equation (3) to $K = 2$ (which also corresponds to the exponentially distributed service case) results in the following:

$$\begin{aligned} C_0 &= \frac{\rho}{\mu(1-\rho)}; \\ C_1 &= \frac{\rho}{\mu(1-\rho)}; \\ \underline{C} &= C_0 = C_1; \\ \overline{C} &= \frac{-3\rho^2 + 4\rho}{4\mu(1-\rho)^2}. \end{aligned}$$

Thus, we see that only unimodal behavior is possible, since the values of C_1 and C_0 are identical, and are also the minimum value of $C_{HPR}(\phi)$ in this case for any ρ. As a result, we have the following theorem:

Theorem 7. *Under the HPR policy with exponentially distributed service, the equilibrium outcomes are as follows:*

- *If $C < \underline{C}$, there is a unique $all - reserve$ equilibrium.*
- *If $\underline{C} < C < \overline{C}$, there are multiple equilibria: two $some - reserve$ and one $none - reserve$.*
- *If $C > \overline{C}$, there is a unique $none - reserve$ equilibrium.*

Appendix B.3. Case III: High Service Variance

Applying the definitions in Equation (3) with $K > 2$ results in the following:

$$\begin{aligned} C_0 &= \frac{\rho}{\mu(1-\rho)}; \\ C_1 &= \frac{K\rho}{2\mu(1-\rho)}; \\ \underline{C} &= C_0; \\ \overline{C} &= \begin{cases} C_1 & \text{for} \quad K > 2, \ \rho \in \left(0, \frac{K-2}{2K-2}\right); \\ \frac{(4-8K)\rho^2 + (12K-8)\rho + (K-2)^2}{8K\mu(1-\rho)^2} & \text{otherwise.} \end{cases} \end{aligned}$$

And the following theorem results from this definition:

Theorem 8. *Under the HPR policy with high service variance, the equilibrium outcomes are as follows:*
Case I; $\rho < (K-2)/(2K-2)$:

- *If If $C < \underline{C}$, there is a unique $all - reserve$ equilibrium.*

- *If* $\underline{C} < C < \overline{C}$*, there is a unique* $some - reserve$ *equilibrium.*
- *If* $C > \overline{C}$*, there is a unique* $none - reserve$ *equilibrium.*

Case II; $\rho > (K-2)/(2K-2)$*:*

- *If If* $C < \underline{C}$*, there is a unique* $all - reserve$ *equilibrium.*
- *If* $\underline{C} < C < C_1$*, there is a unique* $some - reserve$ *equilibrium.*
- *If* $C_1 < C < \overline{C}$*, there are multiple equilibria: two* $some - reserve$ *and one* $none - reserve$*.*
- *If* $C > \overline{C}$*, there is a unique* $none - reserve$ *equilibrium.*

Comparing this to Theorems 1 and 2, under high service variance the equilibrium structure under the HPR system is most similar to that of the NP and PR systems. However, the HPR has a region of costs where an $all - reserve$ equilibrium is possible, unlike NP and PR.

Appendix C. Evaluating the behavior of the revenue function under HPR

We derive an explicit expression for R^*_{HPR} is in certain cases, as well as characterize the behavior of R_{HPR} in general. We do so here under general distribution. The case $K = 1$ yields the result for the $M|D|1$ system.

Lemma 5. *If* $K \in [1, 4)$ *and* $\rho \in [0, (3 - \sqrt{2K+1})/2]$*,* $R_{HPR}(\phi)$ *is monotone decreasing. Otherwise,* $R_{HPR}(\phi)$ *is unimodal with a unique maximum at some* $\phi \in [0, (1-\rho)/(2-\rho))$ *satisfying the following:*

$$K = \frac{2(1-\phi)(1-\rho)(1-\rho(1-\phi))(2-\rho(1-\phi))}{1-2\phi-\rho(1-\phi)}.$$

Proof. As we are asserting increasing/decreasing behavior of the revenue function, we compute the derivative and determine the intervals where the derivative is positive or negative. Computing the derivative, we arrive at the following expression:

$$\frac{dR_{HPR}}{d\phi} = \frac{\rho^2}{2(1-\rho)(1-\rho(1-\phi))^3}\Big[K\left(1-\rho-\phi(2-\rho)\right) - 2(1-\phi)(1-\rho)(1-\rho(1-\phi))(2-\rho(1-\phi))\Big]$$

The sign of the derivative will determine the intervals over which the function is increasing or decreasing. As $\rho \in (0,1), \phi \in [0,1]$, we observe that $2(1-\rho)(1-\rho(1-\phi))^3$ will always be positive, thus the denominator of the derivative is positive, and the sign depends on the numerator.

As $\rho^2 > 0$ follows by the same restriction on ρ, the sign of the derivative depends on the sign of the expression within the square brackets:

$$K\left(1-\rho-\phi(2-\rho)\right) - 2(1-\phi)(1-\rho)(1-\rho(1-\phi))(2-\rho(1-\phi)). \quad \text{(C.1)}$$

We consider the expression as the sum of $K\left(1-\rho-\phi(2-\rho)\right)$ and $-2(1-\phi)(1-\rho)(1-\rho(1-\phi))(2-\rho(1-\phi))$. Because of the restrictions on ρ and ϕ, $-2(1-\phi)(1-\rho)(1-\rho(1-\phi))(2-\rho(1-\phi))$ will always be non-positive.

However, solving the expression

$$K\left(1-\rho-\phi(2-\rho)\right) > 0,$$

we find that it is positive if $\phi \in [0, (1-\rho)/(2-\rho))$ and is non-positive otherwise, for any $\rho \in (0,1)$.

Thus, for the derivative to be positive, and thus $R_{HPR}(\phi)$ to be increasing, $\phi \in [0, (1-\rho)/(2-\rho))$ is a necessary condition. However, it is not sufficient. For such ϕ we are adding together a positive and non-positive value, so whether the sum is positive depends on the values of K and ρ. To determine a condition on K, we solve for where the expression in Equation (C.1) is positive:

$$\frac{2(1-\phi)(1-\rho)(1-\rho(1-\phi))(2-\rho(1-\phi))}{1-\rho-\phi(2-\rho)} < K. \quad \text{(C.2)}$$

The conditions on ϕ and K combined yield two possible behaviors for the derivative:

- The derivative is negative everywhere, thus $R_{HPR}(\phi)$ is monotone decreasing.
- The derivative transitions from positive to negative at some $\phi^* \in [0, (1-\rho)/(2-\rho))$, thus $R_{HPR}(\phi)$ is unimodal with a unique maximum at ϕ^*.

In order to show that these are the only possible behaviors, we claim that the expression on the right hand side of (C.2) is monotone increasing for $\phi \in [0, (1-\rho)/(2-\rho))$ and fixed ρ. To prove this claim, we compute the derivative of the expression with respect to ϕ, yielding the following:

$$\frac{4(1-\rho)\left[1-3(1-\phi)\phi\rho - 2(1-\phi)^3\rho^2 + (1-\phi)^3\rho^3\right]}{\left(1-\rho-\phi(2-\rho)\right)^2}.$$

As the denominator is a square term, and the value of ϕ which makes the denominator equal to 0 has been excluded from the domain, the denominator is always positive over $\phi \in [0, (1-\rho)/(2-\rho))$. Thus, the expression's behavior depends on the sign of the numerator.

$\rho \in (0,1)$ yields $4(1-\rho) > 0$. Thus, the sign of the numerator depends on the remaining term. If $\phi = 0$, then this term reduces to $1 + \rho^2(\rho - 2)$, which is positive for $\rho \in (0,1)$. As ϕ approaches $(1-\rho)/(2-\rho)$, the term approaches

$$\frac{8 - 17\rho + 13\rho^2 - 4\rho^3}{(2-\rho)^3}$$

And this is also always positive for $\rho \in (0,1)$. Therefore the expression on the right hand side of Equation (C.2) is monotone increasing.

Returning to the question of the behavior of $R_{HPR}(\phi)$. We now determine the conditions under which each case prevails, and the resulting maximum revenue.

Case I, Monotone Decreasing $R_{HPR}(\phi)$: If Equation (C.2) is never satisfied, then the derivative of R_{HPR} is negative for all ϕ in our domain, and therefore the function is monotone decreasing.

If (C.2) fails to hold, then certainly for all $\phi \in [0, (1-\rho)/(2-\rho))$ the following must be true for fixed K, ρ and:

$$\frac{2(1-\phi)(1-\rho)(1-\rho(1-\phi))(2-\rho(1-\phi))}{1-2\phi-\rho(1-\phi)} \geq K.$$

As this is true for all such ϕ, it is certainly true for $\phi = 0$. Making the substitution yields

$$2(1-\rho)(2-\rho) \geq K.$$

Rearranging the inequality in terms of $\rho \in (0,1)$ yields

$$\rho \leq \frac{3-\sqrt{2K+1}}{2}$$

And this expression is valid for $\rho \in (0,1)$ as long as $K \in [1,4)$.

Therefore, if $K \in [1,4)$ and $\rho \in (0, (3-\sqrt{2K+1})/2]$, then Equation (C.2) is never satisfied, $\frac{dR_{HPR}}{d\phi}$ is always negative, and $R_{HPR}(\phi)$ is monotone decreasing. As a result, the maximum revenue occurs at $\phi = 0$.

Case II, Unimodal $R_{HPR}(\phi)$ with unique maximum:

Assume we do not have $K \in [1,4)$ and $\rho \in (0, (3-\sqrt{2K+1})/2]$.Thus, we know that (C.2) holds for some $\phi \in [0, (1-\rho)/(2-\rho))$. Because K is a fixed positive constant and the right hand side of Equation (C.2) is monotone increasing in $\phi \in [0, (1-\rho)/(2-\rho))$, there must be some ϕ^* in the interval where (C.2) transitions from being true to being false. This then results in the derivative transitioning from being positive to negative, and therefore $R_{HPR}(\phi)$ is unimodal with unique maximum at ϕ^*.

To determine the value of ϕ^*, we solve for where K is equal to the expression on the RHS of Equation (C.2):

$$\frac{2(1-\phi^*)(1-\rho)(1-\rho(1-\phi^*))(2-\rho(1-\phi^*))}{1-\rho-\phi^*(2-\rho)} = K.$$

Solving for ϕ^* yields three possible solutions. These solutions are of the following form, for $n = 0, 1, 2$, where j is the square root of negative 1:

$$\phi^* = \frac{\rho-1}{\rho} + \frac{(1+j\sqrt{3})^n\,(2(1+K)-(2+K)\rho)}{\sqrt[3]{6\Phi}} - \frac{(1-j\sqrt{3})^n\,\sqrt[3]{\Phi}}{\sqrt[3]{36}(\rho-1)\rho^2},$$

and Φ is equal to the expression

$$18K(\rho-1)^3\rho^3 + \Bigg[6(1-\rho)^3\rho^6\Bigg(K^3(\rho-2)^3 + 12K(\rho-2)(\rho-1)^2 + 8(\rho-1)^3 - 6K^2(\rho-1)(2\rho-1)(4\rho-5)\Bigg)\Bigg]^{\frac{1}{2}}.$$

The specific value of ϕ^{opt} such that the solution is both real and in the bound $[0,(1-\rho)/(2-\rho))$ depends upon the values of ρ and K. Thus, while we can show how to compute ϕ^*, and we have also proven that it must be bounded between $[0,(1-\rho)/(2-\rho))$, we cannot have a simplified closed form expression for R^*_{HPR} in this case, as the valid value of ϕ^* that leads to R^*_{HPR} depends on the specific values of ρ and K we are working with. □

Letting $K=1$, corresponding to the $M|D|1$ queue, we find that $R_{HPR}(\phi)$ is monotone decreasing if $\rho \in \big(0,(3-\sqrt{3})/2\big]$, and unimodal otherwise, which is the assertion originally made in Section 4.4.

Appendix D. Extension of Lemma 2 under a M/G/1 queue

We extend Lemma 2 to the General service case as follows:

Lemma 6. *For fixed λ, μ, and K, and given $\phi \in [0,1]$, $C_{HPR}(\phi) \geq C_{PR}(\phi) \geq C_{NP}(\phi)$.*

Proof. As before, fixing λ and μ fixes ρ. As $(1/\rho)C_{NP}(\phi) = C_{PR}(\phi)$ continues to hold under $M|G|1$, we immediately conclude that $C_{PR}(\phi) \geq C_{NP}(\phi)$ holds.

To show $C_{HPR}(\phi) \geq C_{PR}$ still holds, we show that $C_{HPR}(\phi) - C_{PR} \geq 0$ must be the case regardless of the value of K. The common denominator of the two equations remains $2\mu(1-\rho)(1-\rho(1-\phi))^2$. Multiplying both equations by this quantity, $C_{HPR}(\phi) - C_{PR} \geq 0$ holds if the following holds:

$$2(1-\rho)(1-\rho(1-\phi)) + K\rho\phi - 2(1-\rho)(1-\rho(1-\phi))^2 - K\rho\phi \geq 0.$$

The $K\rho\phi$ terms cancel out, and thus the remainder of the proof is as in the $M|D|1$ case. Because there are no remaining K terms, we conclude that the relation holds for any $K \geq 1$. □

Appendix E. Maximum Guaranteed Revenue

In Section 4.4, we compared the maximum revenues under each policy, under the assumption that the provider is able to steer to the system to its preferred equilibrium if not already present. However, in a situation where the cost leading to the maximum revenue is associated to multiple equilibria, it is not guaranteed that the provider achieves its optimum outcome. In fact, in every multiple equilibria situation that exists, as seen in Theorems 1,2, 6, 7 and 8, a *none − reserve* is one of the equilibria possible.

To avoid a situation in which the system may migrate to the $none-reserve$ equilibrium, the provider may set the cost at the level that maximizes its revenue under the constraint that the resulting cost corresponds to a unique equilibrium. Additionally, we note from Theorems 1, 2, 6, 7 and 8 that if the cost is less than C_1, then the resulting equilibrium is either a unique $all-reserve$ or a unique $some-reserve$. This leads to the following definition:

Definition 4. *We denote the* maximum guaranteed revenue *as R^{**}. This quantity is defined with respect to the relationship between C_1 and $\underline{C}$:*

$$R^{**} = \begin{cases} \max_{\phi\in(0,1)} \lambda(1-\phi)C(\phi) \quad s.t. \quad C(\phi) < C_1 & \text{if } \underline{C} \neq C_1; \\ \lambda C_1 & \text{if } \underline{C} = C_1. \end{cases}$$

For brevity, we consider the analysis specifically for the $M|D|1$ case consistent with the analysis in Section 4. The results are extendable to $M|G|1$, but due to the exact threshold for the maximum revenue under HPR itself requiring a numerical solution except under special circumstances as noted in Appendix C, determining in general whether the maximum revenue under HPR is guaranteed or not in general requires extra steps and a numerical analysis to validate.

Under the NP and PR policies, the maximum revenue R^* occurs at the threshold $\phi^* = (1-\rho)/(2-\rho)$. To determine whether this threshold corresponds to a cost with a unique equilibrium, we solve $C_{NP}((1-\rho)/(2-\rho)) < C_1$, and $C_{PR}((1-\rho)/(2-\rho)) < C_1$, and find that these both hold so long as $\rho < 2/3$. Otherwise, the maximum revenue occurs at a cost associated to multiple equilibria.

To determine the threshold leading to the maximum guaranteed revenue for $\rho \geq 2/3$, we solve $C(\phi) = C_1$ under each policy, and arrive at a threshold of $\phi^{**} = ((1-\rho)/(\rho))^2$. Thus, the maximum revenue a provider is guaranteed occurs when the cost is set ϵ below $C(\phi^{**})$.

The resulting maximum guaranteed revenues under the NP and PR policies are

$$R^{**}_{NP} \begin{cases} R^*_{NP} & = \frac{\rho^3}{8(1-\rho)^2} \text{ if } \rho < \frac{2}{3}; \\ R_{NP}\left(\left(\frac{1-\rho}{\rho}\right)^2\right) & = \frac{\rho(2\rho-1)}{2(1-\rho)} \text{ if } \rho \geq \frac{2}{3}; \end{cases}$$

$$R^{**}_{PR} \begin{cases} R^*_{PR} & = \frac{\rho^2}{8(1-\rho)^2} \text{ if } \rho < \frac{2}{3}; \\ R_{PR}\left(\left(\frac{1-\rho}{\rho}\right)^2\right) & = \frac{2\rho-1}{2(1-\rho)} \text{ if } \rho \geq \frac{2}{3}. \end{cases}$$

For the HPR policy under $M|D|1$, from Theorem 3 the only single equilibrium region that is not the $none-reserve$ region is the $all-reserve$ equilibrium where $C < C_1$. Therefore, we apply the definition for R^{**} where $\underline{C} = C_1$, and thus the resulting maximum guaranteed revenue is

$$R^{**}_{HPR} = \lambda C_1 = \frac{\rho^2}{2(1-\rho)}.$$

With the maximum guaranteed revenues defined for each policy, we now seek to prove that the provider is best off under the HPR policy even when operating under the constraint of maximum guaranteed revenue.

Theorem 9. *For fixed λ and μ, $R^{**}_{HPR} \geq R^{**}_{PR} \geq R^{**}_{NP}$*

Proof. Fixing λ and μ fixes ρ. From the definitions, we note that regardless of the value of ρ, $R^{**}_{PR} = (1/\rho)R^{**}_{NP}$. Thus, $R^{**}_{PR} \geq R^{**}_{NP}$ follows, and it suffices to show that $R^{**}_{HPR} \geq R^{**}_{PR}$.

If $\rho < 2/3$, we want to show

$$\frac{\rho^2}{2(1-\rho)} \geq \frac{\rho^2}{8(1-\rho)^2},$$

dividing by the right hand side reduces this to solving:

$$4(1-\rho) \geq 1.$$

This holds for $\rho \leq 3/4$, and so certainly $R^{**}_{HPR} \geq R^{**}_{PR}$ holds for $\rho < 2/3$. When $\rho \geq 2/3$, we want to show:

$$\frac{\rho^2}{2(1-\rho)} \geq \frac{2\rho - 1}{2(1-\rho)},$$

multiplying both sides by the common denominator, and subtracting the right hand side reduces this to solving

$$\rho^2 - 2\rho + 1 \geq 0.$$

The left hand side factors to $(\rho - 1)^2$, which is indeed nonnegative. Therefore for $\rho \geq 2/3$, $R^{**}_{HPR} \geq R^{**}_{PR}$ holds. Therefore, $R^{**}_{HPR} \geq R^{**}_{PR}$ holds for all ρ. □

Thus, the provider is best off under the HPR policy even if maximizing revenue subject to a constraint. In fact, the maximum guaranteed revenue under HPR is larger than the maximum possible revenue under NP or PR if $\rho < 2/3$.