https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–17
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

The Cost of Impatience in Dynamic Matching: Scaling Laws and Operating Regimes

Angela Kohlenberg,[a,*] **Itai Gurvich**[a]

[a] Kellogg School of Management, Northwestern University, Evanston, Illinois 60208
*Corresponding author
Contact: angela.kohlenberg@kellogg.northwestern.edu, https://orcid.org/0009-0006-0966-6492 (AK); i-gurvich@kellogg.northwestern.edu, https://orcid.org/0000-0001-9746-7755 (IG)

Received: May 18, 2023
Revised: October 29, 2023; January 10, 2024
Accepted: January 19, 2024
Published Online in Articles in Advance: July 19, 2024

https://doi.org/10.1287/mnsc.2023.01513

Abstract. We study matching queues with abandonment. The simplest of these is the two-sided queue with servers on one side and customers on the other, both arriving dynamically over time and abandoning if not matched by the time their patience elapses. We identify nonasymptotic and universal scaling laws for the matching loss due to abandonment, which we refer to as the "cost of impatience." The scaling laws characterize the way in which this cost depends on the arrival rates and the (possibly different) mean patience of servers and customers. Our characterization reveals four operating regimes identified by an operational measure of patience that brings together mean patience and utilization. The four regimes subsume the regimes that arise in asymptotic (heavy-traffic) approximations. The scaling laws, specialized to each regime, reveal the fundamental structure of the cost of impatience and show that its order of magnitude is fully determined by (i) a "winner-take-all" competition between customer impatience and utilization, and (ii) the ability to accumulate inventory on the server side. Practically important is that when servers are impatient, the cost of impatience is, up to an order of magnitude, given by an insightful expression where only the minimum of the two patience rates appears. Considering the trade-off between abandonment and capacity costs, we characterize the scaling of the optimal safety capacity as a function of costs, arrival rates, and patience parameters. We prove that the ability to hold inventory of servers means that the optimal safety capacity grows logarithmically in abandonment cost and, in turn, slower than the square-root growth in the single-sided queue.

History: Accepted by Baris Ata, stochastic models and simulation.
Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2023.01513.

Keywords: two-sided queue • queue approximations • operational regimes • capacity sizing • abandonment

1. Introduction

In dynamic matching, participants arrive at a matching market over time and wait to be matched. The fundamental tension in dynamic matching is between the quality and efficiency of matches. Delaying matches to "thicken the market" can lead to better matches becoming available, but at the expense of increasing the time to match. This tension is especially pronounced when participants are impatient and leave (abandon) the market if not matched within an amount of time that they deem acceptable. In such cases, delaying matches to thicken the market may have the opposite effect of thinning the market through abandonments.

Abandonments (also called departures) are a key feature of dynamic matching applications. Examples include ride hailing, where both riders and drivers may abandon if the wait time is too long (e.g., Yu et al. 2022), and organ exchanges, where donor-recipient pairs may depart the exchange if the recipient's health deteriorates or if a donor is found outside the exchange (e.g., Ashlagi et al. 2018 and the references therein). Impatience is similarly important in the allocation of perishable inventory to impatient demand (e.g., blood banks (Bar-Lev et al. 2017) and food banks (Prendergast 2017)).

Much of the research on dynamic matching to date assumes infinitely patient participants. This focus makes sense, as the optimal control of these networks is sufficiently complicated, even without impatience (e.g., Kerimov et al. 2021 and the references therein).

Ignoring impatience will lead to inaccurate evaluation of system performance and may result in suboptimal decisions. Yet the extent to which abandonment impacts system performance and how this depends *jointly* on the arrival and impatience rates is not fully understood. This knowledge is critical for optimal decision making.

We make a step toward a fuller understanding of impatience in matching by (re)considering the simplest

of matching models, as shown in Figure 1. There are two types of participants: customers and servers. A match consists of one customer and one server. Customers and servers with finite (random) patience arrive dynamically over time and, if not immediately matched, join their dedicated queue and wait to be matched. There is no fundamental difference between customers and servers; for convenience, we label the participants with the greater arrival rate as servers.

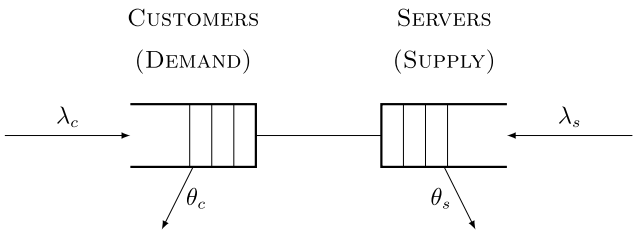
With one possible match, the control that minimizes abandonment is trivial: it is optimal to perform a match whenever there is an available customer and an available server. Under this policy, arriving customers match immediately with waiting servers, and vice versa, so there are either customers waiting or servers waiting, but never both. This is a two-sided (also called "double-ended") queue; see Figure 1.

When participants have infinite patience (do not abandon), the match rate is trivially equal to the minimum of the two arrival rates: $\min\{\lambda_c, \lambda_s\}$; with impatience, this becomes an upper bound on the match rate. The difference between this no-abandonment upper bound and the actual match rate is the *cost of impatience* (CoI), which depends on the abandonment and arrival rates of customers and servers.

Explicit steady-state expressions for the abandonment rate can be derived for this model, but these expressions are not informative beyond allowing for numerical computations. Instead, we establish *scaling laws* that characterize how the cost of impatience changes as a function of the model parameters. Specifically, we derive expressions that both upper and lower bound the true cost of impatience, up to multiplicative constants that do not depend on the parameters. This characterization is *non*asymptotic and holds regardless of any notion of asymptotic regime.

The scaling laws reveal four operating regimes that encompass all combinations of model parameters. These regimes subsume existing asymptotic regimes and offer insights into the performance of a two-sided queue in terms of simple building blocks. The scaling laws identify the key determinants of match loss from impatience. The operating regimes provide a simple framework for identifying when and how settings are fundamentally different in terms of the impact of impatience on match loss. We use these results and scaling laws for the single-sided queue to draw attention to key properties of the two-sided queue with abandonment.

Figure 1. The Simplest Matching Model: A Two-Sided Matching Queue with Abandonment



1.1. Overview of Results

We establish a universal scaling law for the cost of impatience as a function of the model parameters. That is, we identify a function, $\mathcal{S}$, of the arrival and patience rate vectors, $\boldsymbol{\lambda} = (\lambda_c, \lambda_s)$ and $\boldsymbol{\theta} = (\theta_c, \theta_s)$, such that

$$\frac{1}{\Gamma} \le \frac{\mathrm{CoI}(\boldsymbol{\lambda}, \boldsymbol{\theta})}{\mathcal{S}(\boldsymbol{\lambda}, \boldsymbol{\theta})} \le \Gamma, \qquad (\mathrm{CoI} \sim \mathcal{S})$$

for some constant Γ that does not depend on the parameters. The function $\mathcal{S}$ is tractable and exposes four operating regimes that are distinguished by the level of impatience of customers and servers.

The level of impatience is based on a comparison between the amount of time that a participant is willing to wait and the amount of time that they have to wait to match. Informally, customers are "patient" when they are willing to wait longer than their expected time to match; they are "impatient" otherwise. This is determined by a measure of the customer mean patience relative to utilization, $\rho := \lambda_c/\lambda_s$. Servers are similarly patient or impatient, based on a relative measure of the server mean patience and utilization.

Three simple metrics determine whether customers and servers are impatient or patient, which, in turn, specify the operating regime for a matching market. These are the utilization, ρ, and the arrival-to-patience ratios, λ_c/θ_c and λ_s/θ_s.

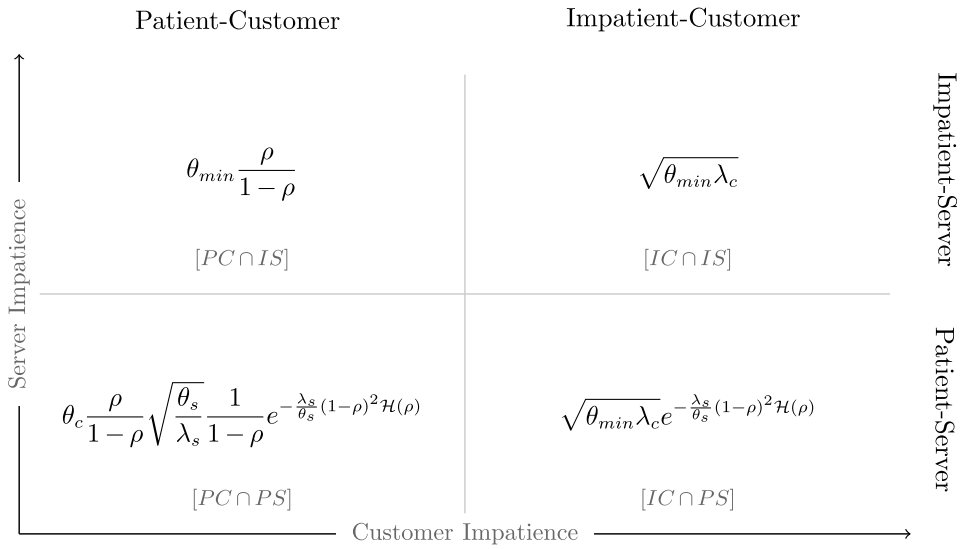
Any two-sided queue with abandonment operates in one of four regimes and has CoI proportional to (upper and lower bounded by) the expression shown for that regime in Figure 2. This nonasymptotic analysis exposes a richer picture of the CoI than asymptotic approximations, which yield a single expression for the CoI when patience rates are fixed and arrival rates are scaled up.

Customer impatience impacts the CoI differently than server impatience (recall that $\lambda_c \le \lambda_s$). To see this, note that the CoI is equal to the expected customer abandonment rate. Customers abandon their queue, Q_c, at a rate of θ_c when the server queue, Q_s, is empty. Therefore,

$$\mathrm{CoI} = \theta_c \mathbb{E}[Q_c \mid Q_s = 0]\mathbb{P}(Q_s = 0).$$

When customers have to wait, the customer queue is conveniently determined by a "winner-take-all" competition between customer impatience and excess capacity, $\lambda_s - \lambda_c$. Only one of customer impatience or excess capacity, but not both, matter for the scaling laws of the customer queue (conditional on customers having to wait). If excess capacity is high enough relative to customer

Figure 2. Cost of Impatience for the Four Operating Regimes Where $\theta_{min} = \min\{\theta_c, \theta_s\}$, $\rho = \lambda_c/\lambda_s$ and $\mathcal{H}(\rho) = \sum_{n=1}^{\infty} \frac{1}{n(n+1)}(1-\rho)^{n-1}$



Note. *PC*, *IC*, *PS*, and *IS* denote Patient-Customer, Impatient-Customer, Patient-Server, and Impatient-Server, respectively, so that $PC \cap IS$ denotes the Patient-Customer, Impatient-Server regime.

impatience, excess capacity "wins," and impatience does not matter for scaling. Otherwise, impatience "wins," and excess capacity does not matter for scaling. Mathematically, this is captured by the fact that the expected customer queue (conditional on customers having to wait) is proportional to the minimum of the expected number in either an $M/M/1$ (no impatience) queue or a critically loaded (no excess capacity) $M/M/1+M$ queue:

$$\mathbb{E}[Q_c|Q_s = 0] \sim \min\left\{\frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_c}}\right\}.$$

When customers are patient, they are willing to wait longer than the expected time to match. Therefore, few customers abandon, and abandonment has a limited impact on the customer queue; the customer queue scales like an $M/M/1$ queue (proportional to $\rho/(1-\rho)$). When customers are impatient, they may have to wait longer than they are willing. Any excess capacity is insufficient to have a scaling effect on the customer queue; the customer queue scales like an $M/M/1+M$ queue with no excess capacity (proportional to $\sqrt{\lambda_c/\theta_c}$).

Server impatience impacts the likelihood that customers have to wait. When servers are patient, sufficient inventory of waiting servers can be accumulated so that most customers are matched immediately and few abandon. When servers are impatient, there will be little or no inventory of servers, and most customers will have to wait to be matched. The significant benefit of server patience is captured by the exponential term in the second row of Figure 2.

Interestingly, only the *minimum* abandonment rate appears when *either* customers or servers are impatient. Recall that $\text{CoI} = \theta_c \mathbb{E}[Q_c|Q_s = 0]\mathbb{P}(Q_s = 0)$. The minimum rate, $\theta_{min} = \{\theta_c, \theta_s\}$ appears, rather than θ_c, in three regimes in Figure 2 is because of the ability to accumulate inventory of waiting servers.

Practically, this means that to decrease the cost of impatience, the focus should be on decreasing the minimum patience rate (the maximum mean patience). It only matters that either customers or servers are patient enough; the patience of the other type has no order-of-magnitude effect. Intuitively, items of one participant type serve as inventory for the other in these matching markets. The ability to accumulate inventory of one of the two types, those who abandon less, creates a buffer protecting against the impatience of the other type. It is only when *both* customers and servers are patient, and the CoI is very low, that both patience rates appear in the CoI scaling.

When servers are impatient, the CoI is proportional to an insightful function of the minimum patience rate, θ_{min}, the utilization, and the smaller of the arrival rates:

$$\text{CoI} \sim \theta_{min} \min\left\{\frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_{min}}}\right\}.$$

Notice that θ_{min} can be either θ_c or θ_s. For example, servers are impatient if either (i) utilization is close to one, or (ii) the server abandonment rate is greater than their

arrival rate ($\theta_s > \lambda_s$). In both cases, θ_{min} can be either θ_c or θ_s.

More generally, the scaling laws identify which parameters have the most impact on match loss from impatience and specify the relative impact of a change in these parameters; this focuses attention on key managerial levers, as we illustrate next.

Table 1 collects examples of dynamic matching applications that fit each of the operating regimes (Online Appendix F reports data supporting the classification of these examples). These settings, and key insights revealed by the scaling laws, are as follows.

- **Blood transfusion ($IC \cap IS$)**. In developing countries, it is common for the supply of blood to be only slightly greater than the demand for blood. Therefore, utilization is close to one, and both customers and servers are impatient. Individuals who need a blood transfusion are the customers, and blood donations are the servers. In the $IC \cap IS$ regime, CoI $\sim \sqrt{\theta_{min}\lambda_c}$. The CoI scales proportionally to the square root of the arrival rate of transfusion patients, λ_c. A small change in the rate at which blood is collected, λ_s, does not have an order-of-magnitude impact on the CoI. Blood donations can be stored for 42 days, but patients may need blood immediately; small changes in the storage life of blood donations, θ_{min}, will have a greater impact on the CoI than small increases in supply (donations).
- **Cadaveric liver transplant ($IC \cap PS$)**. The demand for liver transplants is greater than the supply of donor livers. Livers must be transplanted within 8 to 12 hours, whereas individuals on the transplant waiting list typically wait several months or years. Therefore, livers are the customers, and they are impatient; transplant patients are the servers, and they are patient. In the $IC \cap PS$ regime, CoI $\sim \sqrt{\theta_{min}\lambda_c}e^{-\frac{\lambda_s}{\theta_s}(1-\rho)^2\mathcal{H}(\rho)}$. A small change in the arrival rate of transplant patients, λ_s, has an exponential impact on the CoI, whereas a small change in the storage life of livers (the most impatient type) does not have an order-of-magnitude impact on the CoI. In other words, it is not the perishability of the livers that is the main practical challenge in this setting; it is that—even without abandonment—the number of transplants is limited by the small rate of cadaveric-liver arrivals.
- **Foster care adoption ($PC \cap IS$)**. In the foster care system in the state of Pennsylvania, the number of children becoming available for adoption annually is less than the number of families joining the adoption list. Here, children are the customers and families are the servers. Children wait to be adopted until they reach the age of 18; they are patient. Families have other options available if they are not matched soon enough (e.g., private or international adoption, adoption from a different state, or deciding not to adopt); they are impatient. In the $PC \cap IS$ regime, CoI $\sim \theta_{min}\frac{\rho}{1-\rho}$. Small changes in both the arrival rate of children and families have an impact on the CoI that is proportional to $\rho/(1-\rho)$.
- **Ride hailing ($PC \cap PS$)**. Ride-hailing drivers in the Manhattan Central Business District (CBD) spend approximately 20%–25% of their time in the CBD waiting for a ride request.[1] This means that utilization in the CBD is less than 80%, where passengers are the customers, and drivers are the servers. Passengers are not willing to wait long for each trip to start, but trips start quickly because utilization is relatively low and arrival rates are high. Drivers are at least as patient as passengers. Therefore, both passengers and drivers are patient. In the $PC \cap PS$ regime, the CoI is low, and small changes in the patience of both passengers and drivers have a significant impact on the CoI.

The scaling laws are consistent with the intuition that substantial excess capacity (and, in turn, low utilization) guarantees low CoI. But they also underscore a property of the two-sided queue: the ability to accumulate inventory of servers may result in low CoI, even without substantial excess capacity. We characterize the *safety capacity*, $\lambda_s - \lambda_c$, that balances the trade-off between abandonment and capacity costs. We show that the optimal safety capacity has the form

$$\lambda_s^* - \lambda_c \sim \delta\sqrt{\lambda_c},$$

for $\delta \geq 0$ that does not depend on λ_c but is affected by the relationship between the cost per abandonment, c_a, and the cost per unit of capacity (server arrivals), c_s.[2] In the single-sided $M/M/1+M$ queue, δ scales proportionally to $\sqrt{c_a/c_s}$. As the abandonment cost grows relative to the capacity cost, the safety capacity grows proportionally to the square root of cost growth, with all else fixed. In the two-sided queue, in contrast, δ grows proportionally to $\log(c_a/c_s)$.

This precise characterization of the capacity scaling highlights a key difference between the single-sided and two-sided queue. The ability to build a buffer of

Table 1. Examples of Dynamic Matching Applications That Fit Each of the Four Operating Regimes

Setting	Customers	Servers	Regime
Blood transfusion (in a location with low blood supply)	Patients	Blood donations	$IC \cap IS$
Liver transplant	Donor livers	Patients	$IC \cap PS$
Adoption	Children	Families	$PC \cap IS$
Ride hailing	Riders	Drivers	$PC \cap PS$

waiting servers in the two-sided queue leads to safety capacity that scales slower with the abandonment cost.

Our characterization of the optimal capacity scaling reveals that any of the four operating regimes in Figure 2 can be rationalized from an optimization perspective. It can be optimal to operate with either high or low utilization, depending on the relative costs and patience rates.

Outline of the paper. Section 2 summarizes related literature. The model is described in Section 3. The scaling laws and operating regimes are studied in Section 4. Capacity-sizing results appear in Section 5. We conclude in Section 6. All proofs appear in the Online Appendix.

2. Related Literature

Our matching model is a two-sided, or double-ended, queue with abandonment. When one side is completely impatient, the double-ended queue reduces to a single-server, single-class queue with abandonment. As such, our work speaks to both the dynamic matching literature and the extensive literature on single-class queues with abandonment.

Analysis of matching queues with abandonment. Double-ended queues were introduced as the "taxi model" where taxis queue to wait for a customer, and customers queue to wait for a taxi (e.g., Kendall 1951, Kashyap 1966). Conolly et al. (2002) derive exact analytical results for the transient and steady-state performance of the Markovian two-sided queue with Poisson arrivals and exponential patience times; this two-sided queue is the same as our model. Afèche et al. (2014) and Diamant and Baron (2019) derive closed-form expressions for the steady-state queue-length distribution of a two-sided queue with two types of customers: those who abandon immediately and those with either some or infinite patience. Liu et al. (2015) and Büke and Chen (2017) develop fluid and diffusion approximations for a two-sided queue in heavy traffic. Exact analysis of the two-sided queue and variants thereof appear in the study of organ allocation (Boxma et al. 2011, Elalouf et al. 2018), blood bank allocation (Bar-Lev et al. 2017), and general perishable inventory systems (Perry and Stadje 1999). Beyond a single match, Castro et al. (2020a) derive explicit expressions for the steady-state distributions of a two-customer, two-server network following the first-come first-served policy. Zubeldia et al. (2022) study the stability region for a matching network with two matches operated under a max-weight policy.

Analysis of single-sided queues with abandonment. When either customers or servers (but not both) are infinitely impatient, our model reduces to a single-sided queue with Poisson arrivals, exponential service times, and exponentially distributed patience.

The single-server queue with abandonment has been used to study perishable inventory (Graves 1982), public housing (Kaplan 1986), and organ allocation (Zenios 1999). See Ward (2012) for a survey of results on single-class queues with abandonment. Of immediate relevance to our work, Ward and Glynn (2003) develop diffusion approximations for an $M/M/1+M$ queue under various asymptotic regimes. Our results, specialized to these regimes, align with those in Ward and Glynn (2003); this connection merits (and will receive) further discussion after we introduce our results.

Capacity planning in double- and single-sided queues. Lee and Ward (2019) study joint pricing and capacity sizing for the $M/GI/1+GI$ queue and derive asymptotically optimal policies in a regime where the service distribution is fixed and arrival rates grow along the sequence of queues.

Other levers for controlling supply and demand in two-sided queues are considered, for example, in Nguyen and Stolyar (2018), Chen and Hu (2020), Vaze and Nair (2022), and Varma et al. (2022).

Optimal control of matching queues with abandonment. Recent progress on the optimal control of dynamic matching markets with abandonment includes Collina et al. (2020), Castro et al. (2020b), Aouad and Saritaç (2022), Wang et al. (2022), and Aveklouris et al. (2024), all of whom study control policies for a network of matches with impatient participants and introduce algorithms to determine when to perform matches and which matches to perform.

Analysis and control of matching queues without abandonment. The literature on matching queues without abandonment is relatively mature and includes papers that study performance under specific policies (e.g., Caldentey et al. 2009, Adan et al. 2018 and the references therein), as well as various optimization levers, such as pricing (e.g., Varma and Maguluri 2021), menu design (e.g., Afèche et al. 2022), and dynamic control (e.g., Gurvich and Ward 2014, Özkan and Ward 2020, Kerimov et al. 2021).

In this work, we revisit the simplest matching model with abandonment: the two-sided queue with Poisson arrivals and exponential patience. Our goal is to deepen the understanding of this model on its own and as a necessary building block for networks of matching queues with impatient customers.

3. Model

We consider a matching queue with two types of participants: customers and servers. Customers and servers arrive according to independent Poisson processes with rates λ_c and λ_s, respectively.

Each type has its own dedicated infinite-capacity queue where participants wait to be matched. Matches are made

between one customer and one server according to a first-come-first-served policy. A waiting customer is matched immediately with an arriving server, and vice versa. When a match is performed, the matched customer and server leave the system immediately (there is no processing time).

Participants are never rejected (or blocked) but may choose to abandon after they join their queue. Customers and servers have exponential patience with rates $\theta_c > 0$ and $\theta_s > 0$, respectively. If participants are not matched by the time their patience elapses, they *abandon* the queue. Patience is independent across participants. To avoid trivialities, we assume that either customers or servers have at least some patience. That is, $\theta_{min} = \min\{\theta_c, \theta_s\} < \infty$, or equivalently, $\max\{1/\theta_c, 1/\theta_s\} > 0$. If this assumption is violated, then all arrivals abandon immediately, and because there are no simultaneous arrivals, no match is performed.

In the absence of impatience, the expected long-run average match rate is equal to the minimum of the two arrival rates: $\min\{\lambda_c, \lambda_s\}$. The difference between this upper bound and the actual match rate is the *cost of impatience*.

We label the arrival rates, without loss of generality, so that $\lambda_c \le \lambda_s$. If this is violated, then we replace *customers* with *servers*, and vice versa, in all the following results.

3.1. The Two-Sided Queue

Let $Q_c(t)$ and $Q_s(t)$ denote the number of customers and servers waiting in their queue at time t, respectively. Because customers and servers are matched immediately, only one queue can be positive at any given time. We define the one-dimensional, continuous-time Markov chain, $Q(t) = Q_c(t) - Q_s(t)$, where $Q_c(t) = [Q(t)]^+ = \max\{0, Q(t)\}$ and $Q_s(t) = [Q(t)]^- = \max\{0, -Q(t)\}$, as in Figure 3. With $\theta_c, \theta_s > 0$, $Q = \{Q(t) : t \ge 0\}$ has a steady-state distribution; we omit the time index t when considering the queue in steady state.

Let $A_c(t)$ denote the number of customers that arrive by time t and $R_c(t)$ denote the cumulative number of customers that abandon by time t; $A_s(t)$ and $R_s(t)$ are defined similarly for servers. Let $D(t)$ denote the number of matches performed by time t. At all $t \ge 0$,

$$Q_i(t) = A_i(t) - R_i(t) - D(t)$$

for $i = c, s$. Let $d = \lim_{t\uparrow\infty} \frac{1}{t}\mathbb{E}[D(t)]$ be the expected long-run average match *rate*. It follows that for $i = c, s$,

$$\lim_{t\uparrow\infty} \frac{1}{t}\mathbb{E}[Q_i(t)] = \lim_{t\uparrow\infty} \frac{1}{t}\mathbb{E}[A_i(t) - R_i(t) - D(t)]$$
$$= \lambda_i - \theta_i\mathbb{E}[Q_i] - d = 0.$$

Here, we use the facts that $\mathbb{E}[A_i(t)] = \lambda_i t$ and that $\mathbb{E}[R_i(t)] = \mathbb{E}[\theta_i \int_0^t Q_i(s)ds]$.[3] Hence,

$$d = \lambda_c - \theta_c\mathbb{E}[Q_c] = \lambda_s - \theta_s\mathbb{E}[Q_s]. \tag{1}$$

Recall that the cost of impatience is the difference between the no-abandonment match rate and the actual match rate, d. Because $\lambda_c \le \lambda_s$, the no-abandonment match rate is $\min\{\lambda_c, \lambda_s\} = \lambda_c$. Therefore, from (1), we obtain that the CoI is equal to the expected long-run average rate of customer abandonment:

$$\text{CoI} = \lambda_c - d = \theta_c\mathbb{E}[Q_c]. \tag{CoI}$$

We identify expressions that both lower and upper bound the CoI. We write

$$g(\boldsymbol{\lambda}, \boldsymbol{\theta}) \overset{\mathcal{M}}{\sim} f(\boldsymbol{\lambda}, \boldsymbol{\theta}) \tag{2}$$

when there exists a constant, $\Gamma \ge 1$, that *does not depend on either* $\boldsymbol{\lambda}$ *or* $\boldsymbol{\theta}$, such that

$$\frac{1}{\Gamma} \times f(\boldsymbol{\lambda}, \boldsymbol{\theta}) \le g(\boldsymbol{\lambda}, \boldsymbol{\theta}) \le \Gamma \times f(\boldsymbol{\lambda}, \boldsymbol{\theta}), \text{ for all } (\boldsymbol{\lambda}, \boldsymbol{\theta}) \in \mathcal{M}, \tag{3}$$

where $\mathcal{M}$ is a family of parameters that satisfy certain restrictions.

Definition 1 formalizes the notion of parameter families. The conditions we impose have the same purpose: to guarantee that the customer queue is not negligible and, therefore, that we have a real two-sided queue. If $\lambda_s \gg \lambda_c$, then an arriving customer matches with a server before the next customer arrives. Similarly, if $\theta_c \gg \lambda_c$, then an arriving customer abandons before the next customer arrives. In either case, there is effectively no customer queue, and the two-sided queue "collapses" into a single-sided queue.

Condition (i) restricts our focus to settings where supply (the arrival rate of servers) is not significantly larger than demand (the arrival rate of customers). When λ_s is optimized (see Section 5), condition (i) arises as an outcome under reasonable conditions on the problem parameters.

Definition 1 (Queue Families). Fix $M \ge 1$. Denote by $\mathcal{M}(M)$ the family of primitives $(\boldsymbol{\lambda}, \boldsymbol{\theta})$ such that

i. Nonnegligible demand relative to supply: $\lambda_s \le M\lambda_c$, or equivalently $\rho = (\lambda_c/\lambda_s) \ge (1/M)$, and

ii. Nonnegligible customer patience: $\theta_c \le M\lambda_c$.

Let

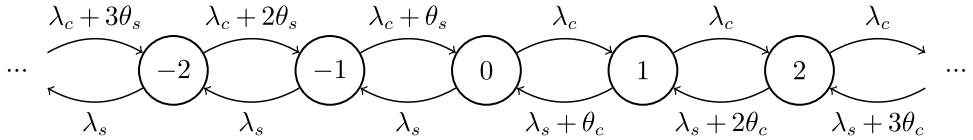
$$\mathcal{M} = \mathcal{M}(M) := \left\{(\lambda, \theta) \ge 0 : \frac{\lambda_s}{\lambda_c} \le M, \ \frac{\theta_c}{\lambda_c} \le M\right\}.$$

The set of parameters that we allow, and relative to which the $\sim$ relationship is evaluated, is $\mathcal{M}(M)$. The constant Γ in Relationship (2) depends on this M.

4. Scaling Laws and Operating Regimes

Our main mathematical result in Theorem 1 characterizes how the cost of impatience scales with model parameters.

Figure 3. The Two-Sided Queue, $Q = Q_c - Q_s$



Theorem 1 (Cost of Impatience Scaling). *Let* $\mathcal{M}' = \{(\boldsymbol{\lambda}, \boldsymbol{\theta}) : \rho = \lambda_c/\lambda_s < 1\}$. *Then,*

$$\mathbb{E}[Q|Q \geq 0] \overset{\mathcal{M}\cap\mathcal{M}'}{\sim} \min\left\{\frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_c}}\right\}$$

and

$$\mathbb{P}(Q \geq 0) \overset{\mathcal{M}\cap\mathcal{M}'}{\sim} \left(1 + \left[1 + \frac{\rho}{1-\rho}\frac{\theta_c}{\lambda_c}\min\left\{\frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_c}}\right\}\right] \sqrt{\frac{\lambda_s}{\theta_s}}(1-\rho)e^{\frac{\lambda_s}{\theta_s}(1-\rho)^2\mathcal{H}(\rho)}\right)^{-1}$$

where $\mathcal{H}(\rho) = \sum_{n=1}^{\infty} \frac{1}{n(n+1)}(1-\rho)^{n-1}$.

The cost of impatience subsequently satisfies

$$\begin{aligned} CoI = \theta_c\mathbb{E}[Q_c] &= \theta_c\mathbb{E}[Q|Q \geq 0]\mathbb{P}(Q \geq 0) \\ &\overset{\mathcal{M}\cap\mathcal{M}'}{\sim} \theta_c\min\left\{\frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_c}}\right\} \\ &\quad \left(1 + \left[1 + \frac{\rho}{1-\rho}\frac{\theta_c}{\lambda_c}\min\left\{\frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_c}}\right\}\right] \sqrt{\frac{\lambda_s}{\theta_s}}(1-\rho)e^{\frac{\lambda_s}{\theta_s}(1-\rho)^2\mathcal{H}(\rho)}\right)^{-1}. \end{aligned}$$

On $\mathcal{M}\backslash\mathcal{M}'$ *(when* $\rho = 1$*), the cost of impatience satisfies*

$$CoI \overset{\mathcal{M}\backslash\mathcal{M}'}{\sim} \sqrt{\theta_{min}\lambda_c}.$$

Theorem 1 identifies how key metrics—utilization, ρ, and the patience-to-arrival ratios, λ_c/θ_c and λ_s/θ_s—jointly determine the performance of a two-sided queue with abandonment. It draws attention to key properties of the expected customer queue, conditional on customers having to wait, $\mathbb{E}[Q|Q \geq 0]$, and the probability that customers have to wait, $\mathbb{P}(Q \geq 0)$. These properties reveal key determinants of the CoI.

A "competition" between customer impatience and excess capacity. Theorem 1 shows that $\mathbb{E}[Q|Q \geq 0] \sim \min\left\{\frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_c}}\right\}$. The quantity $\rho/(1-\rho)$ is the expected number-in-system in an $M(\lambda_c)/M(\lambda_s)/1$ queue (a single-server queue with arrival rate λ_c, service rate λ_s, and no abandonment). The quantity $\sqrt{\lambda_c/\theta_c}$ is, up to a constant multiplier, the expected number-in-system in an $M(\lambda_c)/M(\lambda_c)/1 + M(\theta_c)$ queue (a critically loaded, $\rho = 1$, queue with patience parameter θ_c). Thus, the customer queue, conditional on customers having to wait, is determined by a "winner-take-all" competition between customer impatience and excess capacity, $\lambda_s - \lambda_c$: only one of impatience or excess capacity, but not both, matters for scaling purposes.

Excess capacity determines how fast matches are performed (i.e., the expected wait time in an $M(\lambda_c)/M(\lambda_s)/1$ queue is $\frac{1}{\lambda_s-\lambda_c}$). When $\frac{\rho}{1-\rho} = \frac{\lambda_c}{\lambda_s-\lambda_c} \leq \sqrt{\frac{\lambda_c}{\theta_c}}$, excess capacity is large relative to mean customer impatience, and customers are matched faster than they abandon. In this case, excess capacity "wins," and the customer queue behaves the same as the queue without abandonment (the $M/M/1$ queue). Only excess capacity matters for scaling purposes; impatience has, at most, a constant multiplying effect on the CoI.

Conversely, when $\sqrt{\frac{\lambda_c}{\theta_c}} \leq \frac{\rho}{1-\rho} = \frac{\lambda_c}{\lambda_s-\lambda_c}$, mean customer impatience is large relative to excess capacity. The customer queue behaves the same as a queue with no excess capacity (the critically loaded $M/M/1+M$ queue); only customer impatience matters for scaling purposes, and excess capacity has, at most, a constant multiplying effect on the CoI.

On the proof of Theorem 1. The proof of Theorem 1 is based on expansions of the explicit expressions for the CoI, as well as coupling-based comparisons with simpler queues. The explicit expressions for the steady-state distributions involve infinite sums and products. To derive the CoI approximation (specifically, the approximation for $\mathbb{P}(Q \geq 0)$), we truncate these expressions at carefully chosen thresholds and analyze those truncated expressions. To bound $\mathbb{E}[Q|Q \geq 0]$, we note that the customer queue, conditional on customers having to wait, is equal in distribution to an $M(\lambda_c)/M(\lambda_s)/1 + M(\theta_c)$ queue. It follows from simple coupling arguments that this queue is upper bounded by both an $M(\lambda_c)/M(\lambda_s)/1$ queue (no abandonment) and an $M(\lambda_c)/M(\lambda_c)/1 + M(\theta_c)$ queue (no excess capacity). It is the lower bound where more care is needed. We couple the $M(\lambda_c)/M(\lambda_s)/1 + M(\theta_c)$ queue with an $M(\lambda_c)/M(\lambda_s + \theta_c K)/1/K$ (a queue with finite waiting room and a modified service rate). Here, it is the choice of K that is critical and produces the desired results.

Customer vs. server impatience. Theorem 1 reveals that customer impatience impacts the CoI in a different way than server impatience. Customer impatience acts

on the expected customer queue, $\mathbb{E}[Q|Q \geq 0]$ (θ_s does not appear in this expression). Server impatience acts, together with customer impatience, on the probability that there is a customer queue, $\mathbb{P}(Q \geq 0)$. We introduce four operational regimes that are distinguished by the level of customer and server impatience. When parameters are specialized to one of these regimes, the expressions in Theorem 1 simplify and make explicit how the server impatience, through $\mathbb{P}(Q \geq 0)$, impacts the CoI.

4.1. Operating Regimes

The operating regimes are defined in terms of a relationship between impatience and utilization.

Definition 2 (Operating Regimes). A two-sided matching queue with abandonment has impatient customers (*IC*) if

$$\sqrt{\frac{\lambda_c}{\theta_c}} \leq \frac{\rho}{1-\rho}, \qquad \text{(Impatient-Customer)}$$

and has patient customers (*PC*) otherwise. It has impatient servers (*IS*) if

$$\sqrt{\frac{\lambda_s}{\theta_s}} \leq \frac{1}{1-\rho}, \qquad \text{(Impatient-Server)}$$

and has patient servers (*PS*) otherwise.

We define the corresponding subsets of $\mathcal{M}$:

$$\mathcal{M}_{IC} = \left\{ (\boldsymbol{\lambda}, \boldsymbol{\theta}) \in \mathcal{M} : \sqrt{\frac{\lambda_c}{\theta_c}} \leq \frac{\rho}{1-\rho} \right\} \text{ and}$$

$$\mathcal{M}_{IS} = \left\{ (\boldsymbol{\lambda}, \boldsymbol{\theta}) \in \mathcal{M} : \sqrt{\frac{\lambda_s}{\theta_s}} \leq \frac{1}{1-\rho} \right\}.$$

Customers and servers are patient if the no-impatience case ($\theta_c \downarrow 0$ and $\theta_s \downarrow 0$, respectively) provides a better approximation of a suitable performance metric than the no–excess capacity case ($\lambda_s \downarrow \lambda_c$ and $\lambda_c \uparrow \lambda_s$, respectively); they are impatient otherwise. For customers, the performance metric is the expected customer queue (conditional on customers having to wait), and for servers, it is the fraction of servers who abandon (or remain unused).

Recall that customer impatience acts on the *expected customer queue*, $\mathbb{E}[Q|Q \geq 0]$. As discussed after Theorem 1, the no-impatience case (the $M/M/1$ queue) provides a better approximation of $\mathbb{E}[Q|Q \geq 0]$ than the no-excess capacity case (the critically loaded $M/M/1+M$ queue) when $\frac{\rho}{1-\rho} < \sqrt{\frac{\lambda_c}{\theta_c}}$. Hence, customers are patient if $\frac{\rho}{1-\rho} < \sqrt{\frac{\lambda_c}{\theta_c}}$. Informally, this means that customers are willing to wait for a sufficient amount of time relative to the expected time to match so that abandonments have little impact on the expected customer queue.

Server impatience acts on the probability that customers have to wait, $\mathbb{P}(Q \geq 0)$. This probability is influenced by the *fraction of servers who abandon*. The percentage of servers who abandon is lower bounded by the abandonment in both the case where servers are infinitely patient (no impatience) and the case where the customer arrival rate is increased to the server arrival rate (no excess capacity). At least $\lambda_s - \lambda_c = \lambda_s(1-\rho)$ servers must abandon (or remain unused) because no more than λ_c matches can be performed per unit of time. As $\theta_s \downarrow 0$, keeping all else constant, the percentage of servers who remain unused will converge to $(1-\rho)$; this is the no-impatience bound. The expected number in a critically loaded $M(\lambda_s)/M(\lambda_s)/1+M(\theta_s)$ queue is, up to a multiplicative constant, equal to $\sqrt{\lambda_s/\theta_s}$. Thus, the percentage of servers who abandon in the no-excess capacity case is $\frac{\theta_s}{\lambda_s}\sqrt{\frac{\lambda_s}{\theta_s}} = \sqrt{\frac{\theta_s}{\lambda_s}}$. If $1-\rho > \sqrt{\frac{\theta_s}{\lambda_s}}$, the no-abandonment case provides a better lower bound on the fraction of servers who abandon; hence, servers are patient.

One should not expect that near the boundary of a regime, where the regime condition is held with equality, the performance will vary significantly; it will not. Instead, the point of Definition 2 is that certain "forces" become important as the market parameters transition from one regime to another. These forces have a more pronounced effect, as the parameters are farther into the interior of the regime. For example, when $\sqrt{\frac{\lambda_c}{\theta_c}} \ll \frac{\rho}{1-\rho}$, the effect of customer impatience is more pronounced.

For each of the four regimes, there is a simple approximation for the CoI that works for *all* parameters in that regime; these expressions are reported in Theorem 2.

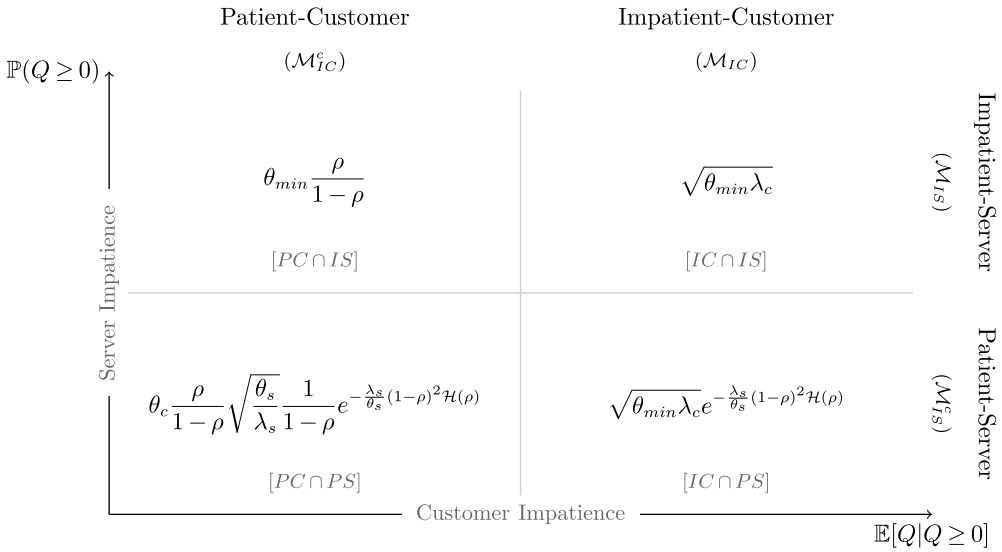
Theorem 2 (Cost-of-Impatience Scaling by Operating Regime). *Figure 4 characterizes the cost of impatience on the parameter set $\mathcal{M}$. In addition, in the Impatient-Server regime,*

$$\mathrm{CoI} \overset{\mathcal{M}_{IS}}{\sim} \theta_{min} \min\left\{ \frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_{min}}} \right\}.$$

At the boundaries between conditions (and corresponding parameter sets), the CoI expressions collapse into one expression. That is, if $(\boldsymbol{\lambda}, \boldsymbol{\theta})$ are such that customers and servers are both "critically" patient, which means that

$$(\boldsymbol{\lambda}, \boldsymbol{\theta}) \in \mathcal{M}_0 := \left\{ (\boldsymbol{\lambda}, \boldsymbol{\theta}) \in \mathcal{M} : \sqrt{\frac{\lambda_c}{\theta_c}} = \frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_s}{\theta_s}} = \frac{1}{1-\rho} \right\}, \qquad \text{(Critical-Impatience)}$$

Figure 4. CoI in the Four Operating Regimes



Notes. Each cell corresponds to an intersection of customer and server regimes and should be read with the appropriate "~" correspondence. For example, the upper-left cell is the statement that $\text{CoI} \overset{\mathcal{M}_{IC}^c \cap \mathcal{M}_{IS}}{\sim} A$, where A is the expression in that cell.

then

$$\text{CoI} \overset{\mathcal{M}_0}{\sim} \sqrt{\theta_{min}\lambda_c}. \tag{4}$$

Some observations are useful at this point.

Waiting servers as inventory. When *either* customers or servers are impatient, the CoI depends only on the *minimum* of the two patience parameters (or the maximum of the mean patience). We show in Theorem 1 that $\text{CoI} = \theta_c \mathbb{E}[Q|Q \geq 0]\mathbb{P}[Q \geq 0] \sim \theta_c \min\left\{\frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_c}}\right\}\mathbb{P}[Q \geq 0]$. The fact that θ_{min} appears in Theorem 2, rather than θ_c, is because of the ability to accumulate inventory of waiting servers.

Notice that θ_{min} can be either θ_c or θ_s, even when customers are patient. For example, if utilization is very low (i.e., $\rho \downarrow 0$) and both customers and servers have low mean patience (i.e., $\theta_c > \lambda_c$ and $\theta_s > \lambda_s$), then customers are patient and servers are impatient. The CoI in the $PC \cap IS$ regime is determined by θ_{min}, which could be either θ_c or θ_s.

If $\theta_c \leq \theta_s$ in the impatient-server regimes ($PC \cap IS$ and $IC \cap IS$), then $\text{CoI} \sim \theta_c \min\left\{\frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_c}}\right\}$; the CoI is proportional to the abandonment rate from the single-sided customer queue. In this case, there is little or no inventory of waiting servers, and the server-side queue does not have a scaling effect on the CoI. Conversely, if $\theta_s < \theta_c$, then the inventory of waiting servers, even if low, reduces the probability that customers must wait. The CoI scales like the abandonment rate from the single-sided customer queue, but with θ_c replaced by the *minimum* abandonment rate, θ_s.

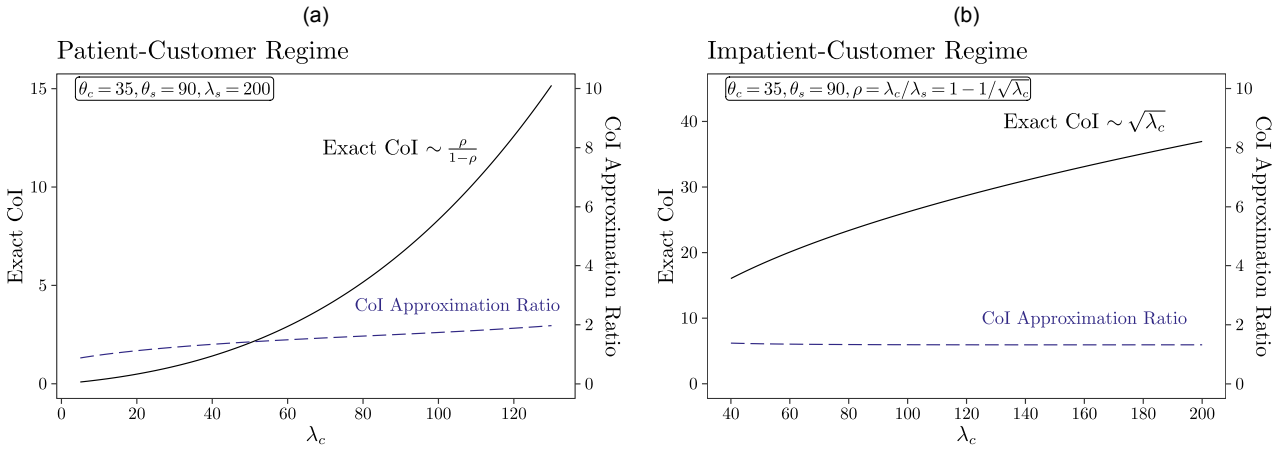
When servers are patient (the $PC \cap PS$ and $IC \cap PS$ regimes), inventory is high, and the CoI is offset (decreases) by a function of the server's level of impatience: $e^{-\frac{\lambda_s}{\theta_s}(1-\rho)^2\mathcal{H}(\rho)}$. This term captures the significant benefit of server inventory. Note that $(1-\rho)^2\mathcal{H}(\rho) = \sum_{n=1}^{\infty}\frac{1}{n(n+1)}(1-\rho)^{n+1}$ is between zero (when $\rho = 1$) and one (when $\rho = 0$). Therefore, $e^{-\frac{\lambda_s}{\theta_s}(1-\rho)^2\mathcal{H}(\rho)} \in [e^{-\frac{\lambda_s}{\theta_s}}, 1]$ decreases when either utilization decreases or the server's arrival-to-patience ratio, λ_s/θ_s, increases.

When *both* customers and servers are patient ($PC \cap PS$ regime), the CoI is very low and is sensitive to small changes in all parameters. Hence, the patience rates of both the customer and the server appear in the CoI scaling.

Overview of the key determinants of the CoI. Theorem 2 exposes the two key determinants of the CoI: (i) the "competition" between customer impatience and excess capacity, represented by the $\rho/(1-\rho)$ term in the PC regimes versus the square-root term in the IC regimes, and (ii) the ability to accumulate inventory of waiting servers, as reflected by the minimum patience rate in three regimes and the exponential term in the PS regimes.

- **Customer impatience versus excess capacity.** Suppose that the patience rates θ_c and θ_s are fixed. In the $PC \cap IS$ regime, $\text{CoI} \sim \frac{\rho}{1-\rho}$; the CoI is proportional to the number of customers in an $M/M/1$ queue and is impacted by small changes in either λ_c or λ_s. In the $IC \cap IS$ regime, $\text{CoI} \sim \sqrt{\lambda_c}$; the CoI scales like the square root of λ_c and is not as sensitive to small changes in λ_s; see Figure 5.

Figure 5. (Color online) With Fixed Patience Rates, CoI ~ $\frac{\rho}{1-\rho}$ When Customers Are Patient and CoI ~ $\sqrt{\lambda_c}$ When Customers Are Impatient



• **The minimum patience rate.** Suppose that the arrival rates, λ_c and λ_s, are fixed. In the $PC \cap IS$ regime, CoI ~ θ_{min}, whereas in the $IC \cap IS$ regime, CoI ~ $\sqrt{\theta_{min}}$. Small changes to the *maximum* patience rate (the minimum mean patience) do not have an order-of-magnitude impact on the CoI; see Figure 6. Small changes in the *minimum* patience rate have a more pronounced impact on the CoI when customers are patient versus impatient; see Figure 7. This is because when the CoI is low, which is the case in the patient-customer regimes ($PC \cap PS$ and $PC \cap IS$) relative to the impatient-customer regimes ($IC \cap PS$ and $IC \cap IS$), the CoI is more sensitive to small changes in the parameters.

• **Inventory of patient servers.** Suppose that the utilization, ρ, and the patience rates, θ_c and θ_s, are fixed; only λ_c and λ_s can change. Then in the $IC \cap IS$ regime, CoI ~ $\sqrt{\lambda_c}$, while in the $PC \cap PS$ regime, CoI ~ $\sqrt{\frac{1}{\lambda_s}}e^{-\kappa\lambda_s}$ for the constant $\kappa = \frac{1}{\theta_s}(1-\rho)^2$. In the $IC \cap IS$ regime, the CoI increases with a proportional increase in both λ_c and λ_s. However, in the $PC \cap PS$ regime, the CoI decreases with a proportional increase in λ_c and λ_s; see Figure 8. The decrease in the CoI when servers are patient is because of the greater inventory of servers when both λ_c and λ_s increase; this benefit is not realized when servers are impatient.

Figures 5–8 confirm that the CoI approximation in Theorem 1 captures the scaling of the true CoI up to a constant. Across all parameters in these examples, the ratio of the CoI approximation to the exact CoI (the CoI approximation ratio) is below two. Recall that the constant Γ in (3) does not depend on $(\boldsymbol{\lambda}, \boldsymbol{\theta})$ but does depend on M, which defines the queue family $\mathcal{M}$ in

Figure 6. (Color online) The Maximum Patience Rate Has a Moderate Impact on the CoI, Whereas the Minimum Patience Rate Has a Significant Impact on the CoI

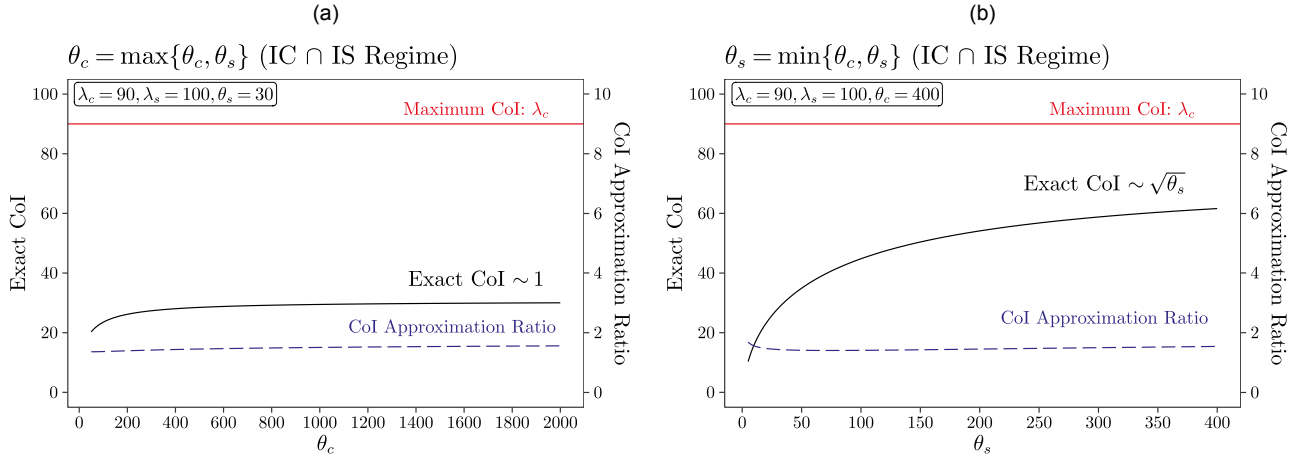
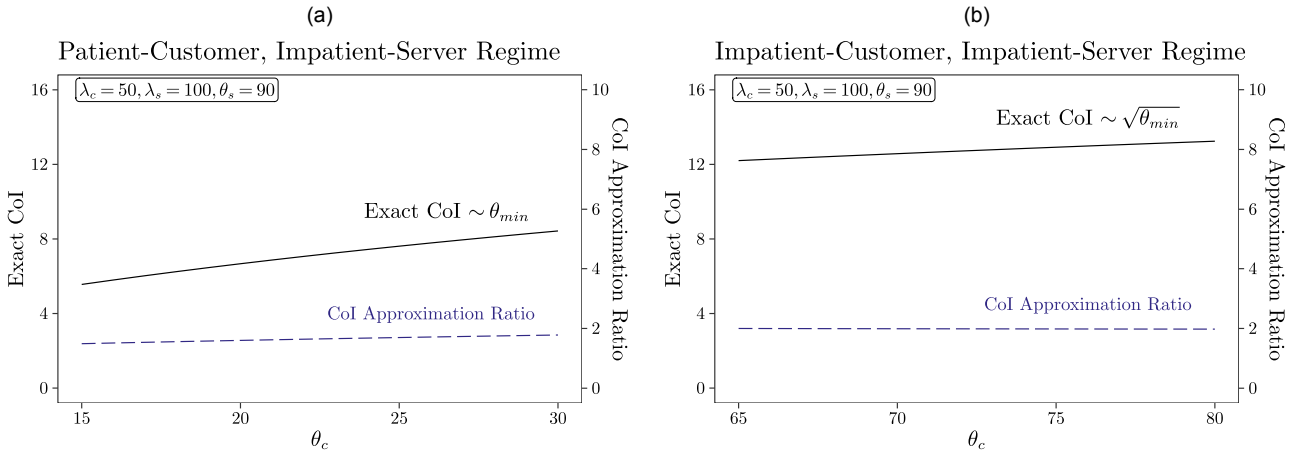


Figure 7. (Color online) The CoI Is More Sensitive to Changes in θ_{min} in the $PC \cap IS$ Regime than in the $IC \cap IS$ Regime



Definition 1. For $M \leq 20$, the approximation ratio is less than four for any combination of parameters $(\boldsymbol{\lambda}, \boldsymbol{\theta})$; see Figure 9.

The four operating regimes vis-à-vis known asymptotic regimes. Consider a sequence of two-sided queues, indexed by n, where $\boldsymbol{\lambda^n} = (\lambda_c^n, \lambda_s^n)$ and $\boldsymbol{\theta^n} = (\theta_c^n, \theta_s^n)$ are the parameters in the nth queue. The heavy-traffic regime studied in Liu et al. (2015) is the one where $\boldsymbol{\theta^n}$ is scaled down, while $\boldsymbol{\lambda^n}$ and $\sqrt{n}(1-\rho^n)$ both approach a constant as n increases: $n\boldsymbol{\theta^n} \to \boldsymbol{\theta}$, $\boldsymbol{\lambda^n} \to \boldsymbol{\lambda}$, and $\sqrt{n}(1-\rho^n) \to \beta \in [0, \infty)$ as $n \uparrow \infty$.[4] This scaling can be equivalently written as

$$\sqrt{\frac{\boldsymbol{\lambda^n}}{\boldsymbol{\theta^n}}}(1-\rho^n) = \sqrt{\frac{\boldsymbol{\lambda^n}}{n\boldsymbol{\theta^n}}}\sqrt{n}(1-\rho^n) \to \sqrt{\frac{\boldsymbol{\lambda}}{\boldsymbol{\theta}}}\beta = \beta' \in [0, \infty).$$

In particular, for all n large enough,

$$\sqrt{\frac{\lambda_c^n}{\theta_c^n}} \approx \frac{\beta'}{\rho^n}\frac{\rho^n}{1-\rho^n} = \beta''\frac{\rho^n}{1-\rho^n} \text{ for } \beta'' \in [0, \infty),$$

$$\text{and } \sqrt{\frac{\lambda_s^n}{\theta_s^n}} \approx \beta'\frac{1}{1-\rho^n}.$$

Under this heavy-traffic assumption, the spectrum of regimes collapses, then, to the Critical-Impatience case in (4) with $\boldsymbol{\theta}$ close to zero and ρ close to one.[5]

4.2. The $M/M/1+M$ Queue as a Special Case

In the special case that either $\theta_c = \infty$ or $\theta_s = \infty$, the two-sided queue becomes a single-sided queue. When $\theta_s = \infty$, the single-sided queue is under- or critically

Figure 8. (Color online) A Proportional Increase in λ_c and λ_s, Keeping ρ and $\boldsymbol{\theta}$ Fixed, Results in a Decrease in the CoI When Servers Are Patient, but an Increase in the CoI When Servers Are Impatient

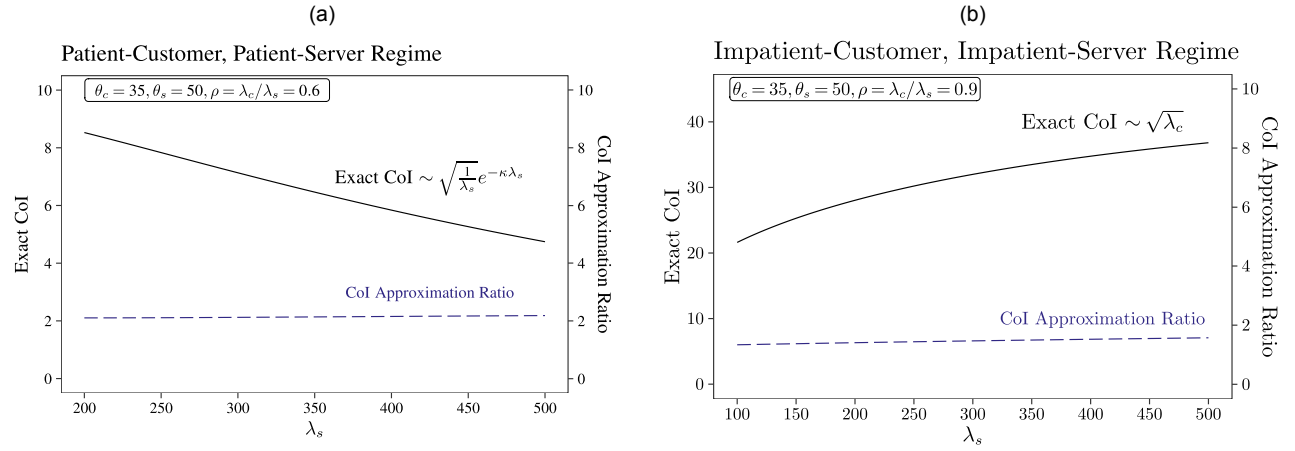
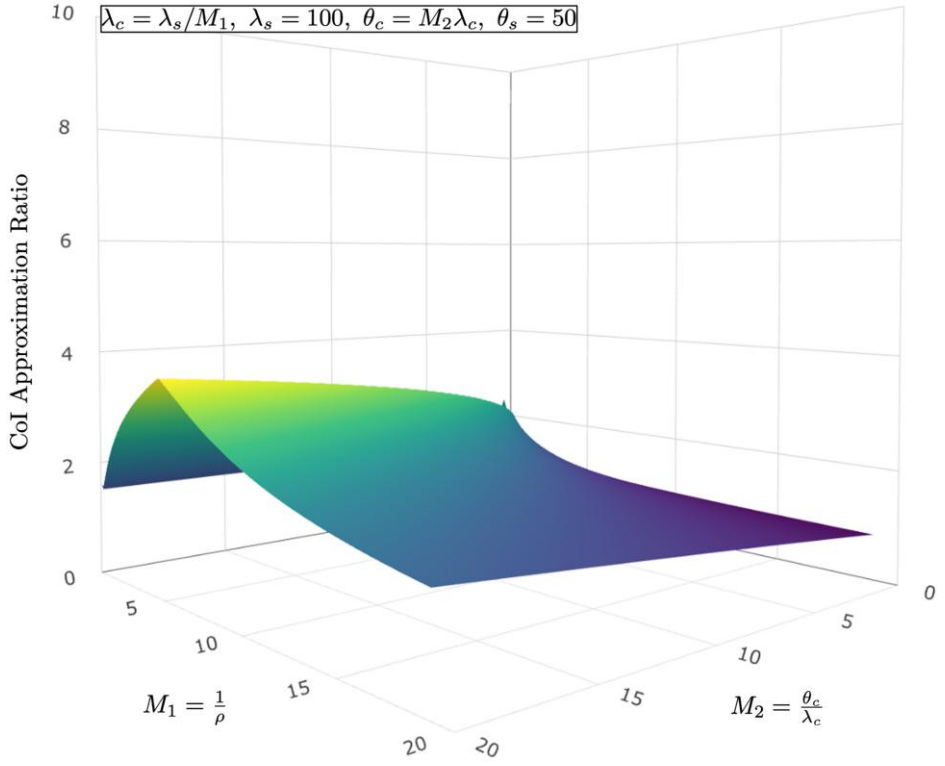


Figure 9. (Color online) The CoI Approximation Ratio (the CoI Approximation from Theorem 1 Divided by the Exact CoI) Is Less Than Four When the Conditions in Definition 1, $\rho \geq 1/M$ and $\theta_c \leq M\lambda_c$, Are Satisfied for $M \in [1, 20]$



loaded (has utilization $\rho \leq 1$); when $\theta_c = \infty$, the single-sided queue is over- or critically loaded (has utilization $1/\rho \geq 1$).

Consider first the case that $\theta_s = \infty$ (servers abandon immediately upon arrival if there are no customers in the queue to match). In this case, $Q_s \equiv 0$ so that $Q = Q_c - Q_s = Q_c$. Let Q_c^+ denote the number of customers in the system in an $M(\lambda_c)/M(\lambda_s)/1 + M(\theta_c)$ queue. The difference between Q_c and Q_c^+ is that Q_c^+ includes up to one customer in service, and the customer in service does not abandon; see Figure 10. Q_c is equal in distribution to the number in the queue (not including in service) in an $M(\lambda_c)/M(\lambda_s)/1 + M(\theta_c)$ queue, conditional on the server being busy: $Q_c \overset{d}{=} Q_c^+ - 1 | Q_c^+ \geq 1$. The abandonment rate from the $M(\lambda_c)/M(\lambda_s)/1 + M(\theta_c)$ queue is

$$\text{CoI} = \theta_c \mathbb{E}[(Q_c^+ - 1)^+].$$

If, instead, $\theta_c = \infty$, the two-sided queue reduces to Q_s. Let Q_s^+ denote the number in the system in an $M(\lambda_s)/M(\lambda_c)/1 + M(\theta_s)$ queue. Because the match rate is $d = \lambda_s - \theta_s \mathbb{E}[(Q_s^+ - 1)^+]$, the CoI for an $M(\lambda_s)/M(\lambda_c)/1 + M(\theta_s)$ queue is the server abandonment rate minus excess capacity:

$$\text{CoI} = \theta_s \mathbb{E}[(Q_s^+ - 1)^+] - (\lambda_s - \lambda_c).$$

To establish the CoI scaling for the $M(\lambda_s)/M(\lambda_c)/1 + M(\theta_s)$ queue, condition (ii) in Definition 1 is replaced with

$$\theta_s \leq M\lambda_s. \tag{1(ii′)}$$

As is the case for condition (ii), (1(ii′)) ensures that the queue does not degenerate (i.e., the patience rate is not significantly greater than the arrival rate).

We define $\mathcal{M}_{\theta_s=\infty}$ as the set of parameters $(\lambda_c, \lambda_s, \theta_c)$ that satisfy conditions (i) and (ii), and $\mathcal{M}_{\theta_c=\infty}$ as the set of parameters $(\lambda_c, \lambda_s, \theta_s)$ that satisfy conditions (i) and 1(ii′):

$$\mathcal{M}_{\theta_s=\infty} := \left\{(\lambda_c, \lambda_s, \theta_c) \geq 0 : \frac{\lambda_s}{\lambda_c} \leq M,\ \frac{\theta_c}{\lambda_c} \leq M\right\},$$
$$\mathcal{M}_{\theta_c=\infty} := \left\{(\lambda_c, \lambda_s, \theta_s) \geq 0 : \frac{\lambda_s}{\lambda_c} \leq M,\ \frac{\theta_s}{\lambda_s} \leq M\right\}.$$

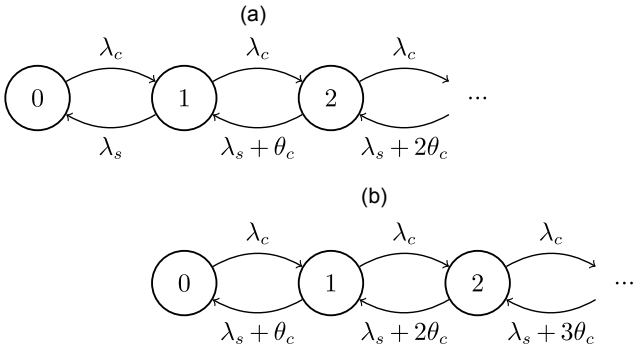
Lemma 1 (Cost-of-Impatience Scaling for the $M/M/1 + M$ Queue). *There exists $\Gamma \geq 1$ (not dependent on $\boldsymbol{\lambda}, \boldsymbol{\theta}$) such that*

$$\frac{1}{\Gamma}\mathcal{S}(\boldsymbol{\lambda}, \boldsymbol{\theta}) - \theta_{min} \leq \text{CoI} \leq \mathcal{S}(\boldsymbol{\lambda}, \boldsymbol{\theta}),$$

where the function $\mathcal{S}(\boldsymbol{\lambda}, \boldsymbol{\theta})$ is, on the parameter sets $\mathcal{M}_{\theta_s=\infty}$ and $\mathcal{M}_{\theta_c=\infty}$, as in Table 2.

Alignment with known results for the $M/M/1 + M$ queue. Ward and Glynn (2003) study the $M/M/1 + M$ queue in various asymptotic regimes. These asymptotic

Figure 10. The Transition-Rate Diagram for the $M(\lambda_c) = M(\lambda_s) = 1 + M(\theta_c)$ Queue and the Single-Sided Truncation of the Two-Sided Queue When $\theta_s = \infty$



regimes apply to settings where the traffic intensity (arrival rate divided by service rate) is close to or greater than one and the patience rate is close to zero.

The heavy-traffic limit for the under-/critically loaded queue in Ward and Glynn (2003) is the same as that of a single-server queue without abandonment when, using our notation, $\sqrt{\theta_c} \ll 1 - \rho$. This is consistent with our result: we prove that universally in λ_c, θ_c, and $\lambda_s \geq \lambda_c$, $\mathbb{E}[Q_c^+] \sim \min\left\{\frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_c}}\right\}$ (see the proof of Lemma 1 in Online Appendix A). Hence, if $\frac{\rho}{1-\rho} \leq \sqrt{\frac{\lambda_c}{\theta_c}}$, then $\mathbb{E}[Q_c^+] \sim \frac{\rho}{1-\rho}$; the expected $M/M/1+M$ queue is proportional to that of a queue without abandonment.

In a suitably overloaded $M/M/1+M$ queue, the queue length process asymptotically centers around $\overline{q} := \frac{\lambda_s - \lambda_c}{\theta_s}$, and the stochastic fluctuations around $\overline{q}$ are of the order of $\sqrt{\lambda_s}$ (Ward and Glynn 2003, theorem 1, case 4). In turn, in this asymptotic setting, this queue rarely visits the empty state (zero). This corresponds to customers almost always finding an available server, which translates into very low CoI. How low is precisely captured by the CoI result for the over-/critically loaded $M/M/1+M$ queue in Table 2. This asymptotic setting corresponds to, using our notation, $\rho^n \to \rho < 1$ as $\boldsymbol{\lambda}^n$ grows large. Our nonasymptotic CoI expression shows that in this setting, the CoI converges to zero at a rate that is subexponential in $\boldsymbol{\lambda}^n$.

Operating regimes for the $M/M/1+M$ queue. There are only two relevant operating regimes for each of the single-sided $M/M/1+M$ queues. The under-/critically loaded $M/M/1+M$ queue operates in either the Patient-Customer or Impatient-Customer regimes in Definition 2. When customers are patient, CoI $\sim \theta_c \frac{\rho}{1-\rho}$; when customers are impatient, CoI $\sim \sqrt{\theta_c \lambda_c}$. Thus, the CoI is completely determined by the competition between customer impatience and excess capacity, as discussed after Theorem 1. The over-/critically loaded $M/M/1+M$ queue operates in either the Patient-Server or Impatient-Server regimes in Definition 2. Note that the arrival rate in this queue is λ_s; the servers in the two-sided queue become the customers in the single-sided queue. In the Patient-Server regime, CoI $\sim \sqrt{\theta_s \lambda_c} e^{-\frac{\lambda_s}{\theta_s}(1-\rho)^2 \mathcal{H}(\rho)}$. In the Impatient-Server regime, $e^{-\frac{\lambda_s}{\theta_s}(1-\rho)^2 \mathcal{H}(\rho)} \sim 1$, and so, CoI $\sim \sqrt{\theta_s \lambda_c}$.[6]

Table 2. $\mathcal{S}(\boldsymbol{\lambda}, \boldsymbol{\theta})$ for the Under-/Critically Loaded and Over-/Critically Loaded $M/M/1+M$ Queues

Under-/Critically Loaded $M(\lambda_c)/M(\lambda_s)/1+M(\theta_c)$ queue	Over-/Critically Loaded $M(\lambda_s)/M(\lambda_c)/1+M(\theta_s)$ queue
$\theta_c \min\left\{\frac{\rho}{1-\rho}, \sqrt{\frac{\lambda_c}{\theta_c}}\right\}$	$\sqrt{\theta_s \lambda_c} e^{-\frac{\lambda_s}{\theta_s}(1-\rho)^2 \mathcal{H}(\rho)}$

5. Optimal Capacity Scaling

The results in Section 4 highlight two key controls for decreasing match loss: decrease the minimum patience rate $\theta_{min} = \min\{\theta_c, \theta_s\}$ and increase supply (thereby decreasing utilization and increasing inventory). In this section, we focus on the latter. An increase in server capacity may decrease customer abandonment, but there is a trade-off between capacity and abandonment costs when both are costly.

We use the scaling laws in Section 4 to study the scaling of optimal capacity in the two-sided queue with abandonment and contrast it with the optimal capacity in a single-sided queue. We are interested in understanding how the optimal capacity scales as the abandonment cost grows relative to the capacity cost. When the abandonment cost is sufficiently large relative to the capacity cost, it is optimal to have supply that is greater than demand. Therefore, we focus on settings where supply exceeds demand.

The optimal capacity, λ_s^*, solves the optimization problem:

$$\lambda_s^* = \arg\min_{\lambda_s \geq \lambda_c} \{c_a \theta_c \mathbb{E}[Q_c] + c_s \lambda_s\},$$

where c_a and c_s denote the per-unit cost of abandonment and capacity, respectively.

For simplicity and focus, we assume in this section that there is a single patience rate, $\theta_\bullet := \theta_c = \theta_s < \infty$. The optimal capacity-scaling results for $\theta_c \neq \theta_s$ appear in Online Appendix D.

Before we proceed, we expand the correspondence $\overset{\mathcal{M}}{\sim}$ to include the cost vector $\boldsymbol{c} = (c_a, c_s)$:

$$g(\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{c}) \overset{\mathcal{M}}{\sim} f(\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{c})$$

when there exists a constant, $\Gamma \geq 1$, that *does not depend on any of* $\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{c}$ such that

$$\frac{1}{\Gamma} \times f(\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{c}) \leq g(\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{c}) \leq \Gamma \times f(\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{c}), \text{ for all } (\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{c}) \in \mathcal{M},$$

where $\mathcal{M}$ is the set of all parameters that satisfy certain restrictions. We replace condition (i) in Definition 1 with

$$\frac{c_a}{c_s} \le M\sqrt{\frac{\lambda_c}{\theta_\bullet}}. \tag{1(i')}$$

In line with (i), (1(i')) restricts our attention to settings where the server capacity is constrained enough to prevent the two-sided queue from "collapsing" into a single-sided queue. This condition guarantees that $\lambda_s^* \le M\lambda_c$ (see Lemma A.7 in Online Appendix B) by ensuring that the abandonment cost and patience rate are not significantly larger than the supply cost and demand and, therefore, that it is not optimal to have capacity that is significantly larger than demand.

Henceforth, we consider the family of parameters

$$\mathcal{M} = \mathcal{M}(M) := \left\{(\lambda_c, \theta_\bullet, c_a, c_s) \ge 0 : \frac{\theta_\bullet}{\lambda_c} \le M, \frac{c_a}{c_s} \le M\sqrt{\frac{\lambda_c}{\theta_\bullet}}\right\}.$$

Finally, we use the abbreviated notation $\beta = (\lambda_c, \theta_\bullet, c_a, c_s)$; $\lambda_s^*(\beta)$ is the optimal capacity when the parameter vector is β.

Lemma 2 (Optimal Capacity Scaling). *The optimal safety capacity for a two-sided matching queue with patience rate $\theta_\bullet = \theta_c = \theta_s$ satisfies*

$$\lambda_s^*(\beta) - \lambda_c \overset{\mathcal{M}}{\sim} \gamma\sqrt{\theta_\bullet \lambda_c}$$

for $\gamma \ge 0$ that does not depend on $\boldsymbol{\lambda}$ or $\boldsymbol{\theta}$ and is characterized by

$$\gamma e^{\gamma^2} = \frac{c_a}{c_s}.$$

In particular,

$$\gamma \le \max\left\{1, \sqrt{\log(c_a/c_s)}\right\}.$$

For contrast, we state the scaling of the optimal safety capacity in the single-sided queue with abandonment. The optimal capacity, λ_s^*, for the $M(\lambda_c)/M(\lambda_s)/1 + M(\theta_c)$ queue solves the optimization problem

$$\lambda_s^* = \arg\min_{\lambda_s \ge \lambda_c}\{c_a\theta_c\mathbb{E}[(Q_c^+ - 1)^+] + c_s\lambda_s\}.$$

Going forward, we use the notation $\beta = (\lambda_c, \theta_c, c_a, c_s)$ and consider the family of parameters that satisfy conditions (ii) and 1(i'):

$$\mathcal{M}_{\theta_s=\infty} := \left\{(\lambda_c, \theta_c, c_a, c_s) \ge 0 : \frac{\theta_c}{\lambda_c} \le M, \ \frac{c_a}{c_s} \le M\sqrt{\frac{\lambda_c}{\theta_c}}\right\}.$$

Lemma 3 (Optimal Capacity Scaling for the $M/M/1+M$ Queue). *The optimal safety capacity for an $M(\lambda_c)/M(\lambda_s)/1 + M(\theta_c)$ queue satisfies*

$$\lambda_s^*(\beta) - \lambda_c \overset{\mathcal{M}_{\theta_s=\infty}}{\sim} \gamma\sqrt{\theta_c \lambda_c},$$

for $\gamma \ge 0$ that does not depend on $\boldsymbol{\lambda}$ or $\boldsymbol{\theta}$. If $c_a \ge c_s$, then

$$\gamma = \sqrt{c_a/c_s}.$$

Slow scaling of capacity in the two-sided queue. In the $M(\lambda)/M(\mu)/1$ queue, the service rate that minimizes total linear waiting and capacity costs takes a square root form. That is, $\mu^* - \lambda = \sqrt{c_w/c_s}\sqrt{\lambda}$, where c_w and c_s denote the waiting and capacity costs, respectively (e.g., Allon and Van Mieghem 2010). Two facts about the $M/M/1$ queue are important here. First, the safety capacity grows proportional to the square root of demand (customer arrival rate) for fixed cost parameters. Second, the safety capacity grows proportional to the square root of the ratio of cost coefficients for fixed demand. Lemma 3 states that the optimal service rate in the $M/M/1+M$ queue also scales proportionally to the square root of the ratio of cost coefficients, $\sqrt{c_a/c_s}$.

Lemma 2 shows that the optimal safety capacity for a two-sided queue scales proportionally to the square root of demand and the square root of the patience rate. But, in contrast to the $M/M/1$ and $M/M/1+M$ queues, it scales proportionally to $\log(c_a/c_s)$ rather than $\sqrt{c_a/c_s}$; see Figure 11. It is because of the ability to accumulate inventory of servers in the two-sided queue that safety capacity is substantially smaller for high abandonment costs.

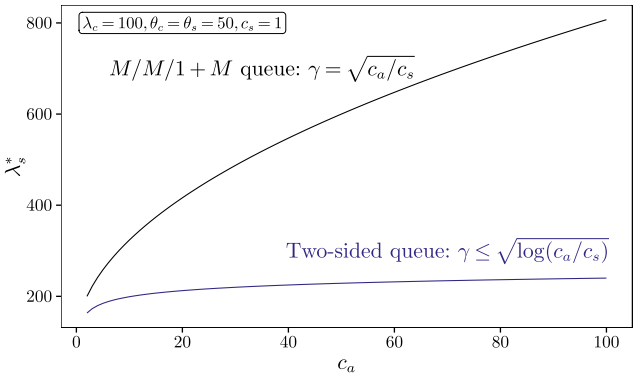
Connection to capacity scaling in the *M/M/N* queue. (Borst et al. 2004) proved that the asymptotically optimal safety capacity for an $M/M/N$ queue (a single-sided queue with N servers) has a logarithmic scaling in the cost ratio (the hourly cost of delay divided by the hourly cost per server). It is not a coincidence that the two-sided queue with abandonment shares this feature with the many-server queue; for a suitable choice of parameters, both models have the same diffusion approximation.

5.1. The Optimal Operating Regime

The optimal operating regime for a given set of parameters, $\lambda_c, \theta_c, \theta_s, c_a, c_s$, can be determined by identifying the optimal capacity, λ_s^*, and the corresponding operating regime using Definition 2. We identify the optimal operating regime for any combination of parameters in Online Appendix E. Any of the four operating regimes can be optimal, depending on the relationship between the cost, abandonment, and demand parameters; see Figure 12.

For instance, it is scale optimal to operate in the *IC* ∩ *IS* regime if $\frac{c_a}{c_s} \le \min\left\{1 + \sqrt{\frac{\theta_c}{\theta_s}}e^{\frac{\theta_c}{\theta_s}}, \sqrt{\frac{\theta_s}{\lambda_c}}e^{\frac{\theta_s}{\lambda_c}}\right\}$. This is consistent with intuition: when servers are very patient ($\theta_s \downarrow 0$) or when the abandonment cost is lower than the supply cost, it is optimal to operate with low safety

Figure 11. (Color online) The Optimal Safety Capacity for a Two-Sided and Single-Sided Queue with Abandonment Has the Form $\lambda_s^* - \lambda_c \sim \gamma\sqrt{\theta_c \lambda_c}$ for $\gamma \geq 0$ That Does Not Depend on $\boldsymbol{\lambda}$ or $\boldsymbol{\theta}$



Note. The ability to hold inventory of servers in the two-sided queue allows the safety capacity to grow logarithmically in abandonment cost c_a/c_s, slower than the square root growth in the single-sided queue.

capacity (high utilization, $IC \cap IS$ regime). However, when customers are very impatient ($\theta_c \uparrow \infty$) and the abandonment cost is very high relative to the supply cost, it is optimal to operate with high safety capacity (low utilization, $IC \cap PS$ regime). When the customer and server patience rates θ_c and θ_s are not too large relative to the customer arrival rate ($\theta_c \approx \theta_s < \lambda_c$), it is optimal to operate with high safety capacity ($PC \cap PS$ regime) if the abandonment cost is high enough relative to the supply cost, but it is optimal to operate with low safety capacity ($PC \cap IS$ regime) if the abandonment cost is not too high relative to the supply cost.

6. Conclusion

In this paper, we establish a universal scaling law for the match loss in a two-sided matching queue with abandonment. The scaling law provides direct insights into how abandonment impacts the match rate for any model parameters. Our results are nonasymptotic and hold for any arrival, utilization, and mean patience.

Any matching queue operates in one of four operating regimes, which are determined by the level of customer and server impatience. The level of impatience is an operational measure that brings together mean patience and utilization. This characterization shows, in simple terms, how relative customer impatience and the ability to accumulate server inventory impact match loss for each operating regime.

There are interesting questions that remain to be studied within the two-sided queue (single-match) setting. We illustrated the use of the scaling laws by studying a simple capacity optimization problem. Pricing in a two-sided queue with abandonment is a natural next step. This is the "dual" problem where one optimizes demand instead of capacity.

The study of control problems in more complex matching networks may benefit from our characterization of operational regimes; some regimes might allow for simpler controls than others. This "regime sensitivity" has precedent in the queueing network literature. For example, the optimal policy for a many-

Figure 12. (Color online) The Optimal Operating Regime for Each Set of Parameters $\lambda_c, \theta_c, \theta_s, c_a, c_s$, Where $\delta \frac{1}{\sqrt{\theta_c}}\left(1 + \sqrt{\frac{\theta_c}{\theta_s}} e^{\delta^2 \frac{1}{\theta_s}}\right) = \frac{c_a}{c_s}$

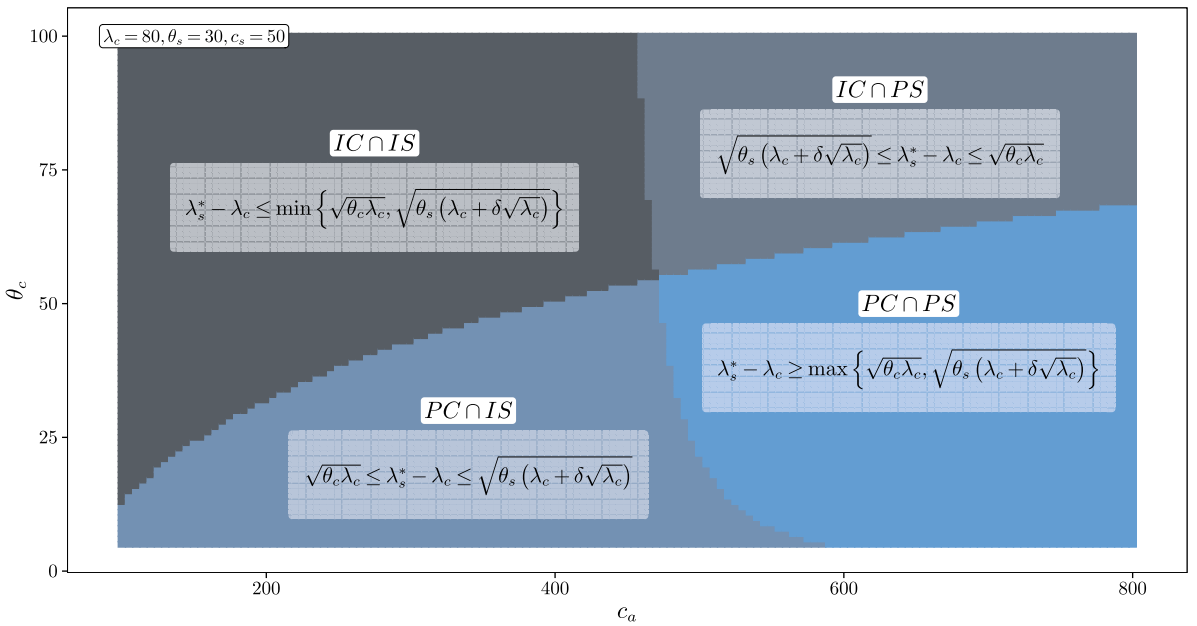
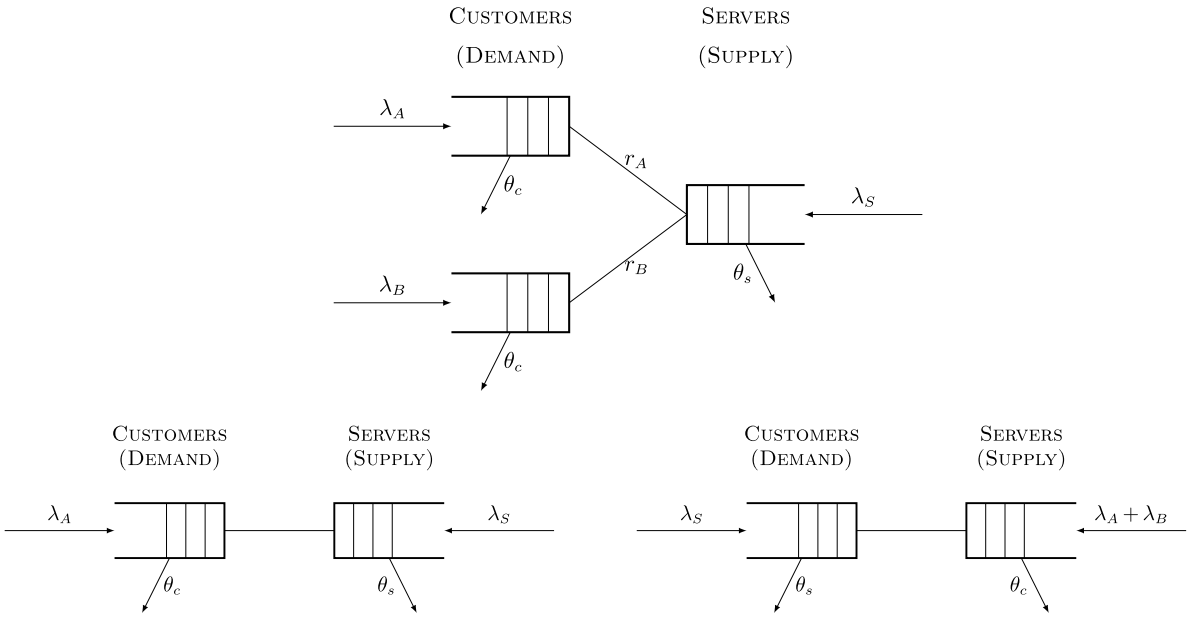


Figure 13. Model with Three Participant Types and Two Matches



server, multiclass queue is different depending on the regime: nondegenerate slowdown (Atar and Gurvich 2014) versus the Halfin-Whitt regime (Atar 2005). Beyond being conceptually useful through the characterization of regimes, the two-sided queue may feed into the mathematical arguments in the study of scale-optimal control of networks, specifically in the characterization of a lower bound on the reward that any matching policy can achieve.

For an informal illustration of this point, consider the simple network in Figure 13. This network has three types of participants: customers A and B and servers S. There are two matches with different rewards, r_A and r_B. Suppose that $r_A \geq r_B$ and $\lambda_A \leq \lambda_S \leq \lambda_A + \lambda_B$. A matching policy specifies when to match and which match to perform. The objective is to identify a policy that maximizes the long-run average reward.

Hidden in this network are two two-sided queues. The one on the bottom left of Figure 13 is composed of only customer A and the servers. In the other (on the bottom right), the servers take the role of customers, and their "servers" are both customers A and B.

A lower bound on match loss due to abandonment can be defined for each of the two-sided queues, per our Theorem 2, based on its operating regime. An overall order-of-magnitude lower bound on the reward loss due to abandonment is the sum of the two lower bounds. No policy can do order of magnitude better than this cumulative lower bound, as long as the parameters remain within the regimes identified. A policy that achieves the established lower bound is scale optimal.

The lower bounds, and the policies that attain them, will take on different forms depending on the underlying regime. As such, our operating regimes and the detailed arguments in our analysis provide a starting point for identifying policies for various network parameters.

Endnotes

[1] When considering ride hailing, our model is best suited for the study of single trips within a single location/neighborhood; hence, we only consider trips that begin and/or end in the Manhattan CBD, and we exclude any driver time spent outside of the CBD.

[2] In the special case that $\theta_s = \theta_c = \theta_\bullet$, $\delta = \gamma\sqrt{\theta_\bullet}$ for γ that depends only on c_a, c_s.

[3] This follows from the fact that $R_i(t) - \theta_i \int_0^t Q_i(s)ds$ is a Martingale; see, for example, Pang et al. (2007).

[4] Liu et al. (2015) allow for $\beta \in (-\infty, \infty)$. In our labeling $\lambda_s \leq \lambda_c$ so $\beta \geq 0$, but this is without loss of generality.

[5] We define Critical-Impatience with $\beta' = \beta'' = 1$, but the implication in (4) is the same for any constants $\beta'' \geq \beta' > 0$.

[6] In the $M/M/1+M$ queues, the CoI is proportional to the expressions in Table 2 when θ_{min} is sufficiently small ($\theta_{min} \leq \frac{1}{\Gamma} S(\boldsymbol{\lambda}, \boldsymbol{\theta})$ for Γ in the proof of Lemma 1); otherwise, the expression in Table 2 is an upper bound on the CoI.

References

Adan I, Kleiner I, Righter R, Weiss G (2018) FCFS parallel service systems and matching models. *Performance Evaluation* 127–128:253–272.

Afèche P, Caldentey R, Gupta V (2022) On the optimal design of a bipartite matching queueing system. *Oper. Res.* 70(1):363–401.

Afèche P, Diamant A, Milner J (2014) Double-sided batch queues with abandonment: Modeling crossing networks. *Oper. Res.* 62(5):1179–1201.

Allon G, Van Mieghem JA (2010) Global dual sourcing: Tailored base-surge allocation to near-and offshore production. *Management Sci.* 56(1):110–124.

Aouad A, Saritaç Ö (2022) Dynamic stochastic matching under limited time. *Oper. Res.* 70(4):2349–2383.

Ashlagi I, Bingaman A, Burq M, Manshadi V, Gamarnik D, Murphey C, Roth AE, Melcher ML, Rees MA (2018) Effect of match-run frequencies on the number of transplants and waiting times in kidney exchange. *Amer. J. Transplantation* 18(5):1177–1186.

Atar R (2005) Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* 15(4):2606–2650.

Atar R, Gurvich I (2014) Scheduling parallel servers in the nondegenerate slowdown diffusion regime: Asymptotic optimality results. *Ann. Appl. Probab.* 24(2):760–810.

Aveklouris A, DeValve L, Stock M, Ward A (2024) Matching impatient and heterogeneous demand and supply. *Oper. Res.*, ePub ahead of print May 15, https://doi.org/10.1287/opre.2022.0005.

Bar-Lev SK, Boxma O, Mathijsen B, Perry D (2017) A blood bank model with perishable blood and demand impatience. *Stochastic Systems* 7(2):237–263.

Borst S, Mandelbaum A, Reiman MI (2004) Dimensioning large call centers. *Oper. Res.* 52(1):17–34.

Boxma OJ, David I, Perry D, Stadje W (2011) A new look at organ transplantation models and double matching queues. *Probab. Engrg. Inform. Sci.* 25(2):135–155.

Büke B, Chen H (2017) Fluid and diffusion approximations of probabilistic matching systems. *Queueing Systems* 86:1–33.

Caldentey R, Kaplan EH, Weiss G (2009) FCFS infinite bipartite matching of servers and customers. *Advances Appl. Probab.* 41(3):695–730.

Castro F, Nazerzadeh H, Yan C (2020a) Matching queues with reneging: A product form solution. *Queueing Systems* 96(3–4):359–385.

Castro F, Frazier P, Ma H, Nazerzadeh H, Yan C (2020b) Matching queues, flexibility and incentives. Preprint, submitted July 17, http://dx.doi.org/10.2139/ssrn.3627920.

Chen Y, Hu M (2020) Pricing and matching with forward-looking buyers and sellers. *Manufacturing Service Oper. Management* 22(4):717–734.

Collina N, Immorlica N, Leyton-Brown K, Lucier B, Newman N (2020) Dynamic weighted matching with heterogeneous arrival and departure rates. Chen X, Gravin N, Hoefer M, Mehta R, eds. *Web and Internet Economics. WINE 2020*, Lecture Notes in Computer Science, vol. 12495 (Springer, Cham, Switzerland).

Conolly B, Parthasarathy P, Selvaraju N (2002) Double-ended queues with impatience. *Comput. Oper. Res.* 29(14):2053–2072.

Diamant A, Baron O (2019) Double-sided matching queues: Priority and impatient customers. *Oper. Res. Lett.* 47(3):219–224.

Elalouf A, Perlman Y, Yechiali U (2018) A double-ended queueing model for dynamic allocation of live organs based on a best-fit criterion. *Appl. Math. Modeling* 60:179–191.

Graves SC (1982) The application of queueing theory to continuous perishable inventory systems. *Management Sci.* 28(4):400–406.

Gurvich I, Ward A (2014) On the dynamic control of matching queues. *Stochastic Systems* 4(2):479–523.

Kaplan EH (1986) Tenant assignment models. *Oper. Res.* 34(6):832–843.

Kashyap BRK (1966) The double-ended queue with bulk service and limited waiting space. *Oper. Res.* 14(5):822–834.

Kendall DG (1951) Some problems in the theory of queues. *J. Royal Statist. Soc. Ser. B Methodological* 13(2):151–173.

Kerimov S, Ashlagi I, Gurvich I (2021) Dynamic matching: Characterizing and achieving constant regret. *Management Sci.* 70(5): 2799–2822.

Kerimov S, Ashlagi I, Gurvich I (2023) On the optimality of greedy policies in dynamic matching. *Oper. Res.*, ePub ahead of print September 12, https://doi.org/10.1287/opre.2021.0596.

Lee C, Ward AR (2019) Pricing and capacity sizing of a service facility: Customer abandonment effects. *Prod. Oper. Management* 28(8):2031–2043.

Liu X, Gong Q, Kulkarni VG (2015) Diffusion models for double-ended queues with renewal arrival processes. *Stochastic Systems* 5(1):1–61.

Nguyen LM, Stolyar AL (2018) A queueing system with on-demand servers: Local stability of fluid limits. *Queueing Systems* 89:243–268.

Özkan E, Ward AR (2020) Dynamic matching for real-time ride sharing. *Stochastic Systems* 10(1):29–70.

Pang G, Talreja R, Whitt W (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surveys* 4:193–267.

Perry D, Stadje W (1999) Perishable inventory systems with impatient demands. *Math. Methods Oper. Res.* 50(1):77–79.

Prendergast C (2017) How food banks use markets to feed the poor. *J. Econom. Perspect.* 31(4):145–162.

Varma SM, Maguluri ST (2021) A heavy traffic theory of two-sided queues. *Performance Evaluation Rev.* 49(3):43–44.

Varma SM, Bumpensanti P, Maguluri ST, Wang H (2022) Dynamic pricing and matching for two-sided queues. *Oper. Res.* 71(1):83–100.

Vaze R, Nair J (2022) Non-asymptotic near optimal algorithms for two sided matchings. *2022 20th Internat. Sympos. Modeling Optim. Mobile Ad hoc Wireless Networks (WiOpt)* (IEEE, Piscataway, NJ), 17–24.

Wang G, Zhang H, Zhang J (2022) On-demand ride-matching in a spatial model with abandonment and cancellation. *Oper. Res.* 72(3):1278–1297.

Ward AR (2012) Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys Oper. Res. Management Sci.* 17(1):1–14.

Ward AR, Glynn PW (2003) A diffusion approximation for a Markovian queue with reneging. *Queueing Systems* 43(1):103–128.

Yu Q, Zhang Y, Zhou YP (2022) Delay information in virtual queues: A large-scale field experiment on a major ride-sharing platform. *Management Sci.* 68(8):5745–5757.

Zenios SA (1999) Modeling the transplant waiting list: A queueing model with reneging. *Queueing Systems* 31:239–251.

Zubeldia M, Jhunjhunwala PR, Maguluri ST (2022) Matching queues with abandonments in quantum switches: Stability and throughput analysis. Preprint, submitted September 25, https://arxiv.org/abs/2209.12324.