

# A Heavy Traffic Theory of Two-Sided Queues

Sushil Mahavir Varma, Siva Theja Maguluri

School of Industrial and Systems Engineering, Georgia Institute of Technology

## ABSTRACT

Motivated by emerging applications in online matching platforms and marketplaces, we study a two-sided queue. Customers and servers that arrive into a two-sided queue depart as soon as they are matched. It is known that a state-dependent control is needed to ensure the stability of a two-sided queue. However, analytically studying the steady-state behaviour of a two-sided queue, in general, is challenging. Therefore, inspired by the heavy-traffic regime in classical queueing theory, we study a two-sided queue in an asymptotic regime where the control decreases to zero. There are two different ways the control can be sent to zero, and we model these using two parameters viz.,  $\epsilon$  that goes to zero and  $\tau$  that goes to infinity. Intuitively,  $\epsilon$  modulates the magnitude of the control and  $\tau$  is the threshold after which we modulate the control.

We show that depending on the relative rates of  $\epsilon$  and  $\tau$ , there is a phase transition in the limiting regime. We christen the regime when  $\epsilon\tau \rightarrow 0$ , the quality-driven regime, and the limiting behaviour is a Laplace distribution. The phase transition starts in the regime when,  $\epsilon\tau$  goes to a nonzero constant when the limiting distribution is a Gibbs distribution, and so we call it the critical regime. When  $\epsilon\tau \rightarrow \infty$ , we conjecture that the limiting distribution is uniform and prove that in a special case. We call this the profit-driven regime. These results are established using two related proof techniques. The first one is a generalization of the characteristic function method, and the second is a novel inverse Fourier transform method.

## 1 Introduction

Since the work of Erlang in telecommunication systems, more than a century ago, queueing theory has emerged as a well-established discipline that has had an impact on a large number of applications including wired and wireless networks, cloud computing, manufacturing and transportation systems, etc. The central building block of queueing theory is a single server queue, which has a fixed server, customers that wait until their service and then depart immediately thereafter. In addition, there is a queue or a waiting space for the customers to wait, and a stochastic model of the arrivals and services. While the single server queue is well-understood when the arrivals and service are memory-less, there is no closed-form expression for the stationary

distribution of the queue length for general distributions. Therefore, queueing systems are studied in various asymptotic regimes, including the heavy-traffic regime, where the arrival rate approaches the service rate.

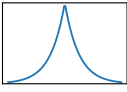
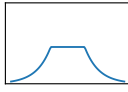
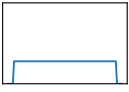
Recent developments in online platforms and matching markets such as ride-hailing, food delivery services, etc have led to an interest in the study of two-sided queues. In a two-sided queue, *both* servers and customers arrive, wait until they are matched, and then immediately depart the system. The behaviour of two-sided queues is different from that of classical queues. In particular, a two-sided queue is never stable without external control. To see this, note that if the arrival rates on the two sides don't match, the system is clearly unstable. But when the rates match, it is analogous to a symmetric random walk on a line, which is null-recurrent. Therefore, two-sided queues have to be always studied under an external control that modulates the arrival rates in a *state-dependent* manner with levers such as prices in online platforms. In contrast to a large amount of literature on classical queues, there is comparatively very little work on two-sided queues.

Analogous to classical queues, except in special cases, it is hard to obtain the exact stationary distribution of queue length in a two-sided queue. The goal of this paper is to develop a heavy-traffic theory of two-sided queues, that will enable us to completely characterize the queueing behaviour in an appropriately defined asymptotic regime. However, unlike in classical queues, there is no natural notion of 'heavy-traffic' here, and we overcome this challenge as follows. Suppose that the uncontrolled arrivals have an equal rate on both sides. Now, we study the system in the regime when the state-dependent control goes to zero. As mentioned before, in the regime when the control is zero, the system is null-recurrent, and the regime is reminiscent of the heavy-traffic regime of a single-server queue.

## 2 Model

We consider a two-sided queue operating in discrete-time with customers and servers both arriving in the system. A waiting customer is matched to a server (and vice versa) as soon as possible, and the pair instantaneously departs from the system. Therefore, both servers and customers cannot be waiting at the same time. The state of the system is the imbalance, i.e. the number of customers minus the number of servers waiting in the system.

Given the imbalance  $z$ , let  $\tilde{\phi}^c(z)$  and  $\tilde{\phi}^s(z)$  be the state-dependent control applied to the arrival rate of the customers and servers respectively. Given a time epoch  $k$  and

	Quality-Driven Regime $\epsilon\tau \rightarrow 0$	Critical Regime $\epsilon\tau \rightarrow (0, \infty)$	Profit-Driven Regime $\epsilon\tau \rightarrow \infty$
Bernoulli Arrivals $\phi^s(x) = -\mathbb{1}\{x < -1\}$ , $\phi^c(x) = -\mathbb{1}\{x > 1\}$ (Illustrative Example)	Laplace 	Hybrid 	Uniform 
General Arrival and General $(\phi^s(\cdot), \phi^c(\cdot))$	Laplace Characteristic Function Method	Gibbs Distribution Inverse Fourier Transform Method	Uniform Open Question

**Table 1: A Summary of Heavy-Traffic Phase Transitions in Two-Sided queues.**

the imbalance of the queue  $z(k) = z$ , the customer and server arrivals are denoted by random variables  $a^c(z, k)$  and  $a^s(z, k)$  with  $\mathbb{E}[a^c(z, k)] = \lambda^* + \tilde{\phi}^c(z)$  and  $\mathbb{E}[a^s(z, k)] = \mu^* + \tilde{\phi}^s(z)$  for all  $z \in \mathbb{Z}$  and  $k \in \mathbb{Z}_+$ . Moreover, the variances are a function of the mean and is denoted by  $\text{Var}[a^c(z, k)] \triangleq \sigma^c(\lambda^* + \tilde{\phi}^c(z))$  and  $\text{Var}[a^s(z, k)] \triangleq \sigma^s(\mu^* + \tilde{\phi}^s(z))$  for some continuous functions  $\sigma^c(\cdot)$  and  $\sigma^s(\cdot)$ . The imbalance is a discrete-time Markov chain (DTMC) with the evolution equation

$$z(k+1) = z(k) + a^c(z(k), k) - a^s(z(k), k).$$

Ideally, one would like to analytically obtain the exact distribution of the imbalance in the steady state. However, this is not possible in general, and so we study the two-sided queue in an asymptotic regime. In particular, we will consider a sequence of pricing policies parametrized by  $\eta$ . Thus, we have a sequence of functions  $(\phi_\eta^c(\cdot), \phi_\eta^s(\cdot))$  and the sequence of policies are such that  $\phi_\eta^c \rightarrow 0$ ,  $\phi_\eta^s \rightarrow 0$  uniformly as  $\eta \rightarrow \infty$ . In particular, we focus on the following family of policies that are characterized by two parameters,  $\epsilon_\eta > 0$  and  $\tau_\eta > 0$ .

$$\lambda_\eta(z) = \lambda^* + \epsilon_\eta \phi^c\left(\frac{z}{\tau_\eta}\right), \quad \mu_\eta(z) = \mu^* + \epsilon_\eta \phi^s\left(\frac{z}{\tau_\eta}\right)$$

for all  $z \in \mathbb{Z}$  and  $\eta > 0$ . Here,  $\phi^c(\cdot)$  and  $\phi^s(\cdot)$  are fixed bounded functions. Lastly, whenever the DTMC  $\{z_\eta(k)\}_{k \in \mathbb{Z}_+}$  is positive recurrent, denote the random variable with distribution same as its stationary distribution by  $\bar{z}_\eta$ .

The parameter  $\epsilon_\eta$  modulates the magnitude of the control, and by picking it such that  $\lim_{\eta \rightarrow \infty} \epsilon_\eta = 0$ , we let the control vanish. The influence of the parameter  $\tau_\eta$  is more subtle. It lets us tune the scale of the imbalance  $z$  at which we apply the control. Thus, if we let  $\lim_{\eta \rightarrow \infty} \tau_\eta = \infty$ , we end up applying no state-dependent control, and so this is equivalent to removing the control. Thus, we will study the two-sided queue when  $\lim_{\eta \rightarrow \infty} \epsilon_\eta = 0$  and  $\lim_{\eta \rightarrow \infty} \tau_\eta = \infty$ . The parameter  $\epsilon_\eta$  is similar to the heavy-traffic parameter in a classical single server queue. The parameter  $\tau_\eta$  is new and it appears because we use state-dependent control.

### 3 Phase Transition in Two-Sided Queues

We first impose mild conditions on  $(\phi^s(\cdot), \phi^c(\cdot))$  and then prove a phase transition result as summarized in Table 1.

**CONDITION 1.** (a) *Negative Drift:* For some  $\delta, K > 0$ ,  $\forall x > K$ ,  $\phi^c(x) - \phi^s(x) < -\delta$ ,  $\forall x < -K$ ,  $\phi^c(x) - \phi^s(x) > \delta$ .  
(b) *Smoothness:*  $\phi^c(\cdot)$ ,  $\phi^s(\cdot)$  are infinitely differentiable.  
(c) *Symmetry:*  $\phi^s(\infty) - \phi^c(\infty) = \phi^c(-\infty) - \phi^s(-\infty) = 1$ .

Let  $\{\epsilon_\eta\}_{\eta > 0}$  and  $\{\tau_\eta\}_{\eta > 0}$  be such that  $\lim_{\eta \rightarrow \infty} \epsilon_\eta \tau_\eta = l$ . We now present the main theorem of our paper below:

**THEOREM 1.** When  $l = 0$ , under Condition 1 (a), (c),

$$\epsilon_\eta \bar{z}_\eta \xrightarrow{D} \text{Laplace}\left(0, \frac{\sigma^c(\lambda^*) + \sigma^s(\mu^*)}{2}\right).$$

Also, when  $l \in (0, \infty)$ , under Condition 1 (a) (b), we have

$$\epsilon_\eta \bar{z}_\eta \xrightarrow{D} \text{Gibbs}(g_{1,l}), \quad \frac{\bar{z}_\eta}{\tau_\eta} \xrightarrow{D} \text{Gibbs}(g_{1,1}).$$

$$\text{where, } g_{b,c}(x) \triangleq \frac{2b}{\sigma^c(\lambda^*) + \sigma^s(\mu^*)} \left( \phi^s\left(\frac{x}{c}\right) - \phi^c\left(\frac{x}{c}\right) \right).$$

Due to technical challenges [2], we present a conjecture for the case  $l = \infty$ , and leave the proof as part of a future work.

**CONJECTURE 1.** Denote by  $\Phi^*$  the set of minimizers of  $\int_0^x (\phi^s(t) - \phi^c(t)) dt$ . When  $l = \infty$ , we have  $\bar{z}_\eta / \tau_\eta \xrightarrow{D} \mathcal{U}(\Phi^*)$

Observe that Gibbs distribution is a mixture of Laplace and Uniform. In particular, when  $l \downarrow 0$ ,  $g_{1,l}$  converges to a function that only depends on the sign of the imbalance which makes it the Laplace distribution and when  $l \rightarrow \infty$ , it converges to a constant everywhere which results in an ill-defined distribution, because it appears to be an infinitely spread uniform distribution. This is because  $\bar{z}_\eta = \Theta(\tau)$  and so, if the imbalance is scaled by  $\tau$  instead of  $\epsilon$ , we obtain a uniform distribution with support  $\Phi^*$ .

*Proof Idea:* Transform method was first introduced in [1], and was used to study queues under static arrival rates. Consequently, [1] directly gets a closed form expression for the characteristic function of the limiting distribution which immediately establishes convergence in distribution. In contrast, due to the dynamic arrivals, we obtain an *implicit* equation. A major methodological contribution of the paper is the introduction of the use of inverse Fourier transform to solve the implicit equation to obtain the limiting distribution. Moreover, due to the implicit equation, we have to separately establish the guarantee that our family converges in distribution. We believe that our proposed method will enable one to use transform techniques in a large class of stochastic network beyond the ones studied in [1].

### 4 References

- [1] D. Hurtado-Lange and S. T. Maguluri. Transform methods for heavy-traffic analysis. *Stochastic Systems*, 10(4):275–309, 2020.
- [2] S. M. Varma and S. T. Maguluri. A heavy traffic theory of two-sided queues, 2021.