Batch latency analysis and phase transitions for a tandem of queues with exponentially distributed service times

Jinho Baik*and Raj Rao Nadakuditi†

June 14, 2012

Abstract

We analyze the latency or sojourn time $L(m, n)$ for the last customer in a batch of n customers to exit from the m-th queue in a tandem of m queues. We consider the setting where the queues are initially empty and where the queues are in equilibrium before the batch of customers arrives at the first queue. We provide an analytical characterization of the distribution of $L(m, n)$ that is exact for every m and n, under the assumption that the queues have unlimited buffers and that each server has customer independent, exponentially distributed service times with an arbitrary, known rate.

We use asymptotic analysis to characterize the limiting distributions and to bring into sharp focus the existence of a phase transition in the latency distribution. The transition separates a regime in which one or more slow or bottleneck servers with service rates below a critical value affect and change the resulting limiting latency distribution from one in which the bottleneck servers are present but the limiting latency distribution is unaffected. We show that this phase transition can also be induced by varying the external arrival rate about the same critical value. This critical value depends, in a manner we make explicit, on the individual service rates, the number of jobs and the number of queues.

1 Introduction

Tandem queues are important models for production systems [22] and communication networks [40]. The fact that the output process of a server is the arrival process for the subsequent queue makes tandem queues difficult to analyze. There are several results in the literature on the waiting time [33] of customers in such queues [16, 37], moment generating functions [15], scaling properties [10], heavy traffic approximations [19] and bounds thereof [39, 26, 27], to list a few. These results capture the behavior of a single randomly selected customer. In contrast, we are interested in the sojourn time of a *batch* of n customers entering a system of m tandem queues with unlimited buffers that are either empty or in equilibrium; this was the setup considered by Glynn and Whitt in their seminal paper [17].

*Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, USA

†Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI, 48109, USA

This setting is considerably more difficult to analyze at a level that captures what happens as the "wave" of customers traverses the network. Of particular relevance to this work, is using the analysis to provide insights on what happens when a few bottleneck servers are particularly slow or when the external arrival process is relatively fast.

In this paper, we leverage recent results from directed last passage percolation (DLPP) and random matrix theory (e.g. [20, 30, 3, 9]) to show that the latency distribution equals that of the largest eigenvalue of a specially constructed random matrix (see Table 1). The connection between batch latency in a tandem of queues, and directed last passage percolation and random matrix theory was also discussed and used, for example, in [28, 29, 12, 23]. An immediate goal in deriving the distribution is to document in this paper a very general result attainable using the underlying theory. An additional goal of ours is to introduce the proof techniques (equivalence mappings in Sections 4, 5 and 6) to theorists who might find it useful for other queuing related questions. To that end, we have presented our results in a manner that unifies the empty queues and queues in equilibrium setting to highlight the generality of the techniques employed.

With exact latency distribution thus characterized, we employ $m, n \to \infty$ asymptotics to uncover a phase transition which separates a regime in which the presence of a one or more slow or bottleneck servers results in the latency different from that of the case with no slow servers. It is shown that, assuming there are finitely many slow servers, they do not affect the latency to the leading order unless their rates are smaller than a critical rate. In addition to the leading order asymptotics, we also consider the next order fluctuations and show that the fluctuations of the latency are normally distributed with variance $O(n)$ when sufficiently slow servers are present but the fluctuations are given by the complex Tracy-Widom distribution with variance $O(n^{2/3})$ when rates of the slow servers are above the critical value. From a rate optimization perspective, this analysis highlights the (asymptotically) vanishing nature of the end-to-end latency reduction realized when optimizing with respect to the service rates of a limited or $o(n)$ number of bottleneck servers. A similar phase transition occurs with respect to the external arrival rate. Our emphasis on uncovering the phase transitions in the presence of slow servers or when the arrival rate is what differentiates this work from related work in the DLPP literature.

The analytical characterization of the latency distribution and the phase transition phenomenon are the main contributions of this paper, which is organized as follows. The problems addressed in this paper are formally stated in Section 2 with the main results described in Section 3. In section 4 we interpret the latency analysis problem as a directed last passage percolation (DLPP) problems. The notion of the solvable DLPP models is introduced in Section 5. These models are called solvable since there are exact formulas for the distribution function of the last passage time. A connection to random matrix theory (RMT) is also discussed here. In section 6 we describe how to map the latency problem considered into a solvable DLPP problem. The proofs of the main results are provided in Section 7.

2 Statement of the problem

Consider a tandem of m queues associated with m servers labeled from left to right as $S_1, \dots, S_m$. We assume that the service time of customer C_j at server S_i is customer-independent and exponentially distributed with rate μ_i. Arrivals at each queue are processed in FIFO order and there are no external arrivals in the system. We assume that the queue buffers are infinitely long.

Suppose that a batch of n customers arrive at the queue 1. Let $L(m, n)$ denote the exit time of customer n from server m where the time is measured from the arrival of the batch of n customers into queue 1. In this paper we characterize the distribution of $L(m, n)$ for this system in both the finite and large n and m settings for the following initial conditions for the system:

- Problem A: The queues are empty. A batch of n customers are placed at the head of previously empty queue 1,

- Problem B: The queues are in equilibrium with a steady-state arrival process at each queue with rate α. A batch of n customers arrives as a part of the Poisson arrival process in the previously at-equilibrium queue 1 .

We assume that $\alpha < \mu_i$ for all i in Problem B so that the queue sizes remain bounded.

3 Main results

3.1 Exact distribution

It is a surprising fact that the exit time $L(m, n)$ is distributed as the largest eigenvalue of a random matrix in both initial conditions as stated next.

Theorem 3.1 (Exact distribution)**.** *For Problem A and B, we have that for all n and m,*

$$L(m, n) \overset{\mathcal{D}}{=} \lambda_{\max}(W),$$

for the random matrix W defined in Table 1. Here $\lambda_{\max}$ denotes the largest eigenvalue and $\overset{\mathcal{D}}{=}$ denotes equality in distribution.

Remark 3.1. *Note that since the normal distribution is rotationally invariant, the distribution of the eigenvalues of W is unchanged even if we re-arrange the ordering of μ_i's in both problems. Hence we find that the distribution of $L(m, n)$ does not depend on the ordering of μ_i's.*

It is well-known that the batch latency for a tandem of queues is related to directed last passage percolation (DLPP) models [17]. Among the DLPP models, there are certain 'solvable' models which are known to be connected to random matrix theory (RMT) [20, 9]. The analysis of these solvable DLPP models has been a very active area of research in probability and statistical

Exact results		
Problem A	$W = \Sigma^{1/2} G G^* \Sigma^{1/2}$ **Theorem 3.1**: $L(m,n) \stackrel{\mathcal{D}}{=} \lambda_{\max}(W)$	$\Sigma = \mathrm{diag}(1/\mu_1, \dots, 1/\mu_m)$. G is an $m \times n$ matrix of i.i.d. $\mathbb{C}\mathcal{N}(0,1)$ entries.
Problem B	$W = \Gamma^{1/2} g g^* \Gamma^{1/2} + \Sigma^{1/2} G G^* \Sigma^{1/2}$ **Theorem 3.1**: $L(m,n) \stackrel{\mathcal{D}}{=} \lambda_{\max}(W)$	$\Sigma = \mathrm{diag}(1/\mu_1, \dots, 1/\mu_m)$ $\Gamma = \mathrm{diag}\left(1/(\mu_1 - \alpha), \dots, 1/(\mu_m - \alpha)\right)$. G and g are, resp., $m \times (n-1)$ matrix and $m \times 1$ vector of i.i.d. $\mathbb{C}\mathcal{N}(0,1)$ entries.

Table 1: Relation between queues and random matrices: exact distributional results.

physics (see e.g. [21, 14, 11]). The Problem A and B are precisely those which map to the solvable DLPP models. Combining these facts we arrive at the above result. See Section 7.1.

The above theorem allows one to use the method from RMT to study the batch latency. Indeed RMT is a subject of rich history and techniques, and has been studied extensively in mathematics, statistics, physics, and engineering [24, 1, 2]. However, for the asymptotic results in the next subsections, we do not use this connection directly. Instead we use the connection to solvable DLPP models mentioned above. The key fact is that there are explicit formulas for the distribution functions of the last passage times for solvable DLPP models. The asymptotic results in the next sections are obtained by analyzing these explicit formulas.

3.2 Asymptotic result I. Leading order

The next theorem is about the asymptotics of $L(m,n)$ in the large n, m regime. We first introduce some definitions. Consider the probability measure (the spectral measure for the service rates)

$$H_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{\mu_i}. \tag{1}$$

In the asymptotic regime, we assume that there is a compactly supported probability measure H such that

$$H_m \to H \qquad \text{weakly} \tag{2}$$

as $m \to \infty$ (i.e. $\int_{\mathbb{R}} g(x) dH_m(x) \to \int_{\mathbb{R}} g(x) dH(x)$ for all bounded continuous functions g.) This includes examples such as (a) $\mu_i = 1$ for all i, (b) $\mu_{i_0} = \mu < 1$ and $\mu_i = 1$ for all $i \neq i_0$, and (c) $\mu_i = F^{-1}(\frac{i}{m})$ for $F(x) := \int_{-\infty}^{x} f(y) dy$ for a continuous, non-negative and compactly support function $f(y)$. We denote the support of H_m by $supp(H_m)$.

Define the function

$$l_m(z) := m \int \frac{dH_m(y)}{y - z} + \frac{n}{z}, \qquad z \notin supp(H_m). \tag{3}$$

In the interval $(0, \inf supp(H_m))$, there is a unique point $z = \lambda_m$ such that $l_m'(z) = 0$, or equivalently

$$m \int \frac{dH_m(y)}{(y-z)^2} - \frac{n}{z^2} = 0, \qquad z = \lambda_m \in (0, \inf supp(H_m)). \tag{4}$$

See the examples in Figure 1.

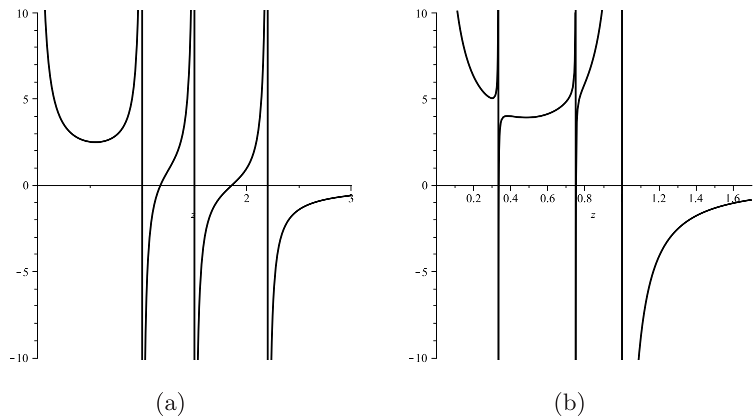


Figure 1: Graphs of $\frac{1}{m} l_m(z)$. (a) is when $m = 3, n = 2$ and $\mu_1 = 1, \mu_2 = 1.5, \mu_3 = 2.2$. (b) is when $m = 100$, $n = 100$, $\mu_1 = \frac{1}{3}$, $\mu_2 = \frac{3}{4}$, and $\mu_i = 1$ for $i = 3, \cdots, 100$. In both cases, λ_m is the argmin of $l_m(z)$ in the interval $(0, \mu_1)$.

Remark 3.2. *We note from* (3) *that* $l_m''(z) > 0$ *for all* $z \in (0, \inf supp(H_m))$ *and*

$$\begin{cases} l_m'(z) < 0, & z \in (0, \lambda_m), \\ l_m'(z) > 0, & z \in (\lambda_m, \inf supp(H_m)). \end{cases} \tag{5}$$

We are now ready to state the first asymptotic result.

Theorem 3.2 (Asymptotic result I. Leading order)**.** *Let α be the arrival rate in Problem B. For Problem A, we set $\alpha = 0$. Let*

$$\mu_{(m)} \le \cdots \le \mu_{(2)} \le \mu_{(1)} \tag{6}$$

be an ordered re-arrangement of the services rates $\mu_1, \cdots, \mu_m$. Assume (2)*. Recall $l_m(z)$ defined in (3) and λ_m in (4). From the definition, $\lambda_m < \mu_{(m)}$. As $m, n \to \infty$ with $m/n \to \gamma \in (0, \infty)$, the following asymptotic result holds in probability.*

(a) If $\alpha < \liminf_{m\to\infty} \lambda_m$ *and* $\liminf_{m\to\infty} (\mu_{(m)} - \lambda_m) > 0$*, then*

$$\frac{L(m,n)}{l_m(\lambda_m)} \to 1. \tag{7}$$

(b) If $\limsup_{m\to\infty} \lambda_m < \alpha$ *(note that necessarily* $\alpha < \liminf_{m\to\infty} \mu_{(m)}$ *by assumption), then*

$$\frac{L(m,n)}{l_m(\alpha)} \to 1. \tag{8}$$

(c) Suppose that $\limsup_{m\to\infty}(\mu_{(m)} - \lambda_m) = 0$ *and there is a fixed* r *(which does not grow in* m*) such that* $\mu_{(m)} = \cdots = \mu_{(m-r+1)}$ *and* $\liminf_{m\to\infty}(\mu_{(m-r)} - \mu_{(m-r+1)}) > 0$*. Furthermore, assume that* $\liminf_{m\to\infty}(\lambda_m^{(r)} - \mu_{(m)}) > 0$ *where* $\lambda_m^{(r)}$ *is the unique real root of the equation* $(l_m^{(r)})'(z) = 0$ *in* $z \in (0, \mu_{(m-r)})$ *where*

$$l_m^{(r)}(z) := m \int \frac{dH_m^{(r)}(y)}{y-z} + \frac{n}{z}, \qquad H_m^{(r)} := \frac{1}{m}\sum_{i=1}^{m-r} \delta_{\mu_{(i)}}. \tag{9}$$

Then

$$\frac{L(m,n)}{l_m^{(r)}(\mu_{(m)})} \to 1. \tag{10}$$

The above theorem can be simplified if we assume that the limit $\mu_{(m)}$ as $m \to \infty$ exists.

Corollary 3.1 (Easier conditions when $\mu_{(m)} \to \mu_{\inf}$)**.** *Suppose that* $\mu_{(m)} \to \mu_{\inf}$ *as* $m \to \infty$ *for some value* $\mu_{\inf}$ *(which is necessarily larger than* α*). Set*

$$l(z) := m \int \frac{dH(y)}{y-z} + \frac{n}{z}, \qquad z \in (0, \inf supp(H)) \tag{11}$$

and let $z = \lambda \in (0, \inf suppH)$ *be the unique solution to the equation*

$$m \int \frac{dH(y)}{(y-z)^2} - \frac{n}{z^2} = 0, \qquad z = \lambda \in (0, \inf supp(H)). \tag{12}$$

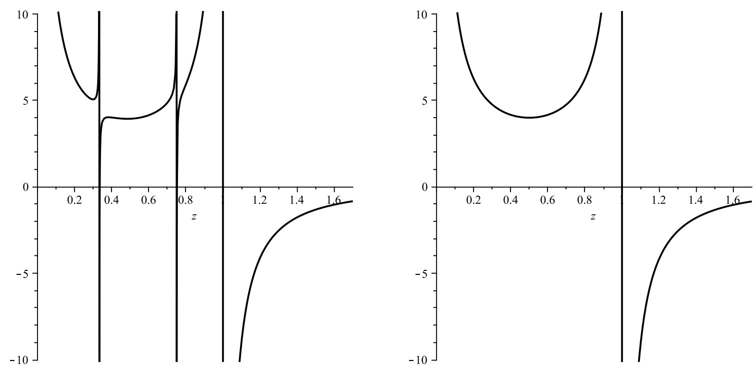
(Hence $l(z)$ *is the analogue of* $l_m(z)$ *with* H_m *replaced by* H *in (2), and* λ *is the analogue of* λ_m*. See Figure 2 for an example of* ℓ_m *and* ℓ*.) Then we have:*

(i) If $\lambda \in (\alpha, \mu_{\inf})$*, then* $\frac{L(m,n)}{l(\lambda)} \to 1$.

(ii) If $\lambda \in (0, \alpha)$*, then* $\frac{L(m,n)}{l(\alpha)} \to 1$.

(iii) If $\lambda > \mu_{\inf}$*, and if* $\mu_{(m)} = \cdots = \mu_{(m-r+1)}$ *for some fixed* r *and* $\liminf_{m\to\infty}(\mu_{(m-r)} - \mu_{\inf}) > 0$*, then* $\frac{L(m,n)}{l(\mu_{\inf})} \to 1$.

Proof. Note that as $H_m \to H$ weakly, $\mu_{(m)} \leq \inf supp(H)$, and hence $\mu_{\inf} \leq \inf supp(H)$. If $\lambda < \mu_{\inf}$, it is easy to check, using the fact that $\mu_{\inf} = \lim_{m\to\infty} \inf supp(H_m)$ and using the analyticity of $l_m(z)$ for $z \in (0, \inf supp(H_m))$, that $\lambda_m \to \lambda$, and also $\frac{1}{m}(l_m(z) - l(\lambda)) \to 0$ for each $z \in (0, \mu_{\inf})$. Hence we are in the case (a) or (b) of Theorem 3.2 depending on $\alpha < \lambda < \mu_{\inf}$ or $\lambda < \alpha$, and the above results for (i) and (ii) follow. On the other hand if $\lambda > \mu_{\inf}$, then it is also easy to check that for any compact interval of $(0, \mu_{\inf})$, there is no zero of the equation (4) for all large enough m. This means that $\limsup_{m\to\infty}(\mu_{(m)} - \lambda_m) \to 0$, and hence we find that the conditions for the case (c) of Theorem 3.2 are satisfied. It is also direct to check that $\frac{1}{m}(l_m^{(r)}(z) - l(z)) \to 0$ for each $z \in (0, \liminf_{m\to\infty}\mu_{(m-r)})$ since $l_m^{(r)}(z)$ is analytic in any closed interval in this interval for all large enough m. Thus the case (iii) follows. □



(a) (b)

Figure 2: We assume that $\mu_1 = \frac{1}{3}$, $\mu_2 = \frac{3}{4}$, and $\mu_i = 1$ for $i \geq 3$. Then $H_m = \frac{1}{m}\delta_{\frac{1}{3}} + \frac{1}{m}\delta_{\frac{3}{4}} + \frac{m-2}{m}\delta_1$ and $H = \delta_1$. (a) is the graph of $\frac{1}{m}l_m(z)$ when $m = n = 100$. (b) is the graph of $\frac{1}{m}l(z)$ when $m = n = 100$.

Example 3.1 (Equal fixed service rates)**.** *Suppose $\mu_1 = \cdots = \mu_m =: \mu$ for a fixed μ. Then clearly Corollary 3.1 applies with $\mu_{\inf} = \mu$ and $H = \delta_\mu$. (We also have $H_m = \delta_\mu$.) The equation* (12) *becomes*

$$\frac{m}{(\mu - z)^2} - \frac{n}{z^2} = 0, \qquad z \in (0, \mu). \tag{13}$$

Solving this algebraic equation we find that

$$\lambda = \frac{\sqrt{n}}{\sqrt{m} + \sqrt{n}}\mu. \tag{14}$$

As $\mu_{\inf} = \mu$, case (iii) does not occur. From the cases (i) and (ii), we have for Problem B,

$$L(m, n) \approx \begin{cases} \frac{(\sqrt{m}+\sqrt{n})^2}{\mu} & \text{if } \alpha < \frac{\sqrt{n}}{\sqrt{m}+\sqrt{n}}\mu, \\ \frac{n}{\alpha} + \frac{m}{\mu-\alpha} & \text{if } \alpha > \frac{\sqrt{n}}{\sqrt{m}+\sqrt{n}}\mu. \end{cases} \tag{15}$$

For Problem A, we set $\alpha = 0$ here and find that $L(m, n) \approx (\sqrt{m} + \sqrt{n})^2\mu$.

Example 3.2 (One slow server)**.** *We consider the case when one of the rate is smaller than the rest which are all equal. From Remark 3.1, we may assume that the slow server is the the first server without loss of generality. Set $\mu_1 = \mu'$ and $\mu_2 = \cdots = \mu_m =: \mu$ where μ, μ' are fixed numbers such that $\mu' < \mu$ and $\mu' > \alpha$. Then as $\mu_{(m)} = \mu'$ is a constant, Corollary 3.1 applies with $\mu_{\inf} = \mu'$.*

We have $H_m = \frac{1}{m}\delta_{\mu'} + (1-\frac{1}{m})\delta_\mu \to H = \delta_\mu$. This is same H as in Example 3.1. Thus we find that $\lambda = \frac{\sqrt{n}}{\sqrt{m}+\sqrt{n}}\mu$ as before. The difference is that since $\mu_{\inf} = \mu'$ is different from μ, the case (iii) can occur. From the corollary, we find that for Problem B

$$L(m,n) \approx \begin{cases} \frac{(\sqrt{m}+\sqrt{n})^2}{\mu} & \text{if } \alpha < \frac{\sqrt{n}}{\sqrt{m}+\sqrt{n}}\mu < \mu', \\ \frac{m}{\mu-\alpha} + \frac{n}{\alpha} & \text{if } \frac{\sqrt{n}}{\sqrt{m}+\sqrt{n}}\mu < \alpha, \\ \frac{m}{\mu-\mu'} + \frac{n}{\mu'} & \text{if } \mu' < \frac{\sqrt{n}}{\sqrt{m}+\sqrt{n}}\mu. \end{cases} \tag{16}$$

The Figure 3 is a simulation of this case. For Problem A, we can simply set $\alpha = 0$ in this result.

Example 3.3 (Two slow servers)**.** *Assume $\mu_1 = \mu'$, $\mu_2 = \mu''$, and $\mu_3 = \cdots = \mu_m =: \mu$ where $\mu' < \mu'' < \mu$ are fixed numbers. Then we can apply Corollary 3.1, and we have $H = \delta_\mu$, $\mu_{\inf} = \mu'$, and $\lambda = \frac{\sqrt{n}}{\sqrt{m}+\sqrt{n}}\mu$. Using Corollary 3.1, we have exactly the same result as* (16) *for Problem B. Note that for the case (iii) in which the slow servers affect the batch latency, the leading order of the exit time depends only on the lowest service rate, not the second lowest service rate. It is easy to check that* (16) *also holds if $\mu' = \mu'' < \mu$ or if there are five slow servers, etc.*

3.3 Discussion

The above result shows a phase transition phenomenon. Recall that α is the arrival rate. Large arrival rate in equilibrium means that when the batch of n customers arrive at queue 1, there are already many customers in the system. This will slow down the n customers of interest.

Another key parameter of the system is $\mu_{(m)}$, the rate of the slowest server. If $\mu_{(m)}$ small, then a server is particularly slow, and hence the customers will be slowed down passing through that particular server.

The above result shows that there is a sharp transitional value of α and $\mu_{(m)}$ that changes the asymptotic value of $L(n,m)$. It happens that the critical values for α and $\mu_{(m)}$ are the same value given by λ_m (or λ in the above examples).

Indeed the case (a) implies that if α is less than λ_m and $\mu_{(m)}$ is larger than λ_m, then $L(m,n)$ has the same asymptotics as the case when the system was empty initially and there are no slow servers (more precisely, the service rates are distributed precisely as H). In this case, the n customers do not feel the already-existing customers and any particularly slow servers up to the leading order asymptotics. (Actually the next theorem shows that this holds up to the second order asymptotics as well.)

The effect of large arrival rate becomes only apparent when α is larger than λ_m as given in the case (b). Similarly, the effect of a particularly slow server becomes only apparent when $\mu_{(m)}$ is smaller than λ_m.

3.4 Asymptotic result II. Second order

The next theorem is about the second order asymptotics. We evaluate the law of the asymptotic fluctuations. Let TW_2 denote the complex Tracy-Widom random variable from random matrix

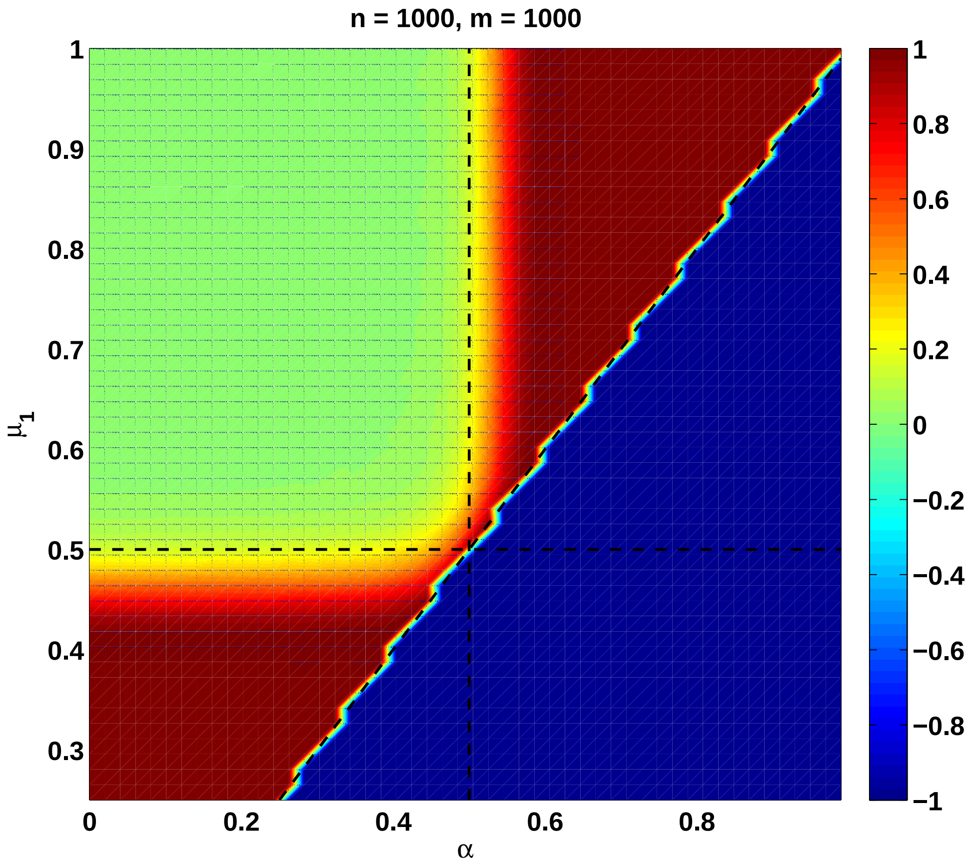


Figure 3: A heat-map of the Kolmogorov-Smirnov (KS) distance between the empirical cdf's of 1000 realizations of the $L(m,n)$ statistic of two systems both of which have $n = m = 1000$ and $m-1$ servers with rates $\mu_2 = \ldots = \mu_m$. The baseline system has $\mu_1 = 1$ and the external arrival rate $\alpha = 0$. The active system has μ_1 and α that are varied in the range specified by the plot. Using the KS distance we can assess if the underlying probability distributions for the $L(m,n)$ statistic for the two systems described differ. A value close to 0 indicates that the distributions are 'near' while a value closer t 1 indicates that they are 'far'. Note the phase transitions that separate regimes where the distributions are similar from regimes where the distributions are different; the dashed lines correspond to the predicted value of the phase transiton (see Example 3.2). The portion of the heat-map where the KS distance equals -1 corresponds to the (inadmissible) setting where $\alpha > \mu$.

theory [38]. There are other Tracy-Widom random variables, TW_β, and the subscript 2 in TW_2 signifies that this random variable is related to the so-called complex case in random matrix theory. The subscript 2 indicates the See (53) below for the explicit formula of TW_2. The standard normal random variable is denoted by $\mathcal{N}(0,1)$. The notation $\xrightarrow{\mathcal{D}}$ means convergence in distribution.

Theorem 3.3 (Asymptotic result II. Second order)**.** *With the same notations and assumptions in Theorem 3.2, we have the following asymptotic result.*

(a) If $\alpha < \liminf_{m\to\infty} \lambda_m$ *and* $\liminf_{m\to\infty}(\mu_{(m)} - \lambda_m) > 0$, *then*

$$\frac{L(m,n) - l_m(\lambda_m)}{(l_m''(\lambda_m)/2)^{1/3}} \xrightarrow{\mathcal{D}} TW_2. \tag{17}$$

(b) If $\limsup_{m\to\infty} \lambda_m < \alpha$, *then*

$$\frac{L(m,n) - l_m(\alpha)}{(l_m'(\alpha))^{1/2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1). \tag{18}$$

(c) Suppose that $\limsup_{m\to\infty}(\mu_{(m)} - \lambda_m) = 0$ *and that* $\liminf_{m\to\infty}(\mu_{(m-1)} - \mu_{(m)}) > 0$. *Furthermore, assume that* $\liminf_{m\to\infty}(\lambda_m^{(1)} - \mu_{(m)}) > 0$ *where* $\lambda_m^{(r)}$ *(and* $l_m^{(r)}$*) is defined in Theorem 3.2 (c). Then*

$$\frac{L(m,n) - l_m^{(1)}(\mu_{(m)})}{(-(l_m^{(1)})'(\mu_{(m)}))^{1/2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1). \tag{19}$$

Note that the denominator in (17) is $O(m^{1/3})$ while the denominators in (18) and (19) are $O(m^{1/2})$. Hence the order of the fluctuations in the case (a) is different from the cases (b) and (c).

The above theorem can be simplified as follows under some extra conditions. The proof of this Corollary is similar to the argument in the proof of Corollary 3.2 and we skip it.

Corollary 3.2 (Easier conditions)**.** *Suppose that* $\mu_{(m)} \to \mu_{\inf}$ *as* $m \to \infty$ *for some value* $\mu_{\inf}$ *as in Corollary 3.2. Furthermore, assume that*

$$l_m(z) - l(z) = o(m^{1/3}) \text{ for any compact interval of } z \in (0, \mu_{(m)}) \tag{20}$$

and

$$l_m^{(1)}(z) - l(z) = o(m^{1/2}) \text{ for any compact interval of } z \in (0, \mu_{(m-1)}) \tag{21}$$

where $l(z)$ *is defined in* (11)*. Let* λ *be defined in* (12)*. Then we have:*

(i) If $\lambda \in (\alpha, \mu_{\inf})$, *then*

$$\frac{L(m,n) - l(\lambda)}{(l''(\lambda)/2)^{1/3}} \xrightarrow{\mathcal{D}} TW_2. \tag{22}$$

(ii) If $\lambda \in (0, \alpha)$, *then*

$$\frac{L(m,n) - l(\alpha)}{(l'(\alpha))^{1/2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1). \tag{23}$$

(iii) If $\lambda > \mu_{\inf}$ and $\liminf_{m\to\infty}(\mu_{(m-1)} - \mu_{\inf}) > 0$, then

$$\frac{L(m,n) - l(\mu_{\inf})}{(-l'(\mu_{\inf}))^{1/2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1). \tag{24}$$

Example 3.4. *For Problem B of Examples 3.2 and 3.3, this result implies that*

$$L(m,n) \approx \begin{cases} \frac{(\sqrt{m}+\sqrt{n})^2}{\mu} + \frac{\sqrt{m}+\sqrt{n}}{\mu}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right)^{1/3} TW_2 & \text{if } \alpha < \frac{\sqrt{n}}{\sqrt{m}+\sqrt{n}}\mu < \mu', \\ \left(\frac{m}{\mu-\alpha} + \frac{n}{\alpha}\right) + \left(\frac{m}{(\mu-\alpha)^2} - \frac{n}{\alpha^2}\right)^{1/2} \mathcal{N}(0,1) & \text{if } \frac{\sqrt{n}}{\sqrt{m}+\sqrt{n}}\mu < \alpha, \\ \left(\frac{m}{\mu-\mu'} + \frac{n}{\mu'}\right) + \left(-\frac{m}{(\mu-\mu')^2} + \frac{n}{(\mu')^2}\right)^{1/2} \mathcal{N}(0,1) & \text{if } \mu' < \frac{\sqrt{n}}{\sqrt{m}+\sqrt{n}}\mu. \end{cases} \tag{25}$$

Remark 3.3. *In the case (c), if $\mu_{(m)} = \cdots = \mu_{(m-r+1)}$ and $\liminf_{m\to\infty}(\mu_{(m-r)} - \mu_{(m-r+1)}) > 0$ for a fixed r, then we have the same scaling but a different limiting distribution. The new distribution is same as the largest eigenvalue of the so-called $r \times r$ matrix from the Gaussian unitary ensemble [24].*

Some special cases of the above asymptotic results (including the second order asymptotics) were proved in several papers in the context of directed last passage percolation (DLPP) models and random matrix theory. The case with equal fixed rates in Example 3.1 was studied in [20] in which the appearance of TW_2 was first obtained. The random matrix model which corresponds to the case (a) of Problem A with general service rates was studied in [13]. The phase transition phenomena was first obtained in [3] for the model which corresponds to Problem A assuming that μ_i are independent of m and all but fixed r service rates are the same. (The Problem A of Examples 3.2, 3.3 and 3.4 belong to this case.) The results of [32] and [9] can also be interpreted as certain cases of Problem B. These previous results are all based on the explicit formula of the distribution function of the solvable DLPP model. The proofs of Theorems 3.2 and 3.3 follow the same approach and we extend the previous asymptotic analysis slightly to include the more general cases.

The study of phase transition phenomenon in random matrix theory is a popular subject. See, for example, [4, 31, 25, 6, 7].

3.5 Non-exponential service time

The connection to DLPP is completely general with any service times but only the exponential service times are known to be related to solvable DLPP models. This means that the analytic tools used for the exponential service times cannot be applied. Nevertheless, for either $m = o(n)$ and $n = o(m)$, some asymptotic results for $L(m,n)$ are known when the service times are identically distributed.

Suppose that the service times are independent identically distributed with mean μ and finite variance σ^2. Then for Problem A with iid service times, it is known that

$$\lim_{\substack{m,n\to\infty \\ n=o(m)}} \frac{L(m,n) - \mu m}{\sigma\sqrt{mn}} = 2. \tag{26}$$

The convergence of the left-hand side of (26) to a universal constant was shown by Glynn and Whitt [17]. Subsequently Seppäläinen showed that the universal constant equals 2 in [34]. We note that the above result also holds with m and n exchanged due to the duality between servers and the customers.

There is also a result for the fluctuations. Assume that the service times are independent and identically distributed with mean μ, variance σ^2, and with a finite pth moment for some $p > 2$. Then it was shown for Problem A that if $n, m \to \infty$ satisfying

$$n = o(m^c), \qquad c < \frac{6}{7}\left(\frac{1}{2} - \frac{1}{p}\right), \tag{27}$$

then

$$\frac{L(m,n) - \mu m - 2\sigma\sqrt{mn}}{\sigma n^{-1/6} m^{1/2}} \xrightarrow{\mathcal{D}} TW_2. \tag{28}$$

This was shown in [5] assuming $p = 4$ and in [8] assuming $p > 2$. These two papers appeared at the same time independently. The case when $p = 3$ was also shown in [36] using a different technique.

It is possible to extend the method of these results to the situation when there are slow servers or for Problem B. But we do not pursue this direction in this paper.

3.6 Heavy-tailed service times

The following result for heavy-tailed service times is due to Hambly and Martin [18]. Let F be a distribution function which regularly varies with index $\alpha \in (0, 2)$ i.e. for all $t > 0$

$$\frac{1 - F(tx)}{1 - F(x)} \to t^{-\alpha} \quad \text{as } x \to \infty. \tag{29}$$

Assume that F is continuous and $F(0) = 0$, and suppose that the service times are iid with common distribution F. Then setting $c_N := F^{-1}(1 - 1/N)$ (then $\frac{\log a_N}{\log N} \to \frac{1}{\alpha}$ as $N \to \infty$) it was shown that

$$\frac{1}{c_{m^2}} L([am], [bm]) \tag{30}$$

converges to a random variable (depending on a, b) as $m \to \infty$ for each $a, b > 0$. The limiting random variable is described as a 'continuous last-passage percolation' models in the unit square (parameterized by α). For the Pareto distribution Pareto(α), $F(x) = 1 - x^{-\alpha}$ and $c_N = N^{1/\alpha}$.

4 Connection with directed last passage percolation (DLPP)

Notation 4.1. *We employ the following notation throughout the rest of the paper:*

- S_i*: Server* $i \in \{1, \dots, m\}$ *where servers labeled from 'left to right',*
- C_j*: Customer* j *where customers are labeled from 'right to left', i.e., in the order that customers exit the system:*
  - $j \in \{1, \dots, n\}$ *denote the* n *'tagged' customers of interest in queue* 1*,*

– $j \in \{-Y+1, \ldots, 0\}$ *denote the* Y *customers already in the system at* $t = 0$ *(for Problem B),*

- $Q_j(0)$*: Queue in which customer* C_j *is present at* $t = 0$*,*
- $L(i,j)$*: Time at which customer* C_j *exits server* S_i*,*
- $w(i,j)$*: Service time for customer* C_j *at server* S_i*.*

We now derive a recursion relationship for $L(i,j)$. Note that $L(i,j-1)$ is the time at which customer C_{j-1} exits server S_i while $L(i-1,j)$ is the time at which customer C_j exits server S_{i-1}. If $L(i,j-1) < L(i-1,j)$, C_j can be served immediately by S_i as soon as C_j exits server S_{i-1}. Conversely, FIFO processing implies that if $L(i,j-1) > L(i-1,j)$, then C_j has to wait in server S_i's queue for C_{j-1} to be processed and exit S_i's queue. Hence we have the desired recursion:

$$L(i,j) = w(i,j) + \begin{cases} L(i-1,j) & \text{when } L(i,j-1) < L(i-1,j), \\ L(i,j-1) & \text{when } L(i,j-1) > L(i-1,j). \end{cases} \tag{31}$$

or equivalently,

$$\boxed{L(i,j) = \max\{L(i-1,j), L(i,j-1)\} + w(i,j)} \tag{32}$$

for all $i, j \in \mathbb{Z}$. This fundamental recursion can be recast as the following 'directed last passage percolation' (DLPP) problem:

$$L(i,j) = \max_{\pi \in P(i,j)} \left(\sum_{(k,\ell) \in \pi} w(k,\ell) \right), \tag{33}$$

where $P(i,j)$ is the set of 'up/right paths' ending at (i,j) *i.e.* $\pi \in P(i,j)$ if $\pi = \{(k_s, \ell_s)\}_{s=-\infty}^{0}$ such that $(k_0, \ell_0) = (i,j)$ and $(k_s, \ell_s) - (k_{s-1}, \ell_{s-1})$ is either $(1,0)$ or $(0,1)$ for all $s \leq 0$. Note that the right-hand-side of (33) satisfies the same recurrence as $L(i,j)$ in (32) since a path in $P(i,j)$ consists of either a path in $P(i-1,j)$ and (i,j), or a path in $P(i,j-1)$ and (i,j). In the context of DLPP, $L(m,n)$ is referred as the 'last passage time' to the site (m,n). DLPP is also related to random growth models and has been extensively studied in both physics and mathematics communities (see, for example, [20], [11], and references therein.)

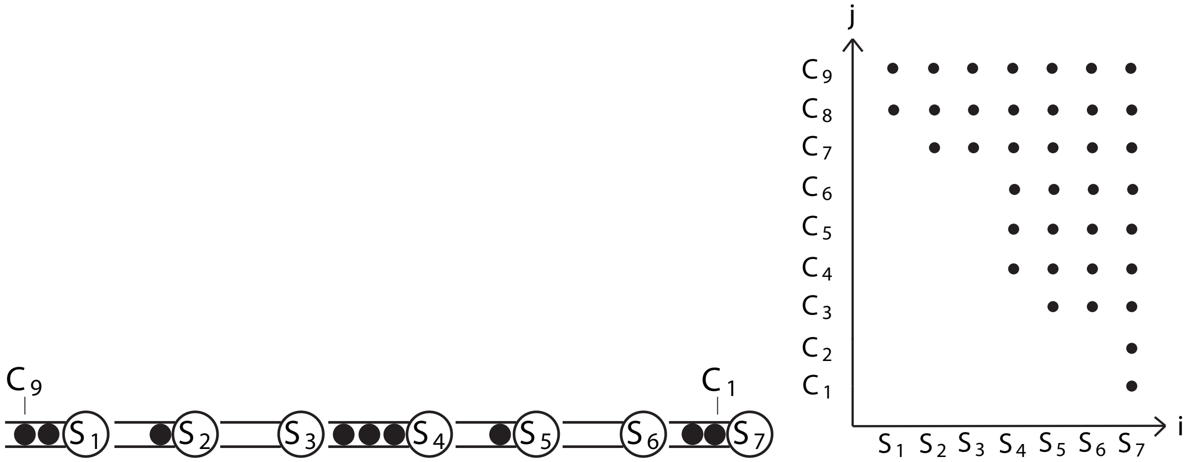
This recursion can be found in the seminar paper by Glynn and Whitt [17] wherein the authors attribute the formulation to Tembe and Wolff [37].

Remark 4.1 (Encoding initial conditions)**.** *Recall that* $Q_j(0)$ *denotes the queue at which customer* j *is present at* $t = 0$*. Customer* C_j *will not require any service from servers* S_i *for* $i < Q_j(0)$*; thus for every* j *we set* $w(i,j) = 0$ *for all integer* $i < Q_j(0)$*. Similarly, since each queue is of finite length, for every* i*, only those entries of* $w(i,j)$ *will be non-zero corresponding to customers* C_j *that have to be processed by* S_i*; we set* $w(i,j) = 0$ *for those customers that are 'ahead' of* S_i *at time* $t = 0$ *and thus, consistent with our labeling scheme, for negatively large enough* j*'s for every* i*.*

Remark 4.2 (Finiteness of the DLPP problem)**.** *One may think that* max *should be* sup *in (33). Remark 4.1 shows why for every negatively large enough* i *and* j*,* $w(i,j) = 0$*. Thus* $P(i,j)$ *is actually a finite set for any realization of the services times.*

Remark 4.3 (Exactness of the DLPP formulation)**.** *The recursion in (32) is an exact, deterministic representation that captures the interconnected dynamics of customers in the tandem of queues. There are no assumptions or restrictions on $w(i,j)$'s (except boundedness); they can be dependent, or not identically distributed.*

Figure 4 illustrates a mapping of initial queue states into the DLPP problem as described in Remark 4.1.



(a) Queue configuration at $t = 0$. (b) Corresponding DLPP encoding.

Figure 4: (a) is a snapshot of the system at $t = 0$ and the corresponding DLPP problem is shown in (b). Here we suppress the sites (i,j) such that $w(i,j) = 0$ and we put a black dot if $w(i,j)$ is non-zero at (i,j). For every C_j, row j contains a non-zero entry for server S_i that it has yet to pass through before exiting the system and a zero for every server that it has already passed through. Equivalently, $w(i,j) = 0$ for $i < Q_j(0)$ for every j. We can then interpret the exit time $L(i,j)$ via (33)

5 Solvable DLPP models and the connection with random matrices

In general, it is not easy to find an explicit formula of cdf of $L(i,j)$. However, during the last ten years or so, a development in DLPP established a computationally easy formula of cdf for a special choice of the waiting times.

Definition 5.1 (Solvable DLPP model)**.** *Let a_i, $i = 1, \cdots, m$ and b_j, $j = 1, \cdots, n$ be the numbers in $(-\infty, \infty]$ such that $a_i + b_j > 0$ for all i, j. Let $w(i,j)$ be independent random variables which is exponentially distribution with rate $a_i + b_j$ (so that the mean is $1/(a_i + b_j)$.) If $a_i + b_j = \infty$, we regard $w(i,j) = 0$. Consider the DLPP problem:*

$$L(m,n) = \max_{\pi \in P} \left(\sum_{(k,\ell)\in\pi} w(k,\ell) \right) \tag{34}$$

where P is the set of all up/right paths starting at $(1,1)$ ending at (m,n) with $w(i,j)$ chosen as above.

The formula of the cdf of $L(m,n)$ for these DLPP models is given as follows. Define the kernel

$$K(\xi,\eta) = \frac{-1}{(2\pi)^2}\int\int\left(\prod_{i=1}^{m}\frac{a_i-w}{a_i-z}\right)\left(\prod_{j=1}^{n}\frac{b_j+z}{b_j+w}\right)\frac{e^{\eta w-\xi z}}{w-z}dzdw, \quad \xi,\eta\in\mathbb{R}, \tag{35}$$

where the contours are simple closed curves in the complex plane $\mathbb{C}$, oriented counter-clockwise, such that the contour of z contains all a_i's inside, the contour of w contains all $-b_j$'s inside, and they do not intersect. Note that the condition $a_i + b_j > 0$ guarantees that such contours exist. Alternatively, we may also take the contours as straight lines parallel to the imaginary axis such that the z-curve is oriented from bottom to top, the w-curve is oriented from top to bottom and they satisfy $\max\{-b_j\} < Re(w) < Re(z) < \min\{a_i\}$. In (35), if $a_i = \infty$ for some i, we interpret that $\frac{a_i-w}{a_i-z} = 1$. The same remark also applies for b_j's. Now let K_x be the integral operator on the $L^2((x,\infty))$ defined by the kernel $K(\xi,\eta)$:

$$(K_x f)(\xi) = \int_x^\infty K(\xi,\eta)f(\eta)d\eta, \qquad \text{for } f\in L^2((x,\infty)). \tag{36}$$

Then we have:

Proposition 5.1. *For $x > 0$,*

$$\mathbb{P}\{L(m,n)\le x\} = \det(1-K_x) \tag{37}$$

where the determinant is the Fredholm determinant of the operator K_x (see, for example, [35] for the definition of Fredholm determinant).

This result when all a_i's and b_j's are the same numbers was obtained by Johansson in [20]. This was done first by noting that the exponential weight is a simple limit of geometric weight and then showing that the DLPP model with iid geometric weights can be mapped to a certain probability measure on the set of partitions. Subsequently, a generalization of this measure on the partitions, so-called the Schur measure, was introduced by Okounkov [30] (see Section 2.2.3). The inverse map from the Schur measure to DLPP yields a DLPP with geometric weights of structure similar to the above solvable DLPP models. The Proposition can then be obtained by taking a limit from the geometric weights to the exponential weights and using the result of [30] on the Schur measure. A good place where this is summarized explcitly is Theorem 3 of [9] (where one should set $r = s = p = 1$).

It was noted (see e.g. [20], [3]) that for the case when b_j's are all the same, the formula of the cdf above of $L(m,n)$ is same as that of the largest eigenvalue of a random matrix (called the complex sample covariant matrix or the complex Wishart ensemble). This observation was further generalized by Borodin and Péché in [9] for the general a_i and b_j as follows:

Proposition 5.2 ([9]). *Let X be an $m \times n$ matrix with independent entries distributed as*

$$X_{ij} \sim \mathbb{C}\mathcal{N}(0, \frac{1}{a_i + b_j}). \tag{38}$$

Let $L(m,n)$ correspond to the DLPP model in Definition 5.1. Then

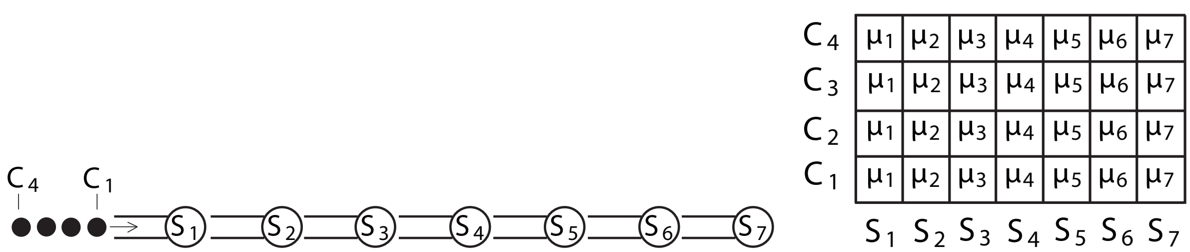
$$L(m,n) \stackrel{\mathcal{D}}{=} \lambda_{\max}(XX^*), \tag{39}$$

where $\lambda_{\max}$ denotes the largest eigenvalue.

6 Mapping queuing problems considered into solvable DLPP problems

6.1 Problem A

Since there are no customers in the queues at time 0, $w(i,j) = 0$ for all $j \leq 0$. Also we only consider the exit time of the nth customer in the batch, we may take $w(i,j) = 0$ for all $j > n$. On the other hand, by the assumption on the service time, $w(i,j)$ is exponential of rate μ_i for $i = 1, \cdots, m$. See Figure 5. The parameters are summarized in Table 2.



(a) Problem A.

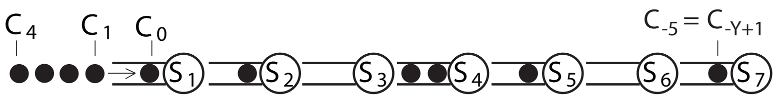
(b) Equivalent 'rectangular' DLPP formulation.

Figure 5: (a) is a snapshot of the queuing system at $t = 0$ and (b) is the corresponding DLPP problem. Here we put the rate of the service time $w(i,j)$ at each site (i,j).

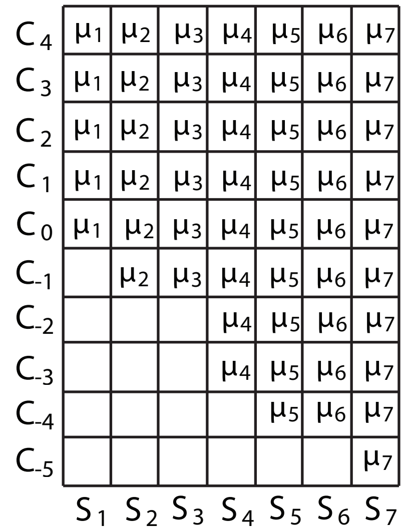
6.2 Problem B

When the batch of n customers arrive at queue 1, there are already Y customers in the queues. Then the corresponding DLPP is, first, of size m by $n+Y$, and second, the boundary of the DLPP is of jagged shape as illustrated in Figure 6. Moreover, Y is random, and the jagged boundary is also random. However, using the fact that the system was in equilibrium, we can effectively change this DLPP with random boundary to a DLPP with a fixed boundary as follow.

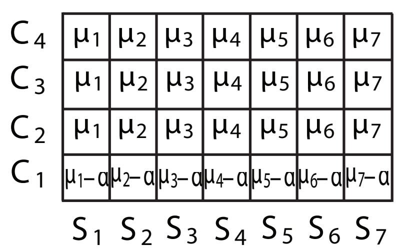
First, the following lemma characterizes the effective service time distribution for a customer when it enters the queue associated with server i.



(a) Problem B.



(b) Equivalent DLPP with 'random jagged boundary' formulation.



(c) Effective rates of DLPP from the view point of customers $1, 2, \cdots$.

Figure 6: (a) is a snapshot of the system at $t = 0$ and (b) is the corresponding DLPP problem. Here we suppressed the sites of zero service time and put the rates of the non-zero service times. (c) represents the effective rates of the DLPP problem for Problem B from the view point of customers $1, 2, \cdots$.

Equivalence relations		
	$w(i,j)$ in (33)	a_i and b_j in (34) and (38)
Problem A	$w(i,j) \sim \mathcal{E}(\mu_i), \quad i=1,\dots,m,\ j=1,\dots,n.$	$a_i = \mu_i \quad i=1,\dots,m,$ $b_j = 0, \quad j=1,\dots,n.$
Problem B	$w(i,j) \sim \mathcal{E}(\mu_i), \quad i=1,\dots,m,\ j=2,\dots,n,$ $w(i,1) \sim \mathcal{E}(\mu_i-\alpha), \quad i=1,\dots,m.$	$a_i = \mu_i \quad i=1,\dots,m,$ $b_1 = -\alpha$ $b_j = 0, \quad j=2,\dots,n.$

Table 2: Correspondence between Problem A/B and solvable DLPP. Here $\mathcal{E}(r)$ denote an exponential random variable of rate r. Thus this random variable has expectation $1/r$.

Lemma 6.1. *Consider a customer arriving via an external Poissonian arrival process of rate α at an $M/M/1$ queue in equilibrium with service rate μ_i (where $\alpha < \mu_i$). The effective service time of the customer is exponentially distributed with rate $\mu_i - \alpha$.*

Proof. By the Burke's theorem for $M/M/1$ queues in equilibrium, the departure rate at a given server is same is the arrival rate at that server. Hence in equilibrium, the arrival rate and the departure rate at each server equal α.

In equilibrium, the stationary distribution of the queue length (including the customer who is being served) at server i, denoted by Q_i, is a 'shifted geometric' random variable with parameter $q_i := \alpha/\mu_i$ so that:

$$\mathbb{P}(Q_i = k) = (1-q_i)^k q_i, \qquad k = 0, 1, \dots$$

When a customer arrives at the server i, due to the equilibrium condition the queue size *right before this customer enters the queue* is the above 'shifted geometric random variable'. Including the customer of interest, there are $\mathcal{N}$ people in the queue where $\mathcal{N}$ is the usual geometric random variable with pmf:

$$\mathbb{P}(\mathcal{N} = k) = (1-q_i)q_i^{k-1}, \qquad k = 1, 2, \dots$$

Since the service time at server i for each customer is exponentially distributed with rate μ_i, the effective service time, denoted by $\mathcal{Z}$, for the customer of interest can be written as $\mathcal{Z} = w_1 + \ldots + w_{\mathcal{N}}$ where w_i are iid $\mathcal{E}(\mu_i)$ random variables and $\mathcal{N}$ an independent geometric random variable above. The characteristic function of $\mathcal{Z}$ is

$$\begin{aligned} E[e^{s\mathcal{Z}}] = E[e^{s(w_1+\ldots+w_{\mathcal{N}})}] &= \sum_{k=1}^{\infty} E[e^{s(w_1+\ldots+w_k)}]\,\mathbb{P}(\mathcal{N}=k) \\ &= \sum_{k=1}^{\infty} \left(\frac{\mu_i}{\mu_i - s}\right)^k (1-q_i)q_i^{k-1} = \frac{(1-q_i)\mu_i}{(1-q_i)\mu_i - s}, \end{aligned}$$

implying that $\mathcal{Z} \sim \mathcal{E}((1-q_i)\mu_i) = \mathcal{E}(\mu_i - \alpha)$. □

Remark 6.1 (Slow customer equivalency in equilibrium)**.** *Lemma 6.1 implies that from the viewpoint of a Poissonian arriving customer j and the customers $j+1,\dots$ that follow, it is as though there are no customers in queue but that the effective service rate at server i for customer j is slowed to $\mu_i - \alpha$ while remaining μ_i for the customers $j+1,\dots$ that follow.*

From this remark, we find that the exist time $L(m,n)$ of Problem B has the same distribution as the exist time of Problem A with the change that for the customer 1, the rate is changed to $\mu_i - \alpha$. See as in Figure 6 (c). Thus the corresponding DLPP problem is solvable and is given by Table 2. This same idea was used in [32] to map an interacting particle system, called the totally asymmetric simple exclusion process, starting in equilibrium to a solvable DLPP problem.

7 Proof of the main theorems

7.1 Proof of Theorem 3.1

The theorem follows from the connection between Problem A/B and the solvable DLPP problems discussed in Section 6 and from the connection between solvable DLPP problems and random matrices stated in Proposition 5.2.

7.2 Formula of cumulative distribution function

We now prove Theorems 3.2 and 3.3. Note that Theorem 3.2 follows from Theorem 3.3, except for the case (c) when $r > 1$ which instead follows from Remark 3.3. For this one, we comment at the end how the proof should be modified. Hence we focus on the proof of Theorem 3.3 for Problem B here. The Problem A is obtained by setting $\alpha = 0$ in Problem B.

From Proposition 5.1,

$$\mathbb{P}\{L(m,n) \le x\} = \det(1 - K_x). \tag{40}$$

where K_x is the operator with kernel (35) and acts on the interval (x,∞). For Problem B, plugging in the choices of the parameters a_i, b_j in Table 2, the kernel becomes

$$K(\xi,\eta) = \frac{-1}{(2\pi)^2}\int_{\Sigma_{\{\alpha,0\}}}\int_{\Sigma_{\{\mu_i\}_{i=1}^m}} \left(\prod_{i=1}^m \frac{\mu_i - w}{\mu_i - z}\right)\frac{z-\alpha}{w-\alpha}\left(\frac{z}{w}\right)^{n-1}\frac{e^{\eta w - \xi z}}{w-z}dzdw \tag{41}$$

where Σ_A denotes a closed, counter-clockwise contour which encloses the points of the set A.

We obtain Theorem 3.3 by evaluating the asymptotics of (41) using the method of steepest-descent after appropriate scaling. Such an asymptotic analysis has been done for similar cases. Especially our analysis can be thought of as a combination of the analysis of [3] and [13] since the cases studied in these two papers are sub-cases of ours. We do not present all the technical details of the analysis here. Instead we give a sketch of the proof and focus on explaining how the centering and scaling of the theorem arise and how the limiting distribution is obtained.

7.3 Scaling and conjugation

We first recall two basic facts about Fredholm determinants. The first is that the determinant is invariant under scalings. Namely, let A be the operator acting on the space $L^2((t,\infty))$ with kernel $A(a,b)$ and let B the scaled operator defined by the kernel $B(a,b) = rA(c+ra, c+rb)$ which acts on $L^2((c+rt,\infty))$. Then $\det(1-A) = \det(1-B)$.

The second basic fact is that the determinant is invariant under conjugations by multiplicative operators. Let A be the operator with kernel $A(a,b)$ acting the space $L^2((t,\infty))$. Let $f(a)$ be a non-vanishing function on (t,∞). Defined the conjugated kernel $C(a,b) = f(a)A(a,b)\frac{1}{f(b)}$. Suppose that the operator C on $L^2((x,\infty))$ defined by the kernel $C(a,b)$ is a bounded (trace-class) operator. Then $\det(1-A) = \det(1-C)$.

7.4 Critical points

Fix $s \in \mathbb{R}$. Recall that we take the limit $m,n \to \infty$ such that $m/n \to \gamma$ for some $\gamma \in (0,\infty)$. For each case of Problem B, we would like to show that for some constants $x = x(\gamma)$ and $p \in (0,1)$ $\mathbb{P}\{L(m,n) \le xm + sm^p\}$ converges to a function in s. In the analysis below, we will determine x and p.

From (40), $\mathbb{P}\{L(m,n) \le xm + sm^p\} = \det(1 - K_{xm+sm^p})$. Note that Hilbert space $L^2((xm + sm^p,\infty))$ of the operator varies in m. From the scale invariance discussed in the previous section, we see that $\det(1 - K_{xm+sm^\gamma}) = \det(1 - \mathcal{K}_x)$ where $\mathcal{K}_x$ is defined by the scaled kernel

$$\mathcal{K}_x(\xi,\eta) = m^p K\big(xm + sm^p + m^p\xi, xm + sm^p + m^p\eta\big). \tag{42}$$

Note that now the Hilbert space $L^2((0,\infty))$ for the operator $\mathcal{K}_x$ does not depend on m. We have

$$\mathbb{P}\left\{\frac{L(m,n) - xm}{m^p} \le s\right\} = \det(1 - \mathcal{K}_x). \tag{43}$$

The kernel (42) equals

$$\mathcal{K}_x(\xi,\eta) = \frac{-m^p}{(2\pi)^2}\int_{\Sigma_{\{\alpha,0\}}}\int_{\Sigma_{\{\mu_{(i)}\}_{i=1}^m}} e^{m(\mathcal{F}_m(z;x) - \mathcal{F}_m(w;x))}\frac{(z-\alpha)we^{m^p((s+\eta)w-(s+\xi)z)}}{(w-\alpha)z(w-z)}dzdw \tag{44}$$

where

$$\mathcal{F}_m(z;x) := -\frac{1}{m}\sum_{i=1}^m \log(\mu_{(i)} - z) + \frac{n}{m}\log z - xz. \tag{45}$$

We evaluate the above double integral asymptotically using using the method of steepest-descent. Note that

$$\begin{aligned}\mathcal{F}_m'(z;x) &= \frac{1}{m}\sum_{i=1}^m \frac{1}{\mu_{(i)} - z} + \frac{n/m}{z} - x = \ell_m(z) - x,\\ \mathcal{F}_m''(z;x) &= \frac{1}{m}\sum_{i=1}^m \frac{1}{(\mu_{(i)} - z)^2} - \frac{n/m}{z^2} = \ell_m'(z)\end{aligned} \tag{46}$$

where $\ell_m(z) := \frac{1}{m} l_m(z)$ with $l(z)$ being defined in (3). Some examples of the graphs of $\ell_m(z)$ are in Figure 1 in Section 3.

It is easy to check that $\mathcal{F}'_m$ is convex in the interval $z \in (0, \mu_{(m)})$. From the definition (4) of λ_m, we see that $\lambda_m = argmin_{z\in(0,\mu_{(m)})} \mathcal{F}'_m(z;x)$. Note that λ_m is independent of x.

The critical points of $\mathcal{F}_m$ (i.e. the roots of $\mathcal{F}'_m$) play the key role in the method of steepest-descent. We note that depending on the sign of $x - \ell_m(\lambda_m)$, the number of real roots of $\mathcal{F}'_m$ in $(0, \mu_{(m)})$ can be 2, 1, or 0 since $\mathcal{F}'_m$ is convex in that interval.

7.5 Case (a)

We are yet to determine x and p. Here is a heuristic argument how we determine x. Note the exponential term $e^{m(\mathcal{F}_m(z;x)-\mathcal{F}_m(w;x))}$ in the double integral (44). If we can deform the contours to the path of steepest-descent for $\mathcal{F}_m(z;x)$ and the path of steepest-descent for $-\mathcal{F}_m(w;x)$, and use the method of steepest-descent, the leading contribution to the double integral becomes $e^{m(\mathcal{F}_m(z_c;x)-\mathcal{F}_m(w_c;x))}$ where z_c and w_c are the critical points. (It can be shown that since we need to deform the original contours to the new contours analytically, the critical points z_c and w_c should be in the strip $0 < Re(z) < \mu_{(m)}$).) Note that the path of steepest-descent for $-\mathcal{F}_m(w;x)$ is the path of steepest-ascent for $\mathcal{F}_m(w;x)$. Hence z_c and w_c are both critical points of the same function $\mathcal{F}_m(z;x)$. Now unless $\mathcal{F}_m(z_c;x) = \mathcal{F}_m(w_c;x)$, the leading term is not of order $O(1)$. This suggests that we should have $z_c = w_c$, which is attained if $\mathcal{F}_m$ has a unique critical point in $(0, \mu_{(m)})$. From the discussion in the last paragraph of the previous section, this happens if we take $x = \ell_m(\lambda_m)$.

We now set

$$x = \ell_m(\lambda_m) \tag{47}$$

and show that the application of the method of steepest-descent to (42) indeed yields the desired asymptotic result. As λ_m is the unique critical point of $\mathcal{F}_m(z;x)$ in the interval $(0, \mu_{(m)})$, $\mathcal{F}'_m(\lambda_m;x) = \mathcal{F}''_m(\lambda_m;x) = 0$. It is straightforward to check that $\mathcal{F}'''_m(z;x) > 0$ in $z \in (0, \mu_{(m)})$, hence especially $\mathcal{F}'''_m(\lambda_m) > 0$. This implies that the path of the steepest-decent of $\mathcal{F}_m(z;x)$, denoted by Γ_1, goes off λ_m at the angles $\pi/3$ and $-\pi/3$ in the complex plane. The steepest-descent of $-\mathcal{F}_m(w;x)$ is the path of steepest-ascent of $\mathcal{F}_m(w;x)$, which we denote by Γ_2. This path goes off the same critical point at the angles $2\pi/3$ and $-2\pi/3$ in the complex plane. The general shapes of the contours Γ_1 are Γ_2 are shown in Figure 7 by the solid curve and the dashed curve, respectively. (These curves can be obtained by finding the integral curves of the vector field $\overline{\mathcal{F}'_m(z)}$.)

We deform the contour $\Sigma_{\{\mu_{(i)}\}_{i=1}^m}$ for z to Γ_1 and the contour $\Sigma_{\{\alpha,0\}}$ for w to Γ_2. (We orient the new contours consistent with the original contours.) During this deformation, we need to be careful of the poles of the integrand. The poles are $z = \mu_{(i)}$, $w = \alpha, 0$, and $z = w$. Since we assume that (as we are in the case (a))

$$\alpha < \liminf_{m\to\infty} \lambda_m, \qquad \liminf_{m\to\infty} (\mu_{(m)} - \lambda_m) > 0, \tag{48}$$

we see that we can deform the contours to Γ_1 and Γ_2 without passing through the poles $z = \mu_{(i)}$ and $w = \alpha, 0$. About the pole $z = w$, even though the new contours meet at $z = w = \lambda_m$, we can modify the contours locally near the critical point as follows. It can be shown that if we take the

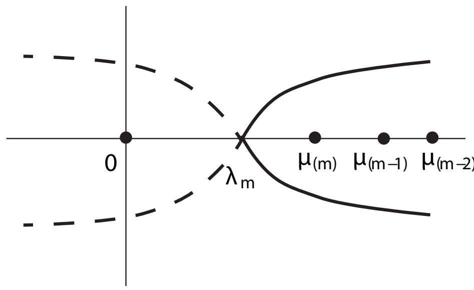


Figure 7: General shape of the paths of steepest-descent (solid) and of steepest-ascent (dashed) of $\mathcal{F}_m$ passing the point $z = \lambda_m$

contours Γ_1 and Γ_2 be $\lambda_m \pm O(m^{-1/3})$, respectively, near the critical point, the pole $z = w$ does not contribute but the method of steepest-descent still applies (see e.g. [3])

Since the critical point $z = \lambda_m$ is away from the poles due to (48), we can show that the main contribution to the double integral comes from a small neighborhood of the critical point. By localizing the integral near the critical point and expanding the exponent in a Taylor series, we find the leading asymptotic term of $\mathcal{K}_x(\xi, \eta)$. Then the kernel becomes

$$\mathcal{K}_x(\xi,\eta) \approx \frac{-m^p}{(2\pi)^2} \int\int e^{\frac{m}{3!}\mathcal{F}_m'''(\lambda_m;x)((z-\lambda_m)^3-(w-\lambda_m)^3)} \frac{(z-\alpha)we^{m^p((s+\eta)w-(s+\xi)z)}}{(w-\alpha)z(w-z)} dzdw. \tag{49}$$

Changing the variables $z \mapsto u$, $w \mapsto v$ by $c_m(z - \lambda_m) =: u$ and $c_m(w - \lambda_m) =: v$ where $c_m := (\frac{m}{2}\mathcal{F}_m'''(\lambda_m; x))^{1/3}$, we find that

$$\mathcal{K}_x(\xi,\eta) \approx e^{-m^p\lambda_m\xi}\left[-\frac{\frac{m^p}{c_m}}{(2\pi)^2} \int\int e^{\frac{1}{3}u^3-\frac{1}{3}v^3} \frac{e^{\frac{m^p}{c_m}((s+\eta)v-(s+\xi)u)}}{u-v} dudv \right] e^{m^p\lambda_m\eta}. \tag{50}$$

Here the contour for u is from $e^{\pi i/3}\infty$ to $e^{-\pi i/3}\infty$, and the contour for v is from $e^{2\pi i/3}\infty$ to $e^{4\pi i/3}\infty$ such that the first is to the right of the second.

Since $c_m = O(m^{1/3})$, the term in the bracket is $O(1)$ if we take $p = \frac{1}{3}$. This implies, together with the invariance of the Fredholm determinant under scaling and conjugation, that

$$\mathbb{P}\left\{ \frac{L(m,n) - m\ell_m(\lambda_m)}{m^{1/3}} \le s \right\} \approx \det(1 - \mathcal{A}_s) \tag{51}$$

where $\mathcal{A}_s$ is the operator on $L^2((0,\infty))$ defined by the kernel

$$\mathcal{A}_s(\xi,\eta) := -\frac{\frac{m^{1/3}}{c_m}}{(2\pi)^2} \int\int e^{\frac{1}{3}u^3-\frac{1}{3}v^3} \frac{e^{\frac{m^{1/3}}{c_m}((s+\eta)v-(s+\xi)u)}}{u-v} dudv. \tag{52}$$

(To be precise, one needs to that the convergence of the operator is in trace norm to be able to conclude that the determinant converges. This can be done by obtaining an estimate of the difference of two operators, but we skip this technical detail here.)

To make the right-hand side of (53) into the standard form, we define $A_s(\xi,\eta)=\beta\mathcal{A}_{\beta s}(\beta\xi,\beta\eta)$ with $\beta:=\frac{c_m}{m^\gamma}=(\frac{1}{2}\mathcal{F}_m'''(\lambda_m;x))^{1/3}$. Hence $A_s(\xi,\eta)$ is same as (52) with the term $\frac{m^{1/3}}{c_m}$ replaced by 1. By changing s to βs and using the invariance of determinants again, we find that

$$\mathbb{P}\left\{\frac{L(m,n)-m\ell_m(\lambda_m)}{(\frac{1}{2}\mathcal{F}_m'''(\lambda_m;x))^{1/3}m^{1/3}}\le s\right\}\approx\det(1-A_s)=\mathbb{P}(TW_2\le s). \tag{53}$$

The last equality is one of the definitions of the Tracy-Widom distribution [38]. Since $\ell_m(z)=\frac{1}{m}l_m$, we conclude that

$$\frac{L(m,n)-l_m(\lambda_m)}{(\frac{1}{2}l_m''(\lambda_m;x))^{1/3}}\xrightarrow{\mathcal{D}}TW_2, \tag{54}$$

and hence Theorem 3.3 (a) is proved.

7.6 Case (b)

In this case the assumption is that

$$\limsup_m \lambda_m<\alpha. \tag{55}$$

We are to determine x and the exponent $p\in(0,1)$ of (43).

If we take x as (47) and take the same new contours as in the previous section, then due to the condition (55), the deformation from the original contour to the new contour for w-variable passes through the pole $w=\alpha$. In this case the leading contribution to the w-integral does not come from the critical point λ_m, but comes from the pole α. Then since $\mathcal{F}_m(\lambda_m,x)\neq\mathcal{F}_m(\alpha,x)$, the leading contribution to the double integral is not $O(1)$. This means that the choice (47) is not suitable for case (b).

Unlike the case (a) where we defined x so that there is a unique critical point of $\mathcal{F}_m$ in $(0,\mu_{(m)})$, we now assume that we have $x>\ell_m(\lambda_m)$ so that there are two real critical values $z_c^-<z_c^+$ in $(0,\mu_{(m)})$. Note that $z_c^\pm$ depend on x. The exact value of x is yet to be determined. As $\mathcal{F}_m'(z;x)=\ell_m(z)-x$ and $\ell_m(z)$ is convex in $(0,\mu_{(m)})$ (see the last three paragraphs of Section 7.4), we see that $z_c^-\in(0,\lambda_m)$ and $z_c^+\in(\lambda_m,\mu_{(m)})$. (Recall that the definition of λ_m does not involve x.) It is easy to check that the path of steepest-descent of $\mathcal{F}_m$ passing the critical point z_c^+ (the future contour for the z-integral) is locally a vertical line near z_c^+ and the path of steepest-ascent of $\mathcal{F}_m$ (the future contour for the w-integral) passing the critical point z_c^- is locally a vertical linear near z_c^-. The general shape of these paths are shown in Figure 8. If we deform the original contours to these contours, the deformation of the w-integral, whose original contour is $\Sigma_{\{0,\alpha\}}$, passes the pole α due to the condition (55).

With this in mind, before applying the method of steepest-descent, we first deform the contours of the double integral (43) so that so that $w=\alpha$ is outside of the contour for w. By evaluating the residue at $w=\alpha$, we find

$$\begin{aligned}\mathcal{K}_x(\xi,\eta)&=\left[\frac{-m^p\alpha}{2\pi i}\int_{\Sigma_{\{\mu_{(i)}\}_{i=1}^m}}e^{m\mathcal{F}_m(z;x)}\frac{e^{-m^p(s+\xi)z}}{z}dz\right]e^{-m\mathcal{F}_m(\alpha;x)+m^p(s+\eta)\alpha}\\&+\frac{-m^p}{(2\pi)^2}\int_{\Sigma_{\{0\}}}\int_{\Sigma_{\{\mu_{(i)}\}_{i=1}^m}}e^{m(\mathcal{F}_m(z;x)-\mathcal{F}_m(w;x))}\frac{(z-\alpha)we^{m^p((s+\eta)w-(s+\xi)z)}}{(w-\alpha)z(w-z)}dzdw.\end{aligned} \tag{56}$$

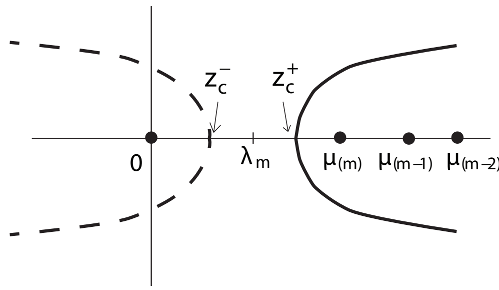


Figure 8: General shape of the path of steepest-descent of $\mathcal{F}_m$ passing the point z_c^+(solid) and the general shape of the path of steepest-ascent passing the point z_c^- (dashed)

Note that the new contour in w in the last double integral contains 0 inside, but α outside. Now we deform the contours to the paths of steepest-descent/steepest-ascent described above. Then the leading term of the double integral in (56) is $e^{m(\mathcal{F}_m(z_c^+(x);x)-\mathcal{F}_m(z_c^-(x);x))}$. But since $\mathcal{F}_m'$ is convex in $(0,\mu_{(m)})$, we see that $\mathcal{F}_m(z_c^+(x);x)-\mathcal{F}_m(z_c^-(x);x)<0$. Thus the double integral is exponentially small for any choice of $x>\ell_m(\lambda_m)$. On the other hand, the term in the bracket (56) has the leading term $e^{m\mathcal{F}_m(z_c^+(x);x)}$. If this term is same as $e^{-m\mathcal{F}_m(\alpha;x)}$, then the first term in (56) is of order $O(1)$. This is achieved if $z_c^+(x)=\alpha$, which means that α is one of the critical points of $\mathcal{F}_m(z;x)$, i.e. $\ell_m(\alpha)=x$. We take this as the choice of x:

$$x=\ell_m(\alpha). \tag{57}$$

With the above choice of x, the application of the method of steepest-descent yields that the integral in the bracket can be localized near the point $z=z_c^+=\alpha$, and we find that the term in the bracket is approximately

$$\frac{-m^p\alpha e^{m\mathcal{F}_m(\alpha;x)}}{2\pi i}\int e^{\frac{1}{2}m\mathcal{F}_m''(\alpha;x)(z-\alpha)^2}\frac{e^{-m^p(s+\xi)z}}{z}dz. \tag{58}$$

Changing the variables $\sqrt{m\mathcal{F}_m''(\alpha;x)}(z-\alpha)=u$, we see that we need to take

$$p=\frac{1}{2}. \tag{59}$$

With this choice of p, we find that the first term of (56) is approximately

$$e^{-m^{1/2}\alpha\xi}\left[\frac{-1}{2\pi i\beta}\int e^{\frac{1}{2}u^2}e^{-\frac{1}{\beta}(s+\xi)u}du\right]e^{m^{1/2}\alpha\eta},\quad \beta:=\sqrt{\mathcal{F}_m''(\alpha;x)}=\sqrt{\ell_m'(\alpha)}. \tag{60}$$

Here the integral is the vertical line from $i\infty$ to $-i\infty$. After evaluating the Gaussian integral, we find that

$$\mathcal{K}_x(\xi,\eta)\approx e^{-m^{1/2}\alpha\xi}\left[\frac{1}{\sqrt{2\pi}\beta}e^{-\frac{1}{2\beta^2}(s+\xi)^2}\right]e^{m^{1/2}\alpha\eta}. \tag{61}$$

Therefore, using the invariance of determinants under conjugations, we find that

$$\mathbb{P}\left\{\frac{L(m,n)-m\ell_m(\alpha)}{m^{1/2}}\le s\right\}=\det(1-\mathcal{K}_x)\approx\det(1-\mathcal{G}) \tag{62}$$

where $\mathcal{G}$ is an operator on $L^2((0,\infty))$ defined by the kernel $\mathcal{G}(\xi,\eta) = g(\xi) = \frac{1}{\sqrt{2\pi}\beta} e^{-\frac{1}{2\beta^2}(s+\xi)^2}$. Since $\mathcal{G}(\xi,\eta)$ does not depend on η, it is easy to check that the only eigenfunction of $\mathcal{G}$ is $g(\eta)$ with the eigenvalue $\int_0^\infty g(\eta)d\eta$. Hence the determinant

$$\det(1-\mathcal{G}) = 1 - \int_0^\infty g(\eta)d\eta = \frac{1}{\sqrt{2\pi}\beta}\int_{-\infty}^{s} e^{-\frac{1}{2\beta^2}\eta^2}d\eta. \tag{63}$$

Therefore, changing $s \mapsto \beta s$, we obtain

$$\frac{L(m,n) - l_m(\alpha)}{\sqrt{l'_m(\alpha)}} \approx \mathcal{N}(0,1) \tag{64}$$

and Theorem 3.3 (b) is proved.

7.7 Case (c)

The assumptions are

$$\limsup_{m\to\infty}(\mu_{(m)} - \lambda_m) = 0, \qquad \liminf_{m\to\infty}(\mu_{(m-1)} - \mu_{(m)}) > 0, \tag{65}$$

and

$$\liminf_{m\to\infty}(\lambda_m^{(1)} - \mu_{(m)}) > 0. \tag{66}$$

The case when $\mu_{(m)} = \cdots = \mu_{(m-r+1)}$ in Remark 3.3 will be discussed at the end.

Due to the condition (65), the (larger) real critical point (assuming that $x > \ell_m(\lambda_m)$ again) becomes close to the pole $\mu_{(m)}$ and hence we cannot argue that the main contribution cannot be localized near the critical point. In this case, we first change the contour $\Sigma_{\{\mu_{(i)}\}_{i=1}^m}$ for the z-integral in (44) to $\Sigma_{\{\mu_{(i)}\}_{i=1}^{m-1}}$ which excludes the pole $\mu_{(m)}$. (Recall that in the previous section, we changed the contour $\Sigma_{\{0,\alpha\}}$ for the w-integral to $\Sigma_{\{0\}}$ (see (56)).) Evaluating the residue at $z = \mu_{(m)}$, we obtain

$$\begin{aligned} \mathcal{K}_x(\xi,\eta) &= e^{m\mathcal{F}_m^{(1)}(\mu_{(m)};x) - m^p(s+\xi)\mu_{(m)}}\left[\frac{m^p(\mu_{(m)}-\alpha)}{2\pi i\mu_{(m)}}\int_{\Sigma_{\{0,\alpha\}}} e^{-m\mathcal{F}_m^{(1)}(w;x)}\frac{we^{m^p(s+\eta)w}}{w-\alpha}dw\right] \\ &+ \frac{-m^p}{(2\pi)^2}\int_{\Sigma_{\{0,\alpha\}}}\int_{\Sigma_{\{\mu_{(i)}:i\geq 2\}}} e^{m(\mathcal{F}_m^{(1)}(z;x) - \mathcal{F}_m^{(1)}(w;x))}\frac{(\mu_{(m)}-w)(z-\alpha)we^{m^p((s+\eta)w-(s+\xi)z)}}{(\mu_{(m)}-z)(w-\alpha)z(w-z)}dzdw, \end{aligned} \tag{67}$$

where

$$\mathcal{F}_m^{(1)}(z;x) := -\frac{1}{m}\sum_{i=1}^{m-1}\log(\mu_{(i)} - z) + \frac{n}{m}\log z - xz. \tag{68}$$

We set (see (9))

$$\ell_m^{(1)}(z) := \frac{1}{m}\sum_{i=1}^{m-1}\frac{1}{\mu_{(i)} - z} + \frac{n/m}{z} = \frac{1}{m}l^{(1)}(z). \tag{69}$$

Then

$$(\mathcal{F}_m^{(1)})'(z;x) = \ell_m^{(1)}(z) - x. \tag{70}$$

The function $\ell_m^{(1)}(z)$ is convex in $(0, \mu_{(m-1)})$ and $\lambda_m^{(1)}$ is defined to be the unique solution of $(\ell_m^{(1)})'(z) = 0$ in $(0, \mu_{(m-1)})$. As $\ell_m^{(1)}(z) = \ell_m(z) - \frac{1}{m(\mu_{(m)}-z)}$, we have $(\ell_m^{(1)})'(\lambda_m) < 0$ and hence $\lambda_m^{(1)} > \lambda_m$. Thus from the condition (65), we find that $\liminf_{n\to\infty}(\lambda_m^{(1)} - \mu_{(m)}) \geq 0$. The condition (66) is that here $\geq$ is indeed $>$.

We now set

$$x = \ell_m^{(1)}(\mu_{(m)}). \tag{71}$$

Then from (70) and (66), we find that there are two real roots of $(\mathcal{F}_m^{(1)})'(z;x) = 0$ in $z \in (0, \mu_{(m-1)})$. The smaller root is $\mu_{(m)}$, and we denote the other root by $z_0 \in (\mu_{(m)}, \mu_{(m-1)})$. The path of steepest-descent of $\mathcal{F}_m^{(1)}$ passing z_0 and the path of steepest-ascent passing $\mu_{(m)}$ are of shape in Figure 9. We evaluate the integrals in (67) using these paths.

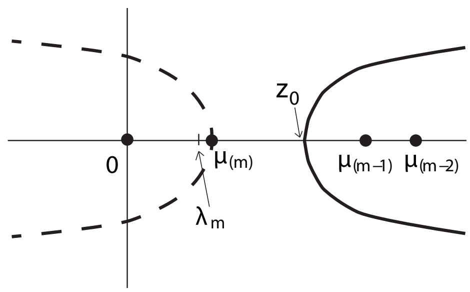


Figure 9: General shape of the path of steepest-descent of $\mathcal{F}_m^{(1)}$ passing the point z_0(solid) and the general shape of the path of steepest-ascent passing the point $\mu_{(m)}$ (dashed)

Since $(\mathcal{F}_m^{(1)})''(z;x) \neq 0$ at $z = \mu_{(m)}$ and at $z = z_0$, we are lead to set

$$p = \frac{1}{2}. \tag{72}$$

The integral in the square bracket in (67) can be evaluated in a matter similar to the case in the previous section and we find that the first term of (67) is approximately

$$e^{-m^{1/2}\mu_{(m)}\xi}\left[\frac{1}{\sqrt{2\pi}\beta}e^{-\frac{1}{2\beta^2}(s+\eta)^2}\right]e^{m^{1/2}\mu_{(m)}\eta}, \qquad \beta := \sqrt{-(\ell_m^{(1)})'(\mu_{(m)})}. \tag{73}$$

On the other hand, the second term in (67) is of order $O(e^{m(\mathcal{F}_m^{(1)}(z_0;x) - \mathcal{F}_m^{(1)}(\mu_{(m)};x))})$. Since $\ell_m^{(1)}$ is convex in $z \in (0, \mu_{(m-1)})$, the exponent is negative and hence the second term is exponentially small. This argument works without any problem if $\liminf_m(\mu_{(m-1)} - z_0) > 0$. (Note that the double integral has a pole at $z = \mu_{(m-1)}$.) Even if $\limsup_m(\mu_{(m-1)} - z_0) = 0$, we can still show that the second term is exponentially small. This is because $\mathcal{F}_m^{(1)}(z;x)$ is decreasing as z increases from $\mu_{(m)}$ to z_0 and hence it is possible to choose a z-contour which passes the real axis in-between $\mu_{(m)}$ and z_0. We skip the details.

Therefore, after a conjugation of the kernel, we find that

$$\mathbb{P}\left\{ \frac{L(m,n) - m\ell_m(\mu_{(m)})}{m^{1/2}} \le s \right\} = \det(1 - \mathcal{K}_x) \approx \det(1 - \mathcal{G}) \tag{74}$$

where $\mathcal{G}$ is an operator on $L^2((0,\infty))$ defined by the kernel $\mathcal{G}(\xi,\eta) = g(\xi) = \frac{1}{\sqrt{2\pi}\beta} e^{-\frac{1}{2\beta^2}(s+\xi)^2}$ as in (62) with the new β defined in (73). From (63), we find, after changing $s \mapsto \beta s$, that

$$\frac{L(m,n) - l_m^{(1)}(\mu_{(m)})}{\sqrt{-(l_m^{(1)})'(\mu_{(m)})}} \approx \mathcal{N}(0,1) \tag{75}$$

and and Theorem 3.3 (c) is proved.

Finally we discuss Remark 3.3. If $\mu_{(m)} = \cdots = \mu_{(m-r)}$ for some r independent of m and $\liminf_{m\to\infty}(\mu_{(r+1)} - \mu_{(r)}) > 0$ as in Remark 3.3, the pole $z = \mu_{(m)}$ in the first step of deriving (67) is not a simple pole but a pole of order r. The rest of the analysis is analogous but in the end we end up with the limit $\det(1 - \mathcal{G}^{(r)})$ where the kernel of the operator $\mathcal{G}^{(r)}$ is a certain rank r generalization of $\mathcal{G}(\xi,\eta)$. This kernel appeared in [3] from which we can see that $\det(1 - \mathcal{G}^{(r)})$ is the distribution function of the largest eigenvalue of $r \times r$ matrix from the Gaussian unitary ensemble.

Acknowledgments

J.B.'s work was support in part by NSF grants DMS-1068646. R.R.N's work was supported in part by NSF grant CCF-1116115.

References

[1] G. Akemann, J. Baik, and P. Di Francesco. The Oxford Handbook of Random Matrix Theory. Oxford Handbooks in Mathematics. Oxford University Press, 2011.

[2] Z. D. Bai and J. W. Silverstein. Spectral Analysis of Large Dimensional Random Matrice. Springer Series in Statistics. Springer, second edition, 2009.

[3] J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. Ann. Probab., 33(5):1643–1697, 2005.

[4] J. Baik and J. W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. J. Multivariate Anal., 97(6):1382–1408, 2006.

[5] J. Baik and T. M. Suidan. A GUE central limit theorem and universality of directed first and last passage site percolation. Int. Math. Res. Not., (6):325–337, 2005.

[6] F. Benaych-Georges, A. Guionnet, and M. Maïda. Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices, 2010. arXiv:1009.0145.

[7] F. Benaych-Georges and R. R. Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. Adv. Math., 227(1):494–521, 2011.

[8] T. Bodineau and J. Martin. A universality property for last-passage percolation paths close to the axis. Electron. Comm. Probab., 10:105–112 (electronic), 2005.

[9] A. Borodin and S. Péché. Airy kernel with two sets of parameters in directed percolation and random matrix theory. J. Stat. Phys., 132(2):275–290, 2008.

[10] F. Ciucu, A. Burchard, and J. Liebeherr. Scaling properties of statistical end-to-end bounds in the network calculus. Information Theory, IEEE Transactions on, 52(6):2300–2312, 2006.

[11] I. Corwin. The Kadar-Parisi-Zhang equation and universality class. Random Matrices: Theory and Applications, 1:1130001, 2012.

[12] M. Draief, J. Mairesse and N. O'Connel. Queues, stores, and tableaux. J. Appl. Probab. 42(4):1145–1167, 2005.

[13] N. El Karoui. Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. Ann. Probab., 35(2):663–714, 2007.

[14] P. Ferrari and H. Spohn. Random Growth Models. In The Oxford Handbook of Random Matrix Theory, edited by G. Akemann, J. Baik, and P. Di Francesco, Oxford University Press, 2011.

[15] M. Fidler. An end-to-end probabilistic network calculus with moment generating functions. In Quality of Service, 2006. IWQoS 2006. 14th IEEE International Workshop on, pages 261–270. IEEE, 2006.

[16] H. Friedman. Reduction methods for tandem queuing systems. Operations Research, 13(1):121–131, 1965.

[17] P. W. Glynn and W. Whitt. Departures from many queues in series. Ann. Appl. Probab., 1(4):546–572, 1991.

[18] B. Hamblyand J. Martin. Heavy tails in last-passage percolation. Probab. Theory Related Fields, 137(1-2):227–275, 2007.

[19] J. Harrison. The diffusion approximation for tandem queues in heavy traffic. Advances in Applied Probability, 10(4):886–905, 1978.

[20] K. Johansson. Shape fluctuations and random matrices. Comm. Math. Phys., 209(2):437–476, 2000.

[21] K. Johansson, Toeplitz determinants, random growth and determinantal processes, in Proceedings of the International Congress of Mathematicians, Vol. III (Beijing, 2002), pages 53–62, Higher Ed. Press, Beijing, 2002.

[22] J. Li and S. Meerkov. Production systems engineering. Springer Verlag, 2008.

[23] J. Martin. Batch queues, reversibility and first-passage percolation. Queueing Syst. 62(4):411–427, 2009.

[24] M. L. Mehta. Random matrices, volume 142 of Pure and Applied Mathematics (Amsterdam). Elsevier/Academic Press, Amsterdam, third edition, 2004.

[25] R. R. Nadakuditi and J. W. Silverstein. Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples. J. Sel. Topics in Signal Proc., 4:468–480, June 2010.

[26] S. Niu. Bounds for the expected delays in some tandem queues. Journal of Applied Probability, 17(3):831–838, 1980.

[27] S. Niu. On the comparison of waiting times in tandem queues. Journal of Applied Probability, 18(3):707–714, 1981.

[28] N. O'Connell. Directed percolation and tandem queues, HP Labs technical report, HPL-BRIMS-2000-28, 2000. http://www.hpl.hp.com/techreports/2000/

[29] N. O'Connell. Random matrices, non-colliding processes and queues, Séminaire de Probabilités, XXXVI, 165–182, Lecture Notes in Math., 1801, Springer, Berlin, 2003.

[30] A. Okounkov. Infinite wedge and random partitions. Selecta Math. (N.S.), 7(1):57–81, 2001.

[31] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. Statist. Sinica, 17(4):1617–1642, 2007.

[32] M. Prähofer and H. Spohn. Current fluctuations for the totally asymmetric simple exclusion process. In In and out of equilibrium (Mambucaba, 2000), volume 51 of Progr. Probab., pages 185–204. Birkhäuser Boston, Boston, MA, 2002.

[33] E. Reich. Waiting times when queues are in tandem. The Annals of Mathematical Statistics, 28(3):768–773, 1957.

[34] T. Seppäläinen. A scaling limit for queues in series. Ann. Appl. Probab., 7(4):855–872, 1997.

[35] B. Simon. Trace ideals and their applications, volume 120 of Mathematical Surveys and Monographs, American Mathematical Society, Providence, RI, second edition, 2005.

[36] T. Suidan. A remark on a theorem of Chatterjee and last passage percolation. J. Phys. A, 39(28):8977–8981, 2006.

[37] S. V. Tembe and R. W. Wolff. The optimal order of service in tandem queues. Operations Res., 22(4):824–832, 1974.

[38] C. A. Tracy and H. Widom. Level-spacing distributions and the Airy kernel. Comm. Math. Phys., 159(1):151–174, 1994.

[39] N. Van Dijk and B. Lamond. Simple bounds for finite single-server exponential tandem queues. Operations research, 36(3):470–477, 1988.

[40] M. Xie and M. Haenggi. Towards an end-to-end delay analysis of wireless multihop networks. Ad Hoc Networks, 7(5):849–861, 2009.