Extreme values for the waiting time in large fork-join queues

Dennis Schol, Maria Vlasiou, and Bert Zwart

September 18, 2023

Abstract

We prove that the scaled maximum steady-state waiting time and the scaled maximum steady-state queue length among N $GI/GI/1$-queues in the N-server fork-join queue, converge to a normally distributed random variable as $N \to \infty$. The maximum steady-state waiting time in this queueing system scales around $\frac{1}{\gamma} \log N$, where γ is determined by the cumulant generating function Λ of the service distribution and solves the Cramér-Lundberg equation with stochastic service times and deterministic inter-arrival times. This value $\frac{1}{\gamma} \log N$ is reached at a certain hitting time. The number of arrivals until that hitting time satisfies the central limit theorem, with standard deviation $\frac{\sigma_A}{\sqrt{\Lambda'(\gamma)\gamma}}$. By using distributional Little's law, we can extend this result to the maximum queue length. Finally, we extend these results to a fork-join queue with different classes of servers.

1 Introduction

Fork-join queues are a useful modeling tool for congestion in complex networks, such as assembly systems, communication networks, and supply chains. Such networks can be large and assembly is only possible upon availability of all parts. Thus, the bottleneck of the system is caused by the slowest production line in the system. This setting motivates us to investigate such delays in a stylized version of a large fork-join queueing system. In this setting, a key quantity of interest is the behavior of the longest queue when the system is in the steady-state situation. Furthermore, we assume that arrival and service processes are general and mutually independent.

As we try to model systems with many servers, we are typically interested in the behavior of this random variable as $N \to \infty$. In [12], it is shown that $\max_{i \leq N}(B_i(s) + B_A(s) - \beta s)$ is in the domain of attraction of the normal distribution:

$$\mathbb{P}\left(\max_{i \leq N}(B_i(s) + B_A(s) - \beta s) > \frac{\sigma^2}{2\beta} \log N + x\sqrt{\log N}\right) \overset{N \to \infty}{\longrightarrow} \mathbb{P}\left(\frac{\sigma \sigma_A}{\sqrt{2}\beta} X > x\right), \tag{1}$$

with $X \overset{d}{=} \mathcal{N}(0,1)$, where $\{B_i(t), t \geq 0\}$ and $\{B_A(t), t \geq 0\}$ are Brownian motions with standard deviations σ and σ_A, respectively. We see from the limit in (1) that $\max_{i \leq N}(B_i(s) + B_A(s) - \beta s)$ centers around $\frac{\sigma^2}{2\beta} \log N$ and deviates with order $\sqrt{\log N}$.

This convergence result provides a prediction of the typical delay. In this study, we aim to extend this result to a more general setting. In particular, we investigate the maximum steady-state waiting time among the N servers with a common arrival process $\max_{i \leq N} W_i(\infty) = \max_{i \leq N} \sup_{k \geq 0} \sum_{j=1}^{k}(S_i(j) - A(j))$. This expression follows from Lindley's recursion. Furthermore, we have that both $(S_i(j), j \geq 1, 1 \leq i \leq N)$ and $(A(j), j \geq 1)$ are i.i.d. and the inter-arrival times and service times are mutually independent. Thus $S_i(j)$ indicates the service time of the j-th customer in queue i, $A(j)$ indicates the inter-arrival time between the $(j-1)$-st and the j-th customer. We see that the maximum steady-state wating time is a maximum of N dependent random variables, due to the common arrival process $(A(j), j \geq 1)$.

The earliest literature on fork-join queues focuses on systems with two service stations. Analytic results, such as asymptotics on limiting distributions, can be found in [2, 6, 9, 17]. However, due to the complexity of fork-join queues, these results cannot be expanded to fork-join queues with more than two service stations. Thus, most of the work on fork-join queues with more than two service stations is focused on finding approximations of performance measures. For example, an approximation of the distribution of the response time in $M/M/s$ fork-join queues is given in Ko and Serfozo [10]. Upper and lower bounds for the mean response time of servers, and other performance measures, are given by Nelson, Tantawi [14] and Baccelli, Makowski [3]. These bounds can be used for fork-join queues with large size, but apart from this, there is not much literature on the convergence of the longest queue length in a fork-join queue as $N \to \infty$. Some results can be found in [12, 15, 16]. In [12], the same convergence

results are given as in this paper, but only for the Brownian fork-join queue, thus this paper extends on these results.

This paper is organized as follows. In Section 2, we present our main results; in Theorem 1 we state that the longest steady-state waiting time satisfies a central limit result; in Theorem 2 we show that a similar result holds for the longest queue length, and in Corollary 1 we present a similar result when the service distributions can differ among the different queues. In Section 3 we give an intuition why the results hold and how we prove these. Section 4 is devoted to proofs.

2 Model

We investigate a fork-join queue with N servers. Each of the N servers has the same arrival stream of jobs and works independently from all other servers but with the same service distribution. In this section, we state the main result for the longest steady-state waiting time in Theorem 1. We also show that a similar result holds for the maximum queue length in Lemma 2 and Theorem 2. Furthermore, we extend the result in Theorem 2 to a heterogeneous model in Corollary 1.

We now specify some properties of the service times and interarrival times in this fork-join queueing system. First, the sequence of non-negative random variables $(S_i(j), i \geq 1, j \geq 1)$ are i.i.d. with $S_i(j) \sim S$, and $S_i(j)$ indicating the service time of the j-th subtask in queue i. Furthermore, the sequence of non-negative random variables $(A(j), j \geq 1)$ are i.i.d. with $A(j) \sim A$, $\mathbb{E}[A(j)] = 1/\lambda$, $\mathrm{Var}(A(j)) = \sigma_A^2$, and $A(j)$ indicating the interarrival time between the $(j-1)$-st and the j-th task. Finally, we have that $\mathbb{E}[S_i(j) - A(j)] = -\mu$, with $\mu > 0$, and $(A(j), j \geq 1)$ and $(S_i(j), i \geq 1, j \geq 1)$ are mutually independent.

We can now write the cumulative distribution function of the longest steady-state waiting time as the cumulative distribution function of the maximum of N all-time suprema of random walks involving the interarrival and service times.

Lemma 1. *For the model given in Section 2 with $W_i(1) = 0$ for all $i \leq N$, we have that the longest waiting time in steady state satisfies*

$$\max_{i \leq N} W_i(\infty) \overset{d}{=} \max_{i \leq N} \sup_{k \geq 0} \sum_{j=1}^{k} (S_i(j) - A(j)). \tag{2}$$

Proof. Using Lindley's recursion [11], we can write the waiting time of tasks in front of server i as

$$W_i(n) = \sup_{0 \leq k \leq n} \sum_{j=k+1}^{n} (S_i(j) - A(j)).$$

Thus, the longest steady steady-state waiting time satisfies.

$$\max_{i \leq N} W_i(n) = \max_{i \leq N} \sup_{0 \leq k \leq n} \sum_{j=k+1}^{n} (S_i(j) - A(j)).$$

We have that

$$\mathbb{P}(\max_{i \leq N} W_i(\infty) \geq x) = \lim_{n \to \infty} \mathbb{P}(\max_{i \leq N} W_i(n) \geq x).$$

Because,

$$\max_{i \leq N} W_i(n) \overset{d}{=} \max_{i \leq N} \sup_{0 \leq k \leq n} \sum_{j=1}^{k} (S_i(j) - A(j)), \tag{3}$$

we obtain the lemma by using the monotone convergence theorem.

$\square$

In order to be able to prove convergence of the longest steady-state waiting time, we need some additional structure for the service-time distribution. We define

$$\Lambda(\theta) := \log(\mathbb{E}[\exp(\theta(S - 1/\lambda)]). \tag{4}$$

Moreover, we write $\mathcal{D}(\Lambda) := \{\theta : \Lambda(\theta) < \infty\}$ and $\mathcal{D}^\circ(\Lambda)$ as the interior of $\mathcal{D}(\Lambda)$.

Assumption 1. *We assume there exists a $\gamma > 0$ such that*

1. $\Lambda(\gamma) = 0$,

2. $\gamma \in \mathcal{D}^\circ(\Lambda)$.

The first assumption indicates that the random variable $S - 1/\lambda$ has a tail that is bounded by an exponential. The second assumption is needed for our proofs. In [4, Ex. 2.2.24], it is namely stated that when $\gamma \in \mathcal{D}^\circ(\Lambda)$, Λ is infinitely differentiable at the point γ. For example, when $S - 1/\lambda$ has density function $f_{S-1/\lambda}(x) = c_1 \exp(-x)/(1 + x^2)$ for $x > 0$, where c_1, λ are chosen such that $\mathbb{P}(S - 1/\lambda < x)$ is a cumulative distribution function and $\gamma = 1$, then the first assumption is satisfied but the second is not, since $\Lambda(\theta)$ is not differentiable at $\theta = \gamma$. Our main result is given in Theorem 1.

Theorem 1. *For the model in Section 2 where the sequence of service times $(S_i(j), i \geq 1, j \geq 1)$ satisfies Assumption 1, we have that*

$$\frac{\max_{i \leq N} W_i(\infty) - \frac{1}{\gamma} \log N}{\sqrt{\log N}} \xrightarrow{d} \frac{\sigma_A}{\sqrt{\Lambda'(\gamma)\gamma}} X, \tag{5}$$

with $X \sim \mathcal{N}(0, 1)$, as $N \to \infty$.

Lemma 2 (Distributional Little's Law). *Let for $t \geq 0$, $\mathbf{N}_A(t)$ indicate the number of arrivals up to time t, where the interarrival times are i.i.d. with $A(j) \sim A$. Then*

$$\max_{i \leq N} Q_i(\infty) \stackrel{d}{=} \mathbf{N}_A\left(\max_{i \leq N} W_i(\infty)\right). \tag{6}$$

Proof. In [8], a short proof is given that for the $GI/GI/1$ queue under the FCFS policy, $Q \stackrel{d}{=} \mathbf{N}_A(W)$. We follow the same steps to prove that $\max_{i \leq N} Q_i(\infty) \stackrel{d}{=} \mathbf{N}_A(\max_{i \leq N} W_i(\infty))$.

First, let $t > 0$ be given such that the system is in steady state. Furthermore, let $\tilde{W}_i(j)$ be the waiting time of the i-th subtask of the j-th task numbered backward in time, beginning at time t. Thus, $\tilde{W}_i(1)$ is the waiting time of the i-th subtask of the last task arriving before time t. Now, let the random variable $T(j)$ be such that $t - T(j)$ is the arrival time of the j-th task numbered backward in time. Then, observe that the event $\{\max_{i \leq N} Q_i(t) \geq j\}$ is equivalent to the event that at least one subtask of the j-th task numbered backward in time is still in the queue at time t. Thus,

$$\left\{\max_{i \leq N} Q_i(t) \geq j\right\} = \left\{\max_{i \leq N} \tilde{W}_i(j) \geq T(j)\right\},$$

for $j \geq 1$. The event $\{T(j) \leq x\}$ is equivalent to the event that the number of arrivals during the period $[t - x, t)$ is larger than or equal to j. The arrival process is a stationary process, thus the event $\{T(j) \leq x\}$ is equivalent to the event $\{\mathbf{N}_A(x) \geq j\}$. Additionally, the random variables $\max_{i \leq N} \tilde{W}_i(j)$ and $T(j)$ are independent. Therefore,

$$\left\{\max_{i \leq N} Q_i(t) \geq j\right\} = \left\{\mathbf{N}_A\left(\max_{i \leq N} \tilde{W}_i(j)\right) \geq j\right\}.$$

As the system is in steady state, we get that

$$\left\{\max_{i \leq N} Q_i(\infty) \geq j\right\} = \left\{\mathbf{N}_A\left(\max_{i \leq N} \tilde{W}_i(\infty)\right) \geq j\right\}.$$

$\square$

Now, combining the result in Lemma 2 with the main result in Theorem 1, we can find a similar convergence result for the maximum queue length in steady state.

Theorem 2. *For the model in Section 2 where the sequence of service times $(S_i(j), i \geq 1, j \geq 1)$ satisfies Assumption 1, we have that*

$$\frac{\max_{i \leq N} Q_i(\infty) - \frac{\lambda}{\gamma} \log N}{\sqrt{\log N}} \xrightarrow{d} \sqrt{\frac{\lambda^2 \sigma_A^2}{\Lambda'(\gamma)\gamma} + \frac{\lambda^3 \sigma_A^2}{\gamma}} X, \tag{7}$$

with $X \sim \mathcal{N}(0, 1)$, as $N \to \infty$.

3

Proof. Let $\hat{A}(j) \sim A$, let $(\hat{A}(j), j \geq 1)$ be mutually independent, and $\hat{A}(j)$ and $\max_{i \leq N} W_i(\infty)$ be mutually independent for all $j \geq 1$. Then, using Lemma 2 and Theorem 1, we get that

$$\mathbb{P}\left(\max_{i \leq N} Q_i(\infty) \leq \frac{\lambda}{\gamma} \log N + x\sqrt{\log N}\right)$$

$$= \mathbb{P}\left(\mathbf{N}_A\left(\max_{i \leq N} W_i(\infty)\right) \leq \left\lfloor \frac{\lambda}{\gamma} \log N + x\sqrt{\log N} \right\rfloor\right)$$

$$= \mathbb{P}\left(\max_{i \leq N} W_i(\infty) \leq \sum_{j=1}^{\left\lfloor \frac{\lambda}{\gamma} \log N + x\sqrt{\log N} \right\rfloor} \hat{A}(j)\right)$$

$$= \mathbb{P}\left(\frac{\max_{i \leq N} W_i(\infty) - \frac{1}{\gamma} \log N}{\sqrt{\log N}} \leq \frac{\sum_{j=1}^{\left\lfloor \frac{\lambda}{\gamma} \log N + x\sqrt{\log N} \right\rfloor} \hat{A}(j) - \frac{1}{\gamma} \log N}{\sqrt{\log N}}\right)$$

$$\overset{N\to\infty}{\longrightarrow} \mathbb{P}\left(\frac{\sigma_A}{\sqrt{\Lambda'(\gamma)\gamma}} X_1 \leq \frac{\sigma_A\sqrt{\lambda}}{\sqrt{\gamma}} X_2 + \frac{x}{\lambda}\right),$$

with X_1, X_2 independent and standard normally distributed, this convergence holds, as $(\hat{A}(j), j \geq 1)$ and $\max_{i \leq N} W_i(\infty)$ are independent. Thus, the theorem follows. $\square$

Until now, we considered the fork-join queueing system where each server has the same service distribution. In Corollary 1, we show that we can extend the convergence of the longest steady-state waiting time to a more heterogeneous setting. We examine a fork-join queueing system with N servers, where each of these N servers belongs to one of K classes. Additionally, we assume that the size of class k with $k \in \{1, \ldots, K\}$ grows as $\alpha_k N$, as N becomes large, with $0 < \alpha_k < 1$.

Corollary 1. *Let $K \in \mathbb{N}$, let $k = 1, \ldots, K$, furthermore, take an increasing sequence of integers given by $M_0^{(N)}, M_1^{(N)}, M_2^{(N)}, \ldots, M_K^{(N)} > 0$ with $M_0^{(N)} = 1$, $M_K^{(N)} = N$, and $M_k^{(N)} - M_{k-1}^{(N)} \in \mathbb{N}$. Moreover, $(M_k^{(N)} - M_{k-1}^{(N)})/N \overset{N\to\infty}{\longrightarrow} \alpha_k \in (0, 1]$ with $\sum_{k=1}^K \alpha_k = 1$. Let $(S_i(j), j \geq 1, M_{k-1}^{(N)} < i \leq M_k^{(N)})$ be i.i.d. with $S_i(j) \sim S_k$, $(A(j), j \geq 1)$ be i.i.d. with $A(j) \sim A$, $\mathbb{E}[A(j)] = 1/\lambda$, $Var(A(j)) = \sigma_A^2$, $\mathbb{E}[S_i(j) - A(j)] = -\mu_k$ with $\mu_k > 0$, $\Lambda_k(\theta) = \log(\mathbb{E}[\exp(\theta(S_k - 1/\lambda)])$, Λ_k satisfies Assumption 1. Furthermore, $S_{i_1}(j_1)$ and $S_{i_2}(j_2)$ are mutually independent for all i_1, i_2, j_1, j_2. Let $K^* = \arg\min\{\gamma_k, k = 1, \ldots, K\}$. We assume that $|K^*| = 1$ and $k^* \in K^*$. Then,*

$$\frac{\max_{i \leq N} W_i(\infty) - \frac{1}{\gamma_{k^*}} \log N}{\sqrt{\log N}} \overset{d}{\longrightarrow} \frac{\sigma_A}{\sqrt{\Lambda'_{k^*}(\gamma_{k^*})\gamma_{k^*}}} X, \tag{8}$$

with $X \sim \mathcal{N}(0, 1)$, as $N \to \infty$.

Proof. We prove this corollary by giving an asymptotically sharp lower and upper bound. First, observe that

$$\max_{i \leq N} W_i(\infty) \geq_{st.} \max_{M_{k^*-1}^{(N)} < i \leq M_{k^*}^{(N)}} \sup_{k \geq 0} \sum_{j=1}^k (S_i(j) - A(j)),$$

with $X \geq_{st.} Y$ meaning that $\mathbb{P}(X \geq x) \geq \mathbb{P}(Y \geq x)$ for all x. Applying the result from Theorem 1 on the lower bound results in (8). By using the union bound we get the following upper bound:

$$\mathbb{P}\left(\max_{i \leq N} W_i(\infty) \geq \frac{1}{\gamma_{k^*}} \log N + x\sqrt{\log N}\right)$$

$$= \sum_{l=1}^K \mathbb{P}\left(\max_{M_{l-1}^{(N)} < i \leq M_l^{(N)}} \sup_{k \geq 0} \sum_{j=1}^k (S_i(j) - A(j)) \geq \frac{1}{\gamma_{k^*}} \log N + x\sqrt{\log N}\right).$$

When $l \neq k^*$, we get after applying the results from Theorem 1 that

$$\mathbb{P}\left(\max_{M_{l-1}^{(N)} < i \leq M_l^{(N)}} \sup_{k \geq 0} \sum_{j=1}^k (S_i(j) - A(j)) \geq \frac{1}{\gamma_l} \log N + x\sqrt{\log N}\right) \overset{N\to\infty}{\longrightarrow} 1 - \Phi\left(\frac{\sqrt{\Lambda'_l(\gamma_l)\gamma_l}}{\sigma_A} x\right),$$

4

with Φ the cumulative distribution function of a standard normal random variable. Because $\gamma_{k^*} < \gamma_l$ we get that

$$\mathbb{P}\left(\max_{M_{l-1}^{(N)} < i \leq M_l^{(N)}} \sup_{k \geq 0} \sum_{j=1}^{k} (S_i(j) - A(j)) \geq \frac{1}{\gamma_{K^*}} \log N + x\sqrt{\log N}\right) \stackrel{N \to \infty}{\longrightarrow} 0.$$

The corollary follows. $\qquad\square$

Remark 1. *In Corollary 1 we assume that $|K^*| = 1$. The situation that $|K^*| > 1$ follows analogously. Assume for instance that $|K^*| = 2$, then we can introduce a new random variable $\tilde{S}$ such that $\tilde{S}_i(j) \sim S_1$ with probability α and $\tilde{S}_i(j) \sim S_2$ with probability $1 - \alpha$, such that $\gamma_1 = \gamma_2 = \gamma_{K^*}$. As N is large enough this fork-join queueing system behaves analogous to the original fork-join queue, and for this system $|K^*| = 1$.*

We give the proofs of the convergence of the longest steady-state waiting time in Section 4. First, we give a heuristic explanation of why the convergence result in Theorem 1 is true, and we illustrate the structure of the proof.

3 Heuristic analysis

To prove Theorem 1, we analyze lower and upper bounds of the tail probability of the longest steady-state waiting time among the N servers $\mathbb{P}(\max_{i \leq N} W_i(\infty) > \frac{1}{\gamma} \log N + x\sqrt{\log N})$ and we show that these lower and upper bounds converge to the same limit as $N \to \infty$. The longest steady-state waiting time has the form $\max_{i \leq N} W_i(\infty) \stackrel{d}{=} \sup_{k \geq 0} \max_{i \leq N} \sum_{j=1}^{k} (S_i(j) - A(j))$. Thus the longest steady-state waiting time is the all-time supremum of the maximum of N random walks. For all processes $(X(t), t \geq 0)$, we have for all $t > 0$

$$\mathbb{P}\left(\sup_{s > 0} X(s) > x\right) \geq \mathbb{P}(X(t) > x). \tag{9}$$

Furthermore, due to the union bound, we have for all $0 < t_1 < t_2$ that

$$\mathbb{P}\left(\sup_{s > 0} X(s) > x\right) \leq \mathbb{P}\left(\sup_{0 < s < t_1} X(s) > x\right) + \mathbb{P}\left(\sup_{t_1 \leq s < t_2} X(s) > x\right) + \mathbb{P}\left(\sup_{s \geq t_2} X(s) > x\right). \tag{10}$$

We use these types of lower and upper bounds to prove Theorem 1. Obviously, not all choices of t, t_1, and t_2 give sharp bounds. We can however make an educated guess about which choices will give the sharpest bounds. Let us first replace the sequence of random variables $(A(j), j \geq 1)$ with their expectation $1/\lambda$. Thus, we look at a simplified fork-join queue with deterministic arrivals. Because the arrivals are deterministic, the waiting times are mutually independent, and we are able to use standard extreme-value theory. We know from the Cramér-Lundberg approximation [1, Ch. XIII, Thm. 5.2] that $\mathbb{P}(\sup_{k \geq 0} \sum_{j=1}^{k} (S_i(j) - 1/\lambda) > x) \sim C \exp(-\gamma x)$, as $x \to \infty$, with $0 < C < 1$. Thus, $\mathbb{P}(\sup_{k \geq 0} \sum_{j=1}^{k} (S_i(j) - 1/\lambda) > \frac{1}{\gamma} \log N) \sim C/N$, as $N \to \infty$. Now we can conclude by using basic extreme-value results; see [7, Thm. 5.4.1, p. 188], that

$$\frac{\max_{i \leq N} \sup_{k \geq 0} \sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda}\right)}{\log N} \stackrel{\mathbb{P}}{\longrightarrow} \frac{1}{\gamma},$$

as $N \to \infty$. Thus, we know that $\max_{i \leq N} \sup_{k \geq 0} \sum_{j=1}^{k} (S_i(j) - 1/\lambda)$ centers around $\frac{1}{\gamma} \log N$. In order to find suitable lower and upper bounds of the form as given in (9) and (10), we need to estimate the hitting time

$$\tau^{(N)} := \inf\left\{k \geq 0 : \max_{i \leq N} \sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda}\right) \geq \frac{1}{\gamma} \log N\right\}.$$

As mentioned before, we have that $\mathbb{P}(\sup_{k \geq 0} \sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda}\right) > \frac{1}{\gamma} \log N) \sim C/N$ as $N \to \infty$. Thus, a good estimate $\hat{\tau}^{(N)}$ for $\tau^{(N)}$ should also satisfy the property that

$$\liminf_{N \to \infty} N \mathbb{P}\left(\sum_{j=1}^{\hat{\tau}^{(N)}} \left(S_i(j) - \frac{1}{\lambda}\right) > \frac{1}{\gamma} \log N\right) > 0 \tag{11}$$

and

$$\limsup_{N\to\infty} N\,\mathbb{P}\left(\sum_{j=1}^{\hat{\tau}^{(N)}}\left(S_i(j)-\frac{1}{\lambda}\right) > \frac{1}{\gamma}\log N\right) < \infty. \tag{12}$$

Now, by using Cramér's theorem and by using the fact that Λ is at least twice differentiable at γ, we know that

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\mathbb{P}\left(\sum_{j=1}^{n}\left(S_i(j)-\frac{1}{\lambda}\right)\geq nx\right)\right) = -\Lambda^*(x), \tag{13}$$

for all $x > \mathbb{E}[S_i(j)-1/\lambda]$ with $\Lambda^*(x) = \sup_{t\in\mathbb{R}}(tx-\Lambda(t))$; see [1, Ch. XIII, Thm. 2.1 (2.3)]. We write $\hat{\tau}^{(N)} = \hat{c}\log N$. Then we can conclude from Equation (13) that

$$\lim_{N\to\infty}\frac{1}{\log N}\log\left(\mathbb{P}\left(\sum_{j=1}^{\lfloor\hat{c}\log N\rfloor}\left(S_i(j)-\frac{1}{\lambda}\right)\geq x\hat{c}\log N\right)\right) = -\Lambda^*(x)\hat{c}. \tag{14}$$

Thus, in order to find a good estimate $\hat{\tau}^{(N)}$ for the hitting time $\tau^{(N)}$ we need to solve two equations. First, $x\hat{c} = 1/\gamma$, because we know that the longest steady-state waiting time under deterministic arrivals is approximately equal to $\frac{1}{\gamma}\log N$. Therefore the expression $x\hat{c}\log N$ in (14) should be the same as $\frac{1}{\gamma}\log N$. Second, $-\Lambda^*(x)\hat{c} = -1$, because we know from (11), (12), and (14) that for large N

$$\mathbb{P}\left(\sum_{j=1}^{\lfloor\hat{c}\log N\rfloor}\left(S_i(j)-\frac{1}{\lambda}\right)\geq x\hat{c}\log N\right) \approx \frac{1}{N} = \exp(-\Lambda^*(x)\hat{c}\log N).$$

Combining these two equations gives $\hat{c} = \frac{1}{\Lambda'(\gamma)\gamma}$ and $x = \Lambda'(\gamma)$. Clearly, $x\hat{c} = 1/\gamma$, and

$$\Lambda^*(x)\hat{c} = \frac{\Lambda^*(\Lambda'(\gamma))}{\gamma\Lambda'(\gamma)}.$$

From [4, Lem. 2.2.5(c)], we know that $\Lambda^*(\Lambda'(\gamma)) = \gamma\Lambda'(\gamma)$, thus indeed, $\Lambda^*(x)\hat{c} = 1$. Finally, we can conclude that $\hat{\tau}^{(N)} = \hat{c}\log N = \frac{1}{\gamma\Lambda'(\gamma)}\log N$. Obviously, in order to be a good estimation for a hitting time we need to have that $\Lambda'(\gamma) > 0$. This is the case because $\Lambda(\theta)$ is convex; see [1, Ch. XIII, Thm. 5.1].

Until this point, we know the first-order scaling of the largest of N steady-state waiting times with deterministic arrivals, and we can give an estimation of the hitting time of this value. Now, we can use these results to obtain a second-order convergence result for the longest steady-state waiting time with stochastic arrivals. Following the analysis above together with the lower bound in (9), we see that

$$\mathbb{P}\left(\frac{\max_{i\leq N} W_i(\infty) - \frac{1}{\gamma}\log N}{\sqrt{\log N}}\geq x\right)$$

$$\geq \mathbb{P}\left(\frac{\max_{i\leq N}\sup_{\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\log N<k<\frac{1}{\Lambda'(\gamma)\gamma}\log N}\sum_{j=1}^{k}(S_i(j)-A(j)) - \frac{1}{\gamma}\log N}{\sqrt{\log N}}\geq x\right), \tag{15}$$

with $\epsilon > 0$ and small. In Lemma 3, we prove that the right-hand side in (15) converges to a function that is close to the tail probability of a normally distributed random variable. Furthermore, we show in Lemmas 4, 5, and 6, that this lower bound is sharp. To achieve this, we first divide the supremum over all positive numbers in the random variable $\max_{i\leq N} W_i(\infty)$ in three parts. After that, we take the supremum over the intervals $\left[0,\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\log N\right]$, $\left(\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\log N,\left(\frac{1}{\Lambda'(\gamma)\gamma}+\epsilon\right)\log N\right]$, and $\left(\left(\frac{1}{\Lambda'(\gamma)\gamma}+\epsilon\right)\log N,\infty\right)$, with $\epsilon > 0$ and small. Consequently, we show that the tail probabilities of the first and third suprema of the maximum of N random walks asymptotically vanish, while

$$\mathbb{P}\left(\max_{i\leq N}\sup_{\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\log N<k<\left(\frac{1}{\Lambda'(\gamma)\gamma}+\epsilon\right)\log N}\sum_{j=1}^{k}(S_i(j)-A(j)) > \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)$$

converges to a limit close to the lower bound as $N\to\infty$.

Remark 2. *The lower bound presented in Equation (9) gives us information about the convergence rate of the result in Theorem 1. From the Berry-Esséen theorem [13], we know that when $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} X \sim \mathcal{N}(0,1)$, the convergence rate is of order $1/\sqrt{n}$. Thus, the lower bound in (9) shows that the convergence rate is of order $1/\sqrt{\log N}$.*

4 Proofs

Lemma 3. *Given the model in Section 2 where the sequence of service times $(S_i(j), i \geq 1, j \geq 1)$ satisfies Assumption 1, $0 < \epsilon < \frac{1}{\Lambda'(\gamma)\gamma}$, $t_1^{(N)} = \left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon\right) \log N$, and $t_2^{(N)} = \frac{1}{\Lambda'(\gamma)\gamma} \log N$, then for all $x \in \mathbb{R}$, we have that*

$$\liminf_{N \to \infty} \mathbb{P}\left(\max_{i \leq N} \sup_{t_1^{(N)} < k < t_2^{(N)}} \sum_{j=1}^{k}(S_i(j) - A(j)) > \frac{1}{\gamma} \log N + x\sqrt{\log N}\right)$$

$$\geq \mathbb{P}\left(\sigma_A \sqrt{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon} X_1 - \sigma_A \sqrt{\epsilon}|X_2| > x\right), \quad (16)$$

with $X_1, X_2 \sim \mathcal{N}(0,1)$ and independent.

Proof. In order to prove this convergence result, we first bound

$$\max_{i \leq N} \sup_{t_1^{(N)} < k < t_2^{(N)}} \sum_{j=1}^{k}(S_i(j) - A(j)) \geq \max_{i \leq N} \sup_{t_1^{(N)} < k < t_2^{(N)}} \sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) + \inf_{t_1^{(N)} < k < t_2^{(N)}} \sum_{j=1}^{k}\left(\frac{1}{\lambda} - A(j)\right).$$

We treat the terms on the right-hand side separately. We first prove that

$$\frac{\inf_{t_1^{(N)} < k < t_2^{(N)}} \sum_{j=1}^{k}\left(\frac{1}{\lambda} - A(j)\right)}{\sqrt{\log N}} \xrightarrow{d} \sigma_A \sqrt{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon} X_1 - \sigma_A \sqrt{\epsilon}|X_2|, \quad (17)$$

as $N \to \infty$. Afterwards, we prove that

$$\frac{\max_{i \leq N} \sup_{t_1^{(N)} < k < t_2^{(N)}} \sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) - \frac{1}{\gamma}\log N}{\sqrt{\log N}} \xrightarrow{\mathbb{P}} 0, \quad (18)$$

as $N \to \infty$.

The first convergence result follows from Donsker's theorem. The left-hand side in (17) is an infimum of a random walk with drift 0. Then for $(B(t), t \geq 0)$ a Brownian motion with drift 0 and standard deviation 1, by using Donsker's theorem [5] and the fact that the infimum is a continuous functional, we obtain that

$$\mathbb{P}\left(\frac{\inf_{t_1^{(N)} < k < t_2^{(N)}} \sum_{j=1}^{k}(\frac{1}{\lambda} - A(j))}{\sqrt{\log N}} > x\right) \xrightarrow{N \to \infty} \mathbb{P}\left(\inf_{\left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon\right) < s < \frac{1}{\Lambda'(\gamma)\gamma}} \sigma_A B(s) > x\right).$$

Furthermore, we can rewrite

$$\inf_{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon < s < \frac{1}{\Lambda'(\gamma)\gamma}} \sigma_A B(s) \stackrel{d}{=} \sigma_A B\left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon\right) - \inf_{0 < s < \epsilon} \sigma_A \tilde{B}(s),$$

where $\tilde{B}$ is an independent copy of B. Obviously, we have that $\sigma_A B\left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon\right) \stackrel{d}{=} \sigma_A \sqrt{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon} X_1$ with $X_1 \sim \mathcal{N}(0,1)$. Because $\inf_{0 < s < \epsilon} \sigma_A \tilde{B}(s) \stackrel{d}{=} \sigma_A \sqrt{\epsilon}|X_2|$, with $X_2 \sim \mathcal{N}(0,1)$, we have that the limit in (17) follows.

In order to prove the second convergence result, we define for $A \in \mathcal{F}_k$, with $\{\mathcal{F}_k, k \geq 1\}$ the natural filtration, the probability measure

$$\mathbb{P}_i(A) := \mathbb{E}\left[\exp\left(\gamma \sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right)\right) \mathbb{1}(A)\right];$$

7

see [1, Ch. XIII, Par. 3]. Now, we know that

$$\mathbb{E}_i\left[S_i(j) - \frac{1}{\lambda}\right] = \mathbb{E}\left[\left(S_i(j) - \frac{1}{\lambda}\right)\exp\left(\gamma\left(S_i(j) - \frac{1}{\lambda}\right)\right)\right] = \Lambda'(\gamma).$$

Thus, by checking the conditions in [1, Ch. XIII, Thm. 5.6], we see that

$$\mathbb{P}\left(\sup_{0\leq k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \geq \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)$$

$$= C\exp\left(-\gamma\left(\frac{1}{\gamma}\log N + x\sqrt{\log N}\right)\right)\Phi\left(-x\frac{\sqrt{\gamma\Lambda'(\gamma)}}{\sqrt{\Lambda''(\gamma)}}\right)(1+o(1)). \quad (19)$$

With the same approach, we get from [1, Ch. XIII, Thm. 5.6] that

$$\mathbb{P}\left(\sup_{0\leq k<t_1^{(N)}}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \geq \frac{1}{\gamma}\log N + x\sqrt{\log N}\right) = o\left(C\exp\left(-\gamma\left(\frac{1}{\gamma}\log N + x\sqrt{\log N}\right)\right)\right), \quad (20)$$

as $N \to \infty$, for all $x \in \mathbb{R}$. By applying the union bound, we get that

$$\mathbb{P}\left(\sup_{0\leq k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \geq \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)$$

$$\leq \mathbb{P}\left(\sup_{0\leq k<t_1^{(N)}}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \geq \frac{1}{\gamma}\log N + x\sqrt{\log N}\right) + \mathbb{P}\left(\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \geq \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)$$

$$\leq \mathbb{P}\left(\sup_{0\leq k<t_1^{(N)}}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \geq \frac{1}{\gamma}\log N + x\sqrt{\log N}\right) + \mathbb{P}\left(\sup_{0\leq k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \geq \frac{1}{\gamma}\log N + x\sqrt{\log N}\right).$$

We can conclude from these bounds, together with (19) and (20) that

$$\mathbb{P}\left(\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \geq \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)$$

$$= C\exp\left(-\gamma\left(\frac{1}{\gamma}\log N + x\sqrt{\log N}\right)\right)\Phi\left(-x\frac{\sqrt{\gamma\Lambda'(\gamma)}}{\sqrt{\Lambda''(\gamma)}}\right)(1+o(1)). \quad (21)$$

By using this expression it is easy to derive that for $x > 0$

$$\mathbb{P}\left(\max_{i\leq N}\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \leq \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)$$

$$= \mathbb{P}\left(\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \leq \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)^{N} \xrightarrow{N\to\infty} 1.$$

Similarly, for $x < 0$,

$$\mathbb{P}\left(\max_{i\leq N}\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \leq \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)$$

$$= \mathbb{P}\left(\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \leq \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)^{N} \xrightarrow{N\to\infty} 0.$$

Combining these two results gives us the limit in (18). Finally, the convergence result in (16) follows from the two limits in (17) and (18). $\qquad\square$

Lemma 4. *Given the model in Section 2 where the sequence of service times $(S_i(j), i \geq 1, j \geq 1)$ satisfies Assumption 1, $t_1^{(N)} = \left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon \right) \log N$, $\delta = \frac{\delta_1}{\Lambda'(\gamma)\gamma} + \delta_2$ with $\delta_{1,2} > 0$ and small, and $\epsilon = \delta^{1/4}$, then for all $x \in \mathbb{R}$, we have that*

$$\mathbb{P}\left(\max_{i \leq N} \sup_{0 \leq k < t_1^{(N)}} \sum_{j=1}^k (S_i(j) - A(j)) > \frac{1}{\gamma} \log N + x\sqrt{\log N} \right) \overset{N \to \infty}{\longrightarrow} 0. \tag{22}$$

Proof. We derive upper bounds for the left-hand side of (22) that converge to 0 as $N \to \infty$.

We get by using the subadditivity property of the sup operator and the union bound that

$$\mathbb{P}\left(\max_{i \leq N} \sup_{0 \leq k < t_1^{(N)}} \sum_{j=1}^k (S_i(j) - A(j)) > \frac{1}{\gamma} \log N \right) \tag{23}$$

$$\leq \mathbb{P}\left(\max_{i \leq N} \sup_{0 \leq k < t_1^{(N)}} \sum_{j=1}^k \left(S_i(j) - \frac{1}{\lambda} + \delta_1 \right) > \left(\frac{1}{\gamma} - \delta_2 \right) \log N \right) \tag{24}$$

$$+ \mathbb{P}\left(\sup_{k \geq 0} \sum_{j=1}^k \left(\frac{1}{\lambda} - \delta_1 - A(j) \right) > \delta_2 \log N + x\sqrt{\log N} \right). \tag{25}$$

First, because $\mathbb{E}[\frac{1}{\lambda} - \delta_1 - A(j)] < 0$, we get that

$$\mathbb{P}\left(\sup_{k \geq 0} \sum_{j=1}^k \left(\frac{1}{\lambda} - \delta_1 - A(j) \right) > \delta_2 \log N + x\sqrt{\log N} \right) \overset{N \to \infty}{\longrightarrow} 0.$$

Second, we can bound the term in (24) as follows;

$$\mathbb{P}\left(\max_{i \leq N} \sup_{0 \leq k < t_1^{(N)}} \sum_{j=1}^k \left(S_i(j) - \frac{1}{\lambda} + \delta_1 \right) > \left(\frac{1}{\gamma} - \delta_2 \right) \log N \right)$$

$$\leq \mathbb{P}\left(\max_{i \leq N} \sup_{0 \leq k < t_1^{(N)}} \sum_{j=1}^k \left(S_i(j) - \frac{1}{\lambda} \right) > \left(\frac{1}{\gamma} - \frac{\delta_1}{\Lambda'(\gamma)\gamma} - \delta_2 \right) \log N \right).$$

Now, we can bound this further to

$$\mathbb{P}\left(\max_{i \leq N} \sup_{0 \leq k < t_1^{(N)}} \sum_{j=1}^k \left(S_i(j) - \frac{1}{\lambda} \right) > \left(\frac{1}{\gamma} - \frac{\delta_1}{\Lambda'(\gamma)\gamma} - \delta_2 \right) \log N \right)$$

$$\leq \sum_{k=0}^{\lfloor t_1^{(N)} \rfloor} N \, \mathbb{P}\left(\sum_{j=1}^k \left(S_i(j) - \frac{1}{\lambda} \right) > \left(\frac{1}{\gamma} - \frac{\delta_1}{\Lambda'(\gamma)\gamma} - \delta_2 \right) \log N \right).$$

By using Chernoff's bound we obtain that for $\Lambda(\theta) < \infty$

$$\sum_{k=0}^{\lfloor t_1^{(N)} \rfloor} N \, \mathbb{P}\left(\sum_{j=1}^k \left(S_i(j) - \frac{1}{\lambda} \right) > \left(\frac{1}{\gamma} - \frac{\delta_1}{\Lambda'(\gamma)\gamma} - \delta_2 \right) \log N \right) \tag{26}$$

$$\leq N \sum_{k=0}^{\lfloor t_1^{(N)} \rfloor} \exp(k\Lambda(\theta)) \exp\left(-\theta \left(\frac{1}{\gamma} - \frac{\delta_1}{\Lambda'(\gamma)\gamma} - \delta_2 \right) \log N \right)$$

$$= N \frac{-1 + \exp\left((\lfloor t_1^{(N)} \rfloor + 1)\Lambda(\theta) \right)}{\exp(\Lambda(\theta)) - 1} \exp\left(-\theta \left(\frac{1}{\gamma} - \frac{\delta_1}{\Lambda'(\gamma)\gamma} - \delta_2 \right) \log N \right). \tag{27}$$

9

Now,

$$\frac{\log\left(N^{\frac{-1+\exp\left((\lfloor t_1^{(N)}\rfloor+1)\Lambda(\theta)\right)}{\exp(\Lambda(\theta))-1}}\exp\left(-\theta\left(\frac{1}{\gamma}-\frac{\delta_1}{\Lambda'(\gamma)\gamma}-\delta_2\right)\log N\right)\right)}{\log N}$$

$$\stackrel{N\to\infty}{\longrightarrow} 1-\left(\theta\left(\frac{1}{\gamma}-\frac{\delta_1}{\Lambda'(\gamma)\gamma}-\delta_2\right)-\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\Lambda(\theta)\right).$$

In order to make the bound in (27) as sharp as possible, we need to choose a convenient θ. The choice of θ that gives the sharpest bound maximizes the function $\theta\left(\frac{1}{\gamma}-\frac{\delta_1}{\Lambda'(\gamma)\gamma}-\delta_2\right)-\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\Lambda(\theta)$. We have that $\delta=\frac{\delta_1}{\Lambda'(\gamma)\gamma}+\delta_2$ and $\epsilon=\delta^{1/4}$. Furthermore, we choose $\theta=\gamma+\sqrt{\delta}$. This gives us a sharp enough bound in (27). We obviously have that

$$\sup_{\eta\in\mathbb{R}}\left(\eta\left(\frac{1}{\gamma}-\delta\right)-\left(\frac{1}{\Lambda'(\gamma)\gamma}-\delta^{1/4}\right)\Lambda(\eta)\right)\geq\left((\gamma+\sqrt{\delta})\left(\frac{1}{\gamma}-\delta\right)-\left(\frac{1}{\Lambda'(\gamma)\gamma}-\delta^{1/4}\right)\Lambda(\gamma+\sqrt{\delta})\right).$$

The first order Taylor series of $\Lambda(\gamma+\sqrt{\delta})$ around γ gives

$$\Lambda(\gamma+\sqrt{\delta})=\Lambda(\gamma)+\sqrt{\delta}\Lambda'(\gamma)+O(\delta)=\sqrt{\delta}\Lambda'(\gamma)+O(\delta).$$

Thus,

$$\left((\gamma+\sqrt{\delta})\left(\frac{1}{\gamma}-\delta\right)-\left(\frac{1}{\Lambda'(\gamma)\gamma}-\delta^{1/4}\right)\Lambda(\gamma+\sqrt{\delta})\right)=1+\delta^{3/4}\Lambda'(\gamma)+O(\delta)>1,$$

for δ small enough. Thus the expression in (27) is upper bounded by the term $N^{-\delta^{3/4}\Lambda'(\gamma)-O(\delta)}\stackrel{N\to\infty}{\longrightarrow}0$. □

Lemma 5. *Given the model in Section 2 where the sequence of service times $(S_i(j),i\geq 1,j\geq 1)$ satisfies Assumption 1, $0<\epsilon<\frac{1}{\Lambda'(\gamma)\gamma}$, $t_1^{(N)}=\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\log N$, and $t_3^{(N)}=\left(\frac{1}{\Lambda'(\gamma)\gamma}+\epsilon\right)\log N$, then for all $x\in\mathbb{R}$, we have that*

$$\limsup_{N\to\infty}\mathbb{P}\left(\max_{i\leq N}\sup_{t_1^{(N)}\leq k<t_3^{(N)}}\sum_{j=1}^{k}(S_i(j)-A(j))>\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)$$

$$\leq\mathbb{P}\left(\sigma_A\sqrt{\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon}X_1+\sigma_A\sqrt{2\epsilon}|X_2|>x\right),\quad(28)$$

with $X_1,X_2\sim\mathcal{N}(0,1)$ and independent.

Proof. In order to prove this lemma, we first rewrite

$$\frac{\max_{i\leq N}\sup_{t_1^{(N)}\leq k<t_3^{(N)}}\sum_{j=1}^{k}(S_i(j)-A(j))-\frac{1}{\gamma}\log N}{\sqrt{\log N}}$$

$$\leq\frac{\max_{i\leq N}\sup_{t_1^{(N)}\leq k<t_3^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)-\frac{1}{\gamma}\log N}{\sqrt{\log N}}+\frac{\sup_{t_1^{(N)}\leq k<t_3^{(N)}}\sum_{j=1}^{k}\left(\frac{1}{\lambda}-A(j)\right)}{\sqrt{\log N}}$$

$$\leq\frac{\max_{i\leq N}\sup_{k\geq 0}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)-\frac{1}{\gamma}\log N}{\sqrt{\log N}}+\frac{\sup_{t_1^{(N)}\leq k<t_3^{(N)}}\sum_{j=1}^{k}\left(\frac{1}{\lambda}-A(j)\right)}{\sqrt{\log N}}.\quad(29)$$

We first look at the first term in (29). This term gives the rescaled longest steady-state waiting time of N i.i.d. $D/G/1$ queues. We know that

$$\mathbb{P}\left(\sup_{k\geq 0}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)>x\right)\sim C\exp(-\gamma x),$$

as $x \to \infty$, with $0 < C < 1$; see [1, Ch. XIII, Thm. 5.2]. Thus for $x > 0$,

$$\mathbb{P}\left(\frac{\max_{i \leq N} \sup_{k \geq 0} \sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda} \right) - \frac{1}{\gamma} \log N}{\sqrt{\log N}} > x \right)$$

$$\sim 1 - \left(1 - C \exp(-\gamma(1/\gamma \log N + x\sqrt{\log N})) \right)^N \overset{N \to \infty}{\longrightarrow} 0.$$

Similarly, for $x < 0$,

$$\mathbb{P}\left(\frac{\max_{i \leq N} \sup_{k \geq 0} \sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda} \right) - \frac{1}{\gamma} \log N}{\sqrt{\log N}} > x \right)$$

$$\sim 1 - \left(1 - C \exp(-\gamma(1/\gamma \log N + x\sqrt{\log N})) \right)^N \overset{N \to \infty}{\longrightarrow} 1.$$

Thus, the first term in (29) converges in probability to 0.

Now, we prove convergence of the tail probability of the second term in (29). This term is a supremum of a random walk with drift 0. Then for $(B(t), t \geq 0)$ a Brownian motion with drift 0 and standard deviation 1, by using Donsker's theorem [5] and the fact that the supremum is a continuous functional, we obtain with a similar analysis as in Lemma 3, that

$$\mathbb{P}\left(\frac{\sup_{t_1^{(N)} \leq k < t_3^{(N)}} \sum_{j=1}^{k} (\frac{1}{\lambda} - A(j))}{\sqrt{\log N}} > x \right) \overset{N \to \infty}{\longrightarrow} \mathbb{P}\left(\sigma_A \sqrt{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon} X_1 + \sigma_A \sqrt{2\epsilon} |X_2| > x \right).$$

$\square$

Lemma 6. *Given the model in Section 2 where the sequence of service times $(S_i(j), i \geq 1, j \geq 1)$ satisfies Assumption 1, $\delta = \frac{\delta_1}{\Lambda'(\gamma)\gamma} + \delta_2$ with $\delta_{1,2} > 0$ and small, $\epsilon = \delta^{1/4}$, and $t_3^{(N)} = \left(\frac{1}{\Lambda'(\gamma)\gamma} + \epsilon \right) \log N$, then for all $x \in \mathbb{R}$, we have that*

$$\mathbb{P}\left(\max_{i \leq N} \sup_{k \geq t_3^{(N)}} \sum_{j=1}^{k} (S_i(j) - A(j)) > \frac{1}{\gamma} \log N + x\sqrt{\log N} \right) \overset{N \to \infty}{\longrightarrow} 0. \tag{30}$$

Proof. As in the proof of Lemma 4, we derive upper bounds for

$$\mathbb{P}\left(\max_{i \leq N} \sup_{k \geq t_3^{(N)}} \sum_{j=1}^{k} (S_i(j) - A(j)) > \frac{1}{\gamma} \log N + x\sqrt{\log N} \right)$$

that converge to 0 as $N \to \infty$.

First, we see that by using subadditivity and the union bound, we obtain

$$\mathbb{P}\left(\max_{i \leq N} \sup_{k \geq t_3^{(N)}} \sum_{j=1}^{k} (S_i(j) - A(j)) > \frac{1}{\gamma} \log N + x\sqrt{\log N} \right)$$

$$\leq \mathbb{P}\left(\max_{i \leq N} \sup_{k \geq t_3^{(N)}} \sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda} + \delta_1 \right) > \left(\frac{1}{\gamma} - \delta_2 \right) \log N \right)$$

$$+ \mathbb{P}\left(\sup_{k \geq 0} \sum_{j=1}^{k} \left(\frac{1}{\lambda} - \delta_1 - A(j) \right) > \delta_2 \log N + x\sqrt{\log N} \right).$$

As in the proof of Lemma 4, we have that

$$\mathbb{P}\left(\sup_{k \geq 0} \sum_{j=1}^{k} \left(\frac{1}{\lambda} - \delta_1 - A(j) \right) > \delta_2 \log N + x\sqrt{\log N} \right) \overset{N \to \infty}{\longrightarrow} 0.$$

11

Furthermore, observe that $\log \mathbb{E}[\exp(\theta(S_i(j) - 1/\lambda + \delta_1))] = \Lambda(\theta) + \theta\delta_1$. Now, as in the proof of Lemma 4, we can bound

$$\mathbb{P}\left(\max_{i \leq N} \sup_{k \geq t_3^{(N)}} \sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda} + \delta_1\right) > \left(\frac{1}{\gamma} - \delta_2\right)\log N\right) \tag{31}$$

$$\leq N \sum_{k=\lfloor t_3^{(N)}\rfloor}^{\infty} \mathbb{P}\left(\sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda} + \delta_1\right) > \left(\frac{1}{\gamma} - \delta_2\right)\log N\right) \tag{32}$$

$$\leq N \sum_{k=\lfloor t_3^{(N)}\rfloor}^{\infty} \exp(k(\Lambda(\theta) + \theta\delta_1)) \exp\left(-\theta\left(\frac{1}{\gamma} - \delta_2\right)\log N\right)$$

$$= N \frac{\exp\left(\lfloor t_3^{(N)}\rfloor(\Lambda(\theta) + \theta\delta_1)\right)}{\exp(\Lambda(\theta) + \theta\delta_1) - 1} \exp\left(-\theta\left(\frac{1}{\gamma} - \delta_2\right)\log N\right), \tag{33}$$

when $\Lambda(\theta) + \theta\delta_1 < 0$. When $\Lambda(\theta) + \theta\delta_1 \geq 0$ the sum in the upper bound diverges to ∞. Now, for the case $\Lambda(\theta) + \theta\delta_1 < 0$, we have that

$$\frac{\log\left(N\frac{\exp\left(\lfloor t_3^{(N)}\rfloor(\Lambda(\theta)+\theta\delta_1)\right)}{\exp(\Lambda(\theta)+\theta\delta_1)-1}\exp\left(-\theta\left(\frac{1}{\gamma}-\delta_2\right)\log N\right)\right)}{\log N} \stackrel{N\to\infty}{\longrightarrow} 1 + \left(\frac{1}{\Lambda'(\gamma)\gamma} + \epsilon\right)(\Lambda(\theta) + \theta\delta_1) - \theta\left(\frac{1}{\gamma} - \delta_2\right).$$

As in the proof of Lemma 4, we have $\delta = \frac{\delta_1}{\Lambda'(\gamma)\gamma} + \delta_2$ and $\epsilon = \delta^{1/4}$. We now get after a similar derivation as in the proof of Lemma 4 that $\theta = \gamma - \sqrt{\delta}$ gives a sharp bound. First, observe that $\Lambda(\gamma - \sqrt{\delta}) = -\sqrt{\delta}\Lambda'(\gamma) + O(\delta)$, thus $\Lambda(\theta) + \theta\delta_1 = -\sqrt{\delta}\Lambda'(\gamma) + (\gamma - \sqrt{\delta})\delta_1 + O(\delta) = -\sqrt{\delta}\Lambda'(\gamma) + O(\delta) < 0$ for δ small enough, thus the upper bound in (33) holds. Second, we see that

$$\sup_{\eta \in \mathbb{R}}\left(\eta\left(\frac{1}{\gamma} - \delta_2\right) - \left(\frac{1}{\Lambda'(\gamma)\gamma} + \epsilon\right)(\Lambda(\eta) + \eta\delta_1)\right)$$

$$\geq (\gamma - \sqrt{\delta})\left(\frac{1}{\gamma} - \delta_2\right) - \left(\frac{1}{\Lambda'(\gamma)\gamma} + \epsilon\right)(\Lambda(\gamma - \sqrt{\delta}) + (\gamma - \sqrt{\delta})\delta_1).$$

So, we can conclude that

$$(\gamma - \sqrt{\delta})\left(\frac{1}{\gamma} - \delta_2\right) - \left(\frac{1}{\Lambda'(\gamma)\gamma} + \delta^{1/4}\right)(\Lambda(\gamma - \sqrt{\delta}) + (\gamma - \sqrt{\delta})\delta_1) = 1 + \delta^{3/4}\Lambda'(\gamma) + O(\delta) > 1$$

for δ small enough, thus the expression in (33) converges to 0 as $N \to \infty$. $\qquad\square$

Proof of Theorem 1. First, to prove a lower bound, we see that

$$\max_{i \leq N} W_i(\infty) \geq_{st.} \max_{i \leq N} \sum_{j=1}^{\left\lfloor \frac{1}{(\Lambda'(\gamma)\gamma)}\log N\right\rfloor} (S_i(j) - A(j)).$$

Thus, combining this inequality with the result from Lemma 3, we see that

$$\liminf_{N\to\infty} \mathbb{P}\left(\max_{i \leq N} W_i(\infty) > \frac{1}{\gamma}\log N + x\sqrt{\log N}\right) \geq \mathbb{P}\left(\sigma_A\sqrt{\frac{1}{\Lambda'(\gamma)\gamma}}X > x\right).$$

Second, by using the union bound of the types as given in (10) and explained in Section 3, we get from Lemmas 4,

12

5, and 6, with $t_1^{(N)} = \left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon\right)\log N$ and $t_3^{(N)} = \left(\frac{1}{\Lambda'(\gamma)\gamma} + \epsilon\right)\log N$, that

$$\limsup_{N\to\infty} \mathbb{P}\left(\max_{i\leq N} W_i(\infty) > \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)$$

$$\leq \limsup_{N\to\infty} \mathbb{P}\left(\max_{i\leq N}\sup_{t_1^{(N)}\leq k<t_3^{(N)}}\sum_{j=1}^k (S_i(j) - A(j)) > \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)$$

$$\leq \mathbb{P}\left(\sigma_A\sqrt{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon}\,X_1 + \sigma_A\sqrt{2\epsilon}\,|X_2| > x\right).$$

Finally, we have that

$$\mathbb{P}\left(\sigma_A\sqrt{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon}\,X_1 + \sigma_A\sqrt{2\epsilon}\,|X_2| > x\right) \xrightarrow{\epsilon\downarrow 0} \mathbb{P}\left(\sigma_A\sqrt{\frac{1}{\Lambda'(\gamma)\gamma}}\,X > x\right).$$

$\square$

References

[1] Søren Asmussen. *Applied probability and queues*, volume 2. Springer, 2003.

[2] François Baccelli. Two parallel queues created by arrivals with two demands: The M/G/2 symmetrical case. *Technical report RR–0426, INRIA*, 1985.

[3] François Baccelli and Armand M. Makowski. Queueing models for systems with synchronization constraints. *Proceedings of the IEEE*, 77(1):138–161, 1989.

[4] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38. Springer Science & Business Media, 2009.

[5] Monroe David Donsker. *An invariance principle for certain probability limit theorems*, volume 6. Memoirs of the American Mathematical Society, 1951.

[6] Leopold Flatto and Sann Hahn. Two parallel queues created by arrivals with two demands I. *SIAM Journal on Applied Mathematics*, 44(5):1041–1053, 1984.

[7] Laurens de Haan and Ana Ferreira. *Extreme value theory: an introduction.* Springer Science & Business Media, 2006.

[8] Rasoul Haji and Gordon F Newell. A relation between stationary queue and waiting time distributions. *Journal of Applied Probability*, 8(3):617–620, 1971.

[9] Stephanus J. de Klein. *Fredholm integral equations in queueing analysis.* PhD thesis, Rijksuniversiteit Utrecht, 1988.

[10] Sung-Seok Ko and Richard F. Serfozo. Response times in M/M/s fork-join networks. *Advances in Applied Probability*, 36(3):854–871, 2004.

[11] David V Lindley. The theory of queues with a single server. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, pages 277–289. Cambridge University Press, 1952.

[12] Mirjam Meijer, Dennis Schol, Willem van Jaarsveld, Maria Vlasiou, and Bert Zwart. Extreme-value theory for large fork-join queues, with applications to high-tech supply chains. *https://arxiv.org/abs/2105.09189*, 2021.

[13] Reinhard Michel. On the constant in the nonuniform version of the Berry-Esséen theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 55(1):109–117, 1981.

[14] Randolph Nelson and Asser N Tantawi. Approximate analysis of fork/join synchronization in parallel queues. *IEEE Transactions on Computers*, 37(6):739–743, 1988.

[15] Dennis Schol, Maria Vlasiou, and Bert Zwart. Large fork-join queues with nearly deterministic arrival and service times. *Mathematics of Operations Research*, 47(2):1335–1364, 2021.

[16] Dennis Schol, Maria Vlasiou, and Bert Zwart. Tail asymptotics for the delay in a brownian fork-join queue. *arXiv preprint arXiv:2208.04796*, 2022.

[17] Paul E. Wright. Two parallel processors with coupled inputs. *Advances in Applied Probability*, 24(4):986–1007, 1992.