Queueing Systems (2025) 109:9
https://doi.org/10.1007/s11134-025-09937-2

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

Dennis Schol[1] · **Maria Vlasiou**[2] · **Bert Zwart**[1,3]

Received: 15 September 2023 / Revised: 15 October 2024 / Accepted: 13 January 2025

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 times distribution and solves the Cramér–Lundberg equation with stochastic service times and deterministic interarrival 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 the distributional form of 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.

Keywords Extreme value theory · Supply chains · Distributional Little's Law · Heterogeneous servers · Tail behaviour

Mathematics Subject Classification 60G70 · 60K25 · 60K30 · 90B22

1 Introduction

Fork-join queues model parallel-processing systems, where incoming jobs consisting of subtasks are forked among different service stations. After completion of the subtasks at the service stations, the subtasks are joined to complete the processing of the whole job. This queueing system has been studied in the context of congestion in complex networks, such as assembly systems, healthcare, communication networks,

✉ Maria Vlasiou
m.vlasiou@utwente.nl

1 Eindhoven University of Technology, Eindhoven, Kingdom of the Netherlands

2 University Twente Faculty EEMCS: Universiteit Twente Faculteit Elektrotechniek Wiskunde en Informatica, Enschede, Kingdom of the Netherlands

3 Centrum Wiskunde en Informatica, Amsterdam, Kingdom of the Netherlands

Published online: 22 February 2025

and supply chains. For example, in supply chains, the arriving jobs represent orders for products arriving to a manufacturer, the service stations represent different suppliers of the manufacturer, and joining the completed subtasks represents the process of manufacturing the components into a final product. Fork-join queues capture the (single) arrival process of jobs (e.g., orders) to the system, and their decomposition into parts processed independently in the network (e.g., by different machines in an assembly system).

Our work is motivated by supply chains in high-tech manufacturing. High-tech companies, such as ASML, Philips, and Boeing, are types of original-equipment manufacturers that assemble thousands of components, each produced using specialized equipment, into complex systems. These supply chains are distinguished by general supply chains by the following characteristics. First, high-tech companies typically have thousands of direct suppliers. For instance, the Dutch lithography equipment manufacturer ASML had 4750 direct suppliers in 2020; see [1, p. 53]. Second, high-tech supply chains often comprise multiple tiers, as suppliers make products involving parts delivered by other suppliers; see [1, p. 57]. Due to this structure, the state supply chain is less observable [27]. Next, suppliers may be highly specialized, and thus, no backup may be possible. As a result, the supply chain is vulnerable and can be easily disrupted; see [1, p. 95]. For instance, of the 4750 suppliers of ASML, 188 are called critical suppliers. Last, in high-tech manufacturing revenue is derived from a relatively small number of products that have high value. For example, the ASML total production in 2020 was 258 machines [1, p. 92], each costing more than 150 million USD [44]. This underlines the cost of potential delays in high-tech supply chains. For further background, see also [40].

Such networks can be large and assembly is only possible upon availability of all parts. If one component is missing, the final product cannot be assembled, giving rise to costly delays. Because of the common arrival process, the total inventory per part, including backlogged jobs, is equal for all parts. However, as a result of variations in the service times, the number of backlogged jobs may vary per part. Thus, the bottleneck of the system is caused by the slowest production line in the system. Hence, a straightforward measure for the performance of high-tech supply chains is the delay of the slowest supplier. The maximum queue length and the longest waiting time are the two performance measures that capture delays and are thus of particular interest to the manufacturer. This setting motivates us to investigate delays in a large fork-join queueing system. We consider a fork-join network of N queues driven by a common arrival process and having independent, identical service processes.

In this setting, a key quantity of interest is the behavior of the longest queue when the system is in steady-state. As we try to model systems with many servers, we are typically interested in the behavior of this random variable as $N \to \infty$. Under Brownian assumptions, in [26], we showed that the maximum queue length is in the domain of attraction of the normal distribution. Specifically, assuming that $\{B_i(t), t \geq 0\}$ and $\{B_A(t), t \geq 0\}$ are Brownian motions with standard deviations σ and σ_A, representing the independent service times and the common arrival process, respectively, it is shown that the maximum queue length $\max_{i \leq N}(B_i(s) + B_A(s) - \beta s)$ satisfies:

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

with $X \overset{d}{=} \mathcal{N}(0,1)$. 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 $\max_{i\leq N} W_i(\infty)$ with a common arrival process $A(\cdot)$. Denote by $S_i(j)$ the service time of the j-th customer in queue i, while $A(j)$ indicates the interarrival time between the $(j-1)$-st and the j-th customer. Assume 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 interarrival times and service times are mutually independent. Then by Lindley's recursion, we have that $\max_{i\leq N} W_i(\infty) \overset{d}{=} \max_{i\leq N}\sup_{k\geq 0}\sum_{j=1}^{k}(S_i(j)-A(j))$, i.e., the maximum steady-state waiting time is a maximum of N dependent random variables due to the common arrival process $(A(j),\, j\geq 1)$. In Theorem 1, we prove that the maximum steady-state waiting time in the N queues satisfies a central limit result: Centered around a term with the order $\log N$ and scaled with a term with the order $\sqrt{\log N}$, it converges to a normally distributed random variable. In Theorem 2, we show that a similar result holds for the maximum queue length. Last, in Corollary 1, we present a similar result for non-homogeneous servers having different service distributions.

This paper is organized as follows. We briefly discuss the literature on fork-join queues in Sect. 2. The main results are gathered in Sect. 3. In Sect. 4, we give an intuitive explanation why the results hold and how to prove them, while Sect. 5 is devoted to proofs.

2 Related literature

Fork-join queues have been extensively studied. The first papers on this topic were focused on two service stations. In [12, 13], the authors consider a Poisson arrival stream and two independently working servers with exponential service times. The authors describe the asymptotic behavior of the joint queue-length probability distribution as the number of tasks in one queue goes to infinity [13, Thm. 7.2]. Furthermore, asymptotic expressions are given for the expectation and distribution of the length of one queue conditioned on the number of tasks in the other queue as the number of tasks in the second queue goes to infinity [12, Thm. 1 & 2]. These results were later extended and generalized in [4, 18, 45, 53]. The fact that both service stations have the same arrival stream complicates the exact analysis of fork-join queues; already for fork-join queues with two servers there are no exact expressions for the joint probability distribution of the number of tasks outside the asymptotic regime of one of the two queues. Before our work [41], there were no asymptotic results known to fork-join queues with more than two servers.

For the fork-join queue with more than two servers, stability has been shown in [21]. The main focus has been traditionally on finding bounds for performance measures, e.g., by using inequalities on the maximum of associated random variables and stochastic orderings [6, 32, 33]. For instance, in [6, Cor. 3.4], the authors report that the fork-join queue with i.i.d. single servers and deterministic arrivals minimizes the expected maximum response time among all possible interarrival distributions with fixed mean. This is a computable lower bound, as the fork-join queue with deterministic arrivals can be seen as a queueing system with N independent parallel queues. Thus, the cumulative distribution function of the longest waiting time of a fork-join queue with deterministic arrivals is known. As seen also in our previous work [41], the maximum queue length of a fork-join queue with light-tailed services is relatively stable in probability and possesses the same first-order behavior as the fork-join queue with deterministic arrivals. Thus, the first-order convergence result is the same as the first-order convergence result of the lower bound in [6] as $N \to \infty$. In [19], the analysis in [6] is extended to fork-join queues with exponential interarrival times and service stations all having s memoryless servers. In [16], an algorithm is described to obtain the average response time when service times are Erlang distributed. Furthermore, in [51] the authors interpolate between light-traffic and heavy-traffic results and obtain approximations for symmetric fork-join queues. In [20], the authors present a closed-form approximation of the sojourn time of a job in a $G/M/1$ fork-join queue. In [38], approximations of the response time distribution are provided. In [52], the fork-join queue is studied under the assumption that the number of subtasks is less than or equal to the number of service stations. It is proven that when the number of subtasks k_N is $o(N^{1/4})$, the queues at any k_N servers are asymptotically independent as $N \to \infty$ [38, Thm. 4.1]. The authors also prove that for each $k_N \leq N$ the longest steady-state waiting time is stochastically dominated by the longest steady-state waiting time of an identical system, but with independent arrivals [38, Thm. 4.3].

Heavy-traffic approximations are discussed in [17, 23–25, 34, 35, 46, 50]. In [50], the author gives a heavy-traffic analysis for fork-join queues and shows weak convergence of several processes, such as the joint queue lengths in front of each server. Furthermore, in [34] it is proven that various emerging limiting processes are in fact multi-dimensional reflected Brownian motions. Nguyen [35] extends this result to a fork-join queue with multiple job types, which is also the setting in [36] that proposes an asymptotically optimal scheduling policy in diffusion scale. Lu and Pang study fork-join networks in [23–25]. In [23], they investigate a fork-join network where each service station has multiple servers under non-exchangeable synchronization and operates in the quality-driven regime. They derive functional central limit theorems for the number of tasks waiting in the waiting buffers for synchronization and for the number of synchronized jobs. In [24], they extend this analysis to a fork-join network with a fixed number of service stations, each having many servers, where the system operates in the Halfin–Whitt regime. In [25], they investigate these heavy-traffic limits for a fixed number of infinite-server stations, where services are dependent and could be disrupted.

Research has also been done on controlling performance measures in the fork-join queue, see, for instance, [3] that investigates the control of a fork-join queue in heavy traffic by using feedback procedures. Using a stochastic gradient projection method

that reacts to queue-length changes, the authors in [37] propose a general framework for the design of robust scheduling policies for flexible fork-join networks when the system parameters are unknown. Other studies are carried out in [28, 29].

Specific results on the interplay between fork-join queues and heavy-tailed services can be found in [39, 54, 55]. In [39, Thm. 2], asymptotic lower and upper bounds for the tail probability of the longest waiting time in steady state are given; however, these bounds are not sharp when N is large.

In [54] and [55], the authors investigate the fork-join queue with blocking. More work on fork-join queues with blocking is presented in [7, 8]. In [11], the fork-join queue, under different execution programs, is studied. In [48], the mean-value approach is used to approximate performance measures. In [49], several bounds for performance measures of the fork-join queue with exponential interarrival and service times are given under a variable number of subtasks. In [47], approximation techniques are derived for the fork-join queue with exponential interarrival and general service times. In [30], a fork-join queueing model is studied where the available computational resources are allocated among the different servers, according to a certain algorithm, with the aim to minimize the maximum queue length.

In previous work [41], we study an N-server fork-join queue with nearly deterministic arrival and service times, deriving a fluid limit for the maximum queue length as $N \to \infty$. The fluid limit is shown to depend on the initial number of tasks. In order to prove these results, we develop extreme-value theory and diffusion approximations for the queue lengths. Further, in [26] we prove the same convergence results as in this paper, but for the special case of symmetric fork-join queues under Brownian assumptions. Here, we extend these results to general light-tailed distributions and allow for non-homogeneous servers. In [42], we derive tail asymptotics for the delay in a Brownian fork-join queue, showing that the behavior is different in three regimes defined by a threshold. Deviations beyond a threshold exhibit a form of asymptotic independence, and small deviations reveal a highly irregular behavior with a clear dependence among the N suprema, while a non-trivial transition occurs on the threshold. Last, in [43], motivated by parallel computing systems we study the maximum waiting time in an N-server fork-join queue with heavy-tailed service times and slowdown as $N \to \infty$, providing sharp convergence results to the supremum of an extremal process with negative drift. Further, we extend the analysis to the sojourn time of tasks and to independent service times (i.e., without the slowdown process common among all servers). An extended discussion of all these results can be found in the thesis [40].

3 Model and main results

Consider 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 the service, interarrival and waiting times, and queue lengths for this system. First, the sequence of nonnegative 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 nonnegative 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. Next, the waiting time of the n-th task in front of server i is denoted by $W_i(n)$, which by Lindley's recursion [22] is known to be given by

$$W_i(n) = \sup_{0 \leq k \leq n} \sum_{j=k+1}^{n} (S_i(j) - A(j))\}, \tag{2}$$

using the convention $\sum_{j=n+1}^{n} [\cdot]_j = 0$ and assuming that the system is empty upon arrival of the first task.

Last, $Q_i(n)$ is the number of tasks waiting in line (excluding the task in service) to be processed at server i upon arrival of the n-th job. Without giving a detailed sample-path representation of $Q_i(n)$, note that

$$\left\{ Q_i(n+k) \geq k\} = \{ \sum_{j=n+1}^{n+k} A(j) \leq W_i(n) \right\}. \tag{3}$$

Under the assumptions above, building on [21], there exist random vectors $(Q_i(\infty), i = 1, \ldots, N)$ and $(W_i(\infty), i = 1, \ldots, N)$ such that $(Q_i(n), i = 1, \ldots, N) \xrightarrow{d} (Q_i(\infty), i = 1, \ldots, N)$ and $(W_i(n), i = 1, \ldots, N) \xrightarrow{d} (W_i(\infty), i = 1, \ldots, N)$ as $n \to \infty$, where $\xrightarrow{d}$ denotes convergence in distribution. Their marginal distributions satisfy distributional Little's law [2, Thm. X.4.3]

$$Q_i(\infty) \stackrel{d}{=} \mathbf{N}_A(W_i(\infty)), \tag{4}$$

where $\mathbf{N}_A(t)$ indicates the number of total arrivals until time t and is independent of the steady-state waiting time $W_i(\infty)$ at queue i.

We can now write the cumulative distribution function of the longest steady-state waiting time $\max_{i \leq N} W_i(\infty)$ 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 Sect. 3 with $W_i(1) = 0$ for all $i \leq N$, we have that the longest waiting time in steady state satisfies*

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

Proof From the definition of waiting times (2), the maximum waiting time of the n-th subtask across all queues is given by

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

Additionally, the steady-state maximum waiting time satisfies

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

Since by reversibility we have that

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

we obtain the lemma by using the monotone convergence theorem. □

Note that this representation is consistent with the well-known [2, Prop. X.1.1] result

$$W_i(\infty) \stackrel{d}{=} \sup_{k\ge 0} \sum_{j=1}^{k} (S_i(j) - A(j)).$$

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

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

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 constant $\gamma > 0$ 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 [9, 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 $\theta = \gamma$.

3.1 Homogeneous servers

For the model described thus far, we prove the convergence of the centered and scaled steady-state maximum waiting time and maximum queue length to an external normally distributed random variable in the following two theorems.

Theorem 1 *Let X be a standard normal random variable independent of everything else. For the model in Sect. 3 where the sequence of service times $(S_i(j), i \geq 1, j \geq 1)$ satisfies Assumption 1, we have that as $N \to \infty$*

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

A heuristic of the proof of this theorem is provided in Sect. 4, while the full details are given in Sect. 5. In order to derive a similar result for the maximum queue length, we link it to the maximum waiting time by proving for our model a distributional Little's law in the lemma below.

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

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

with $\mathbf{N}_A$ independent of $\max_{i\leq N} W_i(\infty)$.

Proof We modify the approach of [2, Thm. X.4.3], where (3) is used to prove (4): Using (3), we obtain

$$\begin{aligned}\left\{\max_{i\leq N} Q_i(n+k) < k\right\} &= \cap_{i=1}^N \{Q_i(n+k) < k\} \\ &= \cap_{i=1}^N \left\{\sum_{j=n+1}^{n+k} A(j) > W_i(n)\right\} \\ &= \left\{\sum_{j=n+1}^{n+k} A(j) > \max_{i\leq N} W_i(n)\right\}. \end{aligned} \tag{5}$$

Taking the complement of the probabilities in (5) and also using that $\sum_{j=n+1}^{n+k} A(j)$ is independent of $\max_{i\leq N} W_i(n)$ and equal in distribution to $\sum_{j=1}^{k} \tilde{A}(j)$, with $\tilde{A}(j)$ an independent copy of $A(j)$, we obtain, after taking $n \to \infty$,

$$\mathbb{P}(\max_{i\leq N} Q_i(\infty) \geq k) = \mathbb{P}(\sum_{j=1}^{k} \tilde{A}(j) \leq \max_{i\leq N} W_i(n)). \tag{6}$$

Since $\mathbf{N}_A(t) = \max\{k : \sum_{j=1}^{k} \tilde{A}(j) \le t\}$, we see that the right-hand side of (6) equals $\mathbb{P}(\mathbf{N}_A\big(\max_{i\le N} W_i(\infty)\big) \ge k)$, completing the proof. □

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 *Let X be a standard normal random variable independent of everything else. For the model in Sect. 3 where the sequence of service times $(S_i(j), i \ge 1, j \ge 1)$ satisfies Assumption 1, we have that as $N \to \infty$*

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

Proof For all $j \ge 1$ take a collection of mutually independent random variables $\widehat{A}(j) \sim A$, with $\widehat{A}(j)$ independent of $\max_{i\le N} W_i(\infty)$. Then, using Theorem 1 and Lemma 2, we get that

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\le N} Q_i(\infty) \le \frac{\lambda}{\gamma}\log N + x\sqrt{\log N}\right)\\
&\quad= \mathbb{P}\left(\mathbf{N}_A\left(\max_{i\le N} W_i(\infty)\right) \le \left\lfloor \frac{\lambda}{\gamma}\log N + x\sqrt{\log N}\right\rfloor\right)\\
&\quad= \mathbb{P}\left(\max_{i\le N} W_i(\infty) \le \sum_{j=1}^{\left\lfloor \frac{\lambda}{\gamma}\log N + x\sqrt{\log N}\right\rfloor} \widehat{A}(j)\right)\\
&\quad= \mathbb{P}\left(\frac{\max_{i\le N} W_i(\infty) - \frac{1}{\gamma}\log N}{\sqrt{\log N}} \le \frac{\sum_{j=1}^{\left\lfloor \frac{\lambda}{\gamma}\log N + x\sqrt{\log N}\right\rfloor} \widehat{A}(j) - \frac{1}{\gamma}\log N}{\sqrt{\log N}}\right)\\
&\xrightarrow{N\to\infty} \mathbb{P}\left(\frac{\sigma_A}{\sqrt{\Lambda'(\gamma)\gamma}} X_1 \le \frac{\sigma_A\sqrt{\lambda}}{\sqrt{\gamma}} X_2 + \frac{x}{\lambda}\right),
\end{aligned}$$

with X_1, X_2 independent and standard normally distributed. Thus, the theorem follows by standard properties of the normal distribution. □

A natural question is whether similar results hold for the maximum sojourn time $\max_{i\le N} W_i(\infty) + S_i$, where S_i is taken to be independent of $W_i(\infty)$. We expect that the extra service time should not play a substantial role in the asymptotic behavior, but we have not found a straightforward proof. For example, the upper bound $\max_{i\le N}(W_i(\infty)+S_i) \le \max_{i\le N} W_i(\infty)+\max_{i\le N} S_i$ (together with the lower bound $\max_{i\le N}(W_i(\infty) + S_i) \ge \max_{i\le N} W_i(\infty)$) is in general not sharp enough to produce a sharp limit theorem for the sojourn time, unless $\max_{i\le N} S_i/\sqrt{\log N} \to 0$, which holds if the distribution of S_i has bounded support.

3.2 Heterogeneous servers

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

Corollary 1 *Let $K \in \mathbb{N}$, let $k \in \{1, \ldots, K\}$, and take an increasing sequence of positive integers $M_0^{(N)}$, $M_1^{(N)}$, $M_2^{(N)}$, ..., $M_K^{(N)}$ with $M_0^{(N)} = 1$, $M_K^{(N)} = N$, and $M_k^{(N)} - M_{k-1}^{(N)} \in \mathbb{N}$. Assume that*

$$\frac{M_k^{(N)} - M_{k-1}^{(N)}}{N} \overset{N\to\infty}{\longrightarrow} \alpha_k \in (0, 1]$$

with $\sum_{k=1}^{K} \alpha_k = 1$.

Further, for i such that $M_{k-1}^{(N)} < i \leq M_k^{(N)}$, let all $S_i(j)$, $j \geq 1$, be i.i.d. with $S_i(j) \sim S_k$. Additionally, $S_{i_1}(j_1)$ and $S_{i_2}(j_2)$ are mutually independent for all i_1, i_2, j_1, j_2. Moreover, for all $j \geq 1$ take $A(j)$ be i.i.d. with $A(j) \sim A$, $\mathbb{E}[A(j)] = 1/\lambda$, $Var(A(j)) = \sigma_A^2$, and $\mathbb{E}[S_i(j) - A(j)] = -\mu_k$ with $\mu_k > 0$.

For all $k \in \{1, \ldots, K\}$, take the functions $\Lambda_k(\theta) = \log(\mathbb{E}[\exp(\theta(S_k - 1/\lambda)])$ and assume that Λ_k satisfies Assumption 1 for a constant γ_k. Let K^* be the set of indices minimizing γ_k. That is, $K^* = \arg\min_k\{\gamma_k\}$. We assume that there exists a unique minimizer k^*, i.e., $|K^*| = 1$ and $k^* \in K^*$.

Then,

$$\frac{\max_{i\leq N} W_i(\infty) - \frac{1}{\gamma_{k^*}} \log N}{\sqrt{\log N}} \xrightarrow{d} \frac{\sigma_A}{\sqrt{\Lambda'_{k^*}(\gamma_{k^*})\gamma_{k^*}}} X, \tag{7}$$

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. Recall the definition of stochastic ordering, namely $X \geq_{st.} Y$ implies that $\mathbb{P}(X \geq x) \geq \mathbb{P}(Y \geq x)$ for all x. Observe that

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

The lower bound of (7) is then derived by Theorem 1.

By using the union bound, we get the following upper bound:

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

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

For $l \neq k^*$ and from Theorem 1, we obtain that

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

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

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

The corollary follows. □

Remark 1 In Corollary 1, we assume that $|K^*| = 1$. The case $|K^*| > 1$ 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 queue behaves analogously to the original fork-join queue, for which $|K^*| = 1$.

In the remainder of the paper, we focus on the proof of Theorem 1, which we first sketch in the next section.

4 Heuristic analysis

To prove Theorem 1, we analyze lower and upper bounds of the tail probability of the steady-state maximum waiting time among the N servers. We namely focus on

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

and we show that the lower and upper bounds we derive converge to the same limit as $N \to \infty$.

As seen from Lemma 1, the steady-state maximum waiting time is the all-time supremum of the maximum of N random walks. Additionally, for any process $(X(t), t \ge 0)$, we have that for all $t > 0$

$$\mathbb{P}\left(\sup_{s>0} X(s) > x \right) \ge \mathbb{P}(X(t) > x), \tag{8}$$

and that, due to the union bound, for all $0 < t_1 < t_2$,

$$\mathbb{P}\left(\sup_{s>0} X(s) > x\right) \le \mathbb{P}\left(\sup_{0<s<t_1} X(s) > x\right) + \mathbb{P}\left(\sup_{t_1\le s<t_2} X(s) > x\right) + \mathbb{P}\left(\sup_{s\ge t_2} X(s) > x\right). \tag{9}$$

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 $(A(j),\ j \ge 1)$ 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 [2, Thm. XIII.5.2] that $\mathbb{P}(\sup_{k\ge0}\sum_{j=1}^{k}(S_i(j)-1/\lambda) > x) \sim C\exp(-\gamma x)$, as $x \to \infty$, with $0 < C < 1$. Thus, $\mathbb{P}(\sup_{k\ge0}\sum_{j=1}^{k}(S_i(j) - 1/\lambda) > \frac{1}{\gamma}\log N) \sim C/N$, as $N \to \infty$. Now, we can conclude by using basic extreme-value results; see [14, Thm. 5.4.1, p. 188], that as $N \to \infty$,

$$\frac{\max_{i\le N}\sup_{k\ge0}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right)}{\log N} \xrightarrow{\mathbb{P}} \frac{1}{\gamma}.$$

We thus prove that in this case, the steady-state maximum waiting time centers around $\frac{1}{\gamma}\log N$; cf. Lemma 1.

Continuing with the case of deterministic arrivals, in order to find suitable lower and upper bounds of the form as in (8) and (9), we need to estimate the hitting time

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

As mentioned above, we have that $\mathbb{P}(\sup_{k\ge0}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) > \frac{1}{\gamma}\log N) \sim C/N$ as $N \to \infty$. Thus, a good estimate $\hat{\tau}^{(N)}$ for $\tau^{(N)}$ should also satisfy the property that

$$\liminf_{N\to\infty} N\,\mathbb{P}\left(\sum_{j=1}^{\hat{\tau}^{(N)}}\left(S_i(j) - \frac{1}{\lambda}\right) > \frac{1}{\gamma}\log N\right) > 0 \tag{10}$$

and

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

Now, by using Cramér's theorem [2, Thm. XIII.5.2] and by using the fact that Λ is at least twice differentiable at γ, we know that

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\mathbb{P}\left(\sum_{j=1}^{n}\left(S_i(j)-\frac{1}{\lambda}\right)\geq nx\right)\right)=-\Lambda^*(x), \tag{12}$$

for all $x > \mathbb{E}[S_i(j) - 1/\lambda]$ with $\Lambda^*(x) = \sup_{t\in\mathbb{R}}(tx - \Lambda(t))$; see [2, Thm. XIII.2.1 (2.3)]. We write $\hat{\tau}^{(N)} = \hat{c}\log N$. Then, we can conclude from Eq. (12) that

$$\lim_{N\to\infty}\frac{1}{\log N}\log\left(\mathbb{P}\left(\sum_{j=1}^{\lfloor\hat{c}\log N\rfloor}\left(S_i(j)-\frac{1}{\lambda}\right)\geq x\hat{c}\log N\right)\right)=-\Lambda^*(x)\hat{c}. \tag{13}$$

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 steady-state maximum waiting time under deterministic arrivals is approximately equal to $\frac{1}{\gamma}\log N$ and therefore the expression $x\hat{c}\log N$ in (13) should be the same as $\frac{1}{\gamma}\log N$. Second, $-\Lambda^*(x)\hat{c} = -1$, because we know from (10), (11), and (13) 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 [9, 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 [2, Thm. XIII.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 (8), we see that

$$\begin{aligned}&\mathbb{P}\left(\frac{\max_{i\leq N}W_i(\infty)-\frac{1}{\gamma}\log N}{\sqrt{\log N}}\geq x\right)\\&\geq\mathbb{P}\left(\frac{\max_{i\leq N}\sup_{\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\log N<k<\frac{1}{\Lambda'(\gamma)\gamma}\log N}\sum_{j=1}^{k}(S_i(j)-A(j))-\frac{1}{\gamma}\log N}{\sqrt{\log N}}\geq x\right),\end{aligned} \tag{14}$$

with $\epsilon > 0$ and small. In Lemma 3, we prove that the right-hand side in (14) 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 of 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 Eq. (8) gives us information about the convergence rate of the result in Theorem 1. From the Berry–Esséen theorem [31], 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 (8) shows that the convergence rate is of order $1/\sqrt{\log N}$.

5 Proof of Theorem 1

Following the path sketched in the previous section and for the model in Sect. 3, we now state and prove all related lemmas and close with the proof of Theorem 1, which is the main goal of this section.

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

$$\begin{aligned}\liminf_{N\to\infty} \mathbb{P}&\left(\max_{i\leq N} \sup_{t_1^{(N)}<k<t_2^{(N)}} \sum_{j=1}^{k}(S_i(j) - A(j)) > \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)\\ &\geq \mathbb{P}\left(\sigma_A\sqrt{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon}X_1 - \sigma_A\sqrt{\epsilon}\,|X_2| > x\right),\end{aligned} \tag{15}$$

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

Proof In order to prove this result, we first observe that the following bound holds trivially:

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

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

$$\frac{\inf_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(\frac{1}{\lambda}-A(j)\right)}{\sqrt{\log N}} \xrightarrow{d} \sigma_A\sqrt{\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon}X_1-\sigma_A\sqrt{\epsilon}\,|X_2|\,, \tag{16}$$

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

$$\frac{\max_{i\le N}\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)-\frac{1}{\gamma}\log N}{\sqrt{\log N}} \xrightarrow{\mathbb{P}} 0, \tag{17}$$

as $N\to\infty$. These two results together yield the lemma.

The first convergence result follows from Donsker's theorem [10]. The left-hand side of (16) is an infimum of a random walk with drift 0. Then, for $(B(t), t\ge 0)$ a Brownian motion with drift 0 and standard deviation 1, by Donsker's theorem and the fact that the infimum is a continuous functional, we obtain that

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

Furthermore, we can rewrite

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

where $\tilde{B}$ is an independent copy of B. Obviously, we have that $\sigma_A B\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\overset{d}{=}\sigma_A\sqrt{\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon}X_1$ with $X_1\sim\mathcal{N}(0,1)$. Because $\inf_{0<s<\epsilon}\sigma_A\tilde{B}(s)\overset{d}{=}\sigma_A\sqrt{\epsilon}|X_2|$, with $X_2\sim\mathcal{N}(0,1)$, we have that the limit in (16) 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

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

see [2, §XIII.3]. Denote by $\mathbb{E}_i$ the expectation associated with this probability measure. Then, we have 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 [2, Thm. XIII.5.6], we see that

$$\begin{aligned}&\mathbb{P}\left(\sup_{0\leq k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \geq \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)\\ &\quad = C\exp\left(-\gamma\left(\frac{1}{\gamma}\log N + x\sqrt{\log N}\right)\right)\Phi\left(-x\frac{\sqrt{\gamma\Lambda'(\gamma)}}{\sqrt{\Lambda''(\gamma)}}\right)(1+o(1)).\end{aligned} \tag{18}$$

With the same approach, we get from [2, Thm. XIII.5.6] that

$$\begin{aligned}&\mathbb{P}\left(\sup_{0\leq k<t_1^{(N)}}\sum_{j=1}^{k}\left(S_i(j) - \frac{1}{\lambda}\right) \geq \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)\\ &\quad = o\left(C\exp\left(-\gamma\left(\frac{1}{\gamma}\log N + x\sqrt{\log N}\right)\right)\right),\end{aligned} \tag{19}$$

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

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

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

We can conclude from these bounds, together with (18) and (19) that

$$\begin{aligned}&\mathbb{P}\left(\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\ge\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\\&\quad=C\exp\left(-\gamma\left(\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\right)\Phi\left(-x\frac{\sqrt{\gamma\Lambda'(\gamma)}}{\sqrt{\Lambda''(\gamma)}}\right)(1+o(1)).\end{aligned}$$

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

$$\begin{aligned}&\mathbb{P}\left(\max_{i\le N}\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\le\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\\&\quad=\mathbb{P}\left(\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\le\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)^N\overset{N\to\infty}{\longrightarrow}1.\end{aligned}$$

Similarly, for $x<0$,

$$\begin{aligned}&\mathbb{P}\left(\max_{i\le N}\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\le\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\\&\quad=\mathbb{P}\left(\sup_{t_1^{(N)}<k<t_2^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)\le\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)^N\overset{N\to\infty}{\longrightarrow}0.\end{aligned}$$

Combining these two results gives us the limit in (17). Thus, the convergence result in (15) follows from the two limits in (16) and (17). □

We proceed to prove that the lower bound in Lemma 3 is sharp by decomposing the supremum in three intervals, which we deal with separately in Lemmas 4, 5, and 6.

Lemma 4 *Given the model in Sect. 3 where the sequence of service times* $(S_i(j), i\ge 1, j\ge 1)$ *satisfies Assumption 1, the time* $t_1^{(N)}$ *that was defined as* $t_1^{(N)}=\left(\frac{1}{\Lambda'(\gamma)\gamma}-\epsilon\right)\log N$, *a constant* $\delta=\frac{\delta_1}{\Lambda'(\gamma)\gamma}+\delta_2$ *with* $\delta_1,\ \delta_2>0$ *and small, and* $\epsilon=\delta^{1/4}$, *then for all* $x\in\mathbb{R}$, *we have that*

$$\mathbb{P}\left(\max_{i\le N}\sup_{0\le k<t_1^{(N)}}\sum_{j=1}^{k}(S_i(j)-A(j))>\frac{1}{\gamma}\log N+x\sqrt{\log N}\right)\overset{N\to\infty}{\longrightarrow}0. \tag{20}$$

Proof We derive upper bounds for the left-hand side of (20) that converge to 0 as $N \to \infty$. We get by using the subadditivity property of the supremum operator and the union bound that

$$\mathbb{P}\left(\max_{i\le N}\sup_{0\le k<t_1^{(N)}}\sum_{j=1}^{k}(S_i(j)-A(j))>\frac{1}{\gamma}\log N\right)$$
$$\le \mathbb{P}\left(\max_{i\le N}\sup_{0\le k<t_1^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}+\delta_1\right)>\left(\frac{1}{\gamma}-\delta_2\right)\log N\right) \tag{21}$$

$$+\,\mathbb{P}\left(\sup_{k\ge 0}\sum_{j=1}^{k}\left(\frac{1}{\lambda}-\delta_1-A(j)\right)>\delta_2\log N+x\sqrt{\log N}\right). \tag{22}$$

First, because $\mathbb{E}[\frac{1}{\lambda}-\delta_1-A(j)]<0$, for the term in (22) we get that

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

Second, we can bound the term in (21) as follows:

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

Now, we can bound this further by

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

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

$$\sum_{k=0}^{\lfloor t_1^{(N)}\rfloor}N\,\mathbb{P}\left(\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}\right)>\left(\frac{1}{\gamma}-\frac{\delta_1}{\Lambda'(\gamma)\gamma}-\delta_2\right)\log N\right) \tag{23}$$

$$\leq N \sum_{k=0}^{\lfloor t_1^{(N)} \rfloor} \exp(k\Lambda(\theta)) \exp\left(-\theta\left(\frac{1}{\gamma} - \frac{\delta_1}{\Lambda'(\gamma)\gamma} - \delta_2\right)\log N\right)$$
$$= N \frac{-1 + \exp\left((\lfloor t_1^{(N)} \rfloor + 1)\Lambda(\theta)\right)}{\exp(\Lambda(\theta)) - 1} \exp\left(-\theta\left(\frac{1}{\gamma} - \frac{\delta_1}{\Lambda'(\gamma)\gamma} - \delta_2\right)\log N\right). \tag{24}$$

Now,

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

In order to make the bound in (24) as sharp as possible, we need to choose a convenient θ. The choice of θ that gives the sharpest bound maximizes the function $\theta\left(\frac{1}{\gamma} - \frac{\delta_1}{\Lambda'(\gamma)\gamma} - \delta_2\right) - \left(\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon\right)\Lambda(\theta)$. Recall that we defined $\delta = \frac{\delta_1}{\Lambda'(\gamma)\gamma} + \delta_2$ and $\epsilon = \delta^{1/4}$. Furthermore, we choose $\theta = \gamma + \sqrt{\delta}$. This gives us a sharp enough bound in (24). 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 (24) is upper bounded by the term $N^{-\delta^{3/4}\Lambda'(\gamma)-O(\delta)} \overset{N\to\infty}{\longrightarrow} 0$. □

Lemma 5 *Given the model in Sect. 3 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*

$$\limsup_{N\to\infty} \mathbb{P}\left(\max_{i\le N} \sup_{t_1^{(N)} \le k < t_3^{(N)}} \sum_{j=1}^{k} (S_i(j) - A(j)) > \frac{1}{\gamma} \log N + x\sqrt{\log N} \right)$$
$$\le \mathbb{P}\left(\sigma_A \sqrt{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon} X_1 + \sigma_A \sqrt{2\epsilon}\, |X_2| > x \right), \tag{25}$$

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

Proof In order to prove this lemma, we first rewrite

$$\frac{\max_{i\le N} \sup_{t_1^{(N)} \le k < t_3^{(N)}} \sum_{j=1}^{k} (S_i(j) - A(j)) - \frac{1}{\gamma}\log N}{\sqrt{\log N}}$$
$$\le \frac{\max_{i\le N} \sup_{t_1^{(N)} \le k < t_3^{(N)}} \sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda}\right) - \frac{1}{\gamma}\log N}{\sqrt{\log N}} + \frac{\sup_{t_1^{(N)} \le k < t_3^{(N)}} \sum_{j=1}^{k} \left(\frac{1}{\lambda} - A(j)\right)}{\sqrt{\log N}}$$
$$\le \frac{\max_{i\le N} \sup_{k\ge 0} \sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda}\right) - \frac{1}{\gamma}\log N}{\sqrt{\log N}} + \frac{\sup_{t_1^{(N)} \le k < t_3^{(N)}} \sum_{j=1}^{k} \left(\frac{1}{\lambda} - A(j)\right)}{\sqrt{\log N}}. \tag{26}$$

We start with the first term in (26). This term gives the rescaled longest steady-state waiting time of N i.i.d. $D/G/1$ queues. We know that

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

as $x \to \infty$, with $0 < C < 1$; see [2, Thm. XIII.5.2]. Thus, for $x > 0$,

$$\mathbb{P}\left(\frac{\max_{i\le N} \sup_{k\ge 0} \sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda}\right) - \frac{1}{\gamma}\log N}{\sqrt{\log N}} > x \right) \sim 1$$
$$- \left(1 - C\exp(-\gamma(1/\gamma \log N + x\sqrt{\log N}))\right)^N \overset{N\to\infty}{\longrightarrow} 0.$$

Similarly, for $x < 0$,

$$\mathbb{P}\left(\frac{\max_{i\le N} \sup_{k\ge 0} \sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda}\right) - \frac{1}{\gamma}\log N}{\sqrt{\log N}} > x \right) \sim 1$$
$$- \left(1 - C\exp(-\gamma(1/\gamma \log N + x\sqrt{\log N}))\right)^N \overset{N\to\infty}{\longrightarrow} 1.$$

Thus, the first term in (26) converges in probability to 0.

Now, we prove convergence of the tail probability of the second term in (26). 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 [10] and the fact that the supremum is a continuous functional, we obtain with a similar analysis as in Lemma 3 that

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

□

Lemma 6 *Given the model in Sect. 3 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* $\delta_1, \delta_2 > 0$ *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*

$$\mathbb{P}\left(\max_{i \leq N} \sup_{k \geq t_3^{(N)}} \sum_{j=1}^{k} (S_i(j) - A(j)) > \frac{1}{\gamma} \log N + x\sqrt{\log N}\right) \overset{N\to\infty}{\longrightarrow} 0. \tag{27}$$

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

$$\begin{aligned}
&\mathbb{P}\left(\max_{i \leq N} \sup_{k \geq t_3^{(N)}} \sum_{j=1}^{k} (S_i(j) - A(j)) > \frac{1}{\gamma} \log N + x\sqrt{\log N}\right) \\
&\leq \mathbb{P}\left(\max_{i \leq N} \sup_{k \geq t_3^{(N)}} \sum_{j=1}^{k} \left(S_i(j) - \frac{1}{\lambda} + \delta_1\right) > \left(\frac{1}{\gamma} - \delta_2\right) \log N\right) \\
&\quad + \mathbb{P}\left(\sup_{k \geq 0} \sum_{j=1}^{k} \left(\frac{1}{\lambda} - \delta_1 - A(j)\right) > \delta_2 \log N + x\sqrt{\log N}\right).
\end{aligned}$$

As in the proof of Lemma 4, we have that

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

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

$$\mathbb{P}\left(\max_{i\leq N}\sup_{k\geq t_3^{(N)}}\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}+\delta_1\right)>\left(\frac{1}{\gamma}-\delta_2\right)\log N\right) \tag{28}$$

$$\leq N\sum_{k=\lfloor t_3^{(N)}\rfloor}^{\infty}\mathbb{P}\left(\sum_{j=1}^{k}\left(S_i(j)-\frac{1}{\lambda}+\delta_1\right)>\left(\frac{1}{\gamma}-\delta_2\right)\log N\right) \tag{29}$$

$$\begin{aligned}&\leq N\sum_{k=\lfloor t_3^{(N)}\rfloor}^{\infty}\exp(k(\Lambda(\theta)+\theta\delta_1))\exp\left(-\theta\left(\frac{1}{\gamma}-\delta_2\right)\log N\right)\\&=N\frac{\exp\left(\lfloor t_3^{(N)}\rfloor(\Lambda(\theta)+\theta\delta_1)\right)}{\exp(\Lambda(\theta)+\theta\delta_1)-1}\exp\left(-\theta\left(\frac{1}{\gamma}-\delta_2\right)\log N\right),\end{aligned} \tag{30}$$

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

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

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 (30) 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

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

for δ small enough, and thus, the expression in (30) converges to 0 as $N \to \infty$. □

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 (9) and explained in Sect. 4, we get from Lemmas 4, 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

$$\begin{aligned}&\limsup_{N\to\infty} \mathbb{P}\left(\max_{i \leq N} W_i(\infty) > \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)\\ &\leq \limsup_{N\to\infty} \mathbb{P}\left(\max_{i \leq N} \sup_{t_1^{(N)} \leq k < t_3^{(N)}} \sum_{j=1}^{k} (S_i(j) - A(j)) > \frac{1}{\gamma}\log N + x\sqrt{\log N}\right)\\ &\leq \mathbb{P}\left(\sigma_A\sqrt{\frac{1}{\Lambda'(\gamma)\gamma} - \epsilon}X_1 + \sigma_A\sqrt{2\epsilon}\,|X_2| > x\right).\end{aligned}$$

Finally, we have that

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

□

References

1. ASML Holding N.V. ASML annual report 2020. https://www.asml.com/en/investors/annual-report/2020, (2021)
2. Asmussen, S.: Applied Probability and Queues, vol. 2. Springer, Cham (2003)
3. Atar, R., Mandelbaum, A., Zviran, A.: Control of fork-join networks in heavy traffic. In: 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 823–830. IEEE, (2012)
4. Baccelli, F.: Two parallel queues created by arrivals with two demands: The M/G/2 symmetrical case. Technical report RR–0426, INRIA, (1985)
5. Baccelli, F., Makowski, A.M.: Queueing models for systems with synchronization constraints. Proc. IEEE **77**(1), 138–161 (1989)
6. Baccelli, F., Makowski, A.M., Shwartz, A.: The fork-join queue and related systems with synchronization constraints: stochastic ordering and computable bounds. Adv. Appl. Probab. **21**(3), 629–660 (1989)
7. Dallery, Y., Liu, Z., Towsley, D.: Equivalence, reversibility, symmetry and concavity properties in fork-join queuing networks with blocking. J. ACM (JACM) **41**(5), 903–942 (1994)
8. Dallery, Y., Liu, Z., Towsley, D.: Properties of fork/join queueing networks with blocking under various operating mechanisms. IEEE Trans. Robot. Autom. **13**(4), 503–518 (1997)
9. Dembo, A., Zeitouni, O.: Large Deviations Techniques and Applications, vol. 38. Springer Science & Business Media, Cham (2009)
10. Donsker, M.D.: An invariance principle for certain probability limit theorems, vol. 6. Memoirs of the American Mathematical Society, (1951)
11. Duda, A., Czachórski, T.: Performance evaluation of fork and join synchronization primitives. Acta Inform. **24**(5), 525–553 (1987)
12. Flatto, L.: Two parallel queues created by arrivals with two demands II. SIAM J. Appl. Math. **45**(5), 861–878 (1985)
13. Flatto, L., Hahn, S.: Two parallel queues created by arrivals with two demands I. SIAM J. Appl. Math. **44**(5), 1041–1053 (1984)
14. de Haan, L., Ferreira, A.: Extreme Value Theory: An Introduction. Springer Science & Business Media, Cham (2006)
15. Haji, R., Newell, G.F.: A relation between stationary queue and waiting time distributions. J. Appl. Probab. **8**(3), 617–620 (1971)
16. Kim, C., Agrawala, A.K.: Analysis of the fork-join queue. IEEE Trans. Comput. **38**(2), 250–255 (1989)
17. Knessl, C.: On the diffusion approximation to a fork and join queueing model. SIAM J. Appl. Math. **51**(1), 160–171 (1991)
18. de Klein, S.J.: Fredholm integral equations in queueing analysis. PhD thesis, Rijksuniversiteit Utrecht, (1988)
19. Ko, S.-S., Serfozo, R.F.: Response times in M/M/s fork-join networks. Adv. Appl. Probab. **36**(3), 854–871 (2004)
20. Ko, S.-S., Serfozo, R.F.: Sojourn times in $G/M/1$ fork-join networks. Naval Res. Logist. (NRL) **55**(5), 432–443 (2008)
21. Konstantopoulos, P., Walrand, J.: Stationary and stability of fork-join networks. J. Appl. Probab. **26**(3), 604–614 (1989)
22. Lindley, D.V.: The theory of queues with a single server. In: Mathematical Proceedings of the Cambridge Philosophical Society, vol. 48, pp. 277–289. Cambridge University Press (1952)

23. Hongyuan, L., Pang, G.: Gaussian limits for a fork-join network with nonexchangeable synchronization in heavy traffic. Math. Oper. Res. **41**(2), 560–595 (2015)
24. Hongyuan, L., Pang, G.: Heavy-traffic limits for a fork-join network in the Halfin-Whitt regime. Stoch. Syst. **6**(2), 519–600 (2017)
25. Hongyuan, L., Pang, G.: Heavy-traffic limits for an infinite-server fork-join queueing system with dependent and disruptive services. Queue. Syst. **85**(1–2), 67–115 (2017)
26. Meijer, M., Schol, D., van Jaarsveld, W., Vlasiou, M., Zwart, B.: Optimization of inventory and capacity in large-scale assembly systems using extreme-value theory. Stoch. Syst. **14**(2), 131–166 (2024). https://doi.org/10.1287/stsy.2022.0014
27. Mena, C., Humphries, A., Choi, T.Y.: Toward a theory of multi-tier supply chain management. J. Supply Chain Manag. **49**(2), 58–77 (2013)
28. Marin, A., Rossi, S.: Dynamic control of the join-queue lengths in saturated fork-join stations. In: International Conference on Quantitative Evaluation of Systems, pp. 123–138. Springer, (2016)
29. Marin, A., Rossi, S.: Power control in saturated fork-join queueing systems. Perform. Eval. **116**, 101–118 (2017)
30. Marin, A., Rossi, S., Sottana, M.: Biased processor sharing in fork-join queues. In: International Conference on Quantitative Evaluation of Systems, pp. 273–288. Springer, (2018)
31. Michel, R.: On the constant in the nonuniform version of the Berry-Esséen theorem. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **55**(1), 109–117 (1981)
32. Nelson, R., Tantawi, A.N.: Approximating task response times in fork/join queues. IBM Thomas J, Watson Research Division (1987)
33. Nelson, R., Tantawi, A.N.: Approximate analysis of fork/join synchronization in parallel queues. IEEE Trans. Comput. **37**(6), 739–743 (1988)
34. Nguyen, V.: Processing networks with parallel and sequential tasks: heavy traffic analysis and Brownian limits. Ann. Appl. Probab. **3**(1), 28–55 (1993)
35. Nguyen, V.: The trouble with diversity: fork-join networks with heterogeneous customer population. Ann. Appl. Probab. **4**(1), 1–25 (1994)
36. Özkan, E.: Control of fork-join processing networks with multiple job types and parallel shared resources. Math. Oper. Res. **47**(2), 1310–1334 (2022)
37. Pedarsani, R., Walrand, J., Zhong, Y.: Robust scheduling for flexible processing networks. Adv. Appl. Probab. **49**(2), 603–628 (2017). https://doi.org/10.1017/apr.2017.14
38. Qiu, Z., Pérez, J.F., Harrison, P.G.: Efficient approximation for response-time tails. Perform. Eval. **91**, 99–116 (2015)
39. Raaijmakers, Y., Borst, S., Boxma, O.: Fork-join and redundancy systems with heavy-tailed job sizes. Queue. Syst. **103**, 131–159 (2023). https://doi.org/10.1007/s11134-022-09856-6
40. Schol, D.: Extreme-value theory for large fork-join queues. PhD thesis, Technische Universiteit Eindhoven, (2023). https://research.tue.nl/files/295195950/20230516_Schol_hf.pdf
41. Schol, D., Vlasiou, M., Zwart, B.: Large fork-join queues with nearly deterministic arrival and service times. Math. Oper. Res. **47**(2), 1335–1364 (2021)
42. Schol, D., Vlasiou, M., Zwart, B.: Tail asymptotics for the delay in a Brownian fork-join queue. Stoch. Process. Appl. **164**, 99–138 (2023)
43. Schol, D., Vlasiou,M., Zwart, B.: Maximum waiting time in heavy-tailed fork-join queues. arxiv:2211.02313
44. Toby, S.: Intel orders ASML system for well over $340 mln in quest for chipmaking edge. https://www.reuters.com/technology/intel-orders-ASML-machine-still-drawing-board-chipmakers-look-an-edge-2022-01-19/. Accessed: 2024-08-26
45. Shwartz, A., Weiss, A.: Induced rare events: analysis via large deviations and time reversal. Adv. Appl. Probab. **25**(3), 667–689 (1993)
46. Tan, X., Knessl, C.: A fork-join queueing model: diffusion approximation, integral representations and asymptotics. Queue. Syst. **22**(3), 287–322 (1996)
47. Thomasian, A., Tantawi, A.N.: Approximate solutions for $M/G/1$ fork/join synchronization. In: Proceedings of Winter Simulation Conference, pp. 361–368. IEEE, (1994)
48. Varki, E.: Mean value technique for closed fork-join networks. ACM SIGMETRICS Perform. Eval. Rev. **27**(1), 103–112 (1999)
49. Varki, E., Merchant, A., Chen, H.: The $M/M/1$ fork-join queue with variable sub-tasks. https://www.cs.unh.edu/~varki/publication/2002-nov-open.pdf, (2002)

50. Varma, S.: Heavy and light traffic approximations for queues with synchronization constraints. PhD thesis, University of Maryland, (1990)
51. Varma, S., Makowski, A.M.: Interpolation approximations for symmetric fork-join queues. Perform. Eval. **20**(1–3), 245–265 (1994)
52. Wang, W., Harchol-Balter, M., Jiang, H., Scheller-Wolf, A., Srikant, R.: Delay asymptotics and bounds for multitask parallel jobs. Queue. Syst. **91**(3), 207–239 (2019)
53. Wright, P.E.: Two parallel processors with coupled inputs. Adv. Appl. Probab. **24**(4), 986–1007 (1992)
54. Xia, C.H., Liu, Z., Towsley, D., Lelarge, M.: Scalability of fork/join queueing networks with blocking. ACM SIGMETRICS Perform. Eval. Rev. **35**(1), 133–144 (2007)
55. Zeng, Y., Tan, J., Xia, C.H.: Fork and join queueing networks with heavy tails: scaling dimension and throughput limit. J. ACM (JACM) **68**(3), 1–30 (2021)