Large fork-join networks with nearly deterministic service times

Large fork-join networks with nearly deterministic service times

Dennis Schol, Maria Vlasiou
Eindhoven University of Technology

Bert Zwart
Eindhoven University of Technology, CWI

In this paper, we study an N server fork-join queueing network with nearly deterministic arrivals and service times. Specifically, we aim to approximate the length of the largest of the N queues in the network. From a practical point of view, this has interesting applications, such as modeling the delays in a large supply chain. We present a fluid limit and a steady-state result for the maximum queue length, as $N \to \infty$. These results have remarkable differences. The steady-state result depends on two model parameters, while the fluid limit only depends on one model parameter. In addition, the fluid limit requires a different spatial scaling than the backlog in steady state. In order to prove these results, we use extreme value theory and diffusion approximations for the queue lengths.

1. Introduction. To manufacture high level technological products, a substantial number of different intermediate components needs to be assembled. Usually, these components are delivered by suppliers. High-tech products are mainly produced in low volume, while each separate product has many components. Due to globalization and production costs reduction, suppliers of these components can be found all over the world. As a result, manufacturers typically have very large, world-wide and complex supply chains.

The complexity of these networks can lead to severe problems [28]. As an example, local unpredictable natural or economic disasters may damage the supply chain, such as the 2010 volcano eruption in Iceland that shut down air traffic in northwest Europe for six days. This affected transportation activities within the supply chain of European manufacturers [24]. Delays on the supply side could lead to delays for the manufacturer. For example, due to a fire in a small production cell of one of Ericsson's suppliers, which was ended in 10 minutes, Ericsson had a loss of $200M [23]. In fact, 85% of the manufacturers globally experience at least one supply chain disruption each year [10]. So, in a network with many suppliers, it is likely that one of the suppliers has a production delay.

If a substantial number of suppliers produce a unique component of the product, the slowest of such suppliers determines the delay of the manufacturer. It is thus possible that all suppliers have a backlog of orders for the component they manufacture. A good measure of the delay of a certain supplier is the number of unfinished components of that supplier. We can model these unfinished components with a queue, where the number of unfinished components corresponds to the size of the queue. We wish to observe the longest queue in this paper, because the longest queue represents the supplier with the largest backlog.

As we are inspired by supply chain networks of high-tech manufacturers, we model such a supply chain network with a fork-join queueing system with the following characteristics: first of all, the system operates in discrete time. We capture with this assumption that some high-tech manufacturers produce a relatively small number of products per year [3]. Secondly, the number of servers N is very large. Thirdly, this system has one arrival stream representing the manufacturer's demands.

Each product is viewed as an arriving task in a fork-join queue, and each component is viewed as a subtask. In other words, tasks arrive in single stream and are divided in N subtasks, and each of these subtasks is allocated to one of the N servers, which all represent different suppliers. We also assume that the service time distribution is the same for each server, and that each server has independent service times. Next, we consider the situation that arrival and service processes are nearly deterministic. This means that the manufacturer's demand is almost the same in each time slot. Similarly, single suppliers are usually able to deliver their components in time. Finally, we consider a particular setting where the entire system operates closely to its full utilization, which captures a supply chain network operating under full capacity. Since each arriving task is split up in N subtasks, we can represent this process with a fork, as can be seen in Figure 1. After completion of the N subtasks, the final product is assembled, so this means that all the subtasks are joined. However, we do not consider this joining process in this paper. We give a visualization of the fork-join queue in Figure 1.

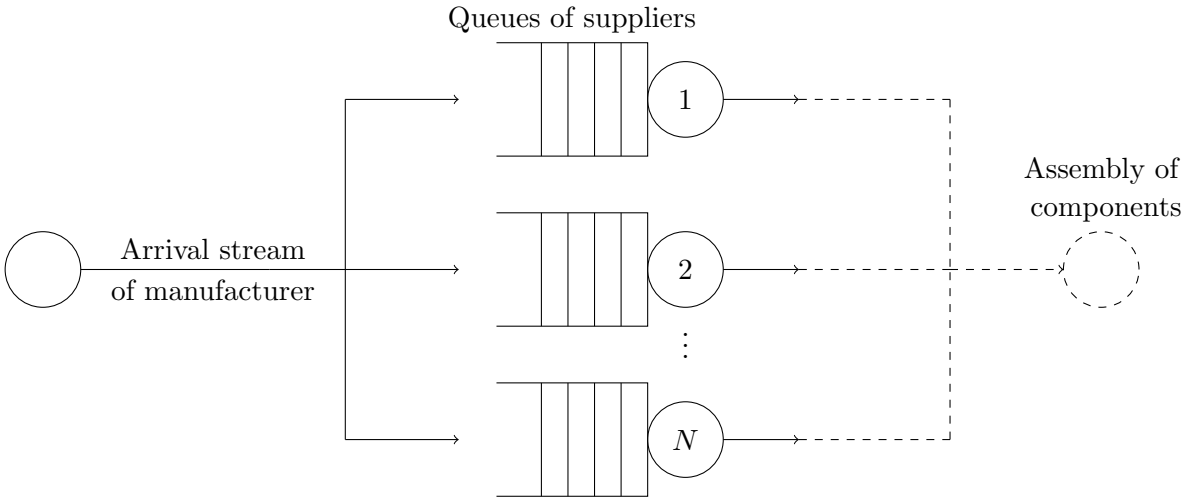


FIGURE 1. Fork-join queue with N servers

We give two results on the longest queue: first, a fluid limit is presented. Secondly, convergence of the steady state is given, as $N \to \infty$. A highly non-standard phenomenon is that the spatial scaling for the steady-state result and the fluid limit at finite times are different. In particular, the size of the longest queue scales with $N \log N$ in steady state, and scales with $N\sqrt{\log N}$ in the finite-time fluid limit. The steady-state result depends on two model parameters, whereas the fluid limit only depends on one model parameter, which is rather surprising, since normally time-dependent solutions are more sensitive to parameters than invariant solutions.

Our setting is related to work done on load balancing in large networks. In this context, the maximum queue length is also considered as a performance measure of a queueing system. Luczak and McDiarmid [18] analyze the size of the longest of N queues, with a load balancing policy where tasks are dispatched to the shortest among $d < N$ randomly chosen queues, for d fixed. They conclude that the size of the longest queue scales with $\log\log N / \log d$ for $d > 1$.

Our work contributes both to the literature on queueing systems with nearly deterministic arrivals and services and to the literature on fork-join queues. The only research line on queueing systems with nearly deterministic service times that we are aware of is Sigman and Whitt, who investigate the G/G/1 [27] and G/D/N [26] many-server queue and establish heavy-traffic results. In the latter paper, they distinguish two cases, one in which $(1-\rho_N)\sqrt{N} \to \beta$ and one in which

$(1-\rho_N)N \to \beta$ with ρ_N the traffic intensity, and β some constant. Apart from investigating a fundamentally different model, we also consider another scaling, namely $(1-\rho_N)N^2 \to \beta$, with $\beta > 0$.

Fork-join queues are extensively studied. However, limiting distributions are only given for fork-join queues with two stations, [6, 12, 13, 30]. Finding instances of tractable steady-state distributions for the maximum queue length seems to be impossible. Time-dependent results are even harder to find.

We derive analytic results on the fork-join queue with many servers in heavy traffic. Our work seems to be the first explicit time-dependent approximation of a large fork-join queue. Most of the work on heavy-traffic analysis in fork-join networks, has mainly been done on networks with a fixed number of stations. Varma [29] and Nguyen [21] investigate a heavy-traffic regime with a fixed number of servers, and derive an approximation involving reflected Brownian motions. In [22], Nguyen considers a fork-join queue with multiple job types. Atar, Mandelbaum and Zviran [5] investigate the control of a fork-join queue in heavy traffic by using feedback procedures. Lu and Pang study fork-join networks in [15, 16, 17]. In [15], they make the distinction between exchangeable synchronization and non-exchangeable synchronization, distinguishing if tasks are unique or not. In [16], the heavy-traffic regime is derived for a fixed number of servers. In [17], they investigate heavy-traffic limits for a fixed number of infinite-server stations. For larger fork-join networks upper and lower bounds for the mean response time of servers in steady state are given by Nelson, Tantawi [20], Baccelli, Makowski, Shwartz [7, 8], Downey [11]. In [11] only Poisson arrivals are considered. Furthermore, in this paper, the bounds are only tight when the expected task times become large in comparison with the expected interarrival times, which is not the case in our setting. Ko and Serfozo [14] give approximations on the distribution of the response time. In particular, in [8] lower and upper bounds for the response time in steady state are considered. We verified the tightness of these steady-state bounds in our setting, it turns out that the lower bound is tight, but the upper bound is not, as argued in Remark 3.1. The bounds in [20] are given for a fork-join queueing system with Poisson arrivals and exponential services. However, the tightness of these bounds is not investigated in that paper, and they use similar techniques as in [8], so these bounds are also not tight in our setting.

To get a heuristic idea of the steady-state result and the fluid limit, we combine ideas from the literature on diffusion approximations for queues and extreme value theory. For each separate queue length, we have a reflected Brownian motion as diffusion approximation. Then, we investigate the maximum of N independent reflected Brownian motions to get an idea of the scaling of the maximum queue length. This is treated in Section 2.1.

The fluid limit we prove holds uniformly on compact intervals. To prove this limit, we derive upper and lower bounds. We need to prove pointwise convergence of the process and tightness of the collection of processes. The tightness proof makes use of properties of reflected Brownian motions. Some non-standard results on extreme value theory are needed as well. Specifically, we define a process which is a scaled maximum. We prove that this process behaves like the scaled maximum of standard normal random variables. To this end, we use the weak convergence result of Anderson, Coles and Hüssler [2] on the maximum of triangular arrays. We use this, together with the result of Michel [19] on the convergence rate of random walks to a normally distributed random variable, to prove convergence of the moments of this process. Pickands' result [25] on convergence of moments of the maximum is not applicable here, since the process we study is a triangular array, which is not covered in Pickands' theorem. We find convergence of the first, second and fourth moment of this scaled maximum, and apply this to prove tightness of the queueing process. Due to this convergence of moments, we can use Markov's inequality to bound the probability that the process makes large jumps, and prove that this probability is small.

Since the queue lengths are dependent random variables, standard extreme value theory results are not directly applicable to prove the steady-state result. Determining an extreme value distribution by using the domain of attraction of the random variables is only possible when those random

variables are independently and identically distributed. In order to deal with this, we derive upper and lower bounds for the scaled queueing process. These bounds can be decomposed in a small dependent part and a large independent part. We show scaling results for these upper and lower bounds and conclude that these upper and lower bounds converge to the same limit as $N \to \infty$, from which the steady-state result follows. The main ingredient we use is the fact that the independent parts of the upper and lower bounds are stochastically dominated by exponentially distributed random variables, which we prove by using the Lundberg inequality [4]. For these exponentially distributed random variables, we can analyze the maximum by using extreme value theory. We also show that the steady-state result approximates the maximum queue length when time is very large but not infinity, and is in fact growing larger than $N^3 \log N$.

The rest of the paper is organized as follows. In Section 2, we describe the fork-join system in more detail. In Section 2.1, we give the arrival and service processes, and we also give a scaled version of the queueing model. In Section 2.2, we present the fluid limit and explain it heuristically. In Section 2.3, we give the steady-state results. We prove these results in Section 3. In Appendix A, we elaborate a bit more on the convergence of the upper and lower bounds that were given in Sections 3.2.1 and 3.2.2. In Appendix B, we prove the lemmas stated in Section 3.1.2.

2. Model description and main results. We consider a fork-join queue with discrete time arrivals and discrete time services. In this queueing system, there is one arrival process. The arriving tasks are divided in N subtasks which are completed by N servers. We assume that both the number of arrivals and services per time step are Bernoulli distributed. The parameters of the Bernoulli random variables depend on the number of servers. This is formalized in Definitions 2.1 and 2.2.

DEFINITION 2.1 (ARRIVAL PROCESS). The random variable $A^{(N)}(n)$ indicates the number of arrivals up to time n and equals

$$A^{(N)}(n) = \sum_{j=1}^{n} X^{(N)}(j)$$

with $X^{(N)}(j)$ indicating whether or not there is an arrival at time j. $X^{(N)}(j)$ is a Bernoulli random variable with parameter $p^{(N)}$. So,

$$X^{(N)}(j) = \begin{cases} 1 & \text{w.p.} \quad p^{(N)}, \\ 0 & \text{w.p.} \quad 1 - p^{(N)}. \end{cases}$$

DEFINITION 2.2 (SERVICE PROCESS i-TH SERVER). The random variable $S_i^{(N)}(n)$ describes the number of potentially completed tasks of the i-th server in the fork-join queue at time n with

$$S_i^{(N)}(n) = \sum_{j=1}^{n} Y_i^{(N)}(j)$$

and where $Y_i^{(N)}(j)$ is a Bernoulli random variable with parameter $q^{(N)}$ indicating whether the i-th server completed a service at time j.

$$Y_i^{(N)}(j) = \begin{cases} 1 & \text{w.p.} \quad q^{(N)}, \\ 0 & \text{w.p.} \quad 1 - q^{(N)}. \end{cases}$$

Observe that $Y_i^{(N)}(j)$ could still be 1 while there are no tasks to be served at server i at time j. Both $p^{(N)}$ and $q^{(N)}$ are taken as functions of N, which we specify in Definition 2.3 below.

We assume that the random variables $X^{(N)}(j)$'s are independent for all j and $Y_i^{(N)}(j)$'s are independent for all j and i. We also assume that an incoming task can be completed in the same time slot as in which the task arrived. Since the processes $S_i^{(N)}$ and $A^{(N)}$ are independent, it is not completely clear how we can express the queue length as a function of these processes. However, by using Lindley's recursion, we can write the queue length of the i-th server at time n as

$$\sup_{0\le k\le n}\left[\left(A^{(N)}(n)-A^{(N)}(k)\right)-\left(S_i^{(N)}(n)-S_i^{(N)}(k)\right)\right],$$

provided that the queue length is 0 at time 0. This is in distribution equal to

$$\sup_{0\le k\le n} A^{(N)}(k)-S_i^{(N)}(k),$$

provided that the queue length is 0 at time 0. As can be seen, the queue lengths of these servers are dependent on each other, since the arrival process is the same.

2.1. Scaling of process. The aim of this study is to investigate the behavior of this system when the number of servers N is very large. The main objective is deriving the distribution of the largest queue, as this represents the slowest supplier, which is the bottleneck for the manufacturer. Furthermore, we explore this model in the heavy-traffic regime. To this end, we let $p^{(N)}$ and $q^{(N)}$ go to 1 at similar rates, so that the arrivals and services are nearly deterministic processes. In this section, we investigate how to choose $p^{(N)}$ and $q^{(N)}$ to get a non-trivial limit for the maximum queue length in the fork-join queue.

Definition 2.3 (Maximum queue length at time n). Let $p^{(N)}=1-\alpha/N-\beta/N^2$ and $q^{(N)}=1-\alpha/N$, with $\alpha,\beta>0$. Let $Q_{(\alpha,\beta)}^{(N)}(n)$ be the maximum queue length of N parallel servers at time n. Then

$$Q_{(\alpha,\beta)}^{(N)}(n)=\max_{i\le N}\sup_{0\le k\le n}\left[\left(A^{(N)}(n)-A^{(N)}(k)\right)-\left(S_i^{(N)}(n)-S_i^{(N)}(k)\right)\right]. \tag{2.1}$$

So,

$$Q_{(\alpha,\beta)}^{(N)}(n)\overset{d}{=}\max_{i\le N}\sup_{0\le k\le n}\left(A^{(N)}(k)-S_i^{(N)}(k)\right) \tag{2.2}$$

under the assumption that $Q_{(\alpha,\beta)}^{(N)}(0)=0$. From these choices of $p^{(N)}$ and $q^{(N)}$, it follows that the throughput ρ_N of a single queue satisfies $(1-\rho_N)N^2\to\beta$, as $N\to\infty$. Thus we derive a heavy-traffic regime combined with time-dependent behavior.

2.2. Fluid limit. Our main result is a fluid approximation for the rescaled queue length process, which is given in Theorem 2.1. We prove that under a certain spatial and temporal scaling the maximum queue length converges to a continuous function, which depends on time t.

Theorem 2.1 **(Process convergence).** *Assume $Q_{(\alpha,\beta)}^{(N)}(0)=0$, then $\forall\, T>0$, we have*

$$\mathbb{P}\left(\sup_{0\le t\le T}\left|\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3)}{N\sqrt{\log N}}-\sqrt{2\alpha t}\right|>\epsilon\right)\overset{N\to\infty}{\longrightarrow}0\ \forall\,\epsilon. \tag{2.3}$$

The fluid limit does not depend on β, which is remarkable. A heuristic justification for this fluid limit, in particular the spatial scaling of $1/(N\sqrt{\log N})$ and the temporal scaling of N^3, can be given by using extreme value theory. In particular, for the spatial scaling we argue as follows: as

we are interested in the convergence of the maximum queue length, we aim to derive a central limit result for each separate queue length, and use the classical result that the scaled maximum of N normal random variables converges to a Gumbel distributed random variable. To argue this, since the arrival and service processes are binomially distributed random variables, we compute the expectation and variance of $\left(A^{(N)}\left(tN^3\right) - S_i^{(N)}\left(tN^3\right)\right)/N$ as

$$\mathbb{E}\left[\frac{1}{N}\left(A^{(N)}\left(tN^3\right) - S_i^{(N)}\left(tN^3\right)\right)\right] = -\beta t, \tag{2.4}$$

and

$$\begin{aligned}
&\operatorname{Var}\left(\frac{1}{N}\left(A^{(N)}\left(tN^3\right) - S_i^{(N)}\left(tN^3\right)\right)\right)\\
=&\frac{1}{N^2}tN^3\left(\left(\frac{\alpha}{N}+\frac{\beta}{N^2}\right)\left(1-\frac{\alpha}{N}-\frac{\beta}{N^2}\right)+\frac{\alpha}{N}\left(1-\frac{\alpha}{N}\right)\right)\\
=&2\alpha t + o(1).
\end{aligned} \tag{2.5}$$

It is easy to see that this leads to a nontrivial scaling limit: observe that $A^{(N)}\left(tN^3\right) - S_i^{(N)}\left(tN^3\right)$ is a sum of independent and identically distributed random variables, so this implies that

$$\frac{1}{N}\left(A^{(N)}\left(tN^3\right) - S_i^{(N)}\left(tN^3\right)\right) \xrightarrow{d} Z \text{ as } N \to \infty,$$

with $Z \sim \mathcal{N}(-\beta t, 2\alpha t)$, which retrieves the result. On the other hand, because $A^{(N)}\left(tN^3\right) - S_i^{(N)}\left(tN^3\right)$ is in fact the difference of two random walks, we also have

$$\sup_{0\le n\le tN^3}\frac{1}{N}\left(A^{(N)}(n) - S_i^{(N)}(n)\right) \xrightarrow{d} R(t) \text{ as } N \to \infty,$$

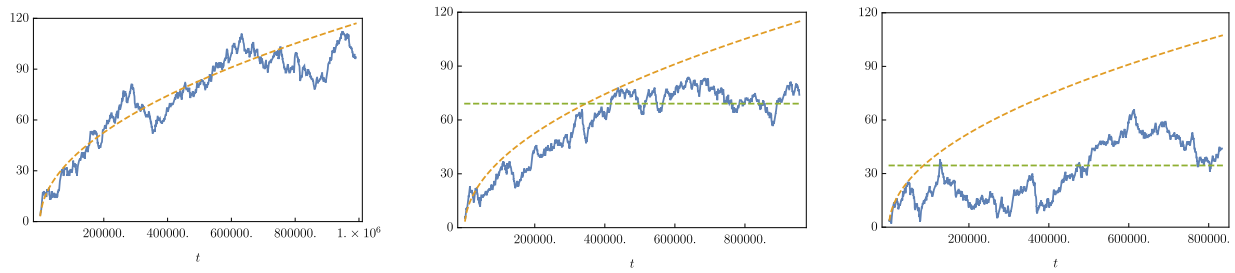
with $R(t)$ a reflected Brownian motion for t fixed. Extreme value results hold for the maximum of N independent reflected Brownian motions, which scales with $\sqrt{\log N}$. This can be deduced from the cumulative distribution function of the reflected Brownian motion which is given in [1]. Concluding, the proper spatial scaling of the fluid limit is $1/(N\sqrt{\log N})$. We see the same scaling in Theorem 2.1. In order to prove Theorem 2.1, we prove pointwise convergence of the process at a fixed time, convergence of the finite-dimensional distributions, and the tightness of $\left(Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)/\left(N\sqrt{\log N}\right), t\in[0,T]\right)$, which is stated in Lemma 3.9.

2.3. Steady state. The fluid limit is increasing in t, and goes to ∞ as t goes to ∞. Thus, this fluid limit cannot be used to give a steady-state estimation. Therefore, a central question is whether we can find a more suitable scaling for $t=\infty$. In fact, when we scale the maximum queue length with $N\log N$ instead of $N\sqrt{\log N}$, we find a steady-state result. This is given in Theorem 2.2. In addition, a slight adaption in the proof of Theorem 2.2 gives a time-dependent result, which is given in Corollary 2.1.

THEOREM 2.2 **(Convergence of the maximum queue length in steady state).** *For α, $\beta>0$ and N the number of queues*

$$\frac{Q_{(\alpha,\beta)}^{(N)}(\infty)}{N\log N} \xrightarrow{\mathbb{P}} \frac{\alpha}{2\beta} \text{ as } N\to\infty. \tag{2.6}$$

In Figure 2, the simulated maximum queue length is plotted together with the fluid approximation and the steady-state approximation, for several choices of α and β. We have $N = 1000$, $p^{(N)} = 1 - \alpha/N - \beta/N^2$, and $q^{(N)} = 1 - \alpha/N$. We see that the fluid limit and the steady-state result approximates the maximum queue length quite well for $N = 1000$. Depending on the choice of β and t, the simulated maximum queue length follows the fluid limit or the steady-state result. The simulated maximum queue length is drawn as a line, the fluid approximation is given as a dashed curve, and the steady-state approximation is given as a dashed straight line.



(a) $\alpha = 1, \beta = 10$ (b) $\alpha = 1, \beta = 50$ (c) $\alpha = 1, \beta = 100$

FIGURE 2. Maximum queue length, fluid limit approximation and steady-state approximation

As stated in [1], the invariant distribution of a reflected Brownian motion with negative drift is exponential. The maximum of N exponentially distributed random variables scales with $\log N$. We see this also appearing in Theorem 2.2. The only difference is that in this result, we have an extra term N, which is due to the spatial scaling of the process.

2.4. Large finite-time intervals. Corollary 2.1 gives an indication when time is large enough such that the maximum queue length can already be approximated by the steady-state result. We scale time with N^3 in the fluid limit, it turns out that when time grows faster than $N^3 \log N$ the maximum queue length can be estimated by the steady-state result.

COROLLARY 2.1 **(Convergence of the maximum queue length over large intervals).** *Let α, β be positive, N be the number of queues, and $t(N)$ the time. If $t(N)/(N^3 \log N) \overset{N\to\infty}{\longrightarrow} \infty$, we get*

$$\frac{Q^{(N)}_{(\alpha,\beta)}(t(N))}{N \log N} \xrightarrow{\mathbb{P}} \frac{\alpha}{2\beta} \text{ as } N \to \infty. \tag{2.7}$$

3. Proofs. Theorem 2.1, Theorem 2.2 and Corollary 2.1 are proven in Sections 3.1, 3.2 and 3.3 respectively. Since each server has the same arrival process, the queue lengths are dependent. The general idea of proving Theorems 2.1 and 2.2 is to find upper and lower bounds for the queue lengths. These bounds are sums of two random variables, one random variable is the same for each server, and the other is an independent and identically distributed random variable. We can use extreme value theory to prove the convergence of the maxima of these bounds, because the dependent parts have a small contribution to the maxima.

3.1. Fluid limit. We gave a heuristic argument in Section 2.2 that the maximum queue length scales with $\sqrt{\log N}$ under a temporal scaling of N^3 and a spatial scaling of $1/N$. In this section, we give a rigorous proof of this. In order to prove the convergence of $Q^{(N)}_{(\alpha,\beta)}(tN^3)/(N\sqrt{\log N})$ uniformly on bounded time intervals, we first prove the pointwise convergence of the process and the weak convergence of its finite-dimensional distributions, which is shown in Section 3.1.3. Afterwards, we invoke a criterion from Billingsley [9] in Section 3.1.4 to prove tightness of the collection of processes. We first give some definitions and preliminary results.

3.1.1. Definitions. For the sake of notation, we use the expression given in Definition 3.1 to prove the tightness.

DEFINITION 3.1. We define the random walk $\tilde{R}_i^{(N)}(n)$ as

$$\tilde{R}_i^{(N)}(n) = \frac{\tilde{A}^{(N)}(n) + \tilde{S}_i^{(N)}(n)}{\sqrt{\log N}},$$

where

$$\tilde{A}^{(N)}(n) = \frac{A^{(N)}(n)}{N} - \left(1 - \frac{\alpha}{N}\right)\frac{n}{N},$$

and

$$\tilde{S}_i^{(N)}(n) = -\frac{S_i^{(N)}(n)}{N} + \left(1 - \frac{\alpha}{N}\right)\frac{n}{N},$$

with $A^{(N)}(n)$ and $S_i^{(N)}(n)$ given in Definitions 2.1 and Definition 2.2 respectively.
Now, by the central limit theorem, $\tilde{S}_i^{(N)}(tN^3)$ and $\tilde{A}^{(N)}(tN^3)$ converge in distribution to normally distributed random variables with variance αt and means 0 and $-\beta t$ respectively. We can deduce from Equation (2.1) that

$$\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3)}{N\sqrt{\log N}} = \max_{i\le N} \sup_{0\le n\le tN^3} \frac{\left(A^{(N)}(tN^3) - A^{(N)}(n)\right) - \left(S_i^{(N)}(tN^3) - S_i^{(N)}(n)\right)}{N\sqrt{\log N}}.$$

Consequently, we can rewrite

$$\begin{aligned}\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3)}{N\sqrt{\log N}} &= \max_{i\le N} \sup_{0\le r\le t} \frac{\tilde{A}^{(N)}(tN^3) - \tilde{A}^{(N)}(rN^3) + \tilde{S}_i^{(N)}(tN^3) - \tilde{S}_i^{(N)}(rN^3)}{\sqrt{\log N}} \\ &= \max_{i\le N} \sup_{0\le r\le t} \left(\tilde{R}_i^{(N)}\left(tN^3\right) - \tilde{R}_i^{(N)}\left(rN^3\right)\right). \end{aligned} \tag{3.1}$$

Furthermore, to prove Theorem 2.1 we use the properties of a scaled maximum of $\tilde{S}_i^{(N)}(tN^3)$. This maximum is given in the following definition.

DEFINITION 3.2. Let $t > 0$ be given. Then the scaled maximal service process $M^{(N)}(t)$ is defined as

$$M^{(N)}(t) = b_N\left(\max_{i\le N} \frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}} - b_N\right),$$

with

$$b_N = \sqrt{2\log N} - \frac{\log(4\pi\log N)}{2\sqrt{2\log N}}.$$

3.1.2. Useful lemmas. In order to prove Theorem 2.1, we need to prove three things. We have to prove the pointwise convergence of the process, the convergence of its finite-dimensional distributions, and the tightness of the process. In order to do this, a few preliminary results are needed. As stated in Definition 3.1, we can write $\tilde{R}_i^{(N)}(n)$ as

$$\frac{\tilde{A}^{(N)}(n) + \tilde{S}_i^{(N)}(n)}{\sqrt{\log N}}.$$

Observe that $\tilde{A}^{(N)}(n)$ does not depend on i, while $\tilde{S}_i^{(N)}(n)$ does. Hence, it is intuitively clear that $\tilde{A}^{(N)}(n)$ pays no contribution to the maximum queue length. Therefore, in order to prove the pointwise convergence of the maximum queue length, we need to analyze $\tilde{S}_i^{(N)}(n)/\sqrt{\log N}$. Specifically, we use the fact that

$$\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t(1-\alpha/N)}} \xrightarrow{d} Z \text{ as } N \to \infty,$$

with Z a standard normal random variable, which can be shown by the central limit theorem. We can use this result to approximate the maximum queue length, because we know that the scaled maximum of N independent and normally distributed random variables converges to a Gumbel distributed random variable. We defined such a scaled maximum $M^{(N)}(t)$ in Definition 3.2. We prove that also $M^{(N)}(t)$ converges in distribution to a Gumbel distributed random variable in Lemma 3.3. In order to prove this, we need to know the rate of convergence of $\tilde{S}_i^{(N)}(tN^3)/\sqrt{\alpha t(1-\alpha/N)}$ to a normally distributed random variable. This rate of convergence is given in Lemma 3.2. From the convergence of $M^{(N)}(t)$ and the fact that $b_N \approx \sqrt{2\log N}$ follows that

$$\max_{i\le N} \frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\log N}} = \frac{\sqrt{\alpha t\left(1-\alpha/N\right)}\left(\frac{M^{(N)}(t)}{b_N}+b_N\right)}{\sqrt{\log N}} \xrightarrow{\mathbb{P}} \sqrt{2\alpha t} \text{ as } N\to\infty.$$

To prove the tightness of the maximum queue length, we have to prove that

$$\frac{1}{\delta}\mathbb{P}\left(\sup_{t\le s\le t+\delta}\left|\frac{Q_{(\alpha,\beta)}^{(N)}(sN^3)}{N\sqrt{\log N}}-\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3)}{N\sqrt{\log N}}\right|>\epsilon\right)$$

is small enough for large N. In Lemma 3.1 a useful upper bound for

$$\left|\frac{Q_{(\alpha,\beta)}^{(N)}(sN^3)}{N\sqrt{\log N}}-\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3)}{N\sqrt{\log N}}\right|$$

is obtained, which we use to prove the tightness of the process.

LEMMA 3.1. *For $t>0$ and $\delta>0$, we have for the queue length that*

$$\begin{aligned}&\sup_{t\le s\le t+\delta}\left|\frac{Q_{(\alpha,\beta)}^{(N)}(sN^3)}{N\sqrt{\log N}}-\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3)}{N\sqrt{\log N}}\right|\\ &\le \sup_{t\le s\le t+\delta}\max_{i\le N}\left(\tilde{R}_i^{(N)}(sN^3)-\tilde{R}_i^{(N)}(tN^3)\right)+2\sup_{t\le s\le t+\delta}\max_{i\le N}\left(\tilde{R}_i^{(N)}(tN^3)-\tilde{R}_i^{(N)}(sN^3)\right).\end{aligned} \tag{3.2}$$

To prove the tightness of the process, we also use Doob's maximal submartingale inequality. From Lemmas 3.2 and 3.3, we know that

$$\max_{i\le N}\frac{\tilde{S}_i^{(N)}(\delta N^3)}{\sqrt{\log N}} \xrightarrow{\mathbb{P}} \sqrt{2\alpha\delta} \text{ as } N\to\infty.$$

LEMMA 3.2. *For $t>0$, we have an upper bound of the rate of convergence of $\tilde{S}_i^{(N)}(tN^3)/\sqrt{\alpha t(1-\alpha/N)}$ to a standard normal random variable given by*

$$\left|\mathbb{P}\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t(1-\alpha/N)}}<y\right)-\Phi(y)\right| \le \frac{c_t}{N\sqrt{N}}y^{-6}, \tag{3.3}$$

with $c_t>0$.

Lemma 3.2 follows from Michel [19, Thm. 1& 2, p. 102 & 103].

LEMMA 3.3. *For $t>0$ and the scaled maximal service process $M^{(N)}(t)$ given by Definition 3.2, we have that $M^{(N)}(t) \xrightarrow{d} G$ as $N\to\infty$ with $G\sim$ Gumbel.*

To be able to use Doob's maximal submartingale inequality, we need that $\mathbb{E}[(\max_{i\leq N}\tilde{S}_i^{(N)}(\delta N^3)/\sqrt{\log N})^4]$ converges to $4\alpha^2\delta^2$. With this result, we can conclude that

$$\frac{1}{\delta}\mathbb{P}\left(\sup_{t\leq s\leq t+\delta}\left|\frac{Q_{(\alpha,\beta)}^{(N)}(sN^3)}{N\sqrt{\log N}}-\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3)}{N\sqrt{\log N}}\right|>\epsilon\right)=O\left(\frac{\mathbb{E}[(\max_{i\leq N}\tilde{S}_i^{(N)}(\delta N^3)/\sqrt{\log N})^4]}{\delta}\right)=O(\delta).$$

To obtain the convergence of the fourth moment of $\max_{i\leq N}\tilde{S}_i^{(N)}(\delta N^3)/\sqrt{\log N}$, we need to prove the convergence of some moments of $M^{(N)}(t)$, which is done in Lemmas 3.4 and 3.5.

LEMMA 3.4. *For $t>0$ and the scaled maximal service process $M^{(N)}(t)$ given by Definition 3.2, we have that $\mathbb{E}\left[M^{(N)}(t)^4\right]\overset{N\to\infty}{\longrightarrow}\mathbb{E}[G^4]$ with $G\sim$ Gumbel.*

LEMMA 3.5. *For $t>0$, and $k=1,2$, or 4,*

$$\mathbb{E}\left[\left(\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\log N}}\right)^k\right]\overset{N\to\infty}{\longrightarrow}\left(\sqrt{2\alpha t}\right)^k. \tag{3.4}$$

The proofs of Lemmas 3.1, 3.2, 3.3, 3.4 and 3.5 can be found in Appendix B.

3.1.3. Pointwise convergence. To prove pointwise convergence of a process to a constant it suffices to show pointwise convergence in distribution. Therefore, we use Lemmas 3.6 and 3.7 below to prove that the upper and lower bound of the cumulative distribution function converge to the same function, which is the cumulative distribution function of the constant $\sqrt{2\alpha t}$.

LEMMA 3.6. *For $\delta>0$*

$$\liminf_{N\to\infty}\mathbb{P}\left(\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3)}{N\sqrt{\log N}}\geq\sqrt{2\alpha t}-\delta\right)=1. \tag{3.5}$$

Proof Let $\delta>0$ be given. We have

$$\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3)}{N\sqrt{\log N}}\overset{d}{=}\sup_{0\leq s\leq t}\max_{i\leq N}\frac{\tilde{A}^{(N)}(sN^3)+\tilde{S}_i^{(N)}(sN^3)}{\sqrt{\log N}}\geq\inf_{0\leq s\leq t}\frac{\tilde{A}^{(N)}(sN^3)}{\sqrt{\log N}}+\sup_{0\leq s\leq t}\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(sN^3)}{\sqrt{\log N}}.$$

Because of the fact that $M^{(N)}(t)\xrightarrow{d}G$ as $N\to\infty$, we know that

$$\mathbb{P}\left(\max_{i\leq N}\sup_{0\leq s\leq t}\frac{\tilde{S}_i^{(N)}(sN^3)}{\sqrt{\log N}}\geq\sqrt{2\alpha t}-\delta\right)\geq\mathbb{P}\left(\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\log N}}\geq\sqrt{2\alpha t}-\delta\right)\overset{N\to\infty}{\longrightarrow}1. \tag{3.6}$$

We also know that $\left|\inf_{0\leq s\leq t}\tilde{A}^{(N)}(sN^3)/\sqrt{\log N}\right|\leq\sup_{0\leq s\leq t}\left|\tilde{A}^{(N)}(sN^3)/\sqrt{\log N}\right|$. Therefore, for all $\epsilon>0$

$$\mathbb{P}\left(\left|\inf_{0\leq s\leq t}\frac{\tilde{A}^{(N)}(sN^3)}{\sqrt{\log N}}\right|>\epsilon\right)\leq\mathbb{P}\left(\sup_{0\leq s\leq t}\left|\frac{\tilde{A}^{(N)}(sN^3)}{\sqrt{\log N}}\right|>\epsilon\right).$$

Furthermore, we know that $\mathbb{E}\left[\tilde{A}^{(N)}(sN^3)\right] = -\beta s$. Therefore, $\tilde{A}^{(N)}(sN^3)/\sqrt{\log N} + \beta s/\sqrt{\log N}$ is a martingale, and $\left|\tilde{A}^{(N)}(sN^3)/\sqrt{\log N} + \beta s/\sqrt{\log N}\right|^2$ is a submartingale. Therefore, by Doob's maximal submartingale inequality, we have

$$\begin{aligned}\mathbb{P}\left(\sup_{0\le s\le t}\left|\frac{\tilde{A}^{(N)}(sN^3)}{\sqrt{\log N}}\right| > \epsilon\right) &\le \mathbb{P}\left(\sup_{0\le s\le t}\left|\frac{\tilde{A}^{(N)}(sN^3)}{\sqrt{\log N}} + \frac{\beta s}{\sqrt{\log N}}\right|^2 > \frac{\epsilon^2}{4}\right) + \mathbb{P}\left(\frac{\beta t}{\sqrt{\log N}} > \frac{\epsilon}{2}\right)\\ &\le \frac{4\mathrm{Var}\left(\tilde{A}^{(N)}(tN^3)/\sqrt{\log N}\right)}{\epsilon^2} + \mathbb{P}\left(\frac{\beta t}{\sqrt{\log N}} > \frac{\epsilon}{2}\right).\end{aligned}$$

Recall that $\mathrm{Var}\left(\tilde{A}^{(N)}(tN^3)\right) = at$. Hence,

$$\frac{4\mathrm{Var}\left(\tilde{A}^{(N)}(tN^3)/\sqrt{\log N}\right)}{\epsilon^2} + \mathbb{P}\left(\frac{\beta t}{\sqrt{\log N}} > \frac{\epsilon}{2}\right) = \frac{4at}{\epsilon^2 \log N} + \mathbb{P}\left(\frac{\beta t}{\sqrt{\log N}} > \frac{\epsilon}{2}\right) \overset{N\to\infty}{\longrightarrow} 0. \tag{3.7}$$

So, from (3.6) and (3.7), (3.5) follows. □

LEMMA 3.7. *For $\delta > 0$*

$$\limsup_{N\to\infty} \mathbb{P}\left(\frac{Q^{(N)}_{(\alpha,\beta)}(tN^3)}{N\sqrt{\log N}} > \sqrt{2\alpha t} + \delta\right) = 0. \tag{3.8}$$

Proof Let $\delta > 0$ be given. We now use the upper bound

$$\sup_{0\le s\le t}\max_{i\le N}\frac{\tilde{A}^{(N)}(sN^3) + \tilde{S}^{(N)}_i(sN^3)}{\sqrt{\log N}} \le \sup_{0\le s\le t}\frac{\tilde{A}^{(N)}(sN^3)}{\sqrt{\log N}} + \sup_{0\le s\le t}\max_{i\le N}\frac{\tilde{S}^{(N)}_i(sN^3)}{\sqrt{\log N}}.$$

From Equation (3.7), we know that $\sup_{0\le s\le t}\left|\tilde{A}^{(N)}(sN^3)/\sqrt{\log N}\right| \xrightarrow{d} 0$ as $N\to\infty$. By Doob's maximal submartingale inequality we find

$$\mathbb{P}\left(\max_{i\le N}\sup_{0\le s\le t}\frac{\tilde{S}^{(N)}_i(sN^3)}{\sqrt{\log N}} > \sqrt{2\alpha t} + \delta\right) \le \mathbb{P}\left(\sup_{0\le s\le t}\max\left(\max_{i\le N}\frac{\tilde{S}^{(N)}_i(sN^3)}{\sqrt{\log N}} - \sqrt{2\alpha t}, 0\right) > \delta\right).$$

We know that $\tilde{S}^{(N)}_i(n)$ is a martingale. Thus,

$$\max\left(\max_{i\le N}\frac{\tilde{S}^{(N)}_i(sN^3)}{\sqrt{\log N}} - \sqrt{2\alpha t}, 0\right)$$

is a non-negative submartingale. By Doob's maximal submartingale inequality, we find that

$$\begin{aligned}\mathbb{P}\left(\sup_{0\le s\le t}\max\left(\max_{i\le N}\frac{\tilde{S}^{(N)}_i(sN^3)}{\sqrt{\log N}} - \sqrt{2\alpha t}, 0\right) > \delta\right) &\le \frac{1}{\delta}\mathbb{E}\left[\max\left(\max_{i\le N}\frac{\tilde{S}^{(N)}_i(tN^3)}{\sqrt{\log N}} - \sqrt{2\alpha t}, 0\right)\right]\\ &\le \frac{1}{\delta}\mathbb{E}\left[\left|\max_{i\le N}\frac{\tilde{S}^{(N)}_i(tN^3)}{\sqrt{\log N}} - \sqrt{2\alpha t}\right|\right].\end{aligned}$$

Now, observe that

$$\begin{aligned}\frac{1}{\delta}\mathbb{E}\left[\left|\max_{i\le N}\frac{\tilde{S}^{(N)}_i(tN^3)}{\sqrt{\log N}} - \sqrt{2\alpha t}\right|\right] &\le \frac{1}{\delta}\mathbb{E}\left[\left|\max_{i\le N}\frac{\tilde{S}^{(N)}_i(tN^3)}{\sqrt{\log N}} - \mathbb{E}\left[\max_{i\le N}\frac{\tilde{S}^{(N)}_i(tN^3)}{\sqrt{\log N}}\right]\right|\right]\\ &\quad + \frac{1}{\delta}\left|\mathbb{E}\left[\max_{i\le N}\frac{\tilde{S}^{(N)}_i(tN^3)}{\sqrt{\log N}}\right] - \sqrt{2\alpha t}\right|.\end{aligned}$$

Because we know from Lemma 3.5 that $\mathbb{E}\left[\max_{i\leq N}\tilde{S}_i^{(N)}(tN^3)/\sqrt{\log N}\right]\overset{N\to\infty}{\longrightarrow}\sqrt{2\alpha t}$ and $\mathbb{E}\left[\left(\max_{i\leq N}\tilde{S}_i^{(N)}(tN^3)/\sqrt{\log N}\right)^2\right]\overset{N\to\infty}{\longrightarrow}2\alpha t$, we can conclude that

$$\begin{aligned}&\frac{1}{\delta}\mathbb{E}\left[\left|\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\log N}}-\mathbb{E}\left[\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\log N}}\right]\right|\right]+\frac{1}{\delta}\left|\mathbb{E}\left[\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\log N}}\right]-\sqrt{2\alpha t}\right|\\ \leq&\frac{1}{\delta}\sqrt{\mathrm{Var}\left(\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\log N}}\right)}+\frac{1}{\delta}\left|\mathbb{E}\left[\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\log N}}\right]-\sqrt{2\alpha t}\right|\\ \overset{N\to\infty}{\longrightarrow}&0.\end{aligned}\tag{3.9}$$

From (3.7) and (3.9), we conclude (3.8). □

Pointwise convergence of $Q_{(\alpha,\beta)}^{(N)}(tN^3)/(N\sqrt{\log N})$ to $\sqrt{2\alpha t}$ for fixed t, as $N\to\infty$, follows from Lemmas 3.6 and 3.7. We can easily extend this result to finite-dimensional distributions.

LEMMA 3.8 **(The finite-dimensional distributions converge).** *For* $(t_1,t_2,\ldots,t_k)$

$$\left(\frac{Q_{(\alpha,\beta)}^{(N)}(t_1N^3)}{N\sqrt{\log N}},\frac{Q_{(\alpha,\beta)}^{(N)}(t_2N^3)}{N\sqrt{\log N}},\ldots,\frac{Q_{(\alpha,\beta)}^{(N)}(t_kN^3)}{N\sqrt{\log N}}\right)\overset{\mathbb{P}}{\longrightarrow}\left(\sqrt{2\alpha t_1},\sqrt{2\alpha t_2},\ldots,\sqrt{2\alpha t_k}\right)\ \textit{as}\ N\to\infty.\tag{3.10}$$

Proof

$$\begin{aligned}&\mathbb{P}\left(\left\|\left(\frac{Q_{(\alpha,\beta)}^{(N)}(t_1N^3)}{N\sqrt{\log N}},\frac{Q_{(\alpha,\beta)}^{(N)}(t_2N^3)}{N\sqrt{\log N}},\ldots,\frac{Q_{(\alpha,\beta)}^{(N)}(t_kN^3)}{N\sqrt{\log N}}\right)-\left(\sqrt{2\alpha t_1},\sqrt{2\alpha t_2},\ldots,\sqrt{2\alpha t_k}\right)\right\|>\epsilon\right)\\ \leq&\mathbb{P}\left(\left|\frac{Q_{(\alpha,\beta)}^{(N)}(t_1N^3)}{N\sqrt{\log N}}-\sqrt{2\alpha t_1}\right|+\cdots+\left|\frac{Q_{(\alpha,\beta)}^{(N)}(t_kN^3)}{N\sqrt{\log N}}-\sqrt{2\alpha t_k}\right|>\epsilon\right)\\ \leq&\mathbb{P}\left(\left|\frac{Q_{(\alpha,\beta)}^{(N)}(t_1N^3)}{N\sqrt{\log N}}-\sqrt{2\alpha t_1}\right|>\frac{\epsilon}{k}\right)+\cdots+\mathbb{P}\left(\left|\frac{Q_{(\alpha,\beta)}^{(N)}(t_kN^3)}{N\sqrt{\log N}}-\sqrt{2\alpha t_k}\right|>\frac{\epsilon}{k}\right),\end{aligned}$$

with $\|\cdot\|$ the Euclidean distance in $\mathbb{R}^k$. Because $Q_{(\alpha,\beta)}^{(N)}(t_iN^3)/(N\sqrt{\log N})\overset{\mathbb{P}}{\longrightarrow}\sqrt{2\alpha t_i}$ as $N\to\infty$, we know that

$$\mathbb{P}\left(\left|\frac{Q_{(\alpha,\beta)}^{(N)}(t_iN^3)}{N\sqrt{\log N}}-\sqrt{2\alpha t_i}\right|>\frac{\epsilon}{k}\right)$$

is small and therefore the finite-dimensional distributions converge in probability. □

3.1.4. Tightness. It is known that when a sequence $\{\mathbb{P}_n\}$ is tight and its finite-dimensional distributions converge, then $\{\mathbb{P}_n\}$ converges to $\mathbb{P}$ on bounded time intervals, cf. [9, Thm. 8.1, p. 54]. From [9, Thm. 8.3, p. 56], we know that $Q_{(\alpha,\beta)}^{(N)}(tN^3)/(N\sqrt{\log N})$ is tight when for all positive η there exists an a such that

$$\mathbb{P}\left(\left|\frac{Q_{(\alpha,\beta)}^{(N)}(0)}{N\sqrt{\log N}}\right|>a\right)\leq\eta,\tag{3.11}$$

and for all $\epsilon>0$ and $\eta>0$, there exists a $0<\delta<1$ and an integer N_0 such that for all $N\geq N_0$

$$\frac{1}{\delta}\mathbb{P}\left(\sup_{t\leq s\leq t+\delta}\left|\frac{Q_{(\alpha,\beta)}^{(N)}(sN^3)}{N\sqrt{\log N}}-\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3)}{N\sqrt{\log N}}\right|>\epsilon\right)\leq\eta.\tag{3.12}$$

We have $Q^{(N)}_{(\alpha,\beta)}(0)/\left(N\sqrt{\log N}\right)=0$, therefore condition (3.11) holds. In Lemma 3.9, condition (3.12) is checked.

LEMMA 3.9 **(The process is tight).** *For $\epsilon>0$, $\eta>0$ and $T>0$, $\exists\, 0<\delta<1$ and an integer N_0 such that $\forall\, N\geq N_0$ and $t\in[0,T]$*

$$\frac{1}{\delta}\mathbb{P}\left(\sup_{t\leq s\leq t+\delta}\left|\frac{Q^{(N)}_{(\alpha,\beta)}\left(sN^3\right)}{N\sqrt{\log N}}-\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\right)}{N\sqrt{\log N}}\right|\geq\epsilon\right)\leq\eta. \tag{3.13}$$

Proof We take $t>0$. From Lemma 3.1, we know that

$$\begin{aligned}&\sup_{t\leq s\leq t+\delta}\left|\frac{Q^{(N)}_{(\alpha,\beta)}\left(sN^3\right)}{N\sqrt{\log N}}-\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\right)}{N\sqrt{\log N}}\right|\\&\leq\sup_{t\leq s\leq t+\delta}\max_{i\leq N}\left(\tilde{R}^{(N)}_i\left(sN^3\right)-\tilde{R}^{(N)}_i\left(tN^3\right)\right)+2\sup_{t\leq s\leq t+\delta}\max_{i\leq N}\left(\tilde{R}^{(N)}_i\left(tN^3\right)-\tilde{R}^{(N)}_i\left(sN^3\right)\right).\end{aligned}$$

Observe that $\tilde{R}^{(N)}_i$ is a random walk. Therefore, due to the duality principle, we have that

$$\begin{aligned}&\sup_{t\leq s\leq t+\delta}\max_{i\leq N}\left(\tilde{R}^{(N)}_i\left(sN^3\right)-\tilde{R}^{(N)}_i\left(tN^3\right)\right)+2\sup_{t\leq s\leq t+\delta}\max_{i\leq N}\left(\tilde{R}^{(N)}_i\left(tN^3\right)-\tilde{R}^{(N)}_i\left(sN^3\right)\right)\\&\overset{d}{=}\sup_{0\leq s\leq\delta}\max_{i\leq N}\tilde{R}^{(N)}_i\left(sN^3\right)+2\sup_{0\leq s\leq\delta}\max_{i\leq N}-\tilde{R}^{(N)}_i\left(sN^3\right).\end{aligned}$$

Concluding,

$$\begin{aligned}&\frac{1}{\delta}\mathbb{P}\left(\sup_{t\leq s\leq t+\delta}\left|\frac{Q^{(N)}_{(\alpha,\beta)}\left(sN^3\right)}{N\sqrt{\log N}}-\frac{Q^{(N)}_{(\alpha,\beta)}\left(tN^3\right)}{N\sqrt{\log N}}\right|\geq\epsilon\right) &(3.14)\\&\leq\frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\max_{i\leq N}\tilde{R}^{(N)}_i\left(sN^3\right)+2\sup_{0\leq s\leq\delta}\max_{i\leq N}-\tilde{R}^{(N)}_i\left(sN^3\right)\geq\epsilon\right) &(3.15)\\&\leq\frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\max_{i\leq N}\tilde{R}^{(N)}_i\left(sN^3\right)\geq\frac{\epsilon}{2}\right)+\frac{1}{\delta}\mathbb{P}\left(2\sup_{0\leq s\leq\delta}\max_{i\leq N}-\tilde{R}^{(N)}_i\left(sN^3\right)\geq\frac{\epsilon}{2}\right). &(3.16)\end{aligned}$$

Now we focus on the first term in (3.16). The analysis of the second term goes analogously.

$$\begin{aligned}\frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\max_{i\leq N}\tilde{R}^{(N)}_i\left(sN^3\right)\geq\frac{\epsilon}{2}\right)&=\frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\max_{i\leq N}\frac{\tilde{A}^{(N)}\left(sN^3\right)+\tilde{S}^{(N)}_i\left(sN^3\right)}{\sqrt{\log N}}\geq\frac{\epsilon}{2}\right) &(3.17)\\&\leq\frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\frac{\tilde{A}^{(N)}\left(sN^3\right)}{\sqrt{\log N}}\geq\frac{\epsilon}{4}\right)+\frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\max_{i\leq N}\frac{\tilde{S}^{(N)}_i\left(sN^3\right)}{\sqrt{\log N}}\geq\frac{\epsilon}{4}\right). &(3.18)\end{aligned}$$

This first term in Equation (3.18) goes to 0 as $N\to\infty$ because it satisfies the inequality

$$\frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\frac{\tilde{A}^{(N)}\left(sN^3\right)}{\sqrt{\log N}}\geq\frac{\epsilon}{4}\right)\leq\frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\left|\frac{\tilde{A}^{(N)}\left(sN^3\right)}{\sqrt{\log N}}+\frac{\beta s}{\sqrt{\log N}}\right|\geq\frac{\epsilon}{8}\right)+\frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\frac{\beta s}{\sqrt{\log N}}\geq\frac{\epsilon}{8}\right).$$

In addition, by Doob's maximal submartingale inequality,

$$\begin{aligned}\frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\left|\frac{\tilde{A}^{(N)}\left(sN^3\right)}{\sqrt{\log N}}+\frac{\beta s}{\sqrt{\log N}}\right|\geq\frac{\epsilon}{8}\right)&=\frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\left(\frac{\tilde{A}^{(N)}\left(sN^3\right)}{\sqrt{\log N}}+\frac{\beta s}{\sqrt{\log N}}\right)^2\geq\frac{\epsilon^2}{64}\right)\\&\leq\frac{64}{\delta\epsilon^2\log N}\mathrm{Var}\left(\tilde{A}^{(N)}\left(\delta N^3\right)\right)\\&=\frac{64}{\delta\epsilon^2\log N}\left(\alpha\delta+o_N(1)\right)\overset{N\to\infty}{\longrightarrow}0.\end{aligned}$$

Furthermore, $\tilde{S}_i^{(N)}(n)$ is a martingale with mean 0. The maximum of independent martingales is a submartingale, therefore, $\left(\max\left(0, \max_{i\leq N} \tilde{S}_i^{(N)}(sN^3)/\sqrt{\log N}\right)\right)^4$ is a non-negative submartingale. Hence, we can use Doob's maximal submartingale inequality for the second term in Equation (3.18) and get

$$\begin{aligned}
\frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(sN^3)}{\sqrt{\log N}}\geq\frac{\epsilon}{4}\right) &\leq \frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\max\left(0,\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(sN^3)}{\sqrt{\log N}}\right)\geq\frac{\epsilon}{4}\right)\\
&\leq \frac{1}{\delta}\mathbb{P}\left(\sup_{0\leq s\leq\delta}\left(\max\left(0,\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(sN^3)}{\sqrt{\log N}}\right)\right)^4\geq\frac{\epsilon^4}{256}\right)\\
&\leq \frac{256}{\epsilon^4\delta}\mathbb{E}\left[\left(\max\left(0,\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(\delta N^3)}{\sqrt{\log N}}\right)\right)^4\right]\\
&\leq \frac{256}{\epsilon^4\delta}\mathbb{E}\left[\left(\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(\delta N^3)}{\sqrt{\log N}}\right)^4\right].
\end{aligned}$$

From Lemma 3.4, it follows that $\mathbb{E}\left[\left(\max_{i\leq N}\tilde{S}_i^{(N)}(\delta N^3)\right)^4\right]\approx 4\alpha^2\delta^2(\log N)^2$. Therefore,

$$\mathbb{E}\left[\left(\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(\delta N^3)}{\sqrt{\log N}}\right)^4\right]\overset{N\to\infty}{\longrightarrow}4\alpha^2\delta^2.$$

Hence, we can choose a $\delta>0$ such that

$$\frac{256}{\epsilon^4\delta}\mathbb{E}\left[\left(\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(\delta N^3)}{\sqrt{\log N}}\right)^4\right]$$

is of $O(\delta)$. Concluding, the terms in (3.16) are small, and therefore $\left(Q_{(\alpha,\beta)}^{(N)}(tN^3)/(N\sqrt{\log N}), t\in[0,T]\right)$ is tight. □

3.2. Steady state. In Section 2.3, we explained that a steady-state approximation of the maximum queue length scales with $N\log N$, which we formalized in Theorem 2.2. To prove Theorem 2.2, we investigate lower and upper bounds for $Q_{(\alpha,\beta)}^{(N)}(\infty)$. To this end, we construct for each server i two auxiliary processes. These two processes are split up in a small part, which is the same for each server, and a large part, which is independent and identically distributed for each server. We use

$$A^{(l,N)}(n)=-A^{(N)}(n)+\left(1-\frac{\alpha}{N}-\frac{\beta}{N^2}-\epsilon(N)\right)n \tag{3.19}$$

and

$$S_i^{(l,N)}(n)=-S_i^{(N)}(n)+\left(1-\frac{\alpha}{N}-\frac{\beta}{N^2}-\epsilon(N)\right)n, \tag{3.20}$$

$$A^{(u,N)}(n)=A^{(N)}(n)-\left(1-\frac{\alpha}{N}-\frac{\beta}{N^2}+\epsilon(N)\right)n \tag{3.21}$$

and

$$S_i^{(u,N)}(n) = -S_i^{(N)}(n) + \left(1 - \frac{\alpha}{N} - \frac{\beta}{N^2} + \epsilon(N)\right) n. \tag{3.22}$$

The superscripts (l) and (u) denote lower and upper bounds, respectively. Moreover, $\epsilon(N)$ is defined as m/N^2 with m a number between 0 and β. We choose this $\epsilon(N)$ to ensure that $A^{(l,N)}(n)$ and $A^{(u,N)}(n)$ have a negative drift. We use this property in the proofs of Lemmas 3.10 and 3.13. The random variables given in Equations (3.19), (3.20), (3.21) and (3.22) satisfy the equation

$$S_i^{(l,N)}(n) - A^{(l,N)}(n) = A^{(N)}(n) - S_i^{(N)}(n) = A^{(u,N)}(n) + S_i^{(u,N)}(n).$$

Due to our construction of (3.19), (3.20), (3.21) and (3.22), we have

$$\begin{aligned} \inf_{0\le k\le n} -A^{(l,N)}(k) + \max_{i\le N} \sup_{0\le k\le n} S_i^{(l,N)}(k) &\le Q_{(\alpha,\beta)}^{(N)}(n) \\ &\le \sup_{0\le k\le n} A^{(u,N)}(k) + \max_{i\le N} \sup_{0\le k\le n} S_i^{(u,N)}(k). \end{aligned} \tag{3.23}$$

We prove in Sections 3.2.1 and 3.2.2 that the lower and upper bound of $Q_{(\alpha,\beta)}^{(N)}(\infty)$ converge after scaling to the same constant, which proves Theorem 2.2. First of all, we investigate the lower bound in Equation (3.23). This lower bound consists of two parts, namely $\inf_{0\le k\le n} -A^{(l,N)}(k)$, which is treated in Lemma 3.10, and $\max_{i\le N} \sup_{0\le k\le n} S_i^{(l,N)}(k)$, for which a result is proven in Lemma 3.11. Finally, we do the same for the upper bound of $Q_{(\alpha,\beta)}^{(N)}(\infty)$ in Lemmas 3.13 and 3.14.

3.2.1. Lower bound.

LEMMA 3.10 **(Lower bound of the arrival process).** *For $n \in (0,\infty]$, we have*

$$\frac{\inf_{0\le k\le n} -A^{(l,N)}(k)}{N\log N} \xrightarrow{\mathbb{P}} 0 \text{ as } N\to\infty. \tag{3.24}$$

Proof We can write

$$A^{(l,N)}(n) = \sum_{j=1}^{n} X^{(l,N)}(j),$$

with

$$X^{(l,N)}(j) = \begin{cases} -\alpha/N - \beta/N^2 - m/N^2 & \text{w.p. } 1-\alpha/N-\beta/N^2, \\ 1-\alpha/N-\beta/N^2-m/N^2 & \text{w.p. } \alpha/N+\beta/N^2, \end{cases}$$

and $X^{(l,N)}(i) \perp X^{(l,N)}(j)$ for all $i \neq j$. Now, $\mathbb{E}\left[X^{(l,N)}(j)\right] = -m/N^2$. If we can find a $\theta_A^{(l,N)} > 0$ such that

$$\begin{aligned} \mathbb{E}\left[e^{\theta_A^{(l,N)} X^{(l,N)}(j)}\right] &= \left(\frac{\alpha}{N} + \frac{\beta}{N^2}\right) \exp\left\{\theta_A^{(l,N)}\left(1 - \frac{\alpha}{N} - \frac{\beta}{N^2} - \frac{m}{N^2}\right)\right\} \\ &+ \left(1 - \frac{\alpha}{N} - \frac{\beta}{N^2}\right) \exp\left\{\theta_A^{(l,N)}\left(-\frac{\alpha}{N} - \frac{\beta}{N^2} - \frac{m}{N^2}\right)\right\} = 1 \end{aligned} \tag{3.25}$$

then $\exp(\theta_A^{(l,N)} A^{(l,N)}(n))$ is martingale. This $\theta_A^{(l,N)}$ exists because $\mathbb{E}[e^{\theta X^{(l,N)}(j)}]$ is continuous in θ,

$$\frac{d}{d\theta}\mathbb{E}\left[e^{\theta X^{(l,N)}(j)}\right]\Big|_{\theta=0} < 0,$$

because the random walk has a negative drift, the moment generating function is finite for all θ, and

$$\lim_{\theta\to\infty}\mathbb{E}\left[e^{\theta X^{(l,N)}(j)}\right]=\infty.$$

We tailor a classical argument to our setting by using Lundberg's inequality, see for more details Asmussen, [4, Ch. 13]. Now, for $n>0$ we can use Doob's maximal submartingale inequality to get

$$\mathbb{P}\left(\sup_{0\le k\le n}A^{(l,N)}(k)\ge x\right)=\mathbb{P}\left(\sup_{0\le k\le n}e^{\theta_A^{(l,N)}A^{(l,N)}(k)}\ge e^{\theta_A^{(l,N)}x}\right)\le\mathbb{E}\left[e^{\theta_A^{(l,N)}A^{(l,N)}(n)}\right]e^{-\theta_A^{(l,N)}x}=e^{-\theta_A^{(l,N)}x}. \tag{3.26}$$

By the monotone convergence theorem, we know that this also holds for $n=\infty$. An estimate of this $\theta_A^{(l,N)}$ can be obtained by looking at the second order Taylor expansion of Equation (3.25) around $\theta=0$. The solution to

$$\theta\,\frac{d}{d\theta}\mathbb{E}\left[e^{\theta X^{(l,N)}(j)}\right]\Big|_{\theta=0}+\frac{\theta^2}{2}\,\frac{d^2}{d\theta^2}\mathbb{E}\left[e^{\theta X^{(l,N)}(j)}\right]\Big|_{\theta=0}=0$$

gives

$$\theta_A^{(l,N)}=\frac{2mN^2}{\alpha N^3+\beta N^2-\alpha^2N^2-2\alpha\beta N-\beta^2+m^2}+O\left(\frac{1}{N^2}\right).$$

A justification for this procedure and a more precise approximation of $\theta_A^{(l,N)}$ is given in Appendix A. We can conclude that $\theta_A^{(l,N)}\approx 2m/(\alpha N)$. Hence, we get that

$$\mathbb{E}\left[e^{\theta_A^{(l,N)}X^{(l,N)}(j)}\right]=1+O\left(\frac{1}{N^3}\right).$$

Therefore, from the inequality in Equation (3.26), we can conclude that for N large, $\sup_{k\ge0}A^{(l,N)}(k)$ is stochastically dominated by an exponentially distributed random variable E with mean $\alpha N/(2m)$. We have

$$\frac{E}{N\log N}\xrightarrow{\mathbb{P}}0\text{ as }N\to\infty.$$

Therefore,

$$\frac{\inf_{0\le k\le n}-A^{(l,N)}(n)}{N\log N}=\frac{-\sup_{0\le k\le n}A^{(l,N)}(n)}{N\log N}\xrightarrow{\mathbb{P}}0\text{ as }N\to\infty,$$

concluding the proof. □

Lemma 3.11 **(Lower bound of the service process).** *For $\delta>0$, we have*

$$\liminf_{N\to\infty}\mathbb{P}\left(\frac{\max_{i\le N}\sup_{n\ge0}S_i^{(l,N)}(n)}{N\log N}\ge\frac{\alpha}{2(\beta+m)}-\delta\right)=1. \tag{3.27}$$

Proof We can write

$$S_i^{(l,N)}(n)=\sum_{j=1}^{n}Y_i^{(l,N)}(j),$$

with

$$Y_i^{(l,N)}(j) = \begin{cases} 1-\alpha/N-\beta/N^2-m/N^2 & \text{w.p.} \quad \alpha/N, \\ -\alpha/N-\beta/N^2-m/N^2 & \text{w.p.} \quad 1-\alpha/N. \end{cases}$$

We have $\mathbb{E}\left[Y_i^{(l,N)}(j)\right] = -\beta/N^2 - m/N^2$. For $\{\mathcal{F}_n, n \geq 1\}$ the natural filtration up to time n, we define for all $A \in \mathcal{F}_n$

$$\tilde{\mathbb{P}}(A) = \mathbb{E}\left[e^{\theta_i^{(l,N)} S_i^{(l,N)}(n)} \mathbb{1}\{A\}\right].$$

For all i, $\theta_i^{(l,N)}$ satisfies

$$\begin{aligned}\mathbb{E}\left[e^{\theta_i^{(l,N)} Y_i^{(l,N)}(j)}\right] &= \frac{\alpha}{N}\exp\left\{\theta_i^{(l,N)}\left(1-\frac{\alpha}{N}-\frac{\beta}{N^2}-\frac{m}{N^2}\right)\right\} \\ &\quad + \left(1-\frac{\alpha}{N}\right)\exp\left\{\theta_i^{(l,N)}\left(-\frac{\alpha}{N}-\frac{\beta}{N^2}-\frac{m}{N^2}\right)\right\} = 1.\end{aligned}$$

We define the stopping time

$$\tau_i^{(l,N)}(x) = \min\left\{n : S_i^{(l,N)}(n) \geq x\right\}.$$

Observe that

$$S_i^{(l,N)}\left(\tau_i^{(l,N)}(x)\right) - x \leq 1,$$

and $S_i^{(l,N)}(n)$ has a positive drift under $\tilde{\mathbb{P}}$. Consequently, from [4, Ch. 13], we know that

$$\begin{aligned}\mathbb{P}\left(\sup_{n\geq 0} S_i^{(l,N)}(n) \geq x\right) &= \mathbb{P}\left(\tau_i^{(l,N)}(x) < \infty\right) = \tilde{\mathbb{E}}\left[e^{-\theta_i^{(l,N)} S_i^{(l,N)}\left(\tau_i^{(l,N)}(x)\right)} \mathbb{1}\left\{\tau_i^{(l,N)}(x) < \infty\right\}\right] \\ &= e^{-\theta_i^{(l,N)} x}\tilde{\mathbb{E}}\left[e^{-\theta_i^{(l,N)}\left(S_i^{(l,N)}\left(\tau_i^{(l,N)}(x)\right)-x\right)} \mathbb{1}\left\{\tau_i^{(l,N)}(x) < \infty\right\}\right] \\ &\geq e^{-\theta_i^{(l,N)}(x+1)}\tilde{\mathbb{P}}\left(\tau_i^{(l,N)}(x) < \infty\right) = e^{-\theta_i^{(l,N)}(x+1)}. \end{aligned} \tag{3.28}$$

When we solve the second order Taylor approximation of $\theta_i^{(l,N)}$, we get

$$\theta_i^{(l,N)} = \frac{2N^2(\beta+m)}{-\alpha^2N^2+\alpha N^3+\beta^2+m^2+2\beta m} + O\left(\frac{1}{N^2}\right).$$

Therefore, for N large $\theta_i^{(l,N)} \approx 2(\beta+m)/(\alpha N)$. From this it follows that $S_i^{(l,N)}(n)$ is stochastically bounded from below by an exponentially distributed random variable $E_i^{(l,N)}$ with mean $\alpha N/(2(\beta+m))$. All these random variables $E_i^{(l,N)}$ are independent for each i. Furthermore, $E_i^{(l,N)}/N$ is independent of N. Moreover,

$$\mathbb{P}\left(\max_{i\leq N}\frac{E_i^{(l,N)}}{N} \leq \frac{\alpha}{2(\beta+m)}(x+\log N)\right) \overset{N\to\infty}{\longrightarrow} e^{-e^{-x}}.$$

Therefore,

$$\frac{\max_{i\leq N} E_i^{(l,N)}}{N\log N} \xrightarrow{\mathbb{P}} \frac{\alpha}{2(\beta+m)} \text{ as } N\to\infty.$$

In conclusion, Equation (3.27) holds. □

LEMMA 3.12. *For $\delta > 0$, we have*

$$\liminf_{N\to\infty} \mathbb{P}\left(\frac{Q^{(N)}_{(\alpha,\beta)}(\infty)}{N\log N} \geq \frac{\alpha}{2\beta} - \delta \right) = 1.$$

Proof We constructed $S_i^{(l,N)}(n)$ and $A^{(l,N)}(n)$ such that

$$\inf_{n\geq 0} -A^{(l,N)}(n) + \max_{i\leq N} \sup_{n\geq 0} S_i^{(l,N)}(n) \leq Q^{(N)}_{(\alpha,\beta)}(\infty).$$

From Lemmas 3.10 and 3.11 we obtain that

$$\liminf_{N\to\infty} \mathbb{P}\left(\frac{1}{N\log N} \left(\inf_{n\geq 0} -A^{(l,N)}(n) + \max_{i\leq N} \sup_{n\geq 0} S_i^{(l,N)}(n) \right) \geq \frac{\alpha}{2(\beta+m)} - \delta \right) = 1.$$

The parameter m in this expression occurred in $\epsilon(N) = m/N^2$. This $m > 0$ was free to choose. Therefore,

$$\liminf_{N\to\infty} \mathbb{P}\left(\frac{Q^{(N)}_{(\alpha,\beta)}(\infty)}{N\log N} \geq \frac{\alpha}{2\beta} - \delta \right) = 1.$$

□

3.2.2. Upper bound. In order to complete the proof of the steady-state result, we also investigate the upper bound in Equation (3.23). We show that this upper bound converges to the same limit as the lower bound.

LEMMA 3.13 **(Upper bound of the arrival process).** *For $n \in (0,\infty]$, we have*

$$\frac{\sup_{0\leq k\leq n} A^{(u,N)}(k)}{N\log N} \xrightarrow{\mathbb{P}} 0 \text{ as } N\to\infty. \tag{3.29}$$

Proof We can again write

$$A^{(u,N)}(n) = \sum_{j=1}^{n} X^{(u,N)}(j)$$

with

$$X^{(u,N)}(j) = \begin{cases} \alpha/N + \beta/N^2 - m/N^2 & \text{w.p. } 1-\alpha/N-\beta/N^2, \\ -1+\alpha/N+\beta/N^2 - m/N^2 & \text{w.p. } \alpha/N+\beta/N^2, \end{cases}$$

hence $\mathbb{E}\left[X^{(u,N)}(j)\right] = -m/N^2$. We use the same argument as in Lemma 3.10 and obtain by Doob's maximal submartingale inequality that

$$\mathbb{P}\left(\sup_{0\leq k\leq n} A^{(u,N)}(k) \geq x \right) \leq e^{-\theta_A^{(u,N)} x},$$

with $\theta_A^{(u,N)}$ the solution to the equation

$$\begin{aligned} \mathbb{E}\left[e^{\theta_A^{(u,N)} X^{(u,N)}(j)} \right] &= \left(\frac{\alpha}{N} + \frac{\beta}{N^2} \right) \exp\left\{ \theta_A^{(u,N)} \left(-1 + \frac{\alpha}{N} + \frac{\beta}{N^2} - \frac{m}{N^2} \right) \right\} \\ &+ \left(1 - \frac{\alpha}{N} - \frac{\beta}{N^2} \right) \exp\left\{ \theta_A^{(u,N)} \left(\frac{\alpha}{N} + \frac{\beta}{N^2} - \frac{m}{N^2} \right) \right\} = 1. \end{aligned}$$

When we compute the second order Taylor approximation of this $\theta^{(A)}$ with respect to 0, we obtain

$$\theta_A^{(u,N)} = \frac{2mN^2}{-\alpha^2N^2 + \alpha N^3 - 2\alpha\beta N - \beta^2 + m^2 + \beta N^2} + O\left(\frac{1}{N^2}\right).$$

Consequently, we have for N large $\theta_A^{(u,N)} \approx 2m/(\alpha N)$. Therefore, by the same reasoning as in Lemma 3.10, we get (3.29). □

LEMMA 3.14 **(Upper bound of the service process).** *For $\delta > 0$, we have*

$$\limsup_{N\to\infty} \mathbb{P}\left(\frac{\max_{i\le N} \sup_{n\ge 0} S_i^{(u,N)}(n)}{N\log N} \ge \frac{\alpha}{2(\beta - m)} + \delta\right) = 0. \tag{3.30}$$

Proof We have

$$S_i^{(u,N)}(n) = \sum_{j=1}^{n} Y_i^{(u,N)}(j)$$

with

$$Y_i^{(u,N)}(j) = \begin{cases} -\alpha/N - \beta/N^2 + m/N^2 & \text{w.p. } 1 - \alpha/N, \\ 1 - \alpha/N - \beta/N^2 + m/N^2 & \text{w.p. } \alpha/N. \end{cases}$$

Hence, $\mathbb{E}\left[Y_i^{(u,N)}(j)\right] = (m-\beta)/N^2$. Because $m < \beta$, $S_i^{(u,N)}(n)$ has a negative drift. As in Lemmas 3.10 and 3.13, we again use Doob's maximal submartingale inequality. We have

$$\mathbb{P}\left(\sup_{n\ge 0} S_i^{(u,N)}(n) \ge x\right) \le e^{-\theta_i^{(u,N)} x}$$

with $\theta_i^{(u,N)}$ the solution to the equation

$$\begin{aligned}\mathbb{E}\left[e^{\theta_i^{(u,N)} Y_i^{(u,N)}(j)}\right] &= \frac{\alpha}{N}\exp\left\{\theta_i^{(u,N)}\left(1 - \frac{\alpha}{N} - \frac{\beta}{N^2} + \frac{m}{N^2}\right)\right\} \\ &\quad + \left(1 - \frac{\alpha}{N}\right)\exp\left\{\theta_i^{(u,N)}\left(-\frac{\alpha}{N} - \frac{\beta}{N^2} + \frac{m}{N^2}\right)\right\} = 1.\end{aligned}$$

The second order Taylor approximation of $\mathbb{E}\left[e^{\theta_i^{(u,N)} Y_i^{(u,N)}(j)}\right]$ with $\theta_i^{(u,N)}$ around 0 gives

$$\theta_i^{(u,N)} = \frac{2N^2(\beta - m)}{-\alpha^2N^2 + \alpha N^3 + (\beta - m)^2} + O\left(\frac{1}{N^2}\right).$$

Thus, for N large, $\theta_i^{(u,N)} \approx 2(\beta - m)/(\alpha N)$. Concluding, $\sup_{n\ge 0} S_i^{(u,N)}(n)$ is stochastically dominated by an exponentially distributed random variable $E_i^{(u,N)}$ with mean $\alpha N/(2(\beta - m))$. Because $\sup_{n\ge 0} S_i^{(u,N)}(n) \perp \sup_{n\ge 0} S_j^{(u,N)}(n)$ for $i \ne j$, we can conclude that also $E_i^{(u,N)} \perp E_j^{(u,N)}$ for $i \ne j$. Therefore,

$$\mathbb{P}\left(\frac{\max_{i\le N} E_i^{(u,N)}}{N} \le \frac{\alpha}{2(\beta - m)}(x + \log N)\right) \overset{N\to\infty}{\longrightarrow} e^{-e^{-x}},$$

and

$$\frac{\max_{i\le N} E_i^{(u,N)}}{N\log N} \overset{\mathbb{P}}{\longrightarrow} \frac{\alpha}{2(\beta - m)} \text{ as } N \to \infty.$$

From this we can conclude (3.30). □

Now, we can combine the results proven in Lemmas 3.13 and 3.14 to prove an upper bound for the maximum queue length in steady state.

LEMMA 3.15. *For $\delta > 0$, we have*

$$\limsup_{N\to\infty} \mathbb{P}\left(\frac{Q^{(N)}_{(\alpha,\beta)}(\infty)}{N \log N} \geq \frac{\alpha}{2\beta} + \delta \right) = 0. \tag{3.31}$$

Proof We know that

$$\frac{Q^{(N)}_{(\alpha,\beta)}(\infty)}{N\log N} \leq \frac{\sup_{n\geq 0} A^{(u,N)}(n)}{N\log N} + \frac{\max_{i\leq N} \sup_{n\geq 0} S_i^{(u,N)}(n)}{N\log N}.$$

Lemmas 3.13 and 3.14 provide us with the result that

$$\limsup_{N\to\infty} \mathbb{P}\left(\frac{Q^{(N)}_{(\alpha,\beta)}(\infty)}{N \log N} \geq \frac{\alpha}{2(\beta - m)} + \delta \right) = 0.$$

Letting $m \downarrow 0$ yields (3.31). □

Proof of Theorem 2.2. Lemmas 3.12 and 3.15 show us that for all $\delta > 0$

$$\liminf_{N\to\infty} \mathbb{P}\left(\frac{Q^{(N)}_{(\alpha,\beta)}(\infty)}{N \log N} \geq \frac{\alpha}{2\beta} - \delta \right) = 1,$$

and

$$\limsup_{N\to\infty} \mathbb{P}\left(\frac{Q^{(N)}_{(\alpha,\beta)}(\infty)}{N \log N} \geq \frac{\alpha}{2\beta} + \delta \right) = 0.$$

Concluding,

$$\frac{Q^{(N)}_{(\alpha,\beta)}(\infty)}{N\log N} \xrightarrow{\mathbb{P}} \frac{\alpha}{2\beta} \text{ as } N \to \infty.$$

□

REMARK 3.1. Baccelli, Makowski and Shwartz [8] give upper and lower bounds on the response times in fork-join queues, cf. Figure 1. They state that a fork-join queue with deterministic arrivals functions as a lower bound for the maximum queue length. It is easy to see by a heuristic argument that their lower bound is tight. In our setting, with the arrival and service processes stated in Definitions 2.1 and 2.2, and by replacing the arrival process with its mean, and scaling time with N^3 and space with $1/N$, we see for the expectation of the queue length that

$$\mathbb{E}\left[\frac{1}{N} \left(\left(1 - \frac{\alpha}{N} - \frac{\beta}{N^2} \right) tN^3 - S_i^{(N)}\left(tN^3\right) \right) \right] = -\beta t,$$

which we also see in the original model, cf. Equation (2.4). However, for the altered model with deterministic arrivals,

$$\mathrm{Var}\left(\frac{1}{N} \left(\left(1 - \frac{\alpha}{N} - \frac{\beta}{N^2} \right) tN^3 - S_i^{(N)}\left(tN^3\right) \right) \right) = \frac{1}{N^2} tN^3 \left(\frac{\alpha}{N} \left(1 - \frac{\alpha}{N} \right) \right) = \alpha t + o(1),$$

instead of $2\alpha t$ as in Equation (2.5). However, when the arrival process is deterministic, the queue lengths are independent non-negative random variables, cf. Definition 2.3 and Equation (2.1).

Therefore, the maximum of N independent random variables behaves like the maximum of N reflected Brownian motions with drift $-\beta$ and variance α in steady state. By using the cumulative distribution function given in [1], we see that the invariant distribution of each queue length is exponential, thus the scaled maximum queue length converges to $\alpha/2\beta$ as in Theorem 2.2.

They also give several upper bounds. First of all, an upper bound is given by assuming N independent arrival processes, which are the same in distribution as our arrival process, cf. Definition 2.3. We now see

$$\mathbb{E}\left[\frac{1}{N}\left(A_i^{(N)}\left(tN^3\right)-S_i^{(N)}\left(tN^3\right)\right)\right]=-\beta t,$$

and

$$\mathrm{Var}\left(\frac{1}{N}\left(A_i^{(N)}\left(tN^3\right)-S_i^{(N)}\left(tN^3\right)\right)\right)=2\alpha t+o(1).$$

This gives an upper bound of $2\alpha/2\beta=\alpha/\beta$ in steady state, since this is a maximum of N reflected Brownian motions with drift $-\beta$ and variance 2α, which is too high by a factor two.

Secondly, an upper bound is given by considering an independent system with the following characteristics: write the interarrival time A as a sum of independent and identically distributed random variables,

$$A\overset{d}{=}\frac{1}{N}\sum_{i=1}A_i.$$

Then the maximum queue length of N independent queues, each having interarrival times A_i, gives an upper bound to the maximum queue length. Since we have geometrically distributed interarrival times with parameter $1-\alpha/N-\beta/N^2$, $A_i\sim N\cdot\mathrm{NegativeBinomial}(1/N,1-\alpha/N-\beta/N^2)$. This random variable takes values in $\{1,N+1,2N+1,\ldots\}$ whose variance converges to a nonzero constant as N goes to ∞. In our original process, the variance converges to 0. Therefore, this upper bound in [8] is also not tight.

3.3. Convergence on large finite-time intervals. In Section 2, we gave a heuristic argument for two heavy-traffic regimes. Namely, in Theorem 2.1 we proved convergence when time is scaled with N^3 while in Theorem 2.2 we did so as time goes to ∞. So one might wonder when the shift between two heavy-traffic regimes happens. In Corollary 2.1 an answer to this question is given by expanding the steady-state result to certain finite-time horizons. The corollary states that the steady-state result still holds when time is of larger order than $N^3\log N$. Since the fluid limit given in Theorem 2.1 holds for time of order N^3, we see that we can point out the border between the two regimes quite well.

Proof of Corollary 2.1. We again use the lower and upper bound given in Equation (3.23). We can repeat the results of Lemmas 3.10, 3.13 and 3.14, since these results were obtained for finite length intervals and for the interval $[0,\infty)$. However, Lemma 3.11 gives a result on

$$\frac{\max_{i\le N}\sup_{n\ge 0}S_i^{(l,N)}(n)}{N\log N},$$

so only on the infinite length interval. Hence, we should investigate

$$\frac{\max_{i\le N}\sup_{0\le n\le t(N)}S_i^{(l,N)}(n)}{N\log N}.$$

We can find a stochastic lower bound for $\sup_{0\le n\le t(N)} S_i^{(l,N)}(n)$, by replacing ∞ with $t(N)$ in (3.28), thus obtaining

$$\begin{aligned}\mathbb{P}\left(\sup_{0\le n\le t(N)} S_i^{(l,N)}(n)\ge x\right) &= \mathbb{P}\left(\tau_i^{(l,N)}(x)<t(N)\right) = \tilde{\mathbb{E}}\left[e^{-\theta_i^{(l,N)}S_i^{(l,N)}\left(\tau_i^{(l,N)}(x)\right)}\mathbb{1}\left\{\tau_i^{(l,N)}(x)<t(N)\right\}\right]\\ &= e^{-\theta_i^{(l,N)}x}\tilde{\mathbb{E}}\left[e^{-\theta_i^{(l,N)}\left(S_i^{(l,N)}\left(\tau_i^{(l,N)}(x)\right)-x\right)}\mathbb{1}\left\{\tau_i^{(l,N)}(x)<t(N)\right\}\right]\\ &\ge e^{-\theta_i^{(l,N)}(x+1)}\tilde{\mathbb{P}}\left(\tau_i^{(l,N)}(x)<t(N)\right).\end{aligned}$$

To work out $\tilde{\mathbb{P}}$, we observe the following. We have

$$\tilde{\mathbb{P}}\left(Y_i^{(l,N)}(j)=1-\frac{\alpha}{N}-\frac{\beta}{N^2}-\frac{m}{N^2}\right)=\frac{\alpha}{N}\exp\left\{\theta_i^{(l,N)}\left(1-\frac{\alpha}{N}-\frac{\beta}{N^2}-\frac{m}{N^2}\right)\right\}$$

and

$$\tilde{\mathbb{P}}\left(Y_i^{(l,N)}(j)=-\frac{\alpha}{N}-\frac{\beta}{N^2}-\frac{m}{N^2}\right)=\left(1-\frac{\alpha}{N}\right)\exp\left\{\theta_i^{(l,N)}\left(-\frac{\alpha}{N}-\frac{\beta}{N^2}-\frac{m}{N^2}\right)\right\}.$$

Therefore,

$$\tilde{\mathbb{E}}\left[Y_i^{(l,N)}(j)\right]=\frac{\alpha}{N}\exp\left\{\theta_i^{(l,N)}\left(1-\frac{\alpha}{N}-\frac{\beta}{N^2}-\frac{m}{N^2}\right)\right\}-\frac{\alpha}{N}-\frac{\beta}{N^2}-\frac{m}{N^2}. \tag{3.32}$$

We have $\theta_i^{(l,N)}\approx 2(\beta+m)/(\alpha N)$. Therefore

$$\tilde{\mathbb{E}}\left[Y_i^{(l,N)}(j)\right]=\frac{\beta+m}{N^2}+O\left(\frac{1}{N^3}\right),$$

which we see by taking the first order Taylor approximation of Equation (3.32). So $S_i^{(l,N)}(n)$ has a positive drift under $\tilde{\mathbb{P}}$. Therefore, the first passage time $\tau_i^{(l,N)}(x)$ of level x has a finite expectation under $\tilde{\mathbb{P}}$. By Wald's equation, we have

$$\tilde{\mathbb{E}}\left[Y_i^{(l,N)}(j)\right]\tilde{\mathbb{E}}\left[\tau_i^{(l,N)}(x)\right]=x.$$

From this it follows that

$$\tilde{\mathbb{E}}\left[\tau_i^{(l,N)}(x)\right]=x\frac{N^2}{\beta+m}+O(N).$$

By Markov's inequality, we obtain

$$\tilde{\mathbb{P}}\left(\tau_i^{(l,N)}(x)>t(N)\right)\le x\frac{N^2}{t(N)(\beta+m)}+O\left(\frac{N}{t(N)}\right).$$

Therefore, for $t(N)/(N^3\log N)\overset{N\to\infty}{\longrightarrow}\infty$, there exists a function $h(N)$ with $h(N)N\log N\overset{N\to\infty}{\longrightarrow}0$ such that

$$\tilde{\mathbb{P}}\left(\tau_i^{(l,N)}(x)<t(N)\right)\ge 1-x\frac{N^2}{t(N)(\beta+m)}+O\left(\frac{N}{t(N)}\right)=1-xh(N)$$

and

$$\mathbb{P}\left(\sup_{0\le n\le t(N)} S_i^{(l,N)}(n)\ge x\right)\ge e^{-\theta_i^{(l,N)}(x+1)}(1-xh(N)).$$

Now, we define a sequence of i.i.d. random variables $E_i^{(N)}$ with $E_i^{(N)}$ independent for all i's such that

$$\mathbb{P}(E_i^{(N)} \geq x) = e^{-\theta_i^{(l,N)}(x+1)}(1 - xh(N)).$$

Thus,

$$\sup_{0\leq n\leq t(N)} S_i^{(l,N)}(n) \geq_{st} E_i^{(N)}.$$

Since $\theta_i^{(l,N)} \approx 2(\beta+m)/(\alpha N)$, we know that $\log N/\theta_i^{(l,N)} \approx N\log N\alpha/2(\beta+m)$. Therefore,

$$\mathbb{P}\left(\sup_{0\leq n\leq t(N)} S_i^{(l,N)}(n) \geq \frac{\log N}{\theta_i^{(l,N)}} + \frac{x}{\theta_i^{(l,N)}}\right) \geq e^{-\theta_i^{(l,N)} - x - \log N}(1 - o(1)).$$

Because $h(N)N\log N \overset{N\to\infty}{\longrightarrow} 0$, we know that

$$h(n)\left(\frac{\log N}{\theta_i^{(l,N)}} + \frac{x}{\theta_i^{(l,N)}}\right) \overset{N\to\infty}{\longrightarrow} 0.$$

Since

$$\begin{aligned}\lim_{N\to\infty} \mathbb{P}\left(\max_{i\leq N} E_i^{(N)} \leq \frac{\log N}{\theta_i^{(l,N)}} + \frac{x}{\theta_i^{(l,N)}}\right) &= \lim_{N\to\infty}\left(1 - e^{-\theta_i^{(l,N)} - x - \log N}(1 - o(1))\right)^N \\ &= e^{-e^{-x}},\end{aligned}$$

we know that $\max_{i\leq N} E_i^{(N)}$ converges to a Gumbel random variable after scaling. From this we can conclude that

$$\liminf_{N\to\infty} \mathbb{P}\left(\frac{\max_{i\leq N}\sup_{0\leq n\leq t(N)} S_i^{(l,N)}(n)}{N\log N} \geq \frac{\alpha}{2(\beta+m)} - \delta\right) = 1,$$

which is the same conclusion as is drawn in Lemma 3.11. Combined with the conclusions from Lemmas 3.10, 3.13 and 3.14, we can conclude

$$\frac{Q_{(\alpha,\beta)}^{(N)}(t(N))}{N\log N} \xrightarrow{\mathbb{P}} \frac{\alpha}{2\beta} \text{ as } N\to\infty.$$

□

4. Conclusion. In this paper, we analyzed a fork-join network with N servers in heavy traffic. We considered the case of nearly-deterministic arrivals and service times, and we derived a fluid limit and a steady-state approximation of the maximum queue length, in Theorems 2.1 and 2.2, as N grows large. These two results have remarkable differences. The first interesting difference is that there is a different scaling needed in order to get a non-trivial limit. Secondly, the steady-state approximation depends on two model parameters, whereas the fluid limit only depends on one model parameter and time.

In this paper, we assumed delays to be memoryless. However, we are confident that these results can be extended to nearly deterministic settings where the delays have general distributions.

Moreover, as Figure 2 shows, a further refinement of the limits could be given. Therefore, it is interesting to look at second order convergence of the maximum queue length. In other words, we

gain more insight in the process when we can find convergence results of $Q^{(N)}_{(\alpha,\beta)}(tN^3)/N - \sqrt{\log N}$ and $Q^{(N)}_{(\alpha,\beta)}(\infty)/N - \log N$. Since we have a dependent arrival process, it is a challenge to find weak convergence to an extreme value distribution. A good starting point could be to replace the arrival and service processes with Brownian motions, and analyze

$$\max_{i\leq N} \sup_{t\geq 0} \sqrt{\alpha} W_0(t) + \sqrt{\alpha} W_i(t) - \beta t$$

with $W_i, i \geq 0$ independent standard Brownian motions.

Appendix A: Taylor expansion of $\theta_A^{(l,N)}$. The parameter $\theta_A^{(l,N)}$ is the strictly positive solution to the equation

$$\begin{aligned}\mathbb{E}\left[e^{\theta_A^{(l,N)} X^{(l,N)}(j)}\right] &= \left(\frac{\alpha}{N} + \frac{\beta}{N^2}\right) \exp\left\{\theta_A^{(l,N)} \left(1 - \frac{\alpha}{N} - \frac{\beta}{N^2} - \epsilon(N)\right)\right\} \\ &\quad + \left(1 - \frac{\alpha}{N} - \frac{\beta}{N^2}\right) \exp\left\{\theta_A^{(l,N)} \left(-\frac{\alpha}{N} - \frac{\beta}{N^2} - \epsilon(N)\right)\right\} = 1,\end{aligned}$$

with $\epsilon(N) = m/N^2$. We found an approximation of $\theta_A^{(l,N)}$, of $2m/(\alpha N)$. To investigate the behavior of $\theta_A^{(l,N)}$ more carefully, we look at the function $\theta(x)$ such that

$$\begin{aligned}f(x,\theta(x)) &= \left(\alpha x + \beta x^2\right) \exp\left\{\theta(x)\left(1 - \alpha x - \beta x^2 - m x^2\right)\right\} \\ &\quad + \left(1 - \alpha x - \beta x^2\right) \exp\left\{\theta(x)\left(-\alpha x - \beta x^2 - m x^2\right)\right\} = 1.\end{aligned}$$

When we set $x_N = 1/N$, we get $f(x_N, \theta(x_N)) = \mathbb{E}\left[e^{\theta_A^{(l,N)} X^{(l,N)}(j)}\right]$. We are interested in the case that N is large, therefore we have to investigate f for x around 0. Since $f(x,\theta(x)) = 1$, we know that $f^{(n)}(0,\theta(0)) = 0$ for all $n \geq 1$. When we solve these equations for θ iteratively, we can find $\theta^{(i)}(0)$ for all $i \geq 0$ and we get a Taylor expansion of $\theta(x)$ around 0. Since $f(x,\theta(x)) = 1$, we know that

$$\frac{d}{dx} f(x,\theta(x))\bigg|_{x=0} = -\alpha + \alpha e^{\theta(0)} - \alpha\theta(0) = 0.$$

Hence, $\theta(0) = 0$. When we look at the second and the third derivative of $f(x,\theta(x))$ around 0, while using that $\theta(0) = 0$, we see

$$\frac{d^2}{dx^2} f(x,\theta(x))\bigg|_{x=0} = 0,$$

and

$$\frac{d^3}{dx^3} f(x,\theta(x))\bigg|_{x=0} = 3\theta'(0)\left(\alpha\theta'(0) - 2m\right).$$

Because we know that $f(x,\theta(x)) = 1$ we solve

$$3\theta'(0)\left(\alpha\theta'(0) - 2m\right) = 0.$$

This gives $\theta'(0) = 0$ or $\theta'(0) = 2m/\alpha$. $\theta'(0) = 0$ indicates the situation that $\theta \equiv 0$. If we now use the information that $\theta'(0) = 2m/\alpha$ and look at the fourth derivative of f we see that

$$\frac{d^4}{dx^4} f(x,\theta(x))\bigg|_{x=0} = 4m\left(3\theta''(0) + \frac{4m\left(-3\alpha^2 + 3\beta + 2m\right)}{\alpha^2}\right) = 0.$$

This gives that

$$\theta''(0) = -\frac{4m\left(-3\alpha^2+3\beta+2m\right)}{3\alpha^2}.$$

Indeed, we can compute each derivative of $\theta(0)$ iteratively. This gives

$$\theta(x) = \frac{2m}{\alpha}x - \frac{4m\left(-3\alpha^2+3\beta+2m\right)}{3\alpha^2}\frac{x^2}{2} + O(x^3).$$

Since the function $f(x,\theta)-1$ is analytic we know by the implicit function theorem that the solution $\theta(x)$ is also analytic. So for $x = 1/N$ and N is large enough we know that

$$\theta_A^{(l,N)} = \frac{2m}{\alpha N} + O\left(\frac{1}{N^2}\right).$$

The analysis of $\theta_A^{(u,N)}, \theta_i^{(l,N)}$ and $\theta_i^{(u,N)}$ leads to the same approximation error.

Appendix B: Proof of Lemmas 3.1, 3.2, 3.3, 3.4 and 3.5.

Proof of Lemma 3.1. We take $s > t > 0$. Due to the defined auxiliary processes in Definition 3.1, we can write the maximum queue length as in Equation (3.1):

$$\begin{aligned}\frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)}{N\sqrt{\log N}} &= \max_{i\le N}\sup_{0\le r\le t}\frac{\left(\tilde{A}^{(N)}\left(tN^3\right)-\tilde{A}^{(N)}\left(rN^3\right)\right)+\left(\tilde{S}_i^{(N)}\left(tN^3\right)-\tilde{S}_i^{(N)}\left(rN^3\right)\right)}{\sqrt{\log N}}\\ &= \max_{i\le N}\sup_{0\le r\le t}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right).\end{aligned}$$

For $s > t$, we prove below that the following upper bound holds:

$$\begin{aligned}\frac{Q_{(\alpha,\beta)}^{(N)}\left(sN^3\right)}{N\sqrt{\log N}} - \frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)}{N\sqrt{\log N}} \le& \max_{i\le N}\left|\tilde{R}_i^{(N)}\left(sN^3\right)-\tilde{R}_i^{(N)}\left(tN^3\right)\right|\\ &+\max_{i\le N}\sup_{t\le r\le s}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right).\end{aligned} \tag{B.1}$$

To deduce this inequality, we first write $Q_{(\alpha,\beta)}^{(N)}\left(sN^3\right)/\left(N\sqrt{\log N}\right) - Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)/\left(N\sqrt{\log N}\right)$ in terms of $\tilde{R}_i^{(N)}$:

$$\begin{aligned}&\frac{Q_{(\alpha,\beta)}^{(N)}\left(sN^3\right)}{N\sqrt{\log N}} - \frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)}{N\sqrt{\log N}}\\ =&\max_{i\le N}\sup_{0\le r\le s}\left(\tilde{R}_i^{(N)}\left(sN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right)-\max_{i\le N}\sup_{0\le q\le t}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(qN^3\right)\right)\\ =&\max_{i\le N}\left[\tilde{R}_i^{(N)}\left(sN^3\right)-\tilde{R}_i^{(N)}\left(tN^3\right)+\sup_{0\le r\le s}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right)\right]\\ &-\max_{i\le N}\sup_{0\le q\le t}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(qN^3\right)\right).\end{aligned}$$

Now, the following upper bounds for $Q_{(\alpha,\beta)}^{(N)}\left(sN^3\right)/\left(N\sqrt{\log N}\right) - Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)/\left(N\sqrt{\log N}\right)$ hold:

$$\begin{aligned}&\frac{Q_{(\alpha,\beta)}^{(N)}\left(sN^3\right)}{N\sqrt{\log N}} - \frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)}{N\sqrt{\log N}}\\ \le&\max_{i\le N}\left(\tilde{R}_i^{(N)}\left(sN^3\right)-\tilde{R}_i^{(N)}\left(tN^3\right)\right)+\max_{i\le N}\sup_{0\le r\le s}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right)\end{aligned}$$

$$
\begin{aligned}
&-\max_{i\leq N}\sup_{0\leq q\leq t}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(qN^3\right)\right)\\
&\leq\max_{i\leq N}\left(\tilde{R}_i^{(N)}\left(sN^3\right)-\tilde{R}_i^{(N)}\left(tN^3\right)\right)\\
&+\max_{i\leq N}\left[\sup_{0\leq r\leq s}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right)-\sup_{0\leq q\leq t}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(qN^3\right)\right)\right].
\end{aligned}
$$

Observe that both $\sup_{0\leq r\leq s}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right)$ and $\sup_{0\leq q\leq t}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(qN^3\right)\right)$ are non-negative random variables. Furthermore,

$$
\begin{aligned}
&\sup_{0\leq r\leq s}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right)-\sup_{0\leq q\leq t}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(qN^3\right)\right)\\
&\leq\sup_{t\leq r\leq s}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right).
\end{aligned}
$$

Now, we can conclude that

$$
\begin{aligned}
&\frac{Q_{(\alpha,\beta)}^{(N)}\left(sN^3\right)}{N\sqrt{\log N}}-\frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)}{N\sqrt{\log N}}\\
&\leq\max_{i\leq N}\left(\tilde{R}_i^{(N)}\left(sN^3\right)-\tilde{R}_i^{(N)}\left(tN^3\right)\right)\\
&+\max_{i\leq N}\left[\sup_{0\leq r\leq s}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right)-\sup_{0\leq q\leq t}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(qN^3\right)\right)\right]\\
&\leq\max_{i\leq N}\left|\tilde{R}_i^{(N)}\left(sN^3\right)-\tilde{R}_i^{(N)}\left(tN^3\right)\right|+\max_{i\leq N}\sup_{t\leq r\leq s}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right),
\end{aligned}
$$

and hence the inequality in Equation (B.1) is satisfied. We can similarly deduce the lower bound

$$
\frac{Q_{(\alpha,\beta)}^{(N)}\left(sN^3\right)}{N\sqrt{\log N}}-\frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)}{N\sqrt{\log N}}\geq-\max_{i\leq N}\left|\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(sN^3\right)\right|. \tag{B.2}
$$

To show this, we write

$$
\begin{aligned}
&\frac{Q_{(\alpha,\beta)}^{(N)}\left(sN^3\right)}{N\sqrt{\log N}}-\frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)}{N\sqrt{\log N}}\\
&=\max_{i\leq N}\sup_{0\leq r\leq s}\left(\tilde{R}_i^{(N)}\left(sN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right)-\max_{i\leq N}\sup_{0\leq q\leq t}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(qN^3\right)\right)\\
&=\max_{i\leq N}\sup_{0\leq r\leq s}\left(\tilde{R}_i^{(N)}\left(sN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right)\\
&-\max_{i\leq N}\left[\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(sN^3\right)+\sup_{0\leq q\leq t}\left(\tilde{R}_i^{(N)}\left(sN^3\right)-\tilde{R}_i^{(N)}\left(qN^3\right)\right)\right]\\
&\geq\max_{i\leq N}\sup_{0\leq r\leq s}\left(\tilde{R}_i^{(N)}\left(sN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right)\\
&-\max_{i\leq N}\left(\tilde{R}_i^{(N)}\left(tN^3\right)-\tilde{R}_i^{(N)}\left(sN^3\right)\right)-\max_{i\leq N}\sup_{0\leq q\leq t}\left(\tilde{R}_i^{(N)}\left(sN^3\right)-\tilde{R}_i^{(N)}\left(qN^3\right)\right).
\end{aligned}
$$

Observe that

$$
\sup_{0\leq r\leq s}\left(\tilde{R}_i^{(N)}\left(sN^3\right)-\tilde{R}_i^{(N)}\left(rN^3\right)\right)\geq\sup_{0\leq q\leq t}\left(\tilde{R}_i^{(N)}\left(sN^3\right)-\tilde{R}_i^{(N)}\left(qN^3\right)\right),
$$

because $s>t$, so on the left side of the inequality, the supremum is taken over a larger interval than on the right side of the inequality. From this we can conclude that

$$
\frac{Q_{(\alpha,\beta)}^{(N)}\left(sN^3\right)}{N\sqrt{\log N}}-\frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)}{N\sqrt{\log N}}
$$

$$
\begin{aligned}
&\geq \max_{i\leq N} \sup_{0\leq r\leq s} \left(\tilde{R}_i^{(N)}\left(sN^3\right) - \tilde{R}_i^{(N)}\left(rN^3\right)\right) \\
&- \max_{i\leq N} \left(\tilde{R}_i^{(N)}\left(tN^3\right) - \tilde{R}_i^{(N)}\left(sN^3\right)\right) - \max_{i\leq N} \sup_{0\leq q\leq t} \left(\tilde{R}_i^{(N)}\left(sN^3\right) - \tilde{R}_i^{(N)}\left(qN^3\right)\right) \\
&\geq - \max_{i\leq N} \left(\tilde{R}_i^{(N)}\left(tN^3\right) - \tilde{R}_i^{(N)}\left(sN^3\right)\right) \geq - \max_{i\leq N} \left|\tilde{R}_i^{(N)}\left(tN^3\right) - \tilde{R}_i^{(N)}\left(sN^3\right)\right|,
\end{aligned}
$$

and indeed (B.2) holds. Combining (B.1) and (B.2) gives

$$
\begin{aligned}
&\left|\frac{Q_{(\alpha,\beta)}^{(N)}\left(sN^3\right)}{N\sqrt{\log N}} - \frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)}{N\sqrt{\log N}}\right| \\
&\leq \max_{i\leq N} \left|\tilde{R}_i^{(N)}\left(sN^3\right) - \tilde{R}_i^{(N)}\left(tN^3\right)\right| + \max_{i\leq N} \sup_{t\leq r\leq s} \left(\tilde{R}_i^{(N)}\left(tN^3\right) - \tilde{R}_i^{(N)}\left(rN^3\right)\right).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\sup_{t\leq s\leq t+\delta} \left|\frac{Q_{(\alpha,\beta)}^{(N)}\left(sN^3\right)}{N\sqrt{\log N}} - \frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)}{N\sqrt{\log N}}\right| \\
&\leq \sup_{t\leq s\leq t+\delta} \max_{i\leq N} \left|\tilde{R}_i^{(N)}\left(sN^3\right) - \tilde{R}_i^{(N)}\left(tN^3\right)\right| + \sup_{t\leq s\leq t+\delta} \max_{i\leq N} \left(\tilde{R}_i^{(N)}\left(tN^3\right) - \tilde{R}_i^{(N)}\left(sN^3\right)\right)
\end{aligned}
$$

Since both $\sup_{t\leq s\leq t+\delta}\left(\tilde{R}_i^{(N)}\left(tN^3\right) - \tilde{R}_i^{(N)}\left(sN^3\right)\right)$ and $\sup_{t\leq s\leq t+\delta}\left(\tilde{R}_i^{(N)}\left(sN^3\right) - \tilde{R}_i^{(N)}\left(tN^3\right)\right)$ are non-negative random variables, we have that

$$
\begin{aligned}
\sup_{t\leq s\leq t+\delta} \max_{i\leq N} \left|\tilde{R}_i^{(N)}\left(sN^3\right) - \tilde{R}_i^{(N)}\left(tN^3\right)\right| &\leq \sup_{t\leq s\leq t+\delta} \max_{i\leq N} \left(\tilde{R}_i^{(N)}\left(sN^3\right) - \tilde{R}_i^{(N)}\left(tN^3\right)\right) \\
&+ \sup_{t\leq s\leq t+\delta} \max_{i\leq N} \left(\tilde{R}_i^{(N)}\left(tN^3\right) - \tilde{R}_i^{(N)}\left(sN^3\right)\right).
\end{aligned}
$$

From this, we can conclude that

$$
\begin{aligned}
&\sup_{t\leq s\leq t+\delta} \left|\frac{Q_{(\alpha,\beta)}^{(N)}\left(sN^3\right)}{N\sqrt{\log N}} - \frac{Q_{(\alpha,\beta)}^{(N)}\left(tN^3\right)}{N\sqrt{\log N}}\right| \\
&\leq \sup_{t\leq s\leq t+\delta} \max_{i\leq N} \left|\tilde{R}_i^{(N)}\left(sN^3\right) - \tilde{R}_i^{(N)}\left(tN^3\right)\right| + \sup_{t\leq s\leq t+\delta} \max_{i\leq N} \left(\tilde{R}_i^{(N)}\left(tN^3\right) - \tilde{R}_i^{(N)}\left(sN^3\right)\right) \\
&\leq \sup_{t\leq s\leq t+\delta} \max_{i\leq N} \left(\tilde{R}_i^{(N)}\left(sN^3\right) - \tilde{R}_i^{(N)}\left(tN^3\right)\right) + 2 \sup_{t\leq s\leq t+\delta} \max_{i\leq N} \left(\tilde{R}_i^{(N)}\left(tN^3\right) - \tilde{R}_i^{(N)}\left(sN^3\right)\right).
\end{aligned}
$$

□

Proof of Lemma 3.2. We know by the central limit theorem that

$$
\frac{\tilde{S}_i^{(N)}\left(tN^3\right)}{\sqrt{\alpha t\left(1-\alpha/N\right)}} \xrightarrow{d} Z.
$$

with $Z \sim \mathcal{N}(0,1)$. Michel [19] proved a result on the rate of convergence of the cumulative distribution function

$$
\mathbb{P}\left(\frac{\tilde{S}_i^{(N)}\left(tN^3\right)}{\sqrt{\alpha t(1-\alpha/N)}} < y\right)
$$

to the cumulative distribution function of a standard normal random variable. [19, Thm. 1& 2, p. 102 & 103] states that

$$\left|\mathbb{P}\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t(1-\alpha/N)}}<y\right)-\Phi(y)\right|$$
$$\leq\begin{cases}\frac{b}{(tN^3)^{\frac{\min(c,1)}{2}}}\exp\left(-\frac{(1-\sigma)y^2}{2}\right)+tN^3\,\mathbb{P}\left(\left|\frac{\tilde{S}_i^{(N)}(1)\sqrt{tN^3}}{\sqrt{\alpha t(1-\frac{\alpha}{N})}}\right|>r\sqrt{tN^3}\,|y|\right) & \text{for } y<\sqrt{(c+1)\log(tN^3)},\\ \frac{b}{t^{\frac{c}{2}}N^{\frac{3c}{2}}}y^{-2(c+2)}+tN^3\,\mathbb{P}\left(\left|\frac{\tilde{S}_i^{(N)}(1)\sqrt{tN^3}}{\sqrt{\alpha t(1-\frac{\alpha}{N})}}\right|>r\sqrt{tN^3}\,|y|\right) & \text{for } y>\sqrt{(c+1)\log(tN^3)}.\end{cases}$$

with $\sigma=\min(c,1)/(2(c+1))$ and $b,r>0$. So when we fill in $c=1$, we get

$$\left|\mathbb{P}\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t(1-\alpha/N)}}<y\right)-\Phi(y)\right|$$
$$\leq\begin{cases}\frac{b}{\sqrt{tN^3}}\exp\left(-\frac{3y^2}{8}\right)+tN^3\,\mathbb{P}\left(\left|\frac{\tilde{S}_i^{(N)}(1)\sqrt{tN^3}}{\sqrt{\alpha t(1-\frac{\alpha}{N})}}\right|>r\sqrt{tN^3}\,|y|\right) & \text{for } y<\sqrt{(c+1)\log(tN^3)},\\ \frac{b}{\sqrt{tN^3}}y^{-6}+tN^3\,\mathbb{P}\left(\left|\frac{\tilde{S}_i^{(N)}(1)\sqrt{tN^3}}{\sqrt{\alpha t(1-\frac{\alpha}{N})}}\right|>r\sqrt{tN^3}\,|y|\right) & \text{for } y>\sqrt{(c+1)\log(tN^3)}.\end{cases}$$

Since $\exp(-3y^2/8)<ky^{-6}$ for certain $k>1$, and for all y, we know that for all y,

$$\left|\mathbb{P}\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t(1-\alpha/N)}}<y\right)-\Phi(y)\right|\leq\frac{kb}{\sqrt{tN^3}}y^{-6}+tN^3\,\mathbb{P}\left(\left|\frac{\tilde{S}_i^{(N)}(1)\sqrt{tN^3}}{\sqrt{\alpha t(1-\alpha/N)}}\right|>r\sqrt{tN^3}\,|y|\right).$$

By using Markov's inequality, we get

$$tN^3\,\mathbb{P}\left(\left|\frac{\tilde{S}_i^{(N)}(1)\sqrt{tN^3}}{\sqrt{\alpha t(1-\alpha/N)}}\right|>r\sqrt{tN^3}\,|y|\right)\leq tN^3\frac{\mathbb{E}\left[\left(\tilde{S}_i^{(N)}(1)\sqrt{tN^3}/\sqrt{\alpha t(1-\alpha/N)}\right)^6\right]}{r^6t^3N^9y^6}.$$

Recall from Definition 3.1 that $\tilde{S}_i^{(N)}(1)$ is a scaled Bernoulli distributed random variable minus a number, hence we can compute its sixth moment explicitly. We get

$$tN^3\frac{\mathbb{E}\left[\left(\tilde{S}_i^{(N)}(1)\sqrt{tN^3}/\sqrt{\alpha t(1-\alpha/N)}\right)^6\right]}{r^6t^3N^9y^6}=\frac{1}{(1-\alpha/N)^3}\left(\frac{1}{\alpha^2N^4r^6t^2y^6}-\frac{6}{\alpha N^5r^6t^2y^6}+\frac{15}{N^6r^6t^2y^6}\right.$$
$$\left.-\frac{20\alpha}{N^7r^6t^2y^6}+\frac{15\alpha^2}{N^8r^6t^2y^6}-\frac{5\alpha^3}{N^9r^6t^2y^6}\right).$$

The largest term in this equation is of $O(N^{-4})$. Hence we can find a constant $c_t>0$ such that

$$\frac{kb}{\sqrt{tN^3}}y^{-6}+tN^3\,\mathbb{P}\left(\left|\frac{\tilde{S}_i^{(N)}(1)\sqrt{tN^3}}{\sqrt{\alpha t(1-\alpha/N)}}\right|>r\sqrt{tN^3}\,|y|\right)\leq\frac{c_t}{N\sqrt{N}}y^{-6}.$$

In conclusion,

$$\left|\mathbb{P}\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t(1-\alpha/N)}}<y\right)-\Phi(y)\right|\leq\frac{c_t}{N\sqrt{N}}y^{-6}$$

for a certain $c_t>0$. □

Proof of Lemma 3.3. We know that for i.i.d. $\mathcal{N}(0,1)$ distributed random variables $X_1, \ldots, X_N$, it holds that

$$b_N \left(\max_{i \leq N} X_i - b_N \right) \xrightarrow{d} G \text{ as } N \to \infty$$

with $G \sim$ Gumbel. However, in our situation we have a triangular array $\tilde{S}_i^{(N)}(tN^3) / \sqrt{\alpha t(1-\alpha/N)}$ which converges in distribution to a standard normal random variable. In [2], the maxima of these kind of triangular arrays is studied, and a proof is given (Proposition 2, p. 961) that under mild conditions, the maxima of these scaled triangular arrays also converge in distribution to Gumbel distributed random variables. Proposition 2 in [2] states that a triangular array $S_{n,i}$, with $i = 1, \ldots, n$ has to satisfy the following conditions in order to be in the domain of attraction of the Gumbel distribution:

1. $S_{n,i}$ has a zero mean and unit variance.
2. $S_{n,i}$ is a sum of k_n independent and identically distributed random summands whose moment generating function exists in an open interval containing the origin.
3. $\log n = o\left(k_n^{1/3}\right)$.

If these three conditions hold, then

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{1 \leq i \leq n} S_{n,i} \leq \alpha_n x + \beta_n \right) = e^{-e^{-x}}$$

with α_n and β_n satisfying

$$n\left(1 - \Phi\left(\beta_n\right)\right) \xrightarrow{n \to \infty} 1$$

and

$$n\left(1 - \Phi\left(\alpha_n x + \beta_n\right)\right) \xrightarrow{n \to \infty} e^{-x}.$$

For the triangular array $\tilde{S}_i^{(N)}(tN^3) / \sqrt{\alpha t(1-\alpha/N)}$, the first two conditions hold. Furthermore, $k_N = tN^3$. Hence, the third condition is also true. When we choose $\beta_N = b_N$ as in Definition 3.2 and $\alpha_N = 1/b_N$, the two limits above are satisfied. Consequently,

$$\begin{aligned} \lim_{N \to \infty} \mathbb{P}\left(\max_{i \leq N} \frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t(1-\alpha/N)}} \leq \frac{x}{b_N} + b_N \right) &= \lim_{N \to \infty} \mathbb{P}\left(b_N \left(\max_{i \leq N} \frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t(1-\alpha/N)}} - b_N \right) \leq x \right) \\ &= \lim_{N \to \infty} \mathbb{P}\big(M^{(N)}(t) \leq x\big) = e^{-e^{-x}}. \end{aligned}$$

Concluding,

$$M^{(N)}(t) \xrightarrow{d} G \text{ as } N \to \infty$$

with $G \sim$ Gumbel. □

Proof of Lemma 3.4. In Lemma 3.3, we proved that $M^{(N)}(t) \xrightarrow{d} G$ as $N \to \infty$ with $G \sim$ Gumbel. From this we know that for all $L > 0$

$$\mathbb{E}\big[M^{(N)}(t)^4 \mathbb{1}\left\{M^{(N)}(t)^4 < L\right\}\big] \xrightarrow{N \to \infty} \mathbb{E}\big[G^4 \mathbb{1}\left\{G^4 < L\right\}\big].$$

Therefore, in order to prove that

$$\mathbb{E}\big[M^{(N)}(t)^4\big] \xrightarrow{N \to \infty} \mathbb{E}\big[G^4\big],$$

we have to prove that $M^{(N)}(t)^4$ is uniformly integrable. We use inequality (3.3) proven in Lemma 3.2 to show that $M^{(N)}(t)^4$ is uniformly integrable. The process $M^{(N)}(t)$ satisfies

$$\begin{aligned}
&M^{(N)}(t)^4 \mathbb{1}\left\{M^{(N)}(t)^4 > L\right\} \\
=&\max_{i\le N}\left(\left(b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}}-b_N\right)\right)^4 \mathbb{1}\left\{b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}}-b_N\right) > L^{1/4}\right\}\right) \\
+&\min_{i\le N}\left(\left(b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}}-b_N\right)\right)^4 \mathbb{1}\left\{b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}}-b_N\right) < -L^{1/4}\right\}\right). \qquad \text{(B.3)}
\end{aligned}$$

To evaluate the first term in Equation (B.3) we first observe that

$$\begin{aligned}
&\mathbb{E}\left[\max_{i\le N}\left(\left(b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}}-b_N\right)\right)^4 \mathbb{1}\left\{b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}}-b_N\right) > L^{1/4}\right\}\right)\right] \\
=&L\,\mathbb{P}\left(\max_{i\le N} b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}}-b_N\right) > L^{1/4}\right) \\
+&\int_L^\infty \mathbb{P}\left(\max_{i\le N} b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}}-b_N\right) > s^{1/4}\right) ds. \qquad \text{(B.4)}
\end{aligned}$$

The first term in Equation (B.4) is small enough, because we know from Lemma 3.3 that the scaled maximal service process $M^{(N)}(t)$ converges in distribution to a Gumbel distributed random variable. Therefore,

$$L\,\mathbb{P}\left(\max_{i\le N} b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}}-b_N\right) > L^{1/4}\right) \overset{N\to\infty}{\longrightarrow} L\left(1-e^{-e^{-L^{1/4}}}\right) \overset{L\to\infty}{\longrightarrow} 0.$$

In order to analyze the second term in Equation (B.4) we perform the substitution

$$s^{1/4} \to (y-b_N)b_N.$$

Thus, we can substitute the integral term in Equation (B.4) to

$$\begin{aligned}
&\int_{\frac{L^{1/4}}{b_N}+b_N}^\infty 4b_N^4\,(y-b_N)^3\,\mathbb{P}\left(\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}} > y\right) dy \\
=&\int_{\frac{L^{1/4}}{b_N}+b_N}^\infty 4b_N^4\,(y-b_N)^3\left(1-\mathbb{P}\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}} < y\right)^N\right) dy \\
=&\int_{\frac{L^{1/4}}{b_N}+b_N}^\infty 4b_N^4\,(y-b_N)^3\left(1-\left(1-\mathbb{P}\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}} > y\right)\right)^N\right) dy. \qquad \text{(B.5)}
\end{aligned}$$

Now, we can use the inequality in (3.3) to find an upper bound for Equation (B.5).

$$\begin{aligned}
&\int_{\frac{L^{1/4}}{b_N}+b_N}^\infty 4b_N^4\,(y-b_N)^3\left(1-\left(1-\mathbb{P}\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}} > y\right)\right)^N\right) dy \\
\le&\int_{\frac{L^{1/4}}{b_N}+b_N}^\infty 4b_N^4\,(y-b_N)^3\left(1-\left(\Phi(y)-\frac{c_t}{N\sqrt{N}}y^{-6}\right)^N\right) dy.
\end{aligned}$$

We have for all y that

$$\left(\Phi(y)-\frac{c_t}{N\sqrt{N}}y^{-6}\right)^N \geq \Phi(y)^N+1-\left(1+\frac{c_t}{N\sqrt{N}}y^{-6}\right)^N.$$

Therefore,

$$\begin{aligned}&\int_{\frac{L^{1/4}}{b_N}+b_N}^{\infty} 4b_N^4\left(y-b_N\right)^3\left(1-\left(\Phi(y)-\frac{c_t}{N\sqrt{N}}y^{-6}\right)^N\right)dy\\ \leq&\int_{\frac{L^{1/4}}{b_N}+b_N}^{\infty} 4b_N^4\left(y-b_N\right)^3\left(1-\Phi(y)^N-1+\left(1+\frac{c_t}{N\sqrt{N}}y^{-6}\right)^N\right)dy.\end{aligned} \tag{B.6}$$

Splitting the integral in Equation (B.6) into two terms and considering their limits as $N\to\infty$, we have that

$$\int_{\frac{L^{1/4}}{b_N}+b_N}^{\infty} 4b_N^4\left(y-b_N\right)^3\left(-1+\left(1+\frac{c_t}{N\sqrt{N}}y^{-6}\right)^N\right)dy \overset{N\to\infty}{\longrightarrow} 0.$$

We can see this convergence heuristically by observing that $b_N\sim\sqrt{\log N}$,

$$-1+\left(1+\frac{c_t}{N\sqrt{N}}y^{-6}\right)^N \leq \frac{1}{\sqrt{N}}\left(-1+\left(1+\frac{c_t}{N}y^{-6}\right)^N\right),$$

and that the integral

$$\int_1^{\infty} y^3\left(\exp\left(c_t y^{-6}\right)-1\right)dy<\infty.$$

Furthermore, the second term in which the integral in Equation (B.6) is split yields

$$\lim_{L\to\infty}\limsup_{N\to\infty}\int_{\frac{L^{1/4}}{b_N}+b_N}^{\infty} 4b_N^4\left(y-b_N\right)^3\left(1-\Phi(y)^N\right)dy=0,$$

because the fourth moment of the maximum of standard normal random variables converges to the fourth moment of a Gumbel random variable, by Pickands' theorem in [25]. Concluding, the first term in (B.3) is uniformly integrable. For the second term in (B.3) we have

$$\begin{aligned}&\mathbb{E}\left[\min_{i\leq N}\left(\left(b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\left(1-\alpha/N\right)}}-b_N\right)\right)^4\mathbb{1}\left\{b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\left(1-\alpha/N\right)}}-b_N\right)<-L^{1/4}\right\}\right)\right]\\ =&L\,\mathbb{P}\left(\max_{i\leq N} b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\left(1-\alpha/N\right)}}-b_N\right)<-L^{1/4}\right)\\ +&\int_L^{\infty}\mathbb{P}\left(\max_{i\leq N} b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\left(1-\alpha/N\right)}}-b_N\right)<-s^{1/4}\right)ds.\end{aligned} \tag{B.7}$$

We again use the inequality given in (3.3) to evaluate this equation, and we get for $\epsilon>0$

$$\mathbb{E}\left[\min_{i\leq N}\left(\left(b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\left(1-\alpha/N\right)}}-b_N\right)\right)^4\mathbb{1}\left\{b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\left(1-\alpha/N\right)}}-b_N\right)<-L^{1/4}\right\}\right)\right]$$

$$
\begin{aligned}
&\leq L\,\mathbb{P}\left(\max_{i\leq N} b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}}-b_N\right)<-L^{1/4}\right)\\
&+\int_{-\infty}^{-\epsilon}-4b_N^4\,(y-b_N)^3\left(\Phi(y)+\frac{c_t}{N\sqrt{N}}y^{-6}\right)^N dy\\
&+\int_{\epsilon}^{\frac{-L^{1/4}}{b_N}+b_N}-4b_N^4\,(y-b_N)^3\left(\Phi(y)+\frac{c_t}{N\sqrt{N}}y^{-6}\right)^N dy\\
&+\int_{-\epsilon}^{\epsilon}-4b_N^4\,(y-b_N)^3\,\mathbb{P}\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}}<y\right)^N dy. \qquad \text{(B.8)}
\end{aligned}
$$

By using the result from Lemma 3.3 that $M^{(N)}(t)$ converges in distribution to a Gumbel distributed random variable, we know that the first term in Equation (B.8) remains small, because

$$
L\,\mathbb{P}\left(\max_{i\leq N} b_N\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}}-b_N\right)<-L^{1/4}\right)\overset{N\to\infty}{\longrightarrow} Le^{-e^{L^{1/4}}}\overset{L\to\infty}{\longrightarrow}0.
$$

Furthermore, since $\Phi(y)<1$ for all y, we know that

$$
\left(\Phi(y)+\frac{c_t}{N\sqrt{N}}y^{-6}\right)^N\leq\Phi(y)^N+\left(\frac{c_t}{N\sqrt{N}}y^{-6}+1\right)^N-1.
$$

We have for $\epsilon>0$ that

$$
\int_{-\infty}^{-\epsilon}-4b_N^4\,(y-b_N)^3\left(\left(\frac{c_t}{N\sqrt{N}}y^{-6}+1\right)^N-1\right)dy\overset{N\to\infty}{\longrightarrow}0,
$$

and

$$
\int_{\epsilon}^{\frac{-L^{1/4}}{b_N}+b_N}-4b_N^4\,(y-b_N)^3\left(\left(\frac{c_t}{N\sqrt{N}}y^{-6}+1\right)^N-1\right)dy\overset{N\to\infty}{\longrightarrow}0.
$$

Since Φ is continuous, we have uniform convergence in distribution of $\tilde{S}_i^{(N)}(tN^3)/\sqrt{\alpha t\,(1-\alpha/N)}$ to a normally distributed random variable. Therefore,

$$
\int_{-\epsilon}^{\epsilon}-4b_N^4\,(y-b_N)^3\,\mathbb{P}\left(\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}}<y\right)^N dy\overset{N\to\infty}{\longrightarrow}0.
$$

Moreover,

$$
\lim_{L\to\infty}\limsup_{N\to\infty}\int_{-\infty}^{\frac{-L^{1/4}}{b_N}+b_N}-4b_N^4\,(y-b_N)^3\,\Phi(y)^N dy=0.
$$

Concluding, $M^{(N)}(t)$ is uniformly integrable and $\mathbb{E}\left[M^{(N)}(t)^4\right]\overset{N\to\infty}{\longrightarrow}\mathbb{E}[G^4]$. □

Proof of Lemma 3.5. From Definition 3.2 and Lemma 3.4, we know that

$$
\left|\mathbb{E}\left[\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)}}\right]-b_N\right|=\left|\frac{\mathbb{E}\left[M^{(N)}(t)\right]}{b_N}\right|\leq\sqrt[4]{\frac{\mathbb{E}[M^{(N)}(t)^4]}{b_N^4}}\overset{N\to\infty}{\longrightarrow}0.
$$

Since

$$b_N = \sqrt{2\log N} - \frac{\log(4\pi\log N)}{2\sqrt{2\log N}},$$

we can conclude that

$$\mathbb{E}\left[\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\log N}}\right] \overset{N\to\infty}{\longrightarrow} \sqrt{2\alpha t}.$$

Moreover, to prove (3.4) for $k=2$ we use the upper bound

$$\begin{aligned}
&\left|\mathbb{E}\left[\left(\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}}\right)^2 - 2\right]\right| \\
\le\ &\mathbb{E}\left[\left|\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}} + \sqrt{2}\right|\left|\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}} - \sqrt{2}\right|\right] \\
\le\ &\mathbb{E}\left[\left|\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}} - \mathbb{E}\left[\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}}\right]\right|\left|\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}} - \sqrt{2}\right|\right] \\
+\ &\left|\mathbb{E}\left[\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}}\right] + \sqrt{2}\right|\mathbb{E}\left[\left|\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}} - \sqrt{2}\right|\right].
\end{aligned}$$

We already know from Lemma 3.4 that

$$\begin{aligned}
&\left|\mathbb{E}\left[\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}}\right] + \sqrt{2}\right|\mathbb{E}\left[\left|\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}} - \sqrt{2}\right|\right] \\
\le\ &\left|\mathbb{E}\left[\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}}\right] + \sqrt{2}\right|\sqrt[4]{\mathbb{E}\left[\left(\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}} - \sqrt{2}\right)^4\right]} \overset{N\to\infty}{\longrightarrow} 0.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\mathbb{E}\left[\left|\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}} - \mathbb{E}\left[\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}}\right]\right|\left|\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}} - \sqrt{2}\right|\right] \\
\le\ &\mathbb{E}\left[\left|\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}} - \mathbb{E}\left[\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}}\right]\right|^2\right] \\
+\ &\mathbb{E}\left[\left|\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}} - \sqrt{2}\right|^2\right]. \qquad \text{(B.9)}
\end{aligned}$$

By Lemma 3.4, we also can conclude that the second term in Equation (B.9) remains small. For the first term in Equation (B.9), we get

$$\begin{aligned}
&\mathbb{E}\left[\left|\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}} - \mathbb{E}\left[\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}}\right]\right|^2\right] \\
\le\ &\left(\sqrt{\mathbb{E}\left[\left|\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}} - \sqrt{2}\right|^2\right]} + \sqrt{\left|\sqrt{2} - \mathbb{E}\left[\max_{i\le N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\alpha t\,(1-\alpha/N)\sqrt{\log N}}}\right]\right|^2}\right)^2 \\
\overset{N\to\infty}{\longrightarrow}\ &0.
\end{aligned}$$

Concluding,

$$\mathbb{E}\left[\left(\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\log N}}\right)^2\right]\overset{N\to\infty}{\longrightarrow}2\alpha t.$$

We can use the same argument together with Lemma 3.4 to prove that

$$\mathbb{E}\left[\left(\max_{i\leq N}\frac{\tilde{S}_i^{(N)}(tN^3)}{\sqrt{\log N}}\right)^4\right]\overset{N\to\infty}{\longrightarrow}4\alpha^2t^2.$$

□

Acknowledgments. This research is supported by the Netherlands Organisation for Scientific Research through the programmes Grip on Complexity [Schol: 438.16.121], MEERVOUD [Vlasiou: 632.003.002], and Talent VICI [Zwart: 639.033.413]

References

[1] Abate J, Whitt W (1987) Transient behavior of regulated Brownian motion, I: starting at the origin. *Advances in Applied Probability* 19(3):560–598.

[2] Anderson CW, Coles SG, Hüsler J, et al. (1997) Maxima of Poisson-like variables and related triangular arrays. *The Annals of Applied Probability* 7(4):953–971.

[3] ASML (2018) ASML integrated report 2018. Accessed October 28, 2019, https://www.asml.com/en/investors/financial-results/annual-reports.

[4] Asmussen S (2008) *Applied probability and queues*, volume 51 (Springer Science & Business Media).

[5] Atar R, Mandelbaum A, Zviran A (2012) Control of fork-join networks in heavy traffic. *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 823–830 (IEEE).

[6] Baccelli F (1985) Two parallel queues created by arrivals with two demands: The M/G/2 symmetrical case. *RR-0426, INRIA. ffinria-00076130* .

[7] Baccelli F, Makowski AM (1989) Queueing models for systems with synchronization constraints. *Proceedings of the IEEE* 77(1):138–161.

[8] Baccelli F, Makowski AM, Shwartz A (1989) The fork-join queue and related systems with synchronization constraints: Stochastic ordering and computable bounds. *Advances in Applied Probability* 21(3):629–660.

[9] Billingsley P (1968) *Convergence of probability measures* (John Wiley & Sons).

[10] Business Continuity Institute (2011) Supply chain resilience, 3rd annual survey. Accessed October 22, 2019, https://www.cips.org/Documents/Resources/Knowledge%20Summary/BCI%20Supply%20Chain%20Resilience%202011%20Public%20Version.pdf.

[11] Downey PJ (1995) Bounds and approximations for overheads in the time to join parallel forks. *ORSA Journal on Computing* 7(2):125–139.

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

[13] de Klein SJ (1988) *Fredholm integral equations in queueing analysis*. Ph.D. thesis, Rijksuniversiteit Utrecht.

[14] Ko SS, Serfozo RF (2004) Response times in M/M/s fork-join networks. *Advances in Applied Probability* 36(3):854–871.

[15] Lu H, Pang G (2015) Gaussian limits for a fork-join network with nonexchangeable synchronization in heavy traffic. *Mathematics of Operations Research* 41(2):560–595.

[16] Lu H, Pang G (2017) Heavy-traffic limits for a fork-join network in the Halfin-Whitt regime. *Stochastic Systems* 6(2):519–600.

[17] Lu H, Pang G (2017) Heavy-traffic limits for an infinite-server fork–join queueing system with dependent and disruptive services. *Queueing Systems* 85(1-2):67–115.

[18] Luczak MJ, McDiarmid C, et al. (2006) On the maximum queue length in the supermarket model. *The Annals of Probability* 34(2):493–527.

[19] Michel R (1976) Nonuniform central limit bounds with applications to probabilities of deviations. *The Annals of Probability* 102–106.

[20] Nelson R, Tantawi AN (1988) Approximate analysis of fork/join synchronization in parallel queues. *IEEE transactions on computers* 37(6):739–743.

[21] Nguyen V (1993) Processing networks with parallel and sequential tasks: Heavy traffic analysis and Brownian limits. *The Annals of Applied Probability* 28–55.

[22] Nguyen V (1994) The trouble with diversity: Fork-join networks with heterogeneous customer population. *The Annals of Applied Probability* 1–25.

[23] Norrman A, Jansson U (2004) Ericsson's proactive supply chain risk management approach after a serious sub-supplier accident. *International journal of physical distribution & logistics management* 34(5):434–456.

[24] Parker CF (2015) Complex negative events and the diffusion of crisis: lessons from the 2010 and 2011 Icelandic volcanic ash cloud events. *Geografiska Annaler: Series A, Physical Geography* 97(1):97–108.

[25] Pickands III J (1968) Moment convergence of sample extremes. *The Annals of Mathematical Statistics* 39(3):881–889.

[26] Sigman K, Whitt W (2011) Heavy-traffic limits for nearly deterministic queues. *Journal of Applied Probability* 48(3):657–678.

[27] Sigman K, Whitt W (2011) Heavy-traffic limits for nearly deterministic queues: stationary distributions. *Queueing Systems* 69(2):145.

[28] Stecke KE, Kumar S (2009) Sources of supply chain disruptions, factors that breed vulnerability, and mitigating strategies. *Journal of Marketing Channels* 16(3):193–226.

[29] Varma S (1990) *Heavy and light traffic approximations for queues with synchronization constraints.* Ph.D. thesis, University of Maryland.

[30] Wright PE (1992) Two parallel processors with coupled inputs. *Advances in applied probability* 24(4):986–1007.