

Large fork-join queues with nearly deterministic arrival and service times

Dennis Schol

Eindhoven University of Technology

Maria Vlasiou

Eindhoven University of Technology, University of Twente

Bert Zwart

Eindhoven University of Technology, CWI

In this paper, we study an N server fork-join queue with nearly deterministic arrival and service times. Specifically, we present a fluid limit for the maximum queue length as $N \rightarrow \infty$. This fluid limit depends on the initial number of tasks. In order to prove these results, we develop extreme value theory and diffusion approximations for the queue lengths.

Key words: queueing network; heavy traffic; fluid limit; extreme value theory

MSC2000 subject classification: 60K25

OR/MS subject classification: Primary: Queues: networks; limit theorems; secondary: probability: Markov processes; random walk; stochastic model applications

History: Received December 24, 2019; revised December 15, 2020.

1. Introduction. Fork-join queues are widely studied in many applications, such as communication systems and production processes. However, due to the fact that all service stations see exactly the same arrival process, which is the main characteristic of fork-join queues, these fork-join queues are very challenging to analyze. Hence, there are only a few exact results, which are mainly for systems in stationarity and are restricted to fork-join queues with two service stations.

In this paper, we focus on a fork-join queue where the number of service stations is large. Our objective is to analyze the queue length of the longest queue. We explore a discrete-time fork-join queue where the arrival and service times are nearly deterministic. In addition, we consider a heavily loaded system. That is, we assume that the arrival rate to a queue times the expected service time of that queue, i.e. the traffic intensity per queue ρ_N , depends on the number of service stations N and satisfies $(1 - \rho_N)N^2 \xrightarrow{N \rightarrow \infty} \beta$, with $\beta > 0$. Our main result is a fluid limit of the maximum queue length of the system as N goes to infinity, which holds under very mild conditions on the distribution of the number of jobs at time 0.

Both the model and the scaling studied in this paper are inspired by assembly systems. In particular, we are inspired by problems faced by original equipment manufacturers (OEMs) that assemble thousands of components, each produced using specialized equipment, into complex systems. Examples of such OEMs are Airbus and ASML. If one component is missing, the final product cannot be assembled, giving rise to costly delays. In reality, for some components, OEMs may hedge the shortage risk by investing in capacity or by keeping an inventory of finished components. However, we study the maximum queue length, which is only relevant for components where there is no inventory. As such, our model is a somewhat stylized model of reality.

An interesting question is whether the manufacturer can produce on schedule. To answer this question, we consider a make-to-order system, i.e. suppliers only produce when they have an order,

and we assume that the manufacturer sends orders to all the suppliers at the same time. Now, we can model this process by a fork-join queueing system, where the various servers represent suppliers, jobs in the system represent orders requested by the manufacturer and queue lengths in front of each server represent the number of unfinished components each supplier has. As the slowest supplier determines the delay that the manufacturer observes, we wish to study the longest queue. Additionally, we consider a supply chain network operating under full capacity, which is indeed the situation in this industry. Last, we capture the property that in high-tech manufacturing arrival and service times have a low variance by considering nearly deterministic arrival and service times. A visualization of the fork-join queue as a simple representation of a high-tech supply chain, is given in Figure 1. Note that in this paper we focus on the backlogs of the suppliers and not on the assembly phase.

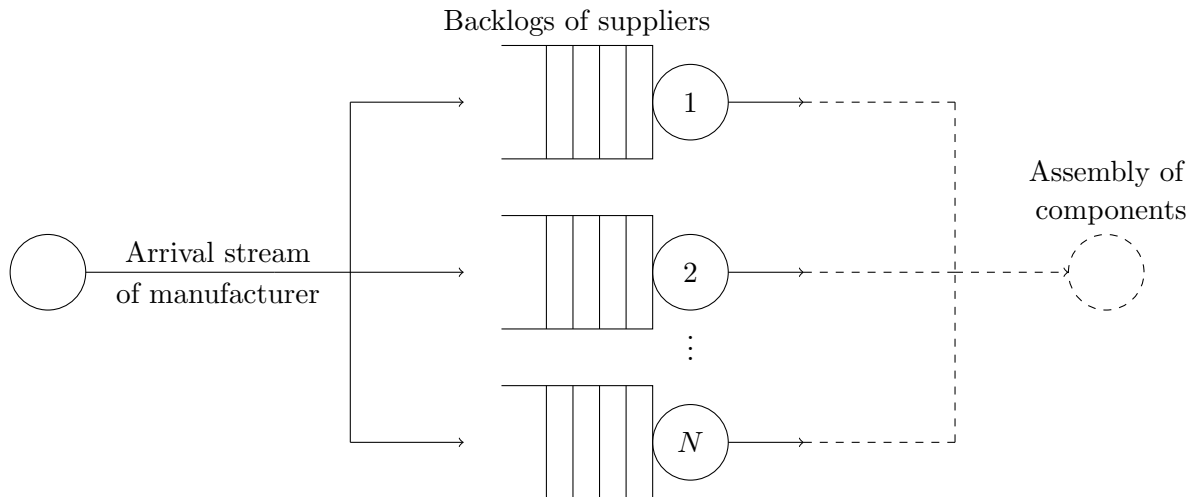


FIGURE 1. Fork-join queue with N servers

We now turn to a survey of related literature. As mentioned, the earliest literature on fork-join queues focuses on systems with two service stations. Analytic results, such as asymptotics on limiting distributions, can be found in [3, 9, 11, 24]. However, due to the complexity of fork-join queues, these results cannot be expanded to fork-join queues with more than two service stations. Thus, most of the work on fork-join queues with more than two service stations is focused on finding approximations of performance measures. For example, an approximation of the distribution of the response time in M/M/s fork-join queues is given in Ko and Serfoso [12]. Upper and lower bounds for the mean response time of servers, and other performance measures, are given by Nelson, Tantawi [17] and Baccelli, Makowski [4].

A common property of the aforementioned classic literature is that it mainly focuses on steady-state distributions or other one-dimensional performance measures. Some work on the heavy-traffic process limit has been done, for example, Varma [23] derives a heavy-traffic analysis for fork-join queues, and shows weak convergence of several processes, such as the joint queue lengths in front of each server. Furthermore, Nguyen [18] proves that various appearing limiting processes are in fact multi-dimensional reflected Brownian motions. In [19], Nguyen extends this result to a fork-join queue with multiple job types. Lu and Pang study fork-join networks in [13, 14, 15]. In [13], they investigate a fork-join network where each service station has multiple servers under non-exchangeable synchronization, and operates in the quality-driven regime. They derive functional central limit theorems for the number of tasks waiting in the waiting buffers for synchronization and for the number of synchronized jobs. In [14], they extend this analysis to a fork-join network

with a fixed number of service stations, each having many servers, where the system operates in the Halfin-Whitt regime. In [15], the authors investigate these heavy-traffic limits for a fixed number of infinite-server stations, where services are dependent and could be disrupted. Finally, we mention Atar, Mandelbaum and Zviran [2], who investigate the control of a fork-join queue in heavy traffic by using feedback procedures. Our work contributes to this literature on process-level analysis of fork-join networks. To be precise, we derive a fluid limit of the stochastic process that keeps track of the largest queue length. This study seems to be the first explicit process-level approximation of a large fork-join queue.

Moreover, our work also adds to the literature on queueing systems with nearly deterministic arrivals and services. The only research line on queueing systems with nearly deterministic service times that we are aware of is Sigman and Whitt [21, 22], who investigate the G/G/1 and G/D/N queues and establish heavy-traffic results on waiting times, queue lengths and other performance measures in stationarity, as well as functional central limit theorems on the waiting time and on other performance measures. In these papers, they distinguish two cases, one in which $(1 - \rho_N)\sqrt{N} \xrightarrow{N \rightarrow \infty} \beta$ and one in which $(1 - \rho_N)N \xrightarrow{N \rightarrow \infty} \beta$, with ρ_N the traffic intensity and β some constant.

We now turn to an overview of the techniques that we use in this paper. Because of the fact that we aim to obtain a fluid limit of a maximum of N queue lengths, we mainly use techniques from extreme value theory in our proofs. This is, however, quite a challenge, since on the hand, the queue lengths of the servers are mutually dependent. On the other hand, most results on extreme values hinge heavily on the assumption of mutual independence. Furthermore, we consider a fork-join queue where the arrival and service probability depend on N , which makes the queue lengths to be triangular arrays with respect to N . This makes our paper also rather unusual, as studies on triangular arrays are rare. One paper on this subject, relevant for us, is Anderson, Coles and Hüsler [1], where they study the maximum of a sum of a large number of triangular arrays.

In order to get fluid limits for the maximum queue lengths, we need to study diffusion limits for the individual queue lengths. We thus combine ideas from the literature on extreme value theory with literature on diffusion approximations, which we show in Section 2.2. In order to be able to analyze the queue lengths through diffusion approximations, we impose a heavy-traffic assumption, namely $(1 - \rho_N)N^2 \rightarrow \beta$. Then, for each separate queue length, we have a reflected Brownian motion as diffusion approximation. By using the well-known formula for the cumulative distribution function of a reflected Brownian motion (cf. Harrison [10, p. 49]), we investigate the maximum of N independent reflected Brownian motions to get an idea of the scaling of the maximum queue length.

Now, we give a brief sketch of how we apply these ideas to prove the fluid limit, we start by considering the slightly simpler scenario that each queue is empty at time 0. Because we want to prove a fluid limit that holds uniformly on compact intervals, we need to prove pointwise convergence of the process and tightness of the collection of processes. Our first step in proving this is by showing that each queue length is in distribution the same as a supremum of an arrival process minus a service process. We then show in Section 2.2 that under a temporal scaling of $tN^3 \log N$ and a spatial scaling of $N \log N$, the arrival process minus a drift term converges to $-\beta t$, as $N \rightarrow \infty$. Furthermore, we derive under that same temporal scaling but under a spatial scaling of $N\sqrt{\log N}$, that the centralized service process satisfies the central limit theorem. This scaled centralized service process is given in Equation (4.4). We use the non-uniform Berry-Esséen inequality, which is described by Michel in [16], to deduce the convergence rate of the cumulative distribution function of this scaled centralized service process to the cumulative distribution function of a normally distributed random variable, which is given in Equation (4.8). It turns out that this convergence rate is fast enough, so that we can replace the scaled centralized service process with a normally distributed random variable in the expression of the maximum queue length in order to get the

same limit. By Pickands' result [20] on convergence of moments of the maximum of N scaled random variables, we know that the expectation of the maximum of standard normally distributed random variables divided by $\sqrt{\log N}$ converges to $\sqrt{2}$, as $N \rightarrow \infty$. This gives us the convergence of the maximum of N scaled centralized service processes. After we have obtained these limiting results for the scaled arrival and service process, we use these, together with Doob's maximal submartingale inequality to prove convergence in probability of the maximum queue length, we show this in Section 4.3. Finally, in Section 4.4 we use Doob's maximal submartingale inequality to bound the probability that the process makes large jumps and prove that this probability is small, so that the maximum queue length is a tight process.

After we have considered the maximum queue length for the process with empty queues at time 0, we then turn to the scenario that the length of each queue at time 0 is identically distributed. In this case, we can use Lindley's recursion to express the maximum queue length as the pairwise maximum of the maximum queue length with empty queues at time 0 and a part depending on the number of jobs at time 0, this formula is given in Equation (2.3). How to prove the fluid limit for the first part is already sketched above. In order to derive a fluid limit for the latter part, we first observe that this part equals the maximum of N times the sum of the number of jobs at time 0 at each server plus the number of arrivals minus the number of services at each server. Following a similar path as earlier, we can prove that the scaled centralized service process at server i behaves like a normally distributed random variable. Thus, we have to analyze a maximum of N pairwise sums of normally distributed random variables and random variables describing the number of jobs at time 0, which is stated in more detail in Lemma 4.9.

In Lemma 4.4 we prove a convergence result of this maximum, this is quite a challenge, because we need to apply extreme value theory on pairwise sums. In order to do this, we use the results from Davis, Mulrow & Resnick [7] and Fisher [8] on convergence of samples of random variables to limiting sets. The authors prove convergence results of the convex hull of $\{(Z_i^{(1)}/b_N, \dots, Z_i^{(k)}/b_N)_{i \leq N}\}$ to a limiting set, as $N \rightarrow \infty$, with $(Z_i^{(j)}, i \leq N)$ i.i.d., $Z_i^{(j)}$ and $Z_m^{(l)}$ are independent and b_N is a proper scaling sequence. We show in the proof of Lemma B.1 that these results can be extended in establishing convergence of extreme values of $\max_{i \leq N} \sum_{j=1}^k Y_i^{(j)}/a_N^{(j)}$, where $a_N^{(l)}$ and $a_N^{(m)}$ are not necessarily the same, which is a stand-alone result of independent interest. We did not find this extension in other literature. The result in Lemma 4.4 follows from Lemma B.1.

The rest of the paper is organized as follows. In Section 2, we describe the fork-join system in more detail; we give a definition of the arrival and service processes and we present a scaled version of the queueing model. In Section 2.1, we introduce the fluid limit and explain it heuristically. We elaborate a bit more on the scaling and the shape of the fluid limit in Sections 2.2 and 2.3. Furthermore, we give some examples and numerical results in Section 2.4. We finish with some concluding remarks in Section 3. The proof of the fluid limit is given in Section 4. In Appendix A, we elaborate on the convergence of the upper bound that was given in Lemma 4.7. We prove in Appendix B a convergence result of $\max_{i \leq N} \sum_{j=1}^k Y_i^{(j)}/a_N^{(j)}$. In Appendix C, we prove the lemmas stated in Section 4.2. An overview of all notation is given in Appendix D.

2. Model description and main results. We now turn to a formal definition of the fork-join queue that we study. We consider a fork-join queue with integer valued arrivals and 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

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

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,

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

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

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

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 .

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

Both $p^{(N)}$ and $q^{(N)}$ are taken as functions of N , which we specify in Definition 2.3 below.

We assume that for all $N \geq 1$ the random variables $(X^{(N)}(j), j \geq 1)$ are mutually independent for all j and $(Y_i^{(N)}(j), j \geq 1, i \leq N)$ are mutually 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. Finally, we assume that $X^{(N)}(j)$ and $Y_i^{(N)}(j)$ are independent, in other words, $Y_i^{(N)}(j)$ could still be 1 while there are no tasks to be served at server i at time j . Due to this assumption, we have on the hand the beneficial situation that $(A^{(N)}(n), n \geq 0)$ and $(S_i^{(N)}(n), n \geq 0)$ are independent processes, but on the other hand we should be careful with defining the queue length. However, it is a well known result that we can use Lindley's recursion, and write the queue length of the i -th server at time n as

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

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

$$\sup_{0 \leq k \leq n} \left(A^{(N)}(k) - S_i^{(N)}(k) \right).$$

As can be seen in this expression, the queue lengths of different servers are mutually dependent, since the arrival process is the same. When at time 0 there are already jobs in queue, then we can, after again applying Lindley's recursion, write the queue length of the i -th server at time n as

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

with $Q_i^{(N)}(0)$ the number of jobs in front of the i -th server at time 0. Observe that the queue length of the i -th server equals the maximum of the queue length when the number of jobs at time 0 would be 0, and a random variable that depends on the initial number of jobs.

The aim of this work is to investigate the behavior of the fork-join queue 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. Therefore, we define in Definition 2.3 a random variable indicating the maximum queue length at time n . 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.

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 , with $Q_{(\alpha, \beta)}^{(N)}(0) = 0$. Then

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

So,

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

under the assumption that $Q_{(\alpha, \beta)}^{(N)}(0) = 0$. From these choices of $p^{(N)}$ and $q^{(N)}$, it follows that the traffic intensity ρ_N of a single queue satisfies $(1 - \rho_N)N^2 \rightarrow \beta$, as $N \rightarrow \infty$. Furthermore, if $Q_i^{(N)}(0) > 0$, the maximum queue length at time n can be written as

$$Q_{(\alpha, \beta)}^{(N)}(n) = \max_{i \leq N} \max \left(\sup_{0 \leq k \leq n} \left[\left(A^{(N)}(n) - A^{(N)}(k) \right) - \left(S_i^{(N)}(n) - S_i^{(N)}(k) \right) \right], Q_i^{(N)}(0) + A^{(N)}(n) - S_i^{(N)}(n) \right). \quad (2.3)$$

Observe that we can interchange the order of the $\max_{i \leq N}$ term and the \max term, and rewrite the expression in (2.3) as the pairwise maximum of two random variables, one random variable is the maximum of N queue lengths with initial condition 0, as given in Equation (2.1), and the other is the maximum of N sums of the queue length at time 0 plus the number of arrivals minus the number of services.

2.1. Fluid limit. As we just have formally defined the fork-join queue that we study, with the particular nearly deterministic setting, we now state and explain the main result of this paper. Our central result is a fluid approximation for the rescaled maximum 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 .

There is, however, not a straightforward procedure in choosing the temporal and spatial scaling, there are namely more possibilities that lead to a non-trivial limit. For instance, when we choose a temporal scaling of N^3 and a spatial scaling of $N\sqrt{\log N}$, we get the fluid limit that is given in Proposition 2.1. Here, we assume that the initial condition is 0.

PROPOSITION 2.1 (Temporal scaling of N^3 and spatial scaling of $N\sqrt{\log N}$). For $Q_{(\alpha, \beta)}^{(N)}(0) = 0$, $\alpha > 0$ and $\beta > 0$, with

1. $p^{(N)} = 1 - \alpha/N - \beta/N^2$,
2. $q^{(N)} = 1 - \alpha/N$, we have

$$\mathbb{P} \left(\sup_{0 \leq s \leq T} \left| \frac{Q_{(\alpha, \beta)}^{(N)}(sN^3)}{N\sqrt{\log N}} - \sqrt{2\alpha s} \right| > \epsilon \right) \xrightarrow{N \rightarrow \infty} 0 \quad \forall \epsilon > 0.$$

However, we can also derive a steady-state limit, which is given in Proposition 2.2.

PROPOSITION 2.2 (Steady-state convergence). *For $\alpha > 0$ and $\beta > 0$, with*

1. $p^{(N)} = 1 - \alpha/N - \beta/N^2$,
2. $q^{(N)} = 1 - \alpha/N$, *we have*

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

As we can see in Proposition 2.2, to obtain a non-trivial steady-state limit, we need a spatial scaling of $N \log N$. Since this is the only choice which leads to a non-trivial limit, it is a natural choice to look for a fluid limit which also has this spatial scaling. Our main result, stated in Theorem 2.1, is such a fluid limit, and it turns out that for establishing this limit, we need a temporal scaling of $N^3 \log N$. In Section 2.2 we explain why these scalings are natural. We omit the proof of Proposition 2.1, but we do explain how Proposition 2.1 is connected to Theorem 2.1 at the end of this section. Furthermore, we give a proof of Proposition 2.2 in Section 4.

We now mention and discuss some assumptions under which our main result holds. First of all, we assume that we have nearly deterministic arrivals and services.

ASSUMPTION 2.1. $p^{(N)} = 1 - \alpha/N - \beta/N^2$ and $q^{(N)} = 1 - \alpha/N$, with $\alpha, \beta > 0$.

Secondly, we have a basic assumption on the initial condition.

ASSUMPTION 2.2. $(Q_i^{(N)}(0), i \leq N)$ are *i.i.d. and non-negative for all N .*

Furthermore, we want to prove a fluid limit with a spatial scaling of $N \log N$. Therefore, we need to assume that the maximum number of jobs at time 0 also scales with $N \log N$. In order to do so, we allow $(Q_i^{(N)}(0), i \leq N, N \geq 1)$ to be a triangular array, i.e. a doubly indexed sequence with $i \leq N$. This is a necessity, because otherwise we would be limited to distributions where the maximum scales like $N \log N$, which would lead us to the family of the heavy-tailed distributions for which we do not have convergence in probability of its maximum. Thus in our setting, $Q_i^{(N)}(0)$ and $Q_i^{(N+1)}(0)$ do not need to be the same. Consequently, we need to have some regularity on $Q_i^{(N)}(0)$ as N increases to be able to prove a limit theorem.

ASSUMPTION 2.3. $Q_{(\alpha, \beta)}^{(N)}(0)/(N \log N) \xrightarrow{\mathbb{P}} q(0)$, with $q(0) \geq 0$, as $N \rightarrow \infty$, with $Q_i^{(N)}(0) = \lfloor r_N U_i \rfloor$, where r_N is a scaling sequence.

Finally, we can distinguish two cases in which Theorem 2.1 holds.

ASSUMPTION 2.4. U_i has a finite right endpoint.

ASSUMPTION 2.5. U_i is a continuous random variable and for all $v \in [0, 1]$,

$$\lim_{t \rightarrow \infty} \frac{-\log(\mathbb{P}(U_i > vt))}{-\log(\mathbb{P}(U_i > t))} = h(v).$$

Before stating the theorem, we would like to give two remarks on Assumption 2.5. First of all, the function h has the property that for all $u, v \in [0, 1]$, $h(uv) = h(u)h(v)$. Thus, if h is continuous, $h(v) = v^a$, with $a > 0$. When h is discontinuous, there are two possibilities: $h(v) = \mathbb{1}(v > 0)$, or $h(v) = \mathbb{1}(v = 1)$, this corresponds to $h(v) = v^a$ with $a = 0$ and $a = \infty$, respectively. Secondly, the assumption of continuity of U_i can be removed, which would lead to more cumbersome proofs.

THEOREM 2.1 (Fluid limit with non-zero initial condition). *If Assumptions 2.1, 2.2 and 2.3 hold, and either Assumption 2.4 or Assumption 2.5 holds, then we have $\forall T > 0$, that*

$$\mathbb{P}\left(\sup_{0 \leq t \leq T} \left| \frac{Q_{(\alpha, \beta)}^{(N)}(tN^3 \log N)}{N \log N} - q(t) \right| > \epsilon\right) \xrightarrow{N \rightarrow \infty} 0 \quad \forall \epsilon > 0, \quad (2.4)$$

with

$$q(t) = \max\left(\left(\sqrt{2\alpha t} - \beta t\right) \mathbb{1}\left(t < \frac{\alpha}{2\beta^2}\right) + \frac{\alpha}{2\beta} \mathbb{1}\left(t \geq \frac{\alpha}{2\beta^2}\right), g(t, q(0)) - \beta t\right). \quad (2.5)$$

The function $g(t, q(0))$ has the following properties:

1. If Assumption 2.4 holds, then

$$g(t, q(0)) = q(0) + \sqrt{2\alpha t}. \quad (2.6)$$

2. If Assumption 2.5 holds, then

$$g(t, q(0)) = \sup_{(u, v)} \{\sqrt{2\alpha t} u + q(0)v | u^2 + h(v) \leq 1, 0 \leq u \leq 1, 0 \leq v \leq 1\}. \quad (2.7)$$

There is a connection between Assumptions 2.4 and 2.5 on U_i and extreme value theory. If Assumption 2.4 holds, then this means that U_i is either a degenerate random variable or is in the domain of attraction of the Weibull distribution. On the other hand, if Assumption 2.5 holds, then U_i is in the domain of attraction of the Gumbel distribution.

In order to allow dependence between the initial number of jobs at different servers, we can also replace Assumptions 2.2 and 2.3 with the following assumption.

ASSUMPTION 2.6. *Let $Q_i^{(N)}(0) = U_i^{(N)} + V_i^{(N)}$, with $U_i^{(N)} = \lfloor r_N U_i \rfloor$, where $(U_i, i \leq N)$ are i.i.d. and non-negative, and satisfy either Assumption 2.4 or 2.5. Furthermore, $V_i^{(N)}$ is non-negative, and $\max_{i \leq N} V_i^{(N)} / (N \log N) \xrightarrow{\mathbb{P}} 0$, as $N \rightarrow \infty$.*

When Assumption 2.6 is satisfied, there may be mutual dependence between $Q_i^{(N)}(0)$ and $Q_j^{(N)}(0)$, because $V_i^{(N)}$ and $V_j^{(N)}$ may be mutually dependent.

As can be seen in Theorem 2.1, the fluid limit has an unusual form, $q(t)$ is namely a maximum of two functions. The first part of this maximum is the fluid limit when the initial number of jobs equals 0 and the second part is caused by the initial number of jobs. We elaborate on this more in Section 2.3. The $\log N$ term in the spatial and temporal scaling of the process is also unusual. We show in Section 2.2 that this is due to the fact that we take a maximum of N random variables, with N large. Scaling terms like $(\log N)^c$ are in this context very natural.

We mentioned earlier that different choices for temporal and spatial scalings lead to a fluid limit. We gave Proposition 2.1 as an example. Since we analyze one and only one system, the two fluid limits that we presented should be connected to each other. An easy way to see this, is by observing that from Theorem 2.1 it follows that when $Q_{(\alpha, \beta)}^{(N)}(0) = 0$,

$$\frac{Q_{(\alpha, \beta)}^{(N)}(tN^3 \log N)}{N \log N} \xrightarrow{\mathbb{P}} \sqrt{2\alpha t} - \beta t \quad \text{as } N \rightarrow \infty,$$

for $t < \alpha/(2\beta^2)$. Thus, for all $t > 0$ and for N large, we expect that $Q_{(\alpha, \beta)}^{(N)}(tN^3) / (N\sqrt{\log N}) \approx \sqrt{2\alpha t} - \beta t / \sqrt{\log N} \xrightarrow{N \rightarrow \infty} \sqrt{2\alpha t}$. This shows heuristically how Proposition 2.1 is connected with Theorem 2.1. The formal proof of Proposition 2.1 is analogous to the proof of Theorem 2.1 and is omitted in this paper.

2.2. Scaling. In Section 2.1, we presented the fluid limit under the rather unusual temporal scaling of $N^3 \log N$ and spatial scaling of $N \log N$. A heuristic justification for these scalings can be given by using extreme value theory and ideas from literature on diffusion approximations. In particular, for the spatial scaling we argue as follows: as we are interested in the convergence of the maximum queue length, we can use a central limit result to replace each separate queue length with a reflected Brownian motion and use extreme value theory to get a heuristic idea of the convergence of the scaled maximum queue length. To argue this, first observe that the arrival and service processes are binomially distributed random variables, and we can compute the expectation and variance of $\left(A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N)\right) / (N\sqrt{\log N})$ as

$$\mathbb{E} \left[\frac{1}{N\sqrt{\log N}} \left(A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N) \right) \right] = -\beta t \sqrt{\log N} + o_N(1), \quad (2.8)$$

and

$$\begin{aligned} & \text{Var} \left(\frac{1}{N\sqrt{\log N}} \left(A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N) \right) \right) \\ &= \frac{1}{N^2 \log N} [tN^3 \log N] \left(\left(\frac{\alpha}{N} + \frac{\beta}{N^2} \right) \left(1 - \frac{\alpha}{N} - \frac{\beta}{N^2} \right) + \frac{\alpha}{N} \left(1 - \frac{\alpha}{N} \right) \right) \\ &= 2\alpha t + o_N(1). \end{aligned} \quad (2.9)$$

From this, a non-trivial scaling limit can be easily deduced: observe that $A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N)$ is a sum of independent and identically distributed random variables, so this implies that

$$\frac{1}{N\sqrt{\log N}} \left(A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N) \right) \stackrel{d}{\approx} Z_i,$$

as N is large, with $Z_i \sim \mathcal{N}(-\beta t \sqrt{\log N}, 2\alpha t)$. Furthermore, because $A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N)$ is in fact the difference of two random walks, we also have

$$\sup_{0 \leq n \leq tN^3 \log N} \frac{1}{N\sqrt{\log N}} \left(A^{(N)}(n) - S_i^{(N)}(n) \right) \stackrel{d}{\approx} R_i(t),$$

as N is large, with $R_i(t)$ a reflected Brownian motion for t fixed. We can apply extreme value theory to show that $\max_{i \leq N} R_i(t)$ scales with $\sqrt{\log N}$. This can be deduced from the cumulative distribution function of the reflected Brownian motion which is given in [10, p. 49]. Concluding, the proper spatial scaling of the fluid limit in Theorem 2.1 is $1/(N \log N)$.

As Equations (2.8) and (2.9) show, the right temporal and spatial scalings are determined by the choice of the arrival and service probability. When we change the arrival probability to $p^{(N)} = 1 - \alpha/N - \beta/N^{1+c}$, with $c \geq 1$, and keep the service probability the same, we can derive in the same manner, that under a different temporal and spatial scaling of the queueing process, the fluid limit result still holds; we state this in Proposition 2.3.

PROPOSITION 2.3 (Other arrival and service probabilities). *For $c \geq 1$, $\alpha > 0$ and $\beta > 0$, with*

1. $p^{(N)} = 1 - \alpha/N - \beta/N^{1+c}$,
2. $q^{(N)} = 1 - \alpha/N$,

and $Q_{(\alpha, \beta)}^{(N)}(0) = O(N^c \log N)$ and satisfies the same assumptions as in Theorem 2.1, then

$$\mathbb{P} \left(\sup_{0 \leq t \leq T} \left| \frac{Q_{(\alpha, \beta)}^{(N)}(tN^{1+2c} \log N)}{N^c \log N} - q(t) \right| > \epsilon \right) \xrightarrow{N \rightarrow \infty} 0 \quad \forall \epsilon > 0.$$

The proof of this proposition is very similar to the proof of Theorem 2.1. Thus we omit it here.

2.3. Shape of the fluid limit. In Section 2.2, we gave a heuristic explanation of the temporal and spatial scaling of the process. Here we do the same for the shape of the fluid limit. First of all, we rewrite the expression in (2.3) and get that the scaled maximum queue length satisfies

$$\begin{aligned} & \frac{Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)}{N \log N} = \\ \max & \left(\max_{i \leq N} \sup_{0 \leq s \leq t} \frac{\left(A^{(N)}(tN^3 \log N) - A^{(N)}(sN^3 \log N) \right) - \left(S_i^{(N)}(tN^3 \log N) - S_i^{(N)}(sN^3 \log N) \right)}{N \log N}, \right. \\ & \left. \max_{i \leq N} \frac{A^{(N)}(tN^3 \log N) + S_i^{(N)}(tN^3 \log N) + Q_i^{(N)}(0)}{N \log N} \right). \end{aligned} \quad (2.10)$$

Now, observe that when $Q_i^{(N)}(0) = 0$ for all i , the pairwise maximum in (2.10) simplifies to the first part of the maximum. Furthermore, it turns out that the first and the second part of this maximum converge to the first and second part of the maximum in (2.5), respectively. To see the first limit heuristically, observe that, due to the central limit theorem,

$$\frac{1}{N\sqrt{\log N}} \left(A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N) \right) \stackrel{d}{\approx} \vartheta_i + \zeta,$$

with $\vartheta_i \sim \mathcal{N}(0, \alpha t)$, independently for all i , and $\zeta \sim \mathcal{N}(-\beta t \sqrt{\log N}, \alpha t)$. We can write $\max_{i \leq N}(\vartheta_i + \zeta) = \max_{i \leq N}(\vartheta_i) + \zeta$. Then, by the basic convergence result that the maximum of N i.i.d. standard normal random variables scales like $\sqrt{2 \log N}$, it is easy to see that $\max_{i \leq N}(\vartheta_i + \zeta) / \sqrt{\log N} \xrightarrow{\mathbb{P}} \sqrt{2\alpha t} - \beta t$ as $N \rightarrow \infty$. Because of the fact that a queue length which is 0 at time 0, can be written as the supremum of the arrival process minus the service process up to time t , the fluid limit yields $\sup_{0 \leq s \leq t}(\sqrt{2\alpha s} - \beta s)$, which equals the first part of the maximum in (2.5).

Similarly, for the second part in (2.10) we observe that

$$\begin{aligned} & \max_{i \leq N} \frac{A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N) + Q_i^{(N)}(0)}{N \log N} \\ = & \frac{A^{(N)}(tN^3 \log N) - (1 - \alpha/N)tN^3 \log N}{N \log N} + \max_{i \leq N} \frac{(1 - \alpha/N)tN^3 \log N - S_i^{(N)}(tN^3 \log N) + Q_i^{(N)}(0)}{N \log N}. \end{aligned} \quad (2.11)$$

It is easy to see that the first term converges to $-\beta t$ as $N \rightarrow \infty$, and we prove later on that the second term converges to $g(t, q(0))$. This explains the second part of the fluid limit in (2.5).

Specific properties of the function g can be deduced. First of all, Assumption 2.4 considers the case that U_i has a finite right endpoint. In this scenario, we have that $Q_i^{(N)}(0)/(N \log N) = \lfloor r_N U_i \rfloor / (N \log N) = \lfloor N \log N U_i \rfloor / (N \log N) \approx U_i$. Now, the theorem says that $g(t, q(0)) = q(0) + \sqrt{2\alpha t}$. This actually means that for large N ,

$$\begin{aligned} & \max_{i \leq N} \left(U_i + \frac{(1 - \alpha/N)tN^3 \log N - S_i^{(N)}(tN^3 \log N)}{N \log N} \right) \\ \approx & \max_{i \leq N} U_i + \max_{i \leq N} \frac{(1 - \alpha/N)tN^3 \log N - S_i^{(N)}(tN^3 \log N)}{N \log N}. \end{aligned}$$

This behavior can be very well explained, because due to the assumption that U_i has a finite right endpoint, there will be many observations of U_i that are close to the right endpoint, as

N becomes large, and thus it will be more and more likely that there is a large observation $\left((1 - \alpha/N) (tN^3 \log N) - S_{i^*}^{(N)} (tN^3 \log N) \right) / \left(N \log N \right)$, for which the observation U_{i^*} will also be large.

Furthermore, when Assumption 2.5 holds, $g(t, q(0))$ can be written as a supremum over a set. To give an idea why this is the case, we first observe that we can write the last term in (2.11) as

$$\max_{i \leq N} \left(\frac{(1 - \alpha/N) (tN^3 \log N) - S_i^{(N)} (tN^3 \log N)}{N \log N} + \frac{Q_i^{(N)}(0)}{N \log N} \right). \quad (2.12)$$

Thus, this maximum can be viewed as a maximum of N pairwise sums of random variables. For any $N > 0$, we can write down all the N pairs of random variables as

$$\left\{ \left(\frac{1}{\sqrt{2\alpha t}} \frac{(1 - \alpha/N) (tN^3 \log N) - S_i^{(N)} (tN^3 \log N)}{N \log N}, \frac{1}{q(0)} \frac{Q_i^{(N)}(0)}{N \log N} \right)_{i \leq N} \right\}. \quad (2.13)$$

Now, the expression in Equation (2.12) can be written as $\sqrt{2\alpha t}u + q(0)v$ with (u, v) in the set in (2.13), such that $\sqrt{2\alpha t}u + q(0)v$ is maximized. Due to the central limit theorem, the first term in (2.13) can be approximated by $\vartheta_i / \sqrt{2\alpha t}$ with $\vartheta_i \sim \mathcal{N}(0, \alpha t)$ when N is large. Therefore, the convex hull of the set in (2.13) looks like the convex hull of the set

$$\left\{ \left(\frac{1}{\sqrt{2\alpha t}} \frac{\vartheta_i}{\sqrt{\log N}}, \frac{1}{q(0)} \frac{Q_i^{(N)}(0)}{N \log N} \right)_{i \leq N} \right\}.$$

The convex hull of this set can be seen as a random variable, and converges, under an appropriate metric, in probability to the limiting set

$$\{(u, v) | u^2 + h(v) \leq 1, -1 \leq u \leq 1, 0 \leq v \leq 1\}, \quad (2.14)$$

in \mathbb{R}^2 , as $N \rightarrow \infty$, cf. [7] and [8] for details on this. Our intuition says that the limit of the expression in (2.12) is attained at the coordinate (u, v) in the closure of the limiting set given in (2.14), such that $\sqrt{2\alpha t}u + q(0)v$ is maximized. We show that this is indeed correct. In fact, we prove this in Lemma 4.4 in a more general context than in [7] and [8]. In [7] and [8], the authors make the assumption that the scaling sequences are the same, so the analysis is restricted to samples of the type $\{(X_i/a_N, Y_i/a_N)_{i \leq N}\}$. However, we show that for proving convergence of the maximum of the pairwise sum, the scaling sequences do not need to be the same.

2.4. Examples and numerics. In Section 2.3, we showed that the shape of the fluid limit depends on the distribution of the number of jobs at time 0. Here, we give some basic examples how the fluid limit is influenced by the distribution of the number of jobs at time 0. We also present and discuss some numerical results.

As a first example, for $U_i = X_i^+$, with $X_i \sim \mathcal{N}(0, 1)$, we can write for $v > 0$, $\mathbb{P}(U_i > v) = \exp(-v^2 L(v))$, such that L is slowly varying. Thus for $v \in [0, 1]$,

$$h(v) = \lim_{t \rightarrow \infty} \frac{-\log(\mathbb{P}(U_i > vt))}{-\log(\mathbb{P}(U_i > t))} = \lim_{t \rightarrow \infty} \frac{(vt)^2 L(vt)}{t^2 L(t)} = v^2.$$

Thus,

$$g(t, q(0)) = \sup_{(u, v)} \{ \sqrt{2\alpha t}u + q(0)v | u^2 + v^2 \leq 1, -1 \leq u \leq 1, 0 \leq v \leq 1 \} = \sqrt{q(0)^2 + 2\alpha t}.$$

Concluding,

$$\begin{aligned} & \max_{i \leq N} \frac{(1 - \alpha/N)(tN^3 \log N) - S_i^{(N)}(tN^3 \log N) + Q_i^{(N)}(0)}{N \log N} \\ &= \max_{i \leq N} \frac{(1 - \alpha/N)(tN^3 \log N) - S_i^{(N)}(tN^3 \log N) + \lfloor q(0)N \log NU_i / \sqrt{2 \log N} \rfloor}{N \log N} \\ &\xrightarrow{\mathbb{P}} \sqrt{q(0)^2 + 2\alpha t} - \beta t \text{ as } N \rightarrow \infty, \end{aligned}$$

where $r_N = q(0)N \log N / \sqrt{2 \log N}$, such that $Q_{(\alpha, \beta)}^{(N)}(0) / (N \log N) \xrightarrow{\mathbb{P}} q(0)$, as $N \rightarrow \infty$.

Another example is, when we assume that U_i is lognormally distributed, we know that $\mathbb{P}(U_i > v) = \mathbb{P}(X_i > \log v)$, with $X_i \sim \mathcal{N}(0, 1)$. Thus, $\mathbb{P}(U_i > v) = \exp(-\mathbb{1}(v > 0) \log(v)^2 L(\log v))$. Then, for $v \in [0, 1]$,

$$h(v) = \lim_{t \rightarrow \infty} \frac{\mathbb{1}(v > 0) \log(vt)^2 L(\log(vt))}{\log(t)^2 L(\log(t))} = \mathbb{1}(v > 0).$$

In this case, we have that

$$g(t, q(0)) = \sup_{(u, v)} \{ \sqrt{2\alpha t} u + q(0)v | u^2 + \mathbb{1}(v > 0) \leq 1, -1 \leq u \leq 1, 0 \leq v \leq 1 \} = \max(q(0), \sqrt{2\alpha t}).$$

We also consider the case $\mathbb{P}(U_i > v) = \exp(1 - \exp(v))$, then for $v \in [0, 1]$,

$$\lim_{t \rightarrow \infty} \frac{-\log(\mathbb{P}(U_i > vt))}{-\log(\mathbb{P}(U_i > t))} = \lim_{t \rightarrow \infty} \frac{\exp(vt) - 1}{\exp(t) - 1} = \mathbb{1}(v = 1).$$

Then,

$$g(t, q(0)) = \sup_{(u, v)} \{ \sqrt{2\alpha t} u + q(0)v | u^2 + \mathbb{1}(v = 1) \leq 1, -1 \leq u \leq 1, 0 \leq v \leq 1 \} = q(0) + \sqrt{2\alpha t}.$$

As a last example, we observe the scenario that $\mathbb{P}(U_i > v) = \exp(-vL(v))$, thus $h(v) = v$. Then,

$$\begin{aligned} g(t, q(0)) &= \sup_{(u, v)} \{ \sqrt{2\alpha t} u + q(0)v | u^2 + v \leq 1, 0 \leq u \leq 1, 0 \leq v \leq 1 \} \\ &= \left(q(0) + \frac{\alpha t}{2q(0)} \right) \mathbb{1} \left(t < \frac{2q(0)^2}{\alpha} \right) + \sqrt{2\alpha t} \mathbb{1} \left(t \geq \frac{2q(0)^2}{\alpha} \right). \end{aligned}$$

We would like to give some extra attention to the case where $q(0) = \alpha/(2\beta)$. Then, it is not difficult to see that $q(t) \equiv \alpha/(2\beta)$. Thus, for these choices of $h(v)$ and $q(0)$, the system starts and stays in steady state. One can show that this limit is only obtained for $h(v) = v$, so this gives us some information on the joint steady-state distribution of *all* the queue lengths in the fork-join system.

Now, we turn to some numerical examples. In Figure 2, the simulated maximum queue length is plotted together with the scaled fluid limit $N \log N q(t) / (N^3 \log N)$, with q given in Theorem 2.1, and $N = 1000$. The queue lengths at time zero in Figures 2a, 2b and 2c are exponentially distributed. These figures show that for $N = 1000$, the maximum queue length is not close to its fluid limit.

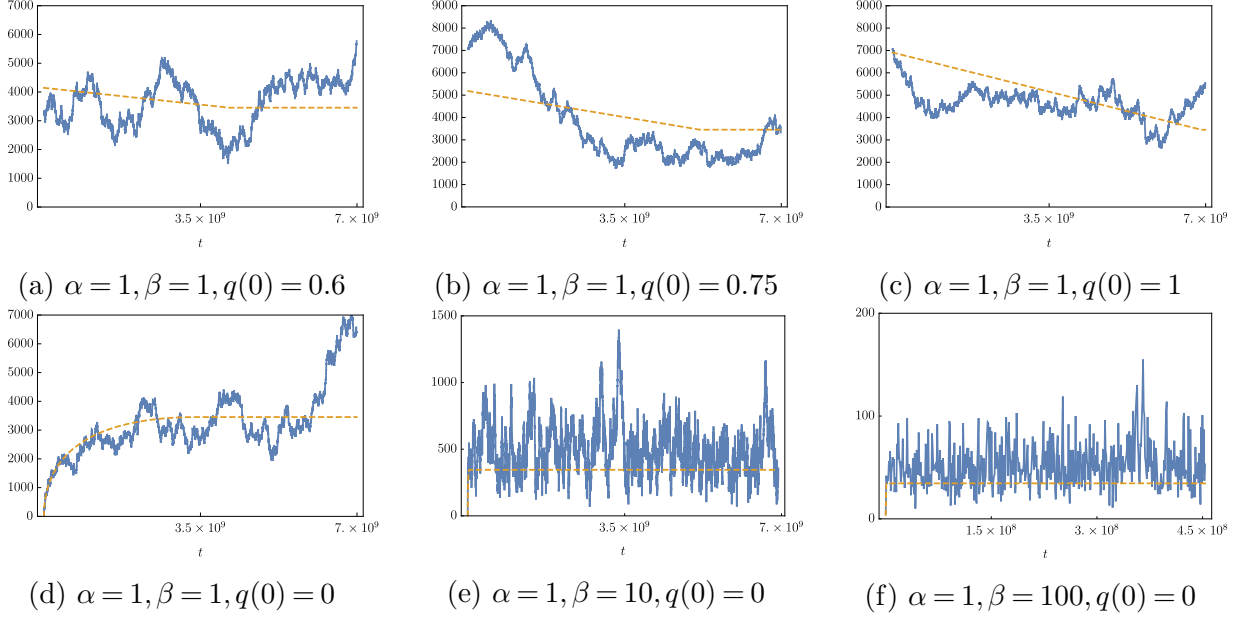


FIGURE 2. Maximum queue length and fluid limit approximation (Thm. 2.1) for $N = 1000$

As these figures show, for $N = 1000$, the variance of the maximum queue length is still high. We could however give some heuristic arguments why these results are not very accurate. As mentioned before, we have that

$$\frac{A^{(N)}(tN^3 \log N) - (1 - \alpha/N)(tN^3 \log N)}{N \log N} \xrightarrow{\mathbb{P}} -\beta t \text{ as } N \rightarrow \infty,$$

which is one building block of the fluid limit.

For $(A^{(N)}(tN^3 \log N) - (1 - \alpha/N)tN^3 \log N)/(N \log N)$, we can compute the standard deviation. We have for $\alpha = \beta = t = 1$ and $N = 1000$ that

$$\begin{aligned} \sqrt{\text{Var} \left(A^{(N)}(tN^3 \log N) - \left(1 - \frac{\alpha}{N}\right)(tN^3 \log N) \right)} &= \sqrt{\left(1 - \frac{\alpha}{N} - \frac{\beta}{N^2}\right) \left(\frac{\alpha}{N} + \frac{\beta}{N^2}\right) [tN^3 \log N]} \\ &= 2628.26. \end{aligned}$$

This is of the order of magnitude of the errors that we see in the figures.

Another way of seeing that there is a significant deviation is by looking at $\max_{i \leq N} \left((1 - \alpha/N)tN^3 \log N - S_i^{(N)}(tN^3 \log N) \right)$. As mentioned in Section 2.3, we have that

$$\frac{(1 - \alpha/N)tN^3 \log N - S_i^{(N)}(tN^3 \log N)}{N\sqrt{\log N}} \stackrel{d}{\approx} \vartheta_i,$$

with $\vartheta_i \sim \mathcal{N}(0, \alpha t)$. Thus, this means that

$$\max_{i \leq N} \left(\left(1 - \frac{\alpha}{N}\right)tN^3 \log N - S_i^{(N)}(tN^3 \log N) \right) \stackrel{d}{\approx} \max_{i \leq N} \vartheta_i N \sqrt{\log N}.$$

When we choose $N = 1000$, $\alpha = t = 1$, and simulate enough samples of $\max_{i \leq N} \vartheta_i N \sqrt{\log N}$, we observe a standard deviation which is higher than 900.

In Figures 2a, 2b and 2c, the high standard deviation is also caused by the distribution of the number of jobs at time 0. For example, for $E_i \sim \text{Exp}(1/N)$, i.i.d. for all i , and $N = 1000$, we have

that $\sqrt{\text{Var}(\max_{i \leq N} E_i)} = 1282.16$, so this is also of the order of magnitude of the errors that we see.

As mentioned, one can prove fluid limits under several temporal and spatial scalings. In Figure 3, the maximum queue length is plotted against the rescaled fluid limit given in Proposition 2.1, which is in orange, and the rescaled steady-state limit, which is in green. In these plots, $N = 1000$. The rescaled fluid limit is $\sqrt{2\alpha t/N^3 N \sqrt{\log N}}$, and the rescaled steady-state limit satisfies $\alpha/(2\beta)N \log N$.

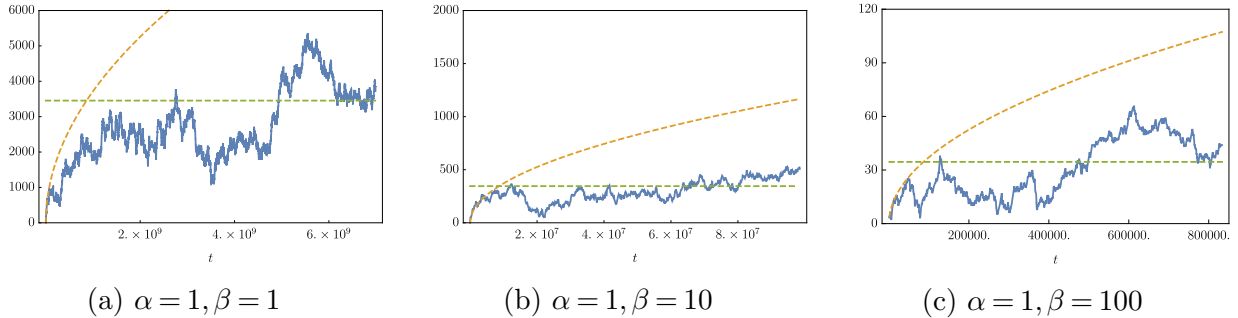


FIGURE 3. Maximum queue length, fluid limit approximation (Prop. 2.1) and steady-state approximation for $N = 1000$

When we observe Figure 3, we see that for small time instances, the maximum queue length follows the fluid limit described in Proposition 2.1 with a negligible deviation, and we also see that, from the point that the fluid limit and steady state have intersected, the maximum queue length follows the steady state, though with a significant deviation. This latter behavior can be very well explained when we plot the same maximum queue lengths together with the fluid limit in Theorem 2.1, this is shown in Figure 4.

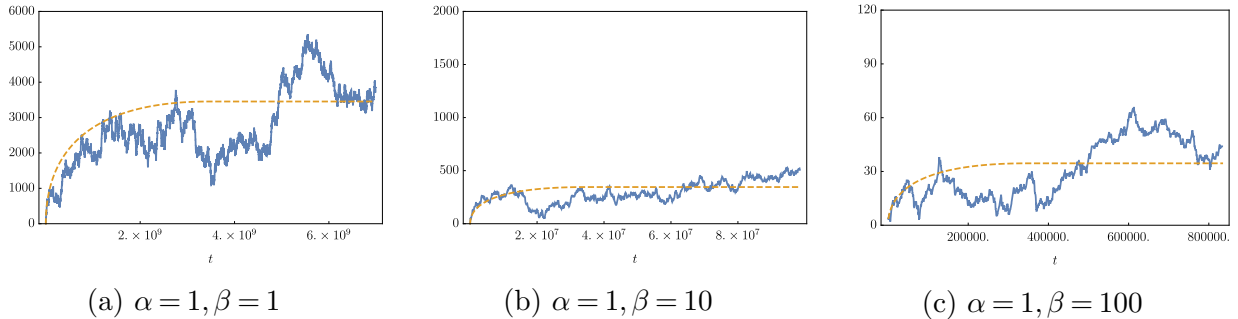


FIGURE 4. Maximum queue length and fluid limit approximation (Thm. 2.1) for $N = 1000$

In Figure 5, we zoom in on the graphs given in Figure 3a and 3b. As these figures show, for small time instances, the maximum queue length follows the fluid limit described in Proposition 2.1 quite well. Again, we can heuristically explain the deviations by approximating the maximum queue length with $\sqrt{1/N^3 N \max_{i \leq N} \vartheta_i}$, with $\vartheta_i \sim \mathcal{N}(0, \alpha t)$, i.i.d. For $\alpha = 1$, and $t = 7 \cdot 10^7$, simulations show that this approximation has a standard deviation around 95, and for $t = 7 \cdot 10^6$, we get a standard deviation around 30, this is of the order of magnitude of the errors in Figure 5a and 5b, respectively.

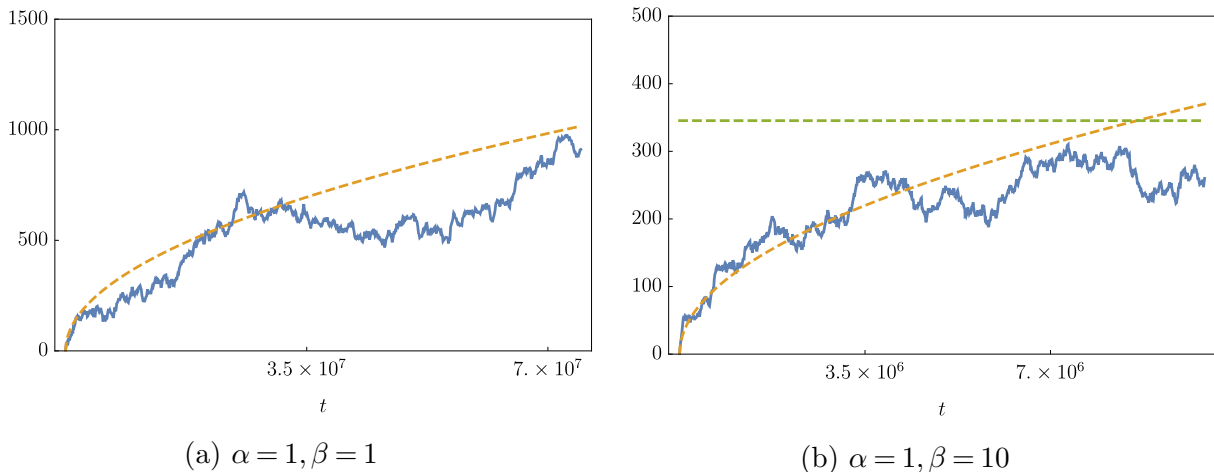


FIGURE 5. Maximum queue length, fluid limit approximation (Prop. 2.1) and steady-state approximation for $N = 1000$

3. Conclusion. In this paper, we analyzed a fork-join queue with N servers in heavy traffic. We considered the case of nearly deterministic arrivals and service times, and we derived a fluid limit of the maximum queue length, in Theorem 2.1, as N grows large.

Furthermore, 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. Another, but less straightforward extension of this result, would be to assume an arrival and service process that are not Markovian.

Moreover, as the figures in Section 2.4 show, it should be possible to derive a more refined limit. Therefore, it is interesting to look at second order convergence of the maximum queue length. We are currently exploring this for the system in steady state. In other words, we try to gain more insight in the process by finding a convergence result of $Q_{(\alpha,\beta)}^{(N)}(\infty)/N - \alpha/(2\beta) \log N$. For the process limit, proving a second order convergence result is much harder and more technical, because the scaled maximum of N independent Brownian motions converges to a Brown-Resnick process [6].

4. Proofs. In this section, we prove Theorem 2.1. Since each server has the same arrival process, the queue lengths are dependent. The general idea of proving Theorem 2.1 is to approximate the scaled centralized service process in (4.4) by a normally distributed random variable. We can use extreme value theory to prove convergence of the maximum of these normally distributed random variables in probability. By using the non-uniform version of the Berry-Esséen theorem, cf. [16], we show that the convergence result of the original process is the same as the convergence result with normally distributed random variables. Furthermore, we prove convergence of the part involving non-zero starting points. This gives us the pointwise convergence of the process, which we prove in Section 4.3. In this section, we also prove convergence of the finite-dimensional distributions. Finally, we prove in Section 4.4 that the process is tight. These three results together prove the theorem.

4.1. Definitions. For the sake of notation, we use the expressions given in Definition 4.1 to prove the tightness.

DEFINITION 4.1. We define the random walk $\tilde{R}_i^{(N)}(n)$ as

$$\tilde{R}_i^{(N)}(n) = \frac{\tilde{A}^{(N)}(n) + \tilde{S}_i^{(N)}(n)}{\log N}, \quad (4.1)$$

where

$$\tilde{A}^{(N)}(n) = \frac{A^{(N)}(n)}{N} - \left(1 - \frac{\alpha}{N}\right) \frac{\lfloor n \rfloor}{N}, \quad (4.2)$$

and

$$\tilde{S}_i^{(N)}(n) = -\frac{S_i^{(N)}(n)}{N} + \left(1 - \frac{\alpha}{N}\right) \frac{\lfloor n \rfloor}{N}. \quad (4.3)$$

Furthermore,

$$M_i^{(N)}(t) = \frac{\tilde{S}_i^{(N)}(tN^3 \log N) \sqrt{tN^3 \log N}}{\sqrt{\alpha t(1 - \alpha/N) \log N} \sqrt{\lfloor tN^3 \log N \rfloor}}, \quad (4.4)$$

with $A^{(N)}(n)$ and $S_i^{(N)}(n)$ given in Definitions 2.1 and Definition 2.2 respectively. As mentioned in Section 2.3, when $Q_{(\alpha,\beta)}^{(N)}(0) = 0$, the quantity in (2.10) simplifies to

$$\begin{aligned} & \frac{Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)}{N \log N} \\ &= \max_{i \leq N} \sup_{0 \leq s \leq t} \frac{\left(A^{(N)}(tN^3 \log N) - A^{(N)}(sN^3 \log N) \right) - \left(S_i^{(N)}(tN^3 \log N) - S_i^{(N)}(sN^3 \log N) \right)}{N \log N}. \end{aligned}$$

Consequently, we can rewrite

$$\begin{aligned} & \frac{Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)}{N \log N} \\ &= \max_{i \leq N} \sup_{0 \leq r \leq t} \frac{\tilde{A}^{(N)}(tN^3 \log N) - \tilde{A}^{(N)}(rN^3 \log N) + \tilde{S}_i^{(N)}(tN^3 \log N) - \tilde{S}_i^{(N)}(rN^3 \log N)}{\log N} \\ &= \max_{i \leq N} \sup_{0 \leq r \leq t} \left(\tilde{R}_i^{(N)}(tN^3 \log N) - \tilde{R}_i^{(N)}(rN^3 \log N) \right). \end{aligned} \quad (4.5)$$

4.2. Useful lemmas. In order to prove Theorem 2.1, a few preliminary results are needed. As stated in Definition 4.1, we can write $\tilde{R}_i^{(N)}(n)$ as

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

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 $\tilde{S}_i^{(N)}(n)/\log N$. Specifically, we use the fact that

$$M_i^{(N)}(t) \xrightarrow{d} Z \text{ as } N \rightarrow \infty,$$

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

$$\lim_{\delta \downarrow 0} \limsup_{N \rightarrow \infty} \frac{1}{\delta} \mathbb{P} \left(\sup_{t \leq s \leq t + \delta} \left| \frac{Q_{(\alpha,\beta)}^{(N)}(sN^3 \log N)}{N \log N} - \frac{Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)}{N \log N} \right| > \epsilon \right) = 0. \quad (4.6)$$

In Lemma 4.1, a useful upper bound for the absolute value in (4.6) is obtained, which we use to prove the tightness of the process.

LEMMA 4.1. For $t > 0$, $\delta > 0$ and $Q_{(\alpha,\beta)}^{(N)}(0) = 0$, we have that

$$\begin{aligned} & \sup_{t \leq s \leq t+\delta} \left| \frac{Q_{(\alpha,\beta)}^{(N)}(sN^3 \log N)}{N \log N} - \frac{Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)}{N \log N} \right| \\ & \leq \sup_{t \leq s \leq t+\delta} \max_{i \leq N} \left(\tilde{R}_i^{(N)}(sN^3 \log N) - \tilde{R}_i^{(N)}(tN^3 \log N) \right) \\ & + 2 \sup_{t \leq s \leq t+\delta} \max_{i \leq N} \left(\tilde{R}_i^{(N)}(tN^3 \log N) - \tilde{R}_i^{(N)}(sN^3 \log N) \right). \end{aligned} \quad (4.7)$$

In our proofs we use the fact that $M_i^{(N)}(t)$ converges in distribution to a normally distributed random variable. To be able to use this convergence result, we prove an upper bound of the convergence rate in Lemma 4.2.

LEMMA 4.2. For $t > 0$, we have that an upper bound of the rate of convergence of $\pm \tilde{S}_i^{(N)}(tN^3 \log N) \sqrt{tN^3 \log N} / \sqrt{\alpha t(1 - \alpha/N) \log N [tN^3 \log N]}$ to a standard normal random variable is given by

$$\left| \mathbb{P}\left(M_i^{(N)}(t) < y\right) - \Phi(y) \right| \leq \frac{c_t}{N \sqrt{\log N}} \frac{1}{1 + |y|^3}, \quad (4.8)$$

with $c_t > 0$.

Lemma 4.2 follows from the main result in [16], where the author proves the non-uniform Berry-Essén inequality. To prove tightness, we need the following lemma:

LEMMA 4.3. For $t > 0$,

$$\limsup_{N \rightarrow \infty} \mathbb{E} \left[\max \left(\max_{i \leq N} \frac{\pm \tilde{S}_i^{(N)}(tN^3 \log N)}{\log N}, 0 \right)^{5/2} \right] \leq (2\alpha t)^{5/4}. \quad (4.9)$$

In order to prove pointwise convergence of the starting position, we show in Lemma 4.9 that

$$\max_{i \leq N} \left(\frac{\tilde{S}_i^{(N)}(tN^3 \log N)}{\log N} + \frac{Q_i^{(N)}(0)}{N \log N} \right) \approx \max_{i \leq N} \left(\frac{\sqrt{\alpha t} X_i}{\sqrt{\log N}} + \frac{Q_i^{(N)}(0)}{N \log N} \right),$$

with $X_i \sim \mathcal{N}(0, 1)$, as N is large.

In Lemma 4.4, we prove the convergence of $\max_{i \leq N} \left(\sqrt{\alpha t} X_i / \sqrt{\log N} + Q_i^{(N)}(0) / (N \log N) \right)$.

LEMMA 4.4 (**Pointwise convergence approximation starting position**).

$$\max_{i \leq N} \left(\frac{\sqrt{\alpha t} X_i}{\sqrt{\log N}} + \frac{Q_i^{(N)}(0)}{N \log N} \right) \xrightarrow{\mathbb{P}} g(t, q(0)) \text{ as } N \rightarrow \infty,$$

with $X_i \sim \mathcal{N}(0, 1)$ i.i.d. and the function g as given in Theorem 2.1.

The proofs of Lemmas 4.1, 4.2, 4.3, and 4.4 can be found in Appendix C. Lemma 4.4 follows from Lemma B.1, where a more general result is proven on $\max_{i \leq N} \sum_{j=1}^k Y_i^{(j)} / a_N^{(j)}$.

4.3. Pointwise convergence. In this section, we prove pointwise convergence of the scaled maximum queue length appearing in Theorem 2.1.

THEOREM 4.1 (Pointwise convergence). For $t > 0$,

$$\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)}{N \log N} \xrightarrow{\mathbb{P}} q(t) \text{ as } N \rightarrow \infty, \quad (4.10)$$

with $q(t)$ given in Equation (2.5).

As Equation (2.10) shows, we can write the scaled maximum queue length as a maximum of two random variables, namely, one pertaining to a system starting empty and one pertaining to a system starting non-empty. We prove the pointwise convergence of the first part of this maximum in Lemma 4.5. In Lemma 4.9 we prove the pointwise convergence of the second part. In order to do so, we need some extra results, which are stated in Lemmas 4.4, 4.6, 4.7, and 4.8.

LEMMA 4.5. For $t > 0$ and $Q_{(\alpha,\beta)}^{(N)}(0) = 0$,

$$\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)}{N \log N} \xrightarrow{\mathbb{P}} \left(\sqrt{2\alpha t} - \beta t \right) \mathbb{1} \left(t < \frac{\alpha}{2\beta^2} \right) + \frac{\alpha}{2\beta} \mathbb{1} \left(t \geq \frac{\alpha}{2\beta^2} \right) \text{ as } N \rightarrow \infty.$$

To prove convergence of sequences of real valued random variables to a constant it suffices to show convergence in distribution. Therefore, we use Lemmas 4.6, 4.7 and 4.8 below to prove that the upper and lower bound of the cumulative distribution function converge to the same function.

LEMMA 4.6. For $\delta > 0$, $t < \alpha/(2\beta^2)$ and $Q_{(\alpha,\beta)}^{(N)}(0) = 0$,

$$\limsup_{N \rightarrow \infty} \mathbb{P} \left(\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)}{N \log N} > \sqrt{2\alpha t} - \beta t + \delta \right) = 0. \quad (4.11)$$

Proof Let $\delta > 0$ be given. Let us assume that $t < \alpha/(2\beta^2)$. We then have that

$$\begin{aligned} & \mathbb{P} \left(\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)}{N \log N} > \sqrt{2\alpha t} - \beta t + \delta \right) \\ &= \mathbb{P} \left(\max_{i \leq N} \sup_{0 \leq s \leq t} \left(\frac{\tilde{A}^{(N)}(sN^3 \log N) + \tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} \right) - \sqrt{2\alpha t} + \beta t > \delta \right). \end{aligned}$$

For $t < \alpha/(2\beta^2)$, $\sqrt{2\alpha t} - \beta t$ is an increasing function. Therefore,

$$\begin{aligned} & \mathbb{P} \left(\max_{i \leq N} \sup_{0 \leq s \leq t} \left(\frac{\tilde{A}^{(N)}(sN^3 \log N) + \tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} \right) - \sqrt{2\alpha t} + \beta t > \delta \right) \\ &\leq \mathbb{P} \left(\max_{i \leq N} \sup_{0 \leq s \leq t} \left(\frac{\tilde{A}^{(N)}(sN^3 \log N) + \tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} - \sqrt{2\alpha s} + \beta s \right) > \delta \right) \\ &= \mathbb{P} \left(\sup_{0 \leq s \leq t} \left(\max_{i \leq N} \frac{\tilde{A}^{(N)}(sN^3 \log N) + \tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} - \sqrt{2\alpha s} + \beta s \right) > \delta \right). \end{aligned}$$

Observe that

$$\begin{aligned} & \mathbb{P} \left(\sup_{0 \leq s \leq t} \left(\max_{i \leq N} \frac{\tilde{A}^{(N)}(sN^3 \log N) + \tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} - \sqrt{2\alpha s} + \beta s \right) > \delta \right) \\ &\leq \mathbb{P} \left(\sup_{0 \leq s \leq t} \left| \max_{i \leq N} \frac{\tilde{A}^{(N)}(sN^3 \log N) + \tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} - \sqrt{2\alpha s} + \beta s \right| > \delta \right) \\ &\leq \mathbb{P} \left(\sup_{0 \leq s \leq t} \left| \frac{\tilde{A}^{(N)}(sN^3 \log N)}{\log N} + \beta s \right| > \frac{\delta}{2} \right) + \mathbb{P} \left(\sup_{0 \leq s \leq t} \left| \frac{\max_{i \leq N} \tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} - \sqrt{2\alpha s} \right| > \frac{\delta}{2} \right). \end{aligned}$$

Moreover, $\tilde{A}^{(N)}(n)/\log N + \beta n/(N^3 \log N)$ is a martingale with mean 0. Therefore, by Doob's maximal submartingale inequality

$$\begin{aligned}
& \mathbb{P} \left(\sup_{0 \leq s \leq t} \left| \frac{\tilde{A}^{(N)}(sN^3 \log N)}{\log N} + \beta s \right| > \frac{\delta}{2} \right) \\
& \leq \mathbb{P} \left(\sup_{0 \leq s \leq t} \left| \frac{\tilde{A}^{(N)}(sN^3 \log N)}{\log N} + \beta \frac{\lfloor sN^3 \log N \rfloor}{N^3 \log N} \right| + \sup_{0 \leq s \leq t} \left| \beta \frac{\lfloor sN^3 \log N \rfloor}{N^3 \log N} - \beta s \right| > \frac{\delta}{2} \right) \\
& \leq \mathbb{P} \left(\sup_{0 \leq s \leq t} \left| \frac{\tilde{A}^{(N)}(sN^3 \log N)}{\log N} + \beta \frac{\lfloor sN^3 \log N \rfloor}{N^3 \log N} \right| > \frac{\delta}{4} \right) + \mathbb{P} \left(\sup_{0 \leq s \leq t} \left| \beta \frac{\lfloor sN^3 \log N \rfloor}{N^3 \log N} - \beta s \right| > \frac{\delta}{4} \right) \\
& \leq \frac{16}{\delta^2} \text{Var} \left(\frac{\tilde{A}^{(N)}(tN^3 \log N)}{\log N} \right) + o_N(1) \\
& = \frac{16}{\delta^2} \left(1 - \frac{\alpha}{N} - \frac{\beta}{N^2} \right) \left(\frac{\alpha}{N} + \frac{\beta}{N^2} \right) \frac{\lfloor tN^3 \log N \rfloor}{N^2 (\log N)^2} + o_N(1) \xrightarrow{N \rightarrow \infty} 0.
\end{aligned} \tag{4.12}$$

Furthermore, in order to have

$$\mathbb{P} \left(\sup_{0 \leq s \leq t} \left| \frac{\max_{i \leq N} \tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} - \sqrt{2\alpha s} \right| > \frac{\delta}{2} \right) \xrightarrow{N \rightarrow \infty} 0, \tag{4.13}$$

we need to have that $(\max_{i \leq N} \tilde{S}_i^{(N)}(sN^3 \log N)/\log N, s \in [0, t])$ converges to $(\sqrt{2\alpha s}, s \in [0, t])$ u.o.c. Thus

$$\lim_{N \rightarrow \infty} \mathbb{P} \left(\left| \frac{\max_{i \leq N} \tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} - \sqrt{2\alpha s} \right| > \epsilon \right) = 0, \tag{4.14}$$

and for all $r \in [0, t]$,

$$\lim_{\eta \downarrow 0} \limsup_{N \rightarrow \infty} \frac{1}{\eta} \mathbb{P} \left(\sup_{r \leq s \leq r+\eta} \left| \frac{\max_{i \leq N} \tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} - \frac{\max_{i \leq N} \tilde{S}_i^{(N)}(rN^3 \log N)}{\log N} \right| > \epsilon \right) = 0. \tag{4.15}$$

To prove the limit in (4.14), we use the result of Lemma 4.2 and observe that for all $\delta > 0$,

$$\begin{aligned}
& \mathbb{P} \left(\frac{\max_{i \leq N} \tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} > \sqrt{2\alpha s} + \delta \right) \\
& = 1 - \mathbb{P} \left(\frac{\tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} < \sqrt{2\alpha s} + \delta \right)^N \\
& = 1 - \mathbb{P} \left(M_i^{(N)}(s) < \frac{\sqrt{2\alpha s} + \delta}{\sqrt{\alpha s(1 - \alpha/N)}} \sqrt{\log N} \frac{\sqrt{sN^3 \log N}}{\sqrt{\lfloor sN^3 \log N \rfloor}} \right)^N \\
& \leq 1 - \left(\Phi \left(\frac{\sqrt{2\alpha s} + \delta}{\sqrt{\alpha s(1 - \alpha/N)}} \sqrt{\log N} \frac{\sqrt{sN^3 \log N}}{\sqrt{\lfloor sN^3 \log N \rfloor}} \right) - \frac{c_s}{N\sqrt{\log N}} \right)^N \\
& \leq 1 - \Phi \left(\frac{\sqrt{2\alpha s} + \delta}{\sqrt{\alpha s(1 - \alpha/N)}} \sqrt{\log N} \frac{\sqrt{sN^3 \log N}}{\sqrt{\lfloor sN^3 \log N \rfloor}} \right)^N + \left(1 + \frac{c_s}{N\sqrt{\log N}} \right)^N - 1 \\
& \xrightarrow{N \rightarrow \infty} 0.
\end{aligned}$$

The proof that

$$\mathbb{P}\left(\frac{\max_{i \leq N} \tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} < \sqrt{2\alpha s} - \delta\right) \xrightarrow{N \rightarrow \infty} 0,$$

goes analogously. To prove the quantity in (4.15), we observe that due to the facts that $\tilde{S}_i^{(N)}(n)$ is a random walk that satisfies the duality principle, $\max_{i \leq N} x_i - \max_{i \leq N} y_i \leq \max_{i \leq N} (x_i - y_i)$, and $\mathbb{P}(|X| > \epsilon) \leq \mathbb{P}(X > \epsilon) + \mathbb{P}(-X > \epsilon)$, we have the upper bound

$$\begin{aligned} & \frac{1}{\eta} \mathbb{P}\left(\sup_{r \leq s \leq r+\eta} \left| \frac{\max_{i \leq N} \tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} - \frac{\max_{i \leq N} \tilde{S}_i^{(N)}(rN^3 \log N)}{\log N} \right| > \epsilon\right) \\ & \leq \frac{1}{\eta} \mathbb{P}\left(\sup_{0 \leq s \leq \eta} \max_{i \leq N} \frac{\tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} > \epsilon\right) + \frac{1}{\eta} \mathbb{P}\left(\sup_{0 \leq s \leq \eta} \max_{i \leq N} \frac{-\tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} > \epsilon\right) + o_N(1). \end{aligned}$$

The $o_N(1)$ term appears since $\lfloor (r+\eta)N^3 \log N \rfloor - \lfloor rN^3 \log N \rfloor \in \{\lfloor \eta N^3 \log N \rfloor, \lfloor \eta N^3 \log N \rfloor + 1\}$. Now, we have that $\pm \tilde{S}_i^{(N)}(n)$ is a martingale with mean 0. The maximum of independent martingales is a submartingale; therefore, $\left(\max\left(0, \max_{i \leq N} \pm \tilde{S}_i^{(N)}(\eta N^3 \log N) / \log N\right)\right)^{5/2}$ is a non-negative submartingale. Hence, by use Doob's maximal submartingale inequality we can conclude that

$$\begin{aligned} & \frac{1}{\eta} \mathbb{P}\left(\sup_{0 \leq s \leq \eta} \max_{i \leq N} \frac{\tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} > \epsilon\right) + \frac{1}{\eta} \mathbb{P}\left(\sup_{0 \leq s \leq \eta} \max_{i \leq N} \frac{-\tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} > \epsilon\right) \\ & \leq \frac{1}{\eta \epsilon^{5/2}} \mathbb{E}\left[\max\left(\max_{i \leq N} \frac{\tilde{S}_i^{(N)}(\eta N^3 \log N)}{\log N}, 0\right)^{5/2}\right] + \frac{1}{\eta \epsilon^{5/2}} \mathbb{E}\left[\max\left(\max_{i \leq N} \frac{-\tilde{S}_i^{(N)}(\eta N^3 \log N)}{\log N}, 0\right)^{5/2}\right]. \end{aligned}$$

By taking the $\limsup_{N \rightarrow \infty}$ in this expression and applying Lemma 4.3, we see that this is upper bounded by $2\eta^{1/4}(2\alpha)^{5/4}/\epsilon^{5/2}$. This can be made as small as possible when η is chosen small enough. We also know that $\max_{i \leq N} \tilde{S}_i^{(N)}(0)/\log N = 0$, and that the finite-dimensional distributions of $\left(\max_{i \leq N} \tilde{S}_i^{(N)}(sN^3 \log N)/\log N, s \in [0, t]\right)$ converge to the finite-dimensional distributions of $(\sqrt{2\alpha s}, s \in [0, t])$, which follows from Theorem 4.2. The lemma follows. \square

Having examined $t \in [0, \alpha/(2\beta^2))$, we now turn to $t \in [\alpha/(2\beta^2), \infty]$.

LEMMA 4.7. *For $\delta > 0$, $\alpha/(2\beta^2) \leq t \leq \infty$ and $Q_{(\alpha, \beta)}^{(N)}(0) = 0$,*

$$\limsup_{N \rightarrow \infty} \mathbb{P}\left(\frac{Q_{(\alpha, \beta)}^{(N)}(tN^3 \log N)}{N \log N} > \frac{\alpha}{2\beta} + \delta\right) = 0.$$

Proof We write

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

with

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

with $0 < m < \beta$. Furthermore, we write

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

with

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

Thus,

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

and

$$\sup_{0 \leq k \leq n} \left(A^{(N)}(k) - S_i^{(N)}(k) \right) \leq \sup_{0 \leq k \leq n} A^{(u,N)}(k) + \sup_{0 \leq k \leq n} S_i^{(u,N)}(k).$$

We obtain by using Doob's maximal submartingale inequality that

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

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 consider the second order Taylor approximation of this expression with $1/N$ around 0, we obtain

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

Consequently, we have for N large $\theta_A^{(u,N)} \approx 2m/(\alpha N)$. By the monotone convergence theorem, we know that

$$\mathbb{P} \left(\sup_{k \geq 0} A^{(u,N)}(k) \geq x \right) \leq e^{-\theta_A^{(u,N)} x} \approx e^{-2m/(\alpha N)x}.$$

In conclusion,

$$\frac{\sup_{k \geq 0} A^{(u,N)}(k)}{N \log N} \xrightarrow{\mathbb{P}} 0 \text{ as } N \rightarrow \infty.$$

Similarly, by using Doob's maximal submartingale inequality, we obtain that

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

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 $1/N$ around 0 gives

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

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

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

and

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

Because,

$$\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)}{N \log N} \leq_{st.} \frac{Q_{(\alpha,\beta)}^{(N)}(\infty)}{N \log N} \leq \frac{\sup_{k \geq 0} A^{(u,N)}(k)}{N \log N} + \frac{\max_{i \leq N} \sup_{k \geq 0} S_i^{(N)}(k)}{N \log N},$$

the lemma follows. \square

LEMMA 4.8. For $\delta > 0$ and $Q_{(\alpha,\beta)}^{(N)}(0) = 0$,

$$\liminf_{N \rightarrow \infty} \mathbb{P} \left(\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)}{N \log N} \geq \left(\sqrt{2\alpha t} - \beta t \right) \mathbb{1} \left(t < \frac{\alpha}{2\beta^2} \right) + \frac{\alpha}{2\beta} \mathbb{1} \left(t \geq \frac{\alpha}{2\beta^2} \right) - \delta \right) = 1. \quad (4.16)$$

Proof Let us first assume that $t \leq \alpha/(2\beta^2)$. We have the lower bound

$$\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)}{N \log N} \geq_{st.} \max_{i \leq N} \frac{A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N)}{N \log N}.$$

By Equations (4.12) and (4.13), we know that

$$\max_{i \leq N} \frac{A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N)}{N \log N} \xrightarrow{\mathbb{P}} \sqrt{2\alpha t} - \beta t \text{ as } N \rightarrow \infty.$$

Let us now assume that $t > \alpha/(2\beta^2)$. We have that

$$\frac{Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)}{N \log N} \geq_{st.} \max_{i \leq N} \frac{A^{(N)}\left(\frac{\alpha}{2\beta^2} N^3 \log N\right) - S_i^{(N)}\left(\frac{\alpha}{2\beta^2} N^3 \log N\right)}{N \log N} \xrightarrow{\mathbb{P}} \frac{\alpha}{2\beta},$$

as $N \rightarrow \infty$, by again using Lemma 4.6. This proves the lemma. \square

Proof of Lemma 4.5 By combining the results of Lemmas 4.6, 4.7 and 4.8, Lemma 4.5 follows. \square

In Lemma 4.9, we connect the convergence of

$$\max_{i \leq N} \frac{A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N) + Q_i^{(N)}(0)}{N \log N}$$

to the convergence of

$$\max_{i \leq N} \left(\frac{\sqrt{\alpha t} X_i}{\sqrt{\log N}} + \frac{Q_i^{(N)}(0)}{N \log N} \right).$$

LEMMA 4.9 (Convergence starting position). *Assume that for X_i i.i.d. standard normally distributed,*

$$\max_{i \leq N} \left(\frac{\sqrt{\alpha t} X_i}{\sqrt{\log N}} + \frac{Q_i^{(N)}(0)}{N \log N} \right) \xrightarrow{\mathbb{P}} g(t, q(0)) \text{ as } N \rightarrow \infty, \quad (4.17)$$

for a certain function g . Then

$$\max_{i \leq N} \frac{A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N) + Q_i^{(N)}(0)}{N \log N} \xrightarrow{\mathbb{P}} g(t, q(0)) - \beta t \text{ as } N \rightarrow \infty.$$

Proof We have

$$\max_{i \leq N} \frac{A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N) + Q_i^{(N)}(0)}{N \log N} \quad (4.18)$$

$$= \frac{A^{(N)}(tN^3 \log N) - (1 - \alpha/N)tN^3 \log N}{N \log N} + \max_{i \leq N} \frac{(1 - \alpha/N)tN^3 \log N - S_i^{(N)}(tN^3 \log N) + Q_i^{(N)}(0)}{N \log N}. \quad (4.19)$$

We already proved in Equation (4.12) that the first term in (4.19) converges to $-\beta t$. Furthermore, we can rewrite the second term as

$$\max_{i \leq N} \left(\frac{\tilde{S}_i^{(N)}(tN^3 \log N)}{\log N} + \frac{Q_i^{(N)}(0)}{N \log N} + O_N \left(\frac{1}{N \log N} \right) \right).$$

We can easily deduce from Lemma 4.2 that

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

with $X_i \sim \mathcal{N}(0, 1)$, and c_t given in Lemma 4.2. Then, it is easy to see that

$$\left| \mathbb{P} \left(\frac{\tilde{S}_i^{(N)}(tN^3 \log N)}{\log N} + \frac{Q_i^{(N)}(0)}{N \log N} < y \right) - \mathbb{P} \left(\frac{\sqrt{\alpha t(1 - \alpha/N)} \sqrt{\lfloor tN^3 \log N \rfloor}}{\sqrt{\log N}} X_i + \frac{Q_i^{(N)}(0)}{N \log N} < y \right) \right| \leq \frac{c_t}{N \sqrt{\log N}}. \quad (4.20)$$

Now, because of the facts that we assume the convergence result in (4.17), and

$$\frac{\sqrt{\alpha t(1-\alpha/N)}}{\sqrt{\log N}} \frac{\sqrt{[tN^3 \log N]}}{\sqrt{tN^3 \log N}} X_i = \frac{\sqrt{\alpha t} X_i}{\sqrt{\log N}} + o_N \left(\frac{1}{\sqrt{\log N}} \right) X_i,$$

it is easy to see that

$$\max_{i \leq N} \left(\frac{\sqrt{\alpha t(1-\alpha/N)}}{\sqrt{\log N}} \frac{\sqrt{[tN^3 \log N]}}{\sqrt{tN^3 \log N}} X_i + \frac{Q_i^{(N)}(0)}{N \log N} \right) \xrightarrow{\mathbb{P}} g(t, q(0)) \text{ as } N \rightarrow \infty.$$

Let $\epsilon > 0$, then because of the bound given in (4.20), and the convergence result in (4.17),

$$\begin{aligned} & \mathbb{P} \left(\max_{i \leq N} \left(\frac{\tilde{S}_i^{(N)}(tN^3 \log N)}{\log N} + \frac{Q_i^{(N)}(0)}{N \log N} \right) < g(t, q(0)) - \epsilon \right) \\ &= \mathbb{P} \left(\frac{\tilde{S}_i^{(N)}(tN^3 \log N)}{\log N} + \frac{Q_i^{(N)}(0)}{N \log N} < g(t, q(0)) - \epsilon \right)^N \\ &\leq \mathbb{P} \left(\frac{\sqrt{\alpha t(1-\alpha/N)}}{\sqrt{\log N}} \frac{\sqrt{[tN^3 \log N]}}{\sqrt{tN^3 \log N}} X_i + \frac{Q_i^{(N)}(0)}{N \log N} < g(t, q(0)) - \epsilon \right)^N + \left(\frac{c_t}{N \sqrt{\log N}} + 1 \right)^N - 1 \\ &\xrightarrow{N \rightarrow \infty} 0. \end{aligned}$$

The proof that

$$\mathbb{P} \left(\max_{i \leq N} \left(\frac{\tilde{S}_i^{(N)}(tN^3 \log N)}{\log N} + \frac{Q_i^{(N)}(0)}{N \log N} \right) > g(t, q(0)) + \epsilon \right) \xrightarrow{N \rightarrow \infty} 0,$$

goes analogously. Hence, the lemma follows. \square

Proof of Theorem 4.1 In Lemmas 4.5 and 4.9 we have proven that both parts in the maximum in (2.10) converge to a limit. The lemma follows. \square

We can easily extend this result to finite-dimensional distributions.

THEOREM 4.2 (The finite-dimensional distributions converge). *If*

$$X^{(N)}(t) \xrightarrow{\mathbb{P}} f(t)$$

for all $t > 0$, then for (t_1, t_2, \dots, t_k)

$$(X^{(N)}(t_1), X^{(N)}(t_2), \dots, X^{(N)}(t_k)) \xrightarrow{\mathbb{P}} (f(t_1), f(t_2), \dots, f(t_k)) \text{ as } N \rightarrow \infty.$$

Proof

$$\begin{aligned} & \mathbb{P} \left(\left\| (X^{(N)}(t_1), X^{(N)}(t_2), \dots, X^{(N)}(t_k)) - (f(t_1), f(t_2), \dots, f(t_k)) \right\| > \epsilon \right) \\ &\leq \mathbb{P} \left(|X^{(N)}(t_1) - f(t_1)| + \dots + |X^{(N)}(t_k) - f(t_k)| > \epsilon \right) \\ &\leq \mathbb{P} \left(|X^{(N)}(t_1) - f(t_1)| > \frac{\epsilon}{k} \right) + \dots + \mathbb{P} \left(|X^{(N)}(t_k) - f(t_k)| > \frac{\epsilon}{k} \right) \xrightarrow{N \rightarrow \infty} 0, \end{aligned}$$

with $\|\cdot\|$ the Euclidean distance in \mathbb{R}^k . \square

4.4. Tightness. It is known that when a sequence of random processes is tight and its finite-dimensional distributions converge, then this sequence converges u.o.c., cf. [5, Thm. 7.1, p. 80]. From [5, Thm. 7.3, p. 82], we know that a process $(X^{(N)}(t), t \in [0, T])$ is tight when for all positive η there exists an a and an integer N_0 such that for all $N \geq N_0$

$$\mathbb{P}(|X^{(N)}(0)| > a) \leq \eta, \quad (4.21)$$

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

$$\frac{1}{\delta} \mathbb{P}\left(\sup_{t \leq s \leq t+\delta} |X^{(N)}(s) - X^{(N)}(t)| > \epsilon\right) \leq \eta. \quad (4.22)$$

The conditions given in Equations (4.21) and (4.22) hold for stochastic processes in the space of continuous functions. The process $(Q_{(\alpha, \beta)}^{(N)}(tN^3 \log N)/(N \log N), t \in [0, T])$ does not lie in this space, because $Q_{(\alpha, \beta)}^{(N)}(n) = Q_{(\alpha, \beta)}^{(N)}(\lfloor n \rfloor)$. However, since $q(t)$ is a continuous function, the conditions in (4.21) and (4.22) do also apply on $(Q_{(\alpha, \beta)}^{(N)}(tN^3 \log N)/(N \log N), t \in [0, T])$, cf. [5, Cor. 13.4, p. 142].

In order to prove tightness for the process given in Theorem 2.1, we need to prove tightness of the maximum of two processes, as Equation (2.10) shows. In Lemma 4.10, we show that it suffices to prove tightness of the two processes separately. Then, in Lemmas 4.11 and 4.12, we prove the tightness of the two parts.

LEMMA 4.10. *Assume that $(X^{(N)}(s), s \in [0, t])$ and $(Y^{(N)}(s), s \in [0, t])$ converge to functions $(k(s), s \in [0, t])$ and $(l(s), s \in [0, t])$ u.o.c., respectively, then $(\max(X^{(N)}(s), Y^{(N)}(s)), s \in [0, t])$ converges to $(\max(k(s), l(s)), s \in [0, t])$ u.o.c.*

Proof The lemma holds because of the fact that

$$\begin{aligned} & \mathbb{P}\left(\sup_{0 \leq s \leq t} |\max(X^{(N)}(s), Y^{(N)}(s)) - \max(k(s), l(s))| > \epsilon\right) \\ & \leq \mathbb{P}\left(\sup_{0 \leq s \leq t} (\max(X^{(N)}(s), Y^{(N)}(s)) - \max(k(s), l(s))) > \epsilon\right) \\ & \quad + \mathbb{P}\left(\sup_{0 \leq s \leq t} (\max(k(s), l(s)) - \max(X^{(N)}(s), Y^{(N)}(s))) > \epsilon\right) \\ & \leq \mathbb{P}\left(\sup_{0 \leq s \leq t} \max(X^{(N)}(s) - k(s), Y^{(N)}(s) - l(s)) > \epsilon\right) \\ & \quad + \mathbb{P}\left(\sup_{0 \leq s \leq t} \max(k(s) - X^{(N)}(s), l(s) - Y^{(N)}(s)) > \epsilon\right) \\ & \leq 2\mathbb{P}\left(\sup_{0 \leq s \leq t} |X^{(N)}(s) - k(s)| > \epsilon\right) + 2\mathbb{P}\left(\sup_{0 \leq s \leq t} |Y^{(N)}(s) - l(s)| > \epsilon\right) \xrightarrow{N \rightarrow \infty} 0. \end{aligned}$$

□

LEMMA 4.11 (Tightness of the first part). *For $\epsilon > 0$, $\eta > 0$, $T > 0$ and $Q_{(\alpha, \beta)}^{(N)}(0) = 0$, $\exists 0 < \delta < 1$ and an integer N_0 such that $\forall N \geq N_0$ and $t \in [0, T]$*

$$\frac{1}{\delta} \mathbb{P}\left(\sup_{t \leq s \leq t+\delta} \left| \frac{Q_{(\alpha, \beta)}^{(N)}(sN^3 \log N)}{N \log N} - \frac{Q_{(\alpha, \beta)}^{(N)}(tN^3 \log N)}{N \log N} \right| \geq \epsilon\right) \leq \eta. \quad (4.23)$$

Proof We take $t > 0$. From Lemma 4.1, and the fact that $\tilde{R}_i^{(N)}$ is a random walk that satisfies the duality principle, we know that for N large enough,

$$\frac{1}{\delta} \mathbb{P} \left(\sup_{t \leq s \leq t+\delta} \left| \frac{Q_{(\alpha,\beta)}^{(N)}(sN^3 \log N)}{N \log N} - \frac{Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)}{N \log N} \right| \geq \epsilon \right) \quad (4.24)$$

$$\leq \frac{1}{\delta} \mathbb{P} \left(\sup_{0 \leq s \leq \delta} \max_{i \leq N} \tilde{R}_i^{(N)}(sN^3 \log N) + 2 \sup_{0 \leq s \leq \delta} \max_{i \leq N} -\tilde{R}_i^{(N)}(sN^3 \log N) \geq \epsilon \right) + o_N(1) \quad (4.25)$$

$$\leq \frac{1}{\delta} \mathbb{P} \left(\sup_{0 \leq s \leq \delta} \max_{i \leq N} \tilde{R}_i^{(N)}(sN^3 \log N) \geq \frac{\epsilon}{2} \right) + \frac{1}{\delta} \mathbb{P} \left(2 \sup_{0 \leq s \leq \delta} \max_{i \leq N} -\tilde{R}_i^{(N)}(sN^3 \log N) \geq \frac{\epsilon}{2} \right) + o_N(1). \quad (4.26)$$

Now we focus on the first term in (4.26). The analysis of the second term goes analogously.

$$\frac{1}{\delta} \mathbb{P} \left(\sup_{0 \leq s \leq \delta} \max_{i \leq N} \tilde{R}_i^{(N)}(sN^3 \log N) \geq \frac{\epsilon}{2} \right) \quad (4.27)$$

$$= \frac{1}{\delta} \mathbb{P} \left(\sup_{0 \leq s \leq \delta} \max_{i \leq N} \frac{\tilde{A}^{(N)}(sN^3 \log N) + \tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} \geq \frac{\epsilon}{2} \right) \quad (4.28)$$

$$\leq \frac{1}{\delta} \mathbb{P} \left(\sup_{0 \leq s \leq \delta} \frac{\tilde{A}^{(N)}(sN^3 \log N)}{\log N} \geq \frac{\epsilon}{4} \right) + \frac{1}{\delta} \mathbb{P} \left(\sup_{0 \leq s \leq \delta} \max_{i \leq N} \frac{\tilde{S}_i^{(N)}(sN^3 \log N)}{\log N} \geq \frac{\epsilon}{4} \right). \quad (4.29)$$

In the proof of Lemma 4.6, we already showed that the second term in (4.29) is small. With a similar proof as in Lemma 4.6, one can also prove that the first term is small. Concluding, $(Q_{(\alpha,\beta)}^{(N)}(tN^3 \log N)/(N \log N), t \in [0, T])$ is tight, when $Q_{(\alpha,\beta)}^{(N)}(0) = 0$. \square

LEMMA 4.12 (Tightness of the second part). *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 $t \in [0, T]$*

$$\frac{1}{\delta} \mathbb{P} \left(\sup_{t \leq s \leq t+\delta} \left| \max_{i \leq N} \frac{A^{(N)}(sN^3 \log N) - S_i^{(N)}(sN^3 \log N) + Q_i^{(N)}(0)}{N \log N} - \max_{i \leq N} \frac{A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N) + Q_i^{(N)}(0)}{N \log N} \right| > \epsilon \right) < \eta. \quad (4.30)$$

Furthermore, for all η there exists an $a > 0$ such that

$$\mathbb{P} \left(\frac{Q_{(\alpha,\beta)}^{(N)}(0)}{N \log N} > a \right) < \eta. \quad (4.31)$$

Proof First of all, we observe that for a random variable X , $\mathbb{P}(|X| > \epsilon) \leq \mathbb{P}(X > \epsilon) + \mathbb{P}(-X > \epsilon)$. Thus, we can remove the absolute values in (4.30) and examine both cases. Since both cases satisfy analogous proofs, we only write down the proof for the first case.

$$\begin{aligned} & \frac{1}{\delta} \mathbb{P} \left(\sup_{t \leq s \leq t+\delta} \left(\max_{i \leq N} \frac{A^{(N)}(sN^3 \log N) - S_i^{(N)}(sN^3 \log N) + Q_i^{(N)}(0)}{N \log N} \right. \right. \\ & \quad \left. \left. - \max_{i \leq N} \frac{A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N) + Q_i^{(N)}(0)}{N \log N} \right) > \epsilon \right) \\ & \leq \\ & \frac{1}{\delta} \mathbb{P} \left(\sup_{t \leq s \leq t+\delta} \left(\max_{i \leq N} \left(\frac{A^{(N)}(sN^3 \log N) - S_i^{(N)}(sN^3 \log N)}{N \log N} \right) \right) \right) \end{aligned}$$

$$\begin{aligned}
& - \frac{A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N)}{N \log N} > \epsilon \Big) \\
& = \frac{1}{\delta} \mathbb{P} \left(\sup_{0 \leq s \leq \delta} \left(\max_{i \leq N} \frac{A^{(N)}(sN^3 \log N) - S_i^{(N)}(sN^3 \log N)}{N \log N} \right) > \epsilon \right) + o_N(1).
\end{aligned}$$

This is the same expression as Equation (4.28). In Lemma 4.11, it is proven that this expression will be small. At $t = 0$, we should choose $a > 0$ such that (4.31) holds for $N \geq N_0$. This is the case, because we know that $Q_{(\alpha, \beta)}^{(N)}(0)/(N \log N) \xrightarrow{\mathbb{P}} q(0)$ as $N \rightarrow \infty$. The lemma follows. \square

COROLLARY 4.1 (Tightness of the process). *The process $(Q_{(\alpha, \beta)}^{(N)}(tN^3 \log N)/(N \log N), t \in [0, T])$ is tight.*

Proof The process $(Q_{(\alpha, \beta)}^{(N)}(tN^3 \log N)/(N \log N), t \in [0, T])$ can be written as a maximum of two processes. In Lemmas 4.11 and 4.12 it is proven that these processes are tight. Then from Lemma 4.10 it follows that $(Q_{(\alpha, \beta)}^{(N)}(tN^3 \log N)/(N \log N), t \in [0, T])$ is tight. \square

Proof of Theorem 2.1 In Theorem 4.1, we proved that for fixed t , the stochastic process converges in probability to a constant, in Theorem 4.2, we proved that the finite-dimensional distributions converge and in Corollary 4.1, we showed that the process is tight. Thus the convergence holds u.o.c. \square

We now prove that the scaled process in steady state converges to the constant $\alpha/(2\beta)$.

Proof of Proposition 2.2. Since we look at the system in steady state, we can assume w.l.o.g. that $Q_{(\alpha, \beta)}^{(N)}(0) = 0$. Then, we have

$$\frac{Q_{(\alpha, \beta)}^{(N)}(\infty)}{N \log N} \geq_{st.} \frac{Q_{(\alpha, \beta)}^{(N)}(\alpha/(2\beta^2)N^3 \log N)}{N \log N},$$

because $Q_{(\alpha, \beta)}^{(N)}(n) \stackrel{d}{=} \max_{i \leq N} \sup_{0 \leq k \leq n} (A^{(N)}(k) - S_i^{(N)}(k))$. We know by Lemma 4.5 that

$$\frac{Q_{(\alpha, \beta)}^{(N)}(\alpha/(2\beta^2)N^3 \log N)}{N \log N} \xrightarrow{\mathbb{P}} \frac{\alpha}{2\beta} \text{ as } N \rightarrow \infty.$$

Furthermore, we know by Lemma 4.7 that for all $\delta > 0$,

$$\limsup_{N \rightarrow \infty} \mathbb{P} \left(\frac{Q_{(\alpha, \beta)}^{(N)}(\infty)}{N \log N} > \frac{\alpha}{2\beta} + \delta \right) = 0.$$

The proposition follows. \square

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

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

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

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

When we set $x_N = 1/N$, we get $f(x_N, \theta(x_N)) = \mathbb{E} \left[e^{\theta_A^{(u,N)} X^{(u,N)}(j)} \right]$. 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

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

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

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

and

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

Because we know that $f(x, \theta(x)) = 1$, we solve $3\theta'(0) (\alpha\theta'(0) - 2m) = 0$. 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

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

This gives that $\theta''(0) = 4m(3\alpha^2 - 3\beta + 2m)/3\alpha^2$. In general, we can compute each derivative of $\theta(0)$ iteratively. This gives

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

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 $\theta_A^{(u,N)} = 2m/(\alpha N) + O(1/N^2)$.

Appendix B: Extreme values of sums of random variables. In this section, we prove a convergence result of the maximum of N sums of n random variables. In order to do so, we use and extend results from [7] and [8].

LEMMA B.1. *Consider sequences of continuous random variables $(Y_i^{(1)}, i \geq 1)$, $(Y_i^{(2)}, i \geq 1)$, \dots , $(Y_i^{(k)}, i \geq 1)$, where all random variables in the sequence $(Y_i^{(j)}, i \geq 1)$ are identically and independently distributed and have infinite right endpoints. Furthermore, $Y_i^{(j)}$ and $Y_m^{(l)}$ are independent for all $j, l \in \{1, 2, \dots, k\}$ and $i, m \geq 1$, and $Y_i^{(j)}$ satisfies Assumption 2.5 with function $h^{(j)}(u^{(j)})$. Finally, we have sequences $(a_N^{(j)}, N \geq 1)$ such that $\mathbb{P}(Y_i^{(j)} \geq a_N^{(j)}) = 1/N$. We assume that the random variables $Y_i^{(j)}$ are relatively stable, thus $\max_{i \leq N} Y_i^{(j)}/a_N^{(j)} \xrightarrow{\mathbb{P}} 1$, as $N \rightarrow \infty$. Then*

$$\max_{i \leq N} \left(\sum_{j=1}^k \frac{Y_i^{(j)}}{a_N^{(j)}} \right) \xrightarrow{\mathbb{P}} \sup_{(u^{(j)}, j \leq k)} \left\{ \sum_{j=1}^k u^{(j)} \mid \sum_{j=1}^k h^{(j)}(u^{(j)}) \leq 1, u^{(j)} \leq 1 \forall j \leq k \right\} \text{ as } N \rightarrow \infty.$$

Proof First of all, let us choose $u^{(1)}, \dots, u^{(k)}$ such that $u^{(j)} \leq 1$ for all j . It is a well-known result that

$$\mathbb{P}\left(\bigcup_{i=1}^N \bigcap_{j=1}^k \left\{Y_i^{(j)} \geq u^{(j)} a_N^{(j)}\right\}\right) \xrightarrow{N \rightarrow \infty} 1 \iff N \mathbb{P}\left(\bigcap_{j=1}^k \left\{Y_i^{(j)} \geq u^{(j)} a_N^{(j)}\right\}\right) \xrightarrow{N \rightarrow \infty} \infty.$$

From this, it follows that

$$\log N + \sum_{j=1}^k \log \left(\mathbb{P}\left(Y_i^{(j)} \geq u^{(j)} a_N^{(j)}\right)\right) \xrightarrow{N \rightarrow \infty} \infty.$$

This is the case when

$$\limsup_{N \rightarrow \infty} \left(\sum_{j=1}^k \frac{-\log \left(\mathbb{P}\left(Y_i^{(j)} \geq u^{(j)} a_N^{(j)}\right)\right)}{\log N} \right) < 1.$$

Similarly,

$$\liminf_{N \rightarrow \infty} \left(\sum_{j=1}^k \frac{-\log \left(\mathbb{P}\left(Y_i^{(j)} \geq u^{(j)} a_N^{(j)}\right)\right)}{\log N} \right) > 1 \Rightarrow \mathbb{P}\left(\bigcup_{i=1}^N \bigcap_{j=1}^k \left\{Y_i^{(j)} \geq u^{(j)} a_N^{(j)}\right\}\right) \xrightarrow{N \rightarrow \infty} 0.$$

Because of the fact that we have $\mathbb{P}\left(Y_i^{(j)} \geq a_N^{(j)}\right) = 1/N$, we can conclude that

$$\lim_{N \rightarrow \infty} \left(\sum_{j=1}^k \frac{-\log \left(\mathbb{P}\left(Y_i^{(j)} \geq u^{(j)} a_N^{(j)}\right)\right)}{\log N} \right) = \lim_{N \rightarrow \infty} \left(\sum_{j=1}^k \frac{-\log \left(\mathbb{P}\left(Y_i^{(j)} \geq u^{(j)} a_N^{(j)}\right)\right)}{-\log \left(\mathbb{P}\left(Y_i^{(j)} \geq a_N^{(j)}\right)\right)} \right) = \sum_{j=1}^k h^{(j)}(u^{(j)}).$$

Let us now call

$$c^* = \sup_{(u^{(j)}, j \leq k)} \left\{ \sum_{j=1}^k u^{(j)} \mid \sum_{j=1}^k h^{(j)}(u^{(j)}) \leq 1, u^{(j)} \leq 1 \forall j \leq k \right\},$$

and let $\epsilon > 0$ be small. Then, we distinguish two scenarios. First of all, we consider the case that $|\{1, \dots, k \mid h^{(j)}(u^{(j)}) = \mathbb{1}(u^{(j)} > 0)\}| \leq k - 2$. Then, there exists a sequence $(u_\epsilon^{(1)}, \dots, u_\epsilon^{(k)})$ such that $\sum_{j=1}^k u_\epsilon^{(j)} = c^* - \epsilon$, and $\sum_{j=1}^k h^{(j)}(u_\epsilon^{(j)}) < 1$. Therefore,

$$\mathbb{P}\left(\max_{i \leq N} \left(\sum_{j=1}^k \frac{Y_i^{(j)}}{a_N^{(j)}} \right) > c^* - \epsilon\right) > \mathbb{P}\left(\bigcup_{i=1}^N \bigcap_{j=1}^k \left\{Y_i^{(j)} \geq u_\epsilon^{(j)} a_N^{(j)}\right\}\right) \xrightarrow{N \rightarrow \infty} 1.$$

If $|\{1, \dots, k \mid h^{(j)}(u^{(j)}) = \mathbb{1}(u^{(j)} > 0)\}| \geq k - 1$, we know that $c^* = 1$, and we know that

$$\max_{i \leq N} \left(\sum_{j=1}^k \frac{Y_i^{(j)}}{a_N^{(j)}} \right) \geq_{st.} \max_{i \leq N} \left(\frac{Y_i^{(1)}}{a_N^{(1)}} \right) + \sum_{j=2}^k \frac{Y_i^{(j)}}{a_N^{(j)}} \xrightarrow{\mathbb{P}} 1 \text{ as } N \rightarrow \infty.$$

Thus, at this moment we can conclude that the limit cannot be smaller than c^* . To prove that

$$\mathbb{P}\left(\max_{i \leq N} \left(\sum_{j=1}^k \frac{Y_i^{(j)}}{a_N^{(j)}} \right) > c^* + \epsilon\right) \xrightarrow{N \rightarrow \infty} 0, \tag{B.1}$$

we first observe that the boundary is given by $\{(u^{(j)}, j \leq k) \mid \sum_{j=1}^k u^{(j)} = c^* + \epsilon\}$. We already know that $N \mathbb{P}\left(Y_i^{(j)} > u^{(j)} a_N\right) \xrightarrow{N \rightarrow \infty} 0$, for $u^{(j)} > 1$. Hence,

$$\limsup_{N \rightarrow \infty} \mathbb{P}\left(\max_{i \leq N} \left(\sum_{j=1}^k \frac{Y_i^{(j)}}{a_N^{(j)}}\right) > c^* + \epsilon\right) > 0,$$

means that there are limiting points in the set $\{(u^{(j)}, j \leq k) \mid \sum_{j=1}^k u^{(j)} = c^* + \epsilon, u^{(j)} \leq 1 \forall j\}$. However, we know that for all $(u^{(j)}, j \leq k)$ with $c^* < \sum_{j=1}^k u^{(j)} < c^* + \epsilon$ that $N \mathbb{P}(\cap_{j=1}^k \{Y_i^{(j)} \geq u^{(j)} a_N^{(j)}\}) \xrightarrow{N \rightarrow \infty} 0$. Thus, we know that there are no limiting points in the positive quadrants with starting points $(u^{(1)}, \dots, u^{(k)})$ with $c^* < \sum_{j=1}^k u^{(j)} < c^* + \epsilon$. The union of a finite number of quadrants covers the set $\{(u^{(j)}, j \leq k) \mid \sum_{j=1}^k u^{(j)} = c^* + \epsilon, u^{(j)} \leq 1 \forall j\}$. For example, in the case that $k = 2$,

$$\{(u, v) \mid u + v \geq c^* + \epsilon, u \in [0, 1], v \in [0, 1]\} \subset \cup_{m=1}^{\lceil 2/\epsilon - 1/2 \rceil + 1} \left\{ (u, v) \mid u \geq c^* - 1 + \frac{m\epsilon}{2}, v \geq 1 + \frac{\epsilon}{4} - \frac{m\epsilon}{2} \right\}.$$

For $k > 2$, an analogous proof can be given. Hence, the limit in (B.1) and the lemma follows. \square

Appendix C: Proof of Lemmas 4.1, 4.2, 4.3, and 4.4.

Proof of Lemma 4.1. We take $s > t > 0$. We write $t^{(N)} = tN^3 \log N$, $s^{(N)} = sN^3 \log N$, etc. We first prove that for $s^{(N)} > t^{(N)}$, the following upper bound holds:

$$\begin{aligned} \frac{Q_{(\alpha, \beta)}^{(N)}(s^{(N)})}{N \log N} - \frac{Q_{(\alpha, \beta)}^{(N)}(t^{(N)})}{N \log N} &\leq \max_{i \leq N} \left| \tilde{R}_i^{(N)}(s^{(N)}) - \tilde{R}_i^{(N)}(t^{(N)}) \right| \\ &\quad + \max_{i \leq N} \sup_{t^{(N)} \leq r \leq s^{(N)}} \left(\tilde{R}_i^{(N)}(t^{(N)}) - \tilde{R}_i^{(N)}(r) \right). \end{aligned} \quad (\text{C.1})$$

Due to the defined auxiliary processes in Definition 4.1, we can write the maximum queue length in terms of $\tilde{R}_i^{(N)}$ as in Equation (4.5). Similarly, we can rewrite $Q_{(\alpha, \beta)}^{(N)}(s^{(N)})/(N \log N) - Q_{(\alpha, \beta)}^{(N)}(t^{(N)})/(N \log N)$ as

$$\begin{aligned} &\max_{i \leq N} \sup_{0 \leq r \leq s^{(N)}} \left(\tilde{R}_i^{(N)}(s^{(N)}) - \tilde{R}_i^{(N)}(r) \right) - \max_{i \leq N} \sup_{0 \leq u \leq t^{(N)}} \left(\tilde{R}_i^{(N)}(t^{(N)}) - \tilde{R}_i^{(N)}(u) \right) \\ &= \max_{i \leq N} \left[\tilde{R}_i^{(N)}(s^{(N)}) - \tilde{R}_i^{(N)}(t^{(N)}) + \sup_{0 \leq r \leq s^{(N)}} \left(\tilde{R}_i^{(N)}(t^{(N)}) - \tilde{R}_i^{(N)}(r) \right) \right] \\ &\quad - \max_{i \leq N} \sup_{0 \leq u \leq t^{(N)}} \left(\tilde{R}_i^{(N)}(t^{(N)}) - \tilde{R}_i^{(N)}(u) \right). \end{aligned}$$

Now, the following upper bounds for $Q_{(\alpha, \beta)}^{(N)}(s^{(N)})/(N \log N) - Q_{(\alpha, \beta)}^{(N)}(t^{(N)})/(N \log N)$ hold:

$$\begin{aligned} &\frac{Q_{(\alpha, \beta)}^{(N)}(s^{(N)})}{N \log N} - \frac{Q_{(\alpha, \beta)}^{(N)}(t^{(N)})}{N \log N} \\ &\leq \max_{i \leq N} \left(\tilde{R}_i^{(N)}(s^{(N)}) - \tilde{R}_i^{(N)}(t^{(N)}) \right) + \max_{i \leq N} \sup_{0 \leq r \leq s^{(N)}} \left(\tilde{R}_i^{(N)}(t^{(N)}) - \tilde{R}_i^{(N)}(r) \right) \\ &\quad - \max_{i \leq N} \sup_{0 \leq u \leq t^{(N)}} \left(\tilde{R}_i^{(N)}(t^{(N)}) - \tilde{R}_i^{(N)}(u) \right) \\ &\leq \max_{i \leq N} \left(\tilde{R}_i^{(N)}(s^{(N)}) - \tilde{R}_i^{(N)}(t^{(N)}) \right) \\ &\quad + \max_{i \leq N} \left[\sup_{0 \leq r \leq s^{(N)}} \left(\tilde{R}_i^{(N)}(t^{(N)}) - \tilde{R}_i^{(N)}(r) \right) - \sup_{0 \leq u \leq t^{(N)}} \left(\tilde{R}_i^{(N)}(t^{(N)}) - \tilde{R}_i^{(N)}(u) \right) \right]. \end{aligned}$$

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

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

Now, we can conclude that

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

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

$$\frac{Q_{(\alpha, \beta)}^{(N)}(s^{(N)})}{N \log N} - \frac{Q_{(\alpha, \beta)}^{(N)}(t^{(N)})}{N \log N} \geq - \max_{i \leq N} \left| \tilde{R}_i^{(N)}(t^{(N)}) - \tilde{R}_i^{(N)}(s^{(N)}) \right|. \quad (\text{C.2})$$

To show this, we write

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

Observe that

$$\sup_{0 \leq r \leq s^{(N)}} \left(\tilde{R}_i^{(N)}(s^{(N)}) - \tilde{R}_i^{(N)}(r) \right) \geq \sup_{0 \leq u \leq t^{(N)}} \left(\tilde{R}_i^{(N)}(s^{(N)}) - \tilde{R}_i^{(N)}(u) \right),$$

because $s^{(N)} > t^{(N)}$, 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

$$\begin{aligned} & \frac{Q_{(\alpha, \beta)}^{(N)}(s^{(N)})}{N \log N} - \frac{Q_{(\alpha, \beta)}^{(N)}(t^{(N)})}{N \log N} \\ & \geq \max_{i \leq N} \sup_{0 \leq r \leq s^{(N)}} \left(\tilde{R}_i^{(N)}(s^{(N)}) - \tilde{R}_i^{(N)}(r) \right) \\ & - \max_{i \leq N} \left(\tilde{R}_i^{(N)}(t^{(N)}) - \tilde{R}_i^{(N)}(s^{(N)}) \right) - \max_{i \leq N} \sup_{0 \leq u \leq t^{(N)}} \left(\tilde{R}_i^{(N)}(s^{(N)}) - \tilde{R}_i^{(N)}(u) \right) \\ & \geq - \max_{i \leq N} \left(\tilde{R}_i^{(N)}(t^{(N)}) - \tilde{R}_i^{(N)}(s^{(N)}) \right) \geq - \max_{i \leq N} \left| \tilde{R}_i^{(N)}(t^{(N)}) - \tilde{R}_i^{(N)}(s^{(N)}) \right|, \end{aligned}$$

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

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

Thus,

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

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

$$\begin{aligned} \sup_{t^{(N)} \leq s \leq t^{(N)} + \delta^{(N)}} \max_{i \leq N} \left| \tilde{R}_i^{(N)}(s) - \tilde{R}_i^{(N)}(t^{(N)}) \right| & \leq \sup_{t^{(N)} \leq s \leq t^{(N)} + \delta^{(N)}} \max_{i \leq N} \left(\tilde{R}_i^{(N)}(s) - \tilde{R}_i^{(N)}(t^{(N)}) \right) \\ & + \sup_{t^{(N)} \leq s \leq t^{(N)} + \delta^{(N)}} \max_{i \leq N} \left(\tilde{R}_i^{(N)}(t^{(N)}) - \tilde{R}_i^{(N)}(s) \right). \quad (\text{C.4}) \end{aligned}$$

Combining the inequalities in (C.3) and (C.4) gives us the desired result. \square

Proof of Lemma 4.2. $\tilde{S}_i^{(N)}(n)$ is a sum of independent and identically distributed random variables with $\mathbb{E} \left[\pm \tilde{S}_i^{(N)}(1) \right] = 0$, and $\text{Var} \left(\pm \tilde{S}_i^{(N)}(1) \right) = (1 - \alpha/N) \alpha / N^3$. So, $\pm M_i^{(N)}(t) = \pm \tilde{S}_i^{(N)}(tN^3 \log N) / \sqrt{tN^3 \log N} / \sqrt{\alpha t(1 - \alpha/N) \log N} / \sqrt{tN^3 \log N}$ has mean 0 and variance 1, and satisfies the central limit theorem. From [16] it follows that for all y ,

$$\left| \mathbb{P} \left(\pm M_i^{(N)}(t) < y \right) - \Phi(y) \right| \leq C \frac{1}{\sqrt{[tN^3 \log N]}} \mathbb{E} \left[\left| \frac{\pm \tilde{S}_i^{(N)}(1)}{\sqrt{\alpha t(1 - \alpha/N) \log N}} \sqrt{tN^3 \log N} \right|^3 \right] \frac{1}{1 + |y|^3}.$$

Observe that for N large enough and $0 < \epsilon < t$, $[tN^3 \log N] > (t - \epsilon)N^3 \log N$. We also have that

$$\begin{aligned} & \mathbb{E} \left[\left| \frac{\pm \tilde{S}_i^{(N)}(1)}{\sqrt{\alpha t(1 - \alpha/N) \log N}} \sqrt{tN^3 \log N} \right|^3 \right] \\ & = \frac{N^4 \sqrt{N}}{\alpha(1 - \alpha/N) \sqrt{\alpha(1 - \alpha/N)} N^3} \left(\left(1 - \frac{\alpha}{N}\right)^3 \frac{\alpha}{N} + \frac{\alpha^3}{N^3} \left(1 - \frac{\alpha}{N}\right) \right) \leq 2\sqrt{N} \frac{(1 + \alpha^2)}{\sqrt{\alpha}}, \end{aligned}$$

which holds for $N > \max(1, 2\alpha)$. Thus, the statement of Lemma 4.2 follows for N large enough, with $c_t = 2C(1 + \alpha^2) / \sqrt{\alpha(t - \epsilon)}$. \square

Proof of Lemma 4.3. We have

$$\mathbb{E} \left[\max \left(0, \frac{\max_{i \leq N} \pm \tilde{S}_i^{(N)}(tN^3 \log N)}{\log N} \right)^{5/2} \right] = \int_0^\infty \mathbb{P} \left(\frac{\max_{i \leq N} \pm \tilde{S}_i^{(N)}(tN^3 \log N)}{\log N} > x^{2/5} \right) dx$$

$$\begin{aligned}
&= \int_0^\infty \mathbb{P} \left(\max_{i \leq N} \pm M_i^{(N)}(t) > x^{2/5} \frac{\log N}{\sqrt{\alpha t (1 - \alpha/N) \log N}} \frac{\sqrt{t N^3 \log N}}{\sqrt{\lfloor t N^3 \log N \rfloor}} \right) dx \\
&= \int_0^\infty 1 - \mathbb{P} \left(\pm M_i^{(N)}(t) < \frac{x^{2/5} \sqrt{N^3 \log N}}{\sqrt{\alpha (1 - \alpha/N) \lfloor t N^3 \log N \rfloor}} \right)^N dx \\
&\leq \int_0^\infty 1 - \left(\Phi \left(\frac{x^{2/5} \sqrt{N^3 \log N}}{\sqrt{\alpha (1 - \alpha/N) \lfloor t N^3 \log N \rfloor}} \right) - \frac{c_t}{N \sqrt{\log N}} \frac{1}{1 + \left(\frac{x^{2/5} \sqrt{N^3 \log N}}{\sqrt{\alpha (1 - \alpha/N) \lfloor t N^3 \log N \rfloor}} \right)^3} \right)^N dx \\
&\leq \int_0^\infty -\Phi \left(\frac{x^{2/5} \sqrt{N^3 \log N}}{\sqrt{\alpha (1 - \alpha/N) \lfloor t N^3 \log N \rfloor}} \right)^N + \left(1 + \frac{c_t}{N \sqrt{\log N}} \frac{1}{1 + \left(\frac{x^{2/5} \sqrt{N^3 \log N}}{\sqrt{\alpha (1 - \alpha/N) \lfloor t N^3 \log N \rfloor}} \right)^3} \right)^N dx \\
&= \mathbb{E} \left[\max \left(0, \frac{\sqrt{\alpha t (1 - \alpha/N)} \max_{i \leq N} X_i}{\sqrt{\log N}} \frac{\sqrt{\lfloor t N^3 \log N \rfloor}}{\sqrt{t N^3 \log N}} \right)^{5/2} \right] \tag{C.5} \\
&+ \int_0^\infty -1 + \left(1 + \frac{c_t}{N \sqrt{\log N}} \frac{1}{1 + \left(\frac{x^{2/5} \sqrt{N^3 \log N}}{\sqrt{\alpha (1 - \alpha/N) \lfloor t N^3 \log N \rfloor}} \right)^3} \right)^N dx, \tag{C.6}
\end{aligned}$$

with X_i standard normally distributed. By Pickands [20, Thm. 3.2, p. 888], we know that the expectation in (C.5) converges to $(2\alpha t)^{5/4}$. Furthermore, the term in (C.6) is upper bounded by

$$\int_0^\infty -1 + \exp \left(\frac{c_t}{\sqrt{\log N} \left(1 + \left(\frac{x^{2/5} \sqrt{N^3 \log N}}{\sqrt{\alpha (1 - \alpha/N) \lfloor t N^3 \log N \rfloor}} \right)^3 \right)} \right) dx. \tag{C.7}$$

We substitute $y = 1 / \left(1 + \left(\frac{x^{2/5} \sqrt{N^3 \log N}}{\sqrt{\alpha (1 - \alpha/N) \lfloor t N^3 \log N \rfloor}} \right)^3 \right)$, then the term in (C.7) can be rewritten as

$$\left(\frac{\lfloor t N^3 \log N \rfloor}{N^3 \log N (\log N)} \right)^{5/4} \int_0^1 \frac{5(\sqrt{\alpha(1-\alpha/N)})^{5/2}}{6(1-y)^{1/6} y^{11/6}} \left(-1 + e^{\frac{c_t}{\sqrt{\log N}} y} \right) dy \xrightarrow{N \rightarrow \infty} 0.$$

The lemma follows. \square

Proof of Lemma 4.4 In order to prove that

$$\max_{i \leq N} \left(\frac{\sqrt{\alpha t} X_i}{\sqrt{\log N}} + \frac{Q_i^{(N)}(0)}{N \log N} \right) \xrightarrow{\mathbb{P}} g(t, q(0)),$$

we first observe that, from the definition of $Q_i^{(N)}(0)$ in Theorem 2.1, it is easy to see that

$$\left| \max_{i \leq N} \left(\frac{\sqrt{\alpha t} X_i}{\sqrt{\log N}} + \frac{Q_i^{(N)}(0)}{N \log N} \right) - \max_{i \leq N} \left(\frac{\sqrt{\alpha t} X_i}{\sqrt{\log N}} + \frac{r_N U_i}{N \log N} \right) \right| \leq \max_{i \leq N} \frac{V_i^{(N)}}{N \log N} + \frac{1}{N \log N} \xrightarrow{\mathbb{P}} 0$$

as $N \rightarrow \infty$. Thus, from this it follows that

$$\max_{i \leq N} \left(\frac{\sqrt{\alpha t} X_i}{\sqrt{\log N}} + \frac{Q_i^{(N)}(0)}{N \log N} \right) \xrightarrow{\mathbb{P}} g(t, q(0)) \iff \max_{i \leq N} \left(\frac{\sqrt{\alpha t} X_i + \frac{r_N U_i}{N \sqrt{\log N}}}{\sqrt{\log N}} \right) \xrightarrow{\mathbb{P}} g(t, q(0)) \text{ as } N \rightarrow \infty.$$

Let us first consider that U_i satisfies Assumption 2.4, thus U_i has a finite right endpoint. Theorem 2.1 says that when U_i has a finite right endpoint, that $g(t, q(0)) = \sqrt{2\alpha t} + q(0)$. To prove this, first observe that $g(t, q(0)) \leq \sqrt{2\alpha t} + q(0)$ because $\max_{i \leq N} \sqrt{\alpha t} X_i / \sqrt{\log N} \xrightarrow{\mathbb{P}} \sqrt{2\alpha t}$ and $Q_{(\alpha, \beta)}^{(N)}(0) / (N \log N) \xrightarrow{\mathbb{P}} q(0)$. Hence, the only thing we need to establish is that for all $\gamma < \sqrt{2\alpha t} + q(0)$,

$$N \mathbb{P} \left(\sqrt{\alpha t} X_i + \frac{r_N U_i}{N \sqrt{\log N}} \geq \gamma \sqrt{\log N} \right) \xrightarrow{N \rightarrow \infty} \infty.$$

When $\gamma < \sqrt{2\alpha t}$, this is obvious, because $U_i > 0$, and $\max_{i \leq N} \sqrt{\alpha t} X_i / \sqrt{\log N} \xrightarrow{\mathbb{P}} \sqrt{2\alpha t}$. So, let us assume that $\sqrt{2\alpha t} \leq \gamma < \sqrt{2\alpha t} + q(0)$. Because U_i has a finite right endpoint, $r_N / (N \sqrt{\log N}) = \sqrt{\log N}$. By convolution, we have that

$$\begin{aligned} & N \mathbb{P} \left(\sqrt{\alpha t} X_i + \sqrt{\log N} U_i \geq \gamma \sqrt{\log N} \right) \\ &= N \mathbb{P} \left(\sqrt{\alpha t} X_i \geq \gamma \sqrt{\log N} \right) + N \int_{-\infty}^{\gamma \sqrt{\log N}} \mathbb{P} \left(\sqrt{\log N} U_i > \gamma \sqrt{\log N} - z \right) \frac{e^{-z^2/(2\alpha t)}}{\sqrt{2\alpha t \pi}} dz \\ &\geq N \int_{-\infty}^{\gamma} \mathbb{P}(U_i > \gamma - v) \frac{N^{-v^2/(2\alpha t)}}{\sqrt{2\alpha t \pi}} \sqrt{\log N} dv = \int_{\gamma - q(0)}^{\gamma} \mathbb{P}(U_i > \gamma - v) \frac{N^{1-v^2/(2\alpha t)}}{\sqrt{2\alpha t \pi}} \sqrt{\log N} dv. \end{aligned}$$

From this it follows, that when $1 - v^2/(2\alpha t) > 0$, this integral converges to ∞ . We chose $\sqrt{2\alpha t} \leq \gamma < \sqrt{2\alpha t} + q(0)$, thus the lower bound $\gamma - q(0)$ in the integral is smaller than $\sqrt{2\alpha t}$ and hence this integral converges to ∞ . Thus $g(t, q(0)) = \sqrt{2\alpha t} + q(0)$.

Let us now consider the scenario described in Assumption 2.5. Then $g(t, q(0))$ satisfies the limit given in (2.7). We have the straightforward limit result that for standard normally distributed X_i , $\lim_{t \rightarrow \infty} -\log(\mathbb{P}(X_i \geq ut)) / -\log(\mathbb{P}(X_i \geq t)) = u^2$. Furthermore, following the assumptions on U_i in Theorem 2.1, we know that $\lim_{t \rightarrow \infty} -\log(\mathbb{P}(U_i \geq vt)) / -\log(\mathbb{P}(U_i \geq t)) = h(v)$. Thus from Lemma B.1, we know that for sequences $(a_N, N \geq 1), (b_N, N \geq 1)$ with $\mathbb{P}(X_i \geq a_N) = \mathbb{P}(U_i \geq b_N) = 1/N$, that

$$\max_{i \leq N} \left(\frac{X_i}{a_N} + \frac{U_i}{b_N} \right) \xrightarrow{\mathbb{P}} \sup_{(u, v)} \{u + v | u^2 + h(v) \leq 1, 0 \leq u \leq 1, 0 \leq v \leq 1\} \text{ as } N \rightarrow \infty.$$

Now, we can use this result to prove that $\max_{i \leq N} (\sqrt{\alpha t} X_i / \sqrt{\log N} + r_N U_i / (N \log N))$ converges to the limit in (2.7). We first observe that

$$\max_{i \leq N} \left(\frac{\sqrt{\alpha t} X_i}{\sqrt{\log N}} + \frac{r_N U_i}{N \log N} \right) = \max_{i \leq N} \left(\sqrt{2\alpha t} \frac{X_i}{\sqrt{2 \log N}} + q(0) \frac{r_N U_i}{q(0) N \log N} \right).$$

We have that $a_N / \sqrt{2 \log N} \xrightarrow{N \rightarrow \infty} 1$, because $\max_{i \leq N} X_i / a_N \xrightarrow{\mathbb{P}} 1$, and $\max_{i \leq N} X_i / \sqrt{2 \log N} \xrightarrow{\mathbb{P}} 1$ as $N \rightarrow \infty$. Analogously, $b_N q(0) N \log N / r_N \xrightarrow{N \rightarrow \infty} 1$. Thus,

$$\left| \max_{i \leq N} \left(\sqrt{2\alpha t} \frac{X_i}{a_N} + q(0) \frac{U_i}{b_N} \right) - \max_{i \leq N} \left(\frac{\sqrt{\alpha t} X_i}{\sqrt{\log N}} + \frac{r_N U_i}{N \log N} \right) \right| \xrightarrow{\mathbb{P}} 0 \text{ as } N \rightarrow \infty.$$

With an analogous proof as before, $\max_{i \leq N} (\sqrt{2\alpha t} X_i / a_N + q(0) U_i / b_N)$ converges to the limit in (2.7). \square

Appendix D: Notation.

- N : the number of servers.
- $A^{(N)}(n)$: the number of arrivals up to time $\lfloor n \rfloor$.
- $X^{(N)}(n)$: Bernoulli random variable indicating a potential arrival at time $n \in \mathbb{N}$.
- $S_i^{(N)}(n)$: the number of finished services of server i up to time $\lfloor n \rfloor$.
- $Y_i^{(N)}(n)$: Bernoulli distributed random variable indicating a potential completed service at server i at time $n \in \mathbb{N}$.
- α, β are system parameters.
- $p^{(N)}$: the arrival probability, $p^{(N)} = 1 - \alpha/N - \beta/N^2$.
- $q^{(N)}$: the service probability, $q^{(N)} = 1 - \alpha/N$.
- $Q_{(\alpha, \beta)}^{(N)}(n)$: the maximum queue length at time $\lfloor n \rfloor$.
- $Q_i^{(N)}(0)$: the number of tasks at time 0 at queue i , $Q_i^{(N)}(0) = U_i^{(N)} + V_i^{(N)}$.
- $U_i^{(N)}$: the independent part of the number of tasks at time 0 at queue i , $U_i^{(N)} = \lfloor r_N U_i \rfloor$.
- $V_i^{(N)}$: the dependent part of the number of tasks at time 0 at queue i .
- U_i : continuously distributed and positive random variable.
- r_N : positive scaling sequence.
- $h(v) = \lim_{t \rightarrow \infty} -\log(\mathbb{P}(U_i > vt)) / -\log(\mathbb{P}(U_i > t))$.
- $q(t)$: fluid limit of the process.
- $g(t, q(0))$: limit of $\max_{i \leq N} (A^{(N)}(tN^3 \log N) - S_i^{(N)}(tN^3 \log N) + Q_i^{(N)}(0)) / (N \log N)$.
- $\tilde{R}_i^{(N)}(n) = (\tilde{A}^{(N)}(n) + \tilde{S}_i^{(N)}(n)) / \log N$.
- $\tilde{A}^{(N)}(n) = A^{(N)}(n) / N - (1 - \alpha/N) \lfloor n \rfloor / N$.
- $\tilde{S}_i^{(N)}(n) = -S_i^{(N)}(n) / N + (1 - \alpha/N) \lfloor n \rfloor / N$.
- $M_i^{(N)}(t) = \tilde{S}_i^{(N)}(tN^3 \log N) \sqrt{tN^3 \log N} / (\sqrt{\alpha t(1 - \alpha/N) \log N} \sqrt{\lfloor tN^3 \log N \rfloor})$.
- $A^{(u, N)}(n) = \sum_{j=1}^n X^{(u, N)}(j)$.
- $X^{(u, N)}(j)$:

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

with $0 < m < \beta$.

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

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

- $\theta_A^{(u, N)}$ solves:

$$\mathbb{E} \left[e^{\theta_A^{(u, N)} X^{(u, N)}(j)} \right] = 1.$$

- $E_i^{(u, N)} \sim \text{Exp}(2(\beta - m)/(\alpha N))$.

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]. We thank the referees for prompting us to investigate the system where the number of jobs at time 0 increases with N . We thank Prof. Balkema for referring us to literature on convergence of samples.

References

- [1] Anderson CW, Coles SG, Hüsler J, et al. (1997) Maxima of Poisson-like variables and related triangular arrays. *Ann. Appl. Probab.* 7(4):953–971.
- [2] Atar R, Mandelbaum A, Zviran A (2012) Control of fork-join networks in heavy traffic. *Proc. 50th Allerton Conf. Comm., Control Comput.* (IEEE, Piscataway, NJ), 823–830.
- [3] Baccelli F (1985) Two parallel queues created by arrivals with two demands: The M/G/2 symmetrical case. Technical report RR-0426, INRIA, Montbonnot-Saint-Martin, France.
- [4] Baccelli F, Makowski AM (1989) Queueing models for systems with synchronization constraints. *Proc. of the IEEE* 77(1):138–161.
- [5] Billingsley P (1968) *Convergence of probability measures*, 2nd ed. (John Wiley & Sons, New York).
- [6] Brown BM, Resnick SI (1977) Extreme values of independent stochastic processes. *J. Appl. Probab.* 14(4):732–739.
- [7] Davis RA, Mulrow E, Resnick SI (1988) Almost sure limit sets of random samples in \mathbb{R}^d . *Adv. Appl. Probab.* 20(3):573–599.
- [8] Fisher L (1969) Limiting sets and convex hulls of samples from product measures. *Ann. Math. Statist.* 40(5):1824–1832.
- [9] Flatto L, Hahn S (1984) Two parallel queues created by arrivals with two demands I. *SIAM J. Appl. Math.* 44(5):1041–1053.
- [10] Harrison JM (1985) *Brownian Motion and Stochastic Flow Systems* (John Wiley & Sons, New York).
- [11] de Klein SJ (1988) *Fredholm integral equations in queueing analysis*. Ph.D. thesis, Rijksuniversiteit Utrecht.
- [12] Ko SS, Serfozo RF (2004) Response times in M/M/s fork-join networks. *Adv. Appl. Probab.* 36(3):854–871.
- [13] Lu H, Pang G (2015) Gaussian limits for a fork-join network with nonexchangeable synchronization in heavy traffic. *Math. Oper. Res.* 41(2):560–595.
- [14] Lu H, Pang G (2017) Heavy-traffic limits for a fork-join network in the Halfin-Whitt regime. *Stochastic Systems* 6(2):519–600.
- [15] 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.
- [16] Michel R (1976) On the constant in the nonuniform version of the Berry-Esséen theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 55(1):109–117.
- [17] Nelson R, Tantawi AN (1988) Approximate analysis of fork/join synchronization in parallel queues. *IEEE Trans. Comput.* 37(6):739–743.
- [18] Nguyen V (1993) Processing networks with parallel and sequential tasks: Heavy traffic analysis and Brownian limits. *Ann. Appl. Probab.* 3(1):28–55.
- [19] Nguyen V (1994) The trouble with diversity: Fork-join networks with heterogeneous customer population. *Ann. Appl. Probab.* 4(1):1–25.
- [20] Pickands III J (1968) Moment convergence of sample extremes. *Ann. Math. Statist.* 39(3):881–889.
- [21] Sigman K, Whitt W (2011) Heavy-traffic limits for nearly deterministic queues. *J. Appl. Probab.* 48(3):657–678.
- [22] Sigman K, Whitt W (2011) Heavy-traffic limits for nearly deterministic queues: stationary distributions. *Queueing Systems* 69(2):145.
- [23] Varma S (1990) *Heavy and light traffic approximations for queues with synchronization constraints*. Ph.D. thesis, University of Maryland.
- [24] Wright PE (1992) Two parallel processors with coupled inputs. *Adv. Appl. Probab.* 24(4):986–1007.