J. Appl. Prob. **44,** 306–320 (2007)
Printed in England

TAIL ASYMPTOTICS FOR MONOTONE-SEPARABLE NETWORKS

MARC LELARGE,* *University College Cork*

Abstract

A network belongs to the monotone separable class if its state variables are homogeneous and monotone functions of the epochs of the arrival process. This framework contains several classical queueing network models, including generalized Jackson networks, max-plus networks, polling systems, multiserver queues, and various classes of stochastic Petri nets. We use comparison relationships between networks of this class with independent and identically distributed driving sequences and the GI/GI/1/1 queue to obtain the tail asymptotics of the stationary maximal dater under light-tailed assumptions for service times. The exponential rate of decay is given as a function of a logarithmic moment generating function. We exemplify an explicit computation of this rate for the case of queues in tandem under various stochastic assumptions.

Keywords: Large deviation; queueing network

2000 Mathematics Subject Classification: Primary 60F10
Secondary 60K25

1. Introduction

Consider the GI/GI/1 single server queue; let $X_n = \sigma_n - \tau_n$, where $\{\sigma_n\}$ and $\{\tau_n\}$ are independent and identically distributed (i.i.d.) nonnegative random variables, σ_n is the amount of service customer n receives, and τ_n is the interarrival time between customer n and $n+1$. Assume that $\mathrm{E}[X_1] < 0$, then the supremum of the random walk $S_n = X_1 + \cdots + X_n$ defined by $M := \sup_{n\geq 1} S_n$ is finite almost surely and has the same distribution as the stationary workload of the single server queue. If we assume moreover that $\mathrm{E}[\exp(\varepsilon X_1)] < \infty$, for some $\varepsilon > 0$, then the following asymptotics is standard:

$$\lim_{x\to\infty} \frac{1}{x} \log \mathrm{P}(M > x) = -\theta^*, \tag{1}$$

where $\theta^* = \sup\{\theta > 0, \log \mathrm{E}[\exp(\theta X_1)] < 0\}$. Motivated by queueing applications, this case has been extensively studied in the literature and much finer estimates are available; see [10] and [12]. The main goal of this paper is to derive analogous results to (1) for networks.

In the context of a network, we consider the maximal dater Z which is the time to empty the network when stopping further arrivals. Clearly in the single server queue, the maximal dater corresponds to the workload. In the case of queues in tandem, it corresponds to the end-to-end delay. Theorem 2, below, gives the logarithmic tail asymptotics for the maximal dater of a monotone separable network. The main difficulty in our task is the absence of closed form formula for Z. The proof of Theorem 2 will proceed by deriving upper and lower bounds for monotone separable networks. This class, which was introduced in [3], contains several

Received 19 July 2006; revision received 16 March 2007.
* Current address: ENS-INRIA, 45 rue d'Ulm, 75005 Paris, France. Email address: marc.lelarge@ens.fr
This work was partially supported by Science Foundation Ireland Research Grant No. SFI 04/RP1/I512.

classical queueing network models like generalized Jackson networks, max-plus networks, polling systems, and multiserver queues. In this paper, we choose to put a particular emphasis on tandem queues that fall in the class of open Jackson networks, and in the class of open (max, +) systems which both belong to the class of monotone separable networks. It serves as a pedagogical example to apply Theorem 2 under various stochastic assumptions and it enables us to link our results with existing asymptotics results from queueing literature.

The paper is structured as follows. In Section 2, we give the precise definition of a monotone separable network and its associated maximal dater. We then give the main result of this paper in Section 2.2. The case of queues in tandem is dealt with in great detail in Section 3. In particular, we show that a kind of phase transition is possible when service times at both stations are dependent. We also link our result to the literature. Finally, technical proofs are deferred to Section 4.

2. Tail asymptotics for monotone-separable networks

In this paper, we consider open stochastic networks with a single input process N, which is a marked point process with points $\{T_n\}$ corresponding to exogenous arrival times and marks $\{\zeta_n\}$ which describe the service times and routing decisions.

More precisely, a stochastic network is described by the following framework (introduced in [3]).

- The network has a single input point process N, with points $\{T_n\}$. That is, for all $m \leq n \in \mathbb{Z}$, let $N_{[m,n]}$ be the $[m, n]$-restriction of N, namely the point process with points $\{T_\ell\}_{m\leq\ell\leq n}$.
- The network has almost sure (a.s.) finite activity for all finite restrictions of N. That is, for all $m \leq n \in \mathbb{Z}$, let $X_{[m,n]}(N)$ be the time of last activity in the network, when this starts empty and is fed by $N_{[m,n]}$. We assume that for all finite m and n as above, $X_{[m,n]}(N)$ is finite.

We assume that there exists a set of functions $\{f_\ell\}$, $f_\ell \colon \mathbb{R}^\ell \times K^\ell \to \mathbb{R}$, such that

$$X_{[m,n]}(N) = f_{n-m+1}\{(T_\ell, \zeta_\ell),\ m \leq \ell \leq n\}, \tag{2}$$

for all n, m, and $N = \{T_n\}$, where the sequence $\{\zeta_n\}$ is that describing service times and routing decisions.

Example 1. Consider a G/G/1/$\infty \to \cdot$/G/1/∞ tandem queue. Denote by $\{\sigma_n^{(i)}\}$ the sequence of service times in station $i = 1, 2$, and denote by $N = \{T_n\}$ the sequence of arrival times at the first station. With the notation introduced above, we have $\zeta_n = (\sigma_n^{(1)}, \sigma_n^{(2)})$, and the time of last activity is given by

$$X_{[m,n]}(N) = \sup_{m\leq k\leq n} \left\{ T_k + \sup_{k\leq i\leq n} \sum_{j=k}^{i} \sigma_j^{(1)} + \sum_{j=i}^{n} \sigma_j^{(2)} \right\}. \tag{3}$$

We refer to Appendix A for an explicit derivation of (3). Here, $X_{[m,n]}(N)$ is simply the last departure time from the network, when only customers $m, m + 1, \ldots, n$ enter the network.

https://doi.org/10.1239/jap/1183667403 Published online by Cambridge University Press

We say that a network described as above is monotone separable if the functions f_n are such that the following properties hold for all input point process N.

1. (Causality.) For all $m \le n$,
$$X_{[m,n]}(N) \ge T_n.$$
2. (External monotonicity.) For all $m \le n$,
$$X_{[m,n]}(N') \ge X_{[m,n]}(N),$$
whenever $N' := \{T'_n\}$ is such that $T'_n \ge T_n$ for all n.
3. (Homogeneity.) For all $c \in \mathbb{R}$ and for all $m \le n$,
$$X_{[m,n]}(N + c) = X_{[m,n]}(N) + c,$$
where $N + c$ is the point process with points $\{T_n + c\}$.
4. (Separability.) For all $m \le \ell < n$, if $X_{[m,\ell]}(N) \le T_{\ell+1}$ then
$$X_{[m,n]}(N) = X_{[\ell+1,n]}(N).$$

We should stress that these four properties are properties satisfied by the functions f_n which define the dynamic of the network. In particular, no stochastic assumption has been made at this stage and so previous properties will hold almost surely in the stochastic framework described in the sequel. Note that the external monotonicity and the homogeneity properties will be valid for any random delay on the T_ns or random shift c (see [2] for more on this).

Remark 1. Clearly, tandem queues belong to the class of monotone-separable networks.

2.1. Stability and stationary maximal daters

In this section, we introduce stochastic assumptions ensuring the stability of the network. More general results can be found in [3] and we refer to it for the statements given in this section without proof.

By definition, for $m \le n$, the $[m, n]$ maximal dater is
$$Z_{[m,n]}(N) := X_{[m,n]}(N) - T_n.$$
Note that $Z_{[m,n]}(N)$ is a function of $\{\zeta_l\}_{m\le\ell\le n}$ and $\{\tau_l\}_{m\le\ell\le n}$ only, where $\tau_n = T_{n+1} - T_n$. In particular, $Z_n := Z_{[n,n]}(N)$ is not a function of N (which makes the notation consistent). When dealing with the maximal dater, we do not loose any generality if we assume that $T_0 = 0$.

Lemma 1. (Internal monotonicity of X and Z [3].) *Under the above conditions, the variables $X_{[m,n]}$ and $Z_{[m,n]}$ satisfy the following internal monotonicity property. For all N and all $m \le n$, we have*
$$X_{[m-1,n]}(N) \ge X_{[m,n]}(N),$$
$$Z_{[m-1,n]}(N) \ge Z_{[m,n]}(N).$$

In particular, the sequence $\{Z_{[-n,0]}(N)\}$ is nondecreasing in n. We define the *stationary maximal dater* as
$$Z := Z_{(-\infty,0]}(N) = \lim_{n\to\infty} Z_{[-n,0]}(N) \le \infty.$$

https://doi.org/10.1239/jap/1183667403 Published online by Cambridge University Press

Example 2. In the case of the tandem queues, the stationary maximal dater is given by

$$Z = \sup_{p \le q \le 0} \left\{ \sum_{k=p}^{q} \sigma_k^{(1)} + \sum_{k=q}^{0} \sigma_k^{(2)} - (T_0 - T_p) \right\}, \tag{4}$$

and Z is the stationary end-to-end delay of the network.

Lemma 2. (Subadditive property of Z [3].) *Under the above conditions, $\{Z_{[m,n]}(N)\}$ satisfies the following subadditive property. For all $m \le \ell < n$ and all N, we have*

$$Z_{[m,n]}(N) \le Z_{[m,\ell]}(N) + Z_{[\ell+1,n]}(N).$$

We assume that the sequence $\{\tau_n, \zeta_n\}_n$ is a sequence of i.i.d. random variables. The following integrability assumptions are also assumed to hold (recall that $Z_n = Z_{[n,n]}(N)$ does not depend on N):

$$\mathrm{E}[\tau_n] := a < \infty, \qquad \mathrm{E}[Z_n] < \infty.$$

Denote by $N^0 = \{T_n^0\}$ the degenerate input process with $T_n^0 = 0$ for all n. This degenerate point process plays a crucial role for the derivation of the stability condition. The following lemma follows from Lemma 2 in which we take the input point process to be N^0 (note that the constant γ defined below is denoted $\gamma(0)$ in [3] to emphasize the fact that the input point process is N^0).

Lemma 3. ([3].) *Under the foregoing stochastic assumption, there exists a nonnegative constant γ such that*

$$\lim_{n\to\infty} \frac{Z_{[-n,0]}(N^0)}{n} = \lim_{n\to\infty} \frac{\mathrm{E}[Z_{[-n,0]}(N^0)]}{n} = \gamma \quad a.s.$$

We now present the main result on the stability region.

Theorem 1. ([3].) *Under the foregoing stochastic assumptions, either $Z = \infty$ a.s. or $Z < \infty$ a.s.*

(a) *If $\gamma < a$ then $Z < \infty$ a.s.*

(b) *If $Z < \infty$ a.s. then $\gamma \le a$.*

A proof is given in Section 4.1, where we derive an upper bound and a lower bound that will be used for the study of large deviations.

Example 3. In the case of tandem queues, the constant γ is easy to compute. We have

$$\lim_{n\to\infty} \sup_{-n \le q \le 0} \frac{\sum_{k=-n}^{q} \sigma_k^{(1)} + \sum_{k=q}^{0} \sigma_k^{(2)}}{n} = \max(\mathrm{E}[\sigma_1^{(1)}], \mathrm{E}[\sigma_1^{(2)}]).$$

Hence, Theorem 1 gives the following standard stability condition: $\max(\mathrm{E}[\sigma_1^{(1)}], \mathrm{E}[\sigma_1^{(2)}]) < \mathrm{E}[\tau_1]$.

https://doi.org/10.1239/jap/1183667403 Published online by Cambridge University Press

2.2. Moment generating function and tail asymptotics

In the rest of the paper, we will make the following assumptions.

Assumption 1. *We assume that the arrival process into the network $\{T_n\}$ is a renewal process independent of the service time and routing sequences $\{\zeta_n\}$.*

Assumption 2. *The sequence $\{\zeta_n\}$ is a sequence of i.i.d. random variables, such that the random variable Z_0 is light tailed, i.e. for θ in a neighbourhood of 0,*

$$\mathrm{E}[\mathrm{e}^{\theta Z_0}] < \infty.$$

Assumption 3. *For stability, we assume that $\gamma < a = \mathrm{E}[T_1 - T_0]$ (see Theorem 1).*

The subadditive property of Z directly implies the following property (which is proved in Lemma 6).

Property 1. For any monotone separable network that satisfies Assumption 2, the following limit exists in $\mathbb{R} \cup \{+\infty\}$ for all θ:

$$\Lambda_Z(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathrm{E}[\exp(\theta Z_{[1,n]}(N^0))]. \tag{5}$$

Note that the subadditive property of Z is valid regardless of the point process N (see Lemma 2). Like in the study of the stability of the network, it turns out that the right quantity to look at is $Z_{[m,n]}(N^0)$, where N^0 is the degenerate input point process with all its point equal to 0. We also define

$$\Lambda_T(\theta) = \log \mathrm{E}[\exp(\theta(T_1 - T_0))].$$

Theorem 2. *Under previous assumptions, the tail asymptotics of the stationary maximal dater is given by*

$$\lim_{x\to\infty} \frac{1}{x} \log \mathrm{P}(Z > x) = -\theta^* < 0,$$

where $\theta^* = \sup\{\theta > 0, \Lambda_T(-\theta) + \Lambda_Z(\theta) < 0\}$.

It is relatively easy to see that, under our light-tailed assumption, the stationary maximal dater Z will be light tailed (see [4, Corollary 3]). The main contribution of Theorem 2 is to give an explicit way of computing the rate of decay of the tail distribution of Z. We refer the interested reader to [11] for more details on the computation of Λ_Z in the case of $(\max, +)$-linear networks. In Section 3, we continue the study of our example and deal with the case of queues in tandem under various stochastic assumptions. The case we study allows us to show a phase transition phenomena and to compare Theorem 2 with results of the literature.

Note that, in the context of heavy-tailed asymptotics, the moment generating function is infinite for all $\theta > 0$. There is no general result for the tail asymptotics of the maximal dater of a monotone separable network. However, the methodology derived in [4] for subexponential distributions allows to get exact asymptotics for $(\max, +)$-linear networks [6] and generalized Jackson networks [5].

3. A case study: queues in tandem

3.1. The impact of dependence

We continue our Example 2 and 3 and consider a stable $\mathrm{G}/\mathrm{G}/1/\infty \to \cdot/\mathrm{G}/1/\infty$ tandem queue where $\{\sigma_n^{(i)}\}$ is the sequence of service times in station i, where $i = 1, 2$ and $\{\tau_n\}$ is the

https://doi.org/10.1239/jap/1183667403 Published online by Cambridge University Press

sequence of interarrival times at the first station. We assume that the sequences $\{\sigma_n^{(1)}, \sigma_n^{(2)}\}$ and $\{\tau_n\}_n$ are sequences of i.i.d. random variables such that $\gamma = \max(\mathrm{E}[\sigma_1^{(1)}], \mathrm{E}[\sigma_1^{(2)}]) < \mathrm{E}[\tau_1]$.

We consider the following two cases.

Case 1. The sequences $\{\sigma_n^{(1)}\}$, $\{\sigma_n^{(2)}\}$, and $\{\tau_n\}$ are independent.

Case 2. The sequences $\{\sigma_n^{(1)}\}$ and $\{\tau_n\}$ are independent and we have $\sigma_n^{(2)} = \sigma_n^{(1)}$.

We let $\Lambda_i(\theta) = \log \mathrm{E}[\exp(\theta\sigma_1^{(i)})]$ and $\delta = \sup\{\theta \geq 0, \mathrm{E}[\exp(\theta\sigma_1^{(1)})] < \infty\}$. A direct application of Theorem 2 gives an extension of the results of [9].

Corollary 1. *The tail asymptotics of the stationary end-to-end delay for two queues in tandem is given by*

$$\lim_{x\to\infty} \frac{1}{x} \log \mathrm{P}(Z > x) = -\theta^*,$$

where in case 1

$$\theta^* = \min(\theta^1, \theta^2), \quad \textit{with } \theta^i = \sup\{\theta > 0, \Lambda_i(\theta) + \Lambda_T(-\theta) < 0\},$$

and in case 2

$$\theta^* = \min\left(\theta^1, \frac{\delta}{2}\right).$$

In case 1, θ^i is the rate of exponential decay for the tail distribution of the stationary workload of a single server queue with interarrival time τ_n and service time $\sigma_n^{(i)}$, and we have $\theta^* = \min(\theta^1, \theta^2)$. It is well known that the stability of such a network is constrained by the 'slowest' component. Here, we see that in a large deviations regime, the 'bad' behaviour of the network is due to a 'bottleneck' component (which is not necessarily the same as the 'slowest' component in average). Note that, in the particular case where the random variables $\sigma_n^{(1)}$, $\sigma_n^{(2)}$, and τ_n are exponentially distributed with means $1/\mu^1$, $1/\mu^2$, and a, respectively, we have $\theta^i = \mu^i - a^{-1}$, and in this case the 'slowest' component in average is also the 'bottleneck' component in the large deviations regime.

In the case where the service times are the same at both stations, Corollary 1 shows that the tail behaviour of the random variable $\sigma_1^{(1)}$ described by δ matters. To simplify this and to get a parametric model, we assume that the arrival process is Poisson with intensity $\lambda := a^{-1}$ and the service times are exponentially distributed with mean $1/\mu$. Then, depending on the intensity of the arrival process λ, the following two situations may occur:

$$\lambda \leq \frac{\mu}{2} \Rightarrow \theta^* = \frac{\mu}{2},$$
$$\lambda > \frac{\mu}{2} \Rightarrow \theta^* = \mu - \lambda.$$

These situations can be expressed as follows.

1. If $\lambda < \mu/2$ then the tail asymptotics of the end-to-end delay is the same as the total service requirement of a single customer.
2. If $\lambda > \mu/2$ then the tail asymptotics of the end-to-end delay is the same as in the independent case.

https://doi.org/10.1239/jap/1183667403 Published online by Cambridge University Press

This shows that the behaviour of tandems differs from that of a single server queue. In particular, it was shown in [1] that, for GI/GI/1 queues, the buildup of large delays can happen in one of the following two ways.

- If the service times have exponential tails, then it involves a large number of customers (whose interarrival and service times differ from their mean values).
- If the service times do not have exponential tails, then large delays are caused by the arrival of a single customer with large service requirement.

We see that the first behaviour is still valid for queues in tandem when the service times are independent at each station or if the intensity of the arrival process is sufficiently large. In contrast, when the service times are the same at both stations, we see that a single customer can create large delays in the network even under the assumption of exponential service times (if the intensity of arrivals is sufficiently small). Note that this phenomena is rather simple and results intrinsically from the fact that the network considered is of dimension greater than 2 (i.e. we cannot get such a phenomena with a single server queue).

Proof of Corollary 1. Recall that we have

$$Z_{[1,n]}(N^0) = \sup_{1\le k\le n} \sum_{i=1}^{k} \sigma_i^{(1)} + \sum_{i=k}^{n} \sigma_i^{(2)}.$$

In case 1, we have

$$\begin{aligned}\log \mathrm{E}[\exp(\theta Z_{[1,n]}(N^0))] &\le \log\left(\sum_{k=1}^{n} \exp(k\Lambda_1(\theta) + (n-k)\Lambda_2(\theta))\right)\\ &\le \log n + n\max(\Lambda_1(\theta), \Lambda_2(\theta)).\end{aligned}$$

Hence, we have $\Lambda_Z(\theta) = \max(\Lambda_1(\theta), \Lambda_2(\theta))$, and the result follows.

In case 2, we have

$$Z_{[1,n]}(N^0) = \sum_{i=1}^{n} \sigma_i^{(1)} + \max_i \sigma_i^{(1)} = \max_i \left(2\sigma_i^{(1)} + \sum_{j\ne i} \sigma_j^{(1)}\right);$$

hence, we have

$$\begin{aligned}&\log \mathrm{E}[\exp(\theta Z_{[1,n]}(N^0))] \ge \max(n\Lambda_1(\theta), \Lambda_1(2\theta)),\\ &\log \mathrm{E}[\exp(\theta Z_{[1,n]}(N^0))] \le (n-1)\Lambda_1(\theta) + \log n + \Lambda_1(2\theta).\end{aligned}$$

It follows that

$$\Lambda_Z(\theta) = \begin{cases} \Lambda_1(\theta), & \theta < \dfrac{\delta}{2},\\ \infty, & \theta > \dfrac{\delta}{2}, \end{cases}$$

which completes the proof.

https://doi.org/10.1239/jap/1183667403 Published online by Cambridge University Press

3.2. Comparison with the literature

In the context of two queues in tandem, if we define

$$Y_n = \sup_{-n \le q \le 0} \sum_{k=-n}^{q} \sigma_k^{(1)} + \sum_{k=q}^{0} \sigma_k^{(2)} - (T_0 - T_{-n}),$$

then, in view of (4), we have $Z = \sup_n Y_n$. The supremum of a stochastic process has been extensively studied in queueing theory but we do not know of any general results that would allow us to derive Corollary 1. To end this section and to make the connection with the existing literature, we state the following result.

Corollary 2. *Consider the system of queues in tandem descried above. Under the assumptions of Theorem 2 and if*

1. *the sequence $\{Y_n/n\}$ satisfies a large deviation principle (LDP) with a good rate function I,*
2. *there exists $\varepsilon > 0$ such that $\Lambda_Z(\theta^* + \varepsilon) < \infty$, where θ^* is defined as in Theorem 2,*

we have

$$\lim_{x \to \infty} \frac{1}{x} \log \mathrm{P}(Z > x) = -\theta^* = -\inf_{\alpha > 0} \frac{I(\alpha)}{\alpha}. \tag{6}$$

This kind of result has been extensively studied in the queueing literature (see [8]). However, we see that considering the moment generating function instead of the rate function allows us to get a more general result than (6) since we do not require assumption (2) of Corollary 2. Indeed, this assumption ensures that the tail asymptotics of $\mathrm{P}(Y_n > nc)$ for a single value of n cannot dominate those of $\mathrm{P}(Z > x)$. In this case, (6) has a nice interpretation: the natural drift of the process Y_n is μn (where $\mu < 0$). The quantity $I(\alpha)$ can be seen as the cost for changing the drift of this process to $\alpha > 0$. Now, in order to reach level x, this drift has to last for a time x/α. Hence, the total cost for reaching level x with drift α is $xI(\alpha)/\alpha$ and the process naturally choose the drift with the minimal associated cost. As already discussed, this heuristic argument is valid only if assumption (2) of Corollary 2 holds. Note also that in our framework, we do not assume any LDP to hold for the sequence $\{Y_n/n\}$. In particular, as shown by Corollary 1, the computation of the moment generating function Λ_Z is much easier than deriving a LDP for $\{Y_n/n\}$. Lastly, we should stress that for general monotone separable networks, the maximal dater Z cannot be expressed as the supremum of a simple stochastic process, in which case the derivation of the tail asymptotics of Z requires new techniques.

Proof of Corollary 2. We have only to show that $\theta^* = \inf_{\alpha>0}(I(\alpha)/\alpha)$. Using Varadhan's integral lemma (see [7, Theorem 4.3.1]), we have

$$\lim_{n \to \infty} \frac{1}{n} \log \mathrm{E}[\exp(\theta Y_n)] = \Lambda(\theta) = \sup_x \{\theta x - I(x)\},$$

for $\theta < \theta^* + \varepsilon$, where $\Lambda(\theta) = \Lambda_Z(\theta) + \Lambda_T(-\theta)$. Then, for $\theta > 0$, the corollary follows from the following observations:

$$\begin{aligned} \theta < \inf_{\alpha > 0} \frac{I(\alpha)}{\alpha} &\iff \theta\alpha - I(\alpha) < 0, \quad \text{for all } \alpha, \\ &\iff \sup_\alpha \{\theta\alpha - I(\alpha)\} = \Lambda(\theta) < 0. \end{aligned}$$

https://doi.org/10.1239/jap/1183667403 Published online by Cambridge University Press

4. Proof of the tail asymptotics

4.1. Upper G/G/1/∞ queue and lower bound for the maximal dater

The material of this subsection is not new and may be found in various references (that are given in what follows). For the sake of completeness, we include all the proofs. We now derive upper and lower bounds for the stationary maximal dater Z. These bounds allow us to prove Theorem 1 and will be the main tools for the study of large deviations.

We first derive a lower bound that can also be found in [2, proof of Theorem 2.11.3].

Proposition 1. *We have the following lower bound:*

$$Z \geq \sup_{n\geq 0}(Z_{[-n,0]}(N^0) + T_{-n} - T_0).$$

Proof. For fixed n, let N^n be the point process with point $T_j^n = T_{-n} - T_0$, for all j. Then

$$\begin{aligned} Z_{[-n,0]} &= X_{[-n,0]}(N) - T_0 \\ &\geq X_{[-n,0]}(N^n) \\ &= X_{[-n,0]}(N^0) + T_{-n} - T_0 \\ &= Z_{[-n,0]}(N^0) + T_{-n} - T_0, \end{aligned}$$

where we used external monotonicity in the first inequality and homogeneity in the second equality.

Proof of Theorem 1(b). Suppose that $\gamma > a$; then we have

$$\liminf_{n\to\infty} \frac{Z_{[-n,0]}(N)}{n} \geq \gamma - a > 0,$$

which concludes the proof.

We assume now that $\gamma < a$. We pick an integer $L \geq 1$ such that

$$\mathrm{E}[Z_{[-L,-1]}(N^0)] < La, \tag{7}$$

which is possible in view of Lemma 3. Without loss of generality, we assume that $T_0 = 0$. Theorem 1(a) follows from the following proposition.

Proposition 2. ([4].) *The stationary maximal dater Z is bounded from above by the stationary response time $\hat{R}$ in the* G/G/1/∞ *queue with service times*

$$\hat{s}_n := Z_{[L(n-1)+1,Ln]}(N^0)$$

and interarrival times $\hat{\tau}_n := T_{Ln} - T_{L(n-1)}$, where L is the integer defined in (7). Since $\mathrm{E}[\hat{s}_1] < \mathrm{E}[\hat{\tau}_1] = La$, this queue is stable. With the convention $\sum_0^{-1} = 0$, we have

$$Z \leq \hat{s}_0 + \sup_{k\geq 0} \sum_{i=-k}^{-1} (\hat{s}_i - \hat{\tau}_{i+1}).$$

https://doi.org/10.1239/jap/1183667403 Published online by Cambridge University Press

To prove Proposition 2, we will need the following two lemmas.

Lemma 4. *Assume that $T_0 = 0$. For any $m < n \le 0$, we have*

$$Z_{[m,0]}(N) \le Z_{[n,0]}(N) + (Z_{[m,n-1]}(N) - \tau_{n-1})^+,$$

where $x^+ = \max(x, 0)$.

Proof. First assume that $Z_{[m,n-1]}(N) - \tau_{n-1} \le 0$, which is exactly $X_{[m,n-1]}(N) \le T_n$. Then, by the separability property, we have

$$Z_{[m,0]}(N) = X_{[m,0]}(N) = X_{[n,0]}(N) = Z_{[n,0]}(N).$$

Now assume that $Z_{[m,n-1]}(N) - \tau_{n-1} > 0$. Let $N' = \{T'_j\}$ be the input process defined as follows:

$$T'_j = \begin{cases} T_j, & \text{for all } j \le n-1, \\ T_j + Z_{[m,n-1]}(N) - \tau_{n-1}, & \text{for all } j \ge n. \end{cases}$$

Then we have $N' \ge N$ and $X_{[m,n-1]}(N') \le T'_n$; hence, by the external monotonicity, the separability, and the homogeneity properties, we have

$$\begin{aligned} Z_{[m,0]}(N) &= X_{[m,0]}(N) \\ &\le X_{[m,0]}(N') \\ &= X_{[n,0]}(N') \\ &= X_{[n,0]}(N) + Z_{[m,n-1]}(N) - \tau_{n-1} \\ &= Z_{[n,0]}(N) + Z_{[m,n-1]}(N) - \tau_{n-1}. \end{aligned}$$

From Lemma 4 we directly derive the following result.

Lemma 5. *Assume that $T_0 = 0$. For any $n < 0$, we have*

$$Z_{[n,0]}(N) \le \sup_{n \le k \le 0} \left(\sum_{i=k}^{-1} (Z_i - \tau_{i+1}) \right) + Z_0,$$

with the convention $\sum_0^{-1} = 0$.

Proof of Proposition 2. To an input process N, we associate the upper bound process $N^+ = \{T_n^+\}$, where $T_n^+ = T_{kL}$ if $n = (k-1)L + 1, \ldots, kL$. Note that $T_n^+ \ge T_n$ for all n. Then, for all n, since we assumed that $T_0 = 0$ and thanks to the external monotonicity, we have

$$X_{[-n,0]}(N) = Z_{[-n,0]}(N) \le X_{[-n,0]}(N^+) = Z_{[-n,0]}(N^+). \tag{8}$$

Applying Lemma 5 to $Z_{[-kL+1,0]}(N^+)$ for all $k \ge 1$, we obtain

$$Z_{[-kL+1,0]}(N^+) \le \hat{s}_0 + \sup_{-k+1 \le i \le 0} \sum_{j=-i}^{-1} (\hat{s}_j - \hat{\tau}_{j+1}). \tag{9}$$

https://doi.org/10.1239/jap/1183667403 Published online by Cambridge University Press

Hence, we have

$$\begin{aligned} Z &= \lim_{k\to\infty} Z_{[-kL+1,0]} \\ &= \sup_{k\geq 0} Z_{[-kL+1,0]}(N) \\ &\leq \sup_{k\geq 0} Z_{[-kL+1,0]}(N^+), \quad \text{using (8)}, \\ &\leq \sup_{k\geq 0}\left(\hat{s}_0 + \sup_{-k+1\leq i\leq 0} \sum_{j=-i}^{-1} (\hat{s}_j - \hat{\tau}_{j+1})\right) = \hat{R}, \quad \text{using (9)}. \end{aligned}$$

4.2. Moment generating function

Lemma 6. *The function $\Lambda_Z(\cdot)$ defined by (5) is a proper convex function with $\Lambda_Z(\theta) < \infty$ for all $\theta < \eta$ and $\Lambda_Z(\theta) = \infty$ for all $\theta > \eta$, where $\eta = \sup\{\theta, \mathrm{E}[\exp(\theta Z_0)] < \infty\}$.*

Proof. Let

$$\Lambda_{Z,n}(\theta) = \log \mathrm{E}\left[\exp\left(\theta \frac{Z_{[1,n]}(N^0)}{n}\right)\right].$$

Thanks to the subadditive property of Z, we have

$$Z_{[1,n+m]}(N^0) \leq Z_{[1,n]}(N^0) + Z_{[n+1,n+m]}(N^0),$$

and $Z_{[1,n]}(N^0)$ and $Z_{[n+1,n+m]}(N^0)$ are independent. Hence, for $\theta \geq 0$, we have

$$\Lambda_{Z,n+m}((n+m)\theta) \leq \Lambda_{Z,n}(n\theta) + \Lambda_{Z,m}(m\theta).$$

Hence, we can define, for any $\theta \geq 0$,

$$\Lambda_Z(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathrm{E}[\exp(\theta Z_{[1,n]}(N^0))] = \lim_{n\to\infty} \frac{\Lambda_{Z,n}(n\theta)}{n} = \inf_{n\geq 1} \frac{\Lambda_{Z,n}(n\theta)}{n}$$

as an extended real number. The fact that Λ_Z is a proper convex function follows from [7, Lemma 2.3.9]. The fact that $\Lambda_Z(\theta) < \infty$ for all $\theta < \eta$ and $\Lambda_Z(\theta) = \infty$ for all $\theta > \eta$, follows from

$$\Lambda_Z(\theta) \leq \log \mathrm{E}[\exp(\theta Z_1)] \quad \text{and} \quad \log \mathrm{E}[\exp(\theta Z_1)] \leq \Lambda_{Z,n}(n\theta), \quad \text{for } \theta \geq 0 \text{ and all } n \geq 1.$$

We define

$$\Lambda(\theta) = \Lambda_T(-\theta) + \Lambda_Z(\theta) \quad \text{and} \quad \Lambda_n(\theta) = \Lambda_T(-\theta) + \Lambda_{Z,n}(\theta).$$

Note that $\Lambda_Z(\cdot)$ and $\Lambda_T(\cdot)$ are proper convex functions; hence, $\Lambda(\cdot)$ is a well-defined convex function. Recall that θ^* is defined as follows:

$$\theta^* = \sup\{\theta > 0, \Lambda(\theta) < 0\}.$$

The following lemma is used repeatedly in what follows.

Lemma 7. *Under the foregoing assumptions, we have $\theta^* > 0$ and*

$$\begin{aligned} &\Lambda(\theta) < 0, \quad \text{if } \theta \in (0, \theta^*), \\ &\Lambda(\theta) > 0, \quad \text{if } \theta > \theta^*. \end{aligned}$$

https://doi.org/10.1239/jap/1183667403 Published online by Cambridge University Press

Proof. Let

$$\theta_n = \sup\{\theta > 0,\ \Lambda_n(n\theta) < 0\}. \tag{10}$$

We fix n such that $\mathrm{E}[Z_{[1,n]}(N^0)] \le na$, which is possible in view of the stability condition.

We first show that $\theta_n > 0$ and

$$\Lambda_n(n\theta) < 0, \quad \text{if } \theta \in (0, \theta_n), \tag{11}$$

$$\Lambda_n(n\theta) > 0, \quad \text{if } \theta > \theta_n. \tag{12}$$

The function $\theta \mapsto \Lambda_n(n\theta)$ is convex, continuous, and differentiable on $[0, \eta)$. Hence, we have

$$\Lambda_n(n\delta) = \delta(\mathrm{E}[Z_{[1,n]}(N^0)] - a) + o(\delta),$$

which is less than zero for sufficiently small $\delta > 0$. Hence, the set over which the supremum in the definition of θ_n is taken is not empty and $\theta_n > 0$. Now (11) and (12) follow from the definition of θ_n, the convexity of $\theta \mapsto \Lambda_n(n\theta)$, and the fact that $\Lambda_n(0) = 0$.

We now show that $\theta_n \to \theta^*$ as $n \to \infty$. For $\theta \ge 0$, we have

$$\lim_{n\to\infty} \frac{\Lambda_n(n\theta)}{n} = \inf_{n\ge 1} \frac{\Lambda_n(n\theta)}{n} = \Lambda(\theta).$$

Hence, for $\theta \ge 0$ we have $\Lambda_n(n\theta)/n \ge \Lambda(\theta)$ and, for all $\theta \in (0, \theta_n)$,

$$\Lambda(\theta) \le \frac{\Lambda_n(n\theta)}{n} < 0.$$

This implies that $\theta^* \ge \theta_n > 0$. If $\theta^* < \infty$ then we can choose $\varepsilon > 0$ such that $\theta^* - \varepsilon > 0$, and then we have $\Lambda_n(n(\theta^* - \varepsilon))/n \to \Lambda(\theta^* - \varepsilon) < 0$. Hence, for sufficiently large n, we have $\Lambda_n(n(\theta^* - \varepsilon))/n < 0$; hence, $\theta^* - \varepsilon \le \theta_n$, and we proved that $\theta_n \to \theta^*$. As $\Lambda(\cdot)$ is a convex function and since $\Lambda(0) = 0$, the lemma follows in this case.

If $\theta^* = \infty$, we still have $\theta_n \to \infty$ (this will be needed in proof of Lemma 9) by the same argument as above with $\theta^* - \varepsilon$ replaced by any real number.

4.3. Lower bound

Lemma 8. *Under previous assumptions, we have*

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathrm{P}(Z > x) \ge -\theta^*.$$

Proof. We have (see Proposition 1)

$$Z \ge \sup_n \{Z_{[-n,0]}(N^0) + T_{-n} - T_0\}. \tag{13}$$

We let $Y_n = Z_{[-n,1]}(N^0) + T_{-n} + T_0$; the lemma follows from the following fact:

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathrm{P}(\sup_n Y_n > x) \ge -\theta^*.$$

Note that we have

$$\lim_{n\to\infty} \frac{1}{n} \log \mathrm{E}[\mathrm{e}^{\theta Y_n}] = \Lambda(\theta).$$

https://doi.org/10.1239/jap/1183667403 Published online by Cambridge University Press

In particular, we are in the setting of the Gärtner–Ellis theorem (see [7, Theorem 2.3.6]), which will be the main tool of the proof.

First note that we only need to consider the case $\theta^* < \infty$. We first consider the case where there exists $\theta > \theta^*$ such that $\Lambda(\theta) < \infty$. First note that the function $\theta \mapsto \Lambda(\theta)$ is convex; hence, the left-hand derivatives $\Lambda'(\theta-)$ and the right-hand derivatives $\Lambda'(\theta+)$ exist for all $\theta > 0$. Moreover, we have $\Lambda'(\theta-) \le \Lambda'(\theta+)$ and the function $\theta \mapsto \frac{1}{2}(\Lambda'(\theta-) + \Lambda'(\theta+))$ is nondecreasing; hence, $\Lambda'(\theta) = \Lambda'(\theta-) = \Lambda'(\theta+)$ except for $\theta \in \Delta$, where Δ is at most countable. Since $\Lambda(\theta) < \infty$ for $\theta > \theta^*$, we have $\Lambda(\theta^*) = 0$ and $\Lambda'(\theta^*+) > 0$. To prove this, assume that $\Lambda'(\theta^*+) = 0$. For $\theta < \theta^*$, using Lemma 7, we have $\Lambda(\theta) < 0$. Choose $\varepsilon > 0$ such that $0 < \Lambda(\theta^* + \varepsilon) < \varepsilon|\Lambda(\theta)|$. We have

$$\frac{\Lambda(\theta^* + \varepsilon)}{\varepsilon} < \frac{-\Lambda(\theta)}{\theta^* - \theta},$$

which contradicts the convexity of $\Lambda(\theta)$. Hence, we can find $t \le \theta^* + \varepsilon$ such that

$$0 < \Lambda(t), \qquad t \notin \Delta.$$

Note that these conditions imply that $t > \theta^*$ and $\Lambda'(t) \ge \Lambda'(\theta^*+) > 0$.

Thanks to the Gärtner–Ellis theorem, we have

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathrm{P}(Y_n > n\alpha) \ge - \inf_{x \in \mathcal{F},\, x > \alpha} \Lambda^*(x), \tag{14}$$

where $\mathcal{F}$ is the set of exposed point of Λ^* and $\Lambda^*(x) = \sup_{\theta \ge 0}(\theta x - \Lambda(\theta))$. Note that, from the monotonicity of $\theta x - \Lambda(\theta)$ in x as θ is fixed, we deduce that Λ^* is nondecreasing. Moreover, take $\alpha = \Lambda'(t)$, then $\Lambda^*(\alpha) = t\alpha - \Lambda(t)$ and $\alpha \in \mathcal{F}$ by [7, Lemma 2.3.9].

Given $x > 0$, define $n = \lceil x/\alpha \rceil$. We have

$$\frac{1}{x} \log \mathrm{P}(\sup_n Y_n > x) \ge \frac{1}{n\alpha} \log \mathrm{P}(Y_n \ge n\alpha),$$

taking the limit in x and n (while $\alpha = \Lambda'(t)$ is fixed) gives, thanks to (14),

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathrm{P}(\sup_n Y_n > x) \ge -\frac{t\alpha - \Lambda(t)}{\alpha} \ge -t \ge -\theta^* - \varepsilon.$$

We now consider the case where, for all $\theta > \theta^*$, we have $\Lambda(\theta) = \infty$, i.e. $\theta^* = \eta$ defined in Lemma 6. Take $K > 0$ and define $\tilde{Z}^K_{[n,m]} = Z_{[n,m]}(N^0) \prod_{i=n}^m \mathbf{1}_{\{Z_i \le K\}}$ and $\tilde{Z}^K = \sup_{n\ge 0}(\tilde{Z}^K_{[-n,0]} + T_{-n})$, where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. By (13), we have $Z \ge \tilde{Z}^K$. It is easy to see that the proof of Lemma 6 is still valid (note that the subadditive property carries over to $\tilde{Z}^K_{[n,m]}$) and the following limit exists:

$$\tilde{\Lambda}^K_Z(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathrm{E}[\exp(\theta \tilde{Z}^K_{[1,n]})] = \inf_n \frac{1}{n} \log \mathrm{E}[\exp(\theta \tilde{Z}^K_{[1,n]})].$$

Moreover, thanks to the subadditive property of Z, we have $\mathrm{P}(\tilde{Z}^K_{[1,n]} \le nK) = 1$, so that $\tilde{\Lambda}^K_Z(\theta) \le \theta K$. Hence, by the first part of the proof, we have

$$\liminf_{x\to\infty} \frac{1}{x} \log \mathrm{P}(\tilde{Z}^K > x) \ge -\tilde{\theta}^K,$$

https://doi.org/10.1239/jap/1183667403 Published online by Cambridge University Press

with $\tilde{\theta}^K = \sup\{\theta > 0, \tilde{\Lambda}_Z^K(\theta) + \Lambda_T(-\theta) < 0\}$. We now prove that $\tilde{\theta}^K \to \eta$ as $K \to \infty$, which will conclude the proof. Note that, for any fixed $\theta \geq 0$, the function $\tilde{\Lambda}_Z^K(\theta)$ is nondecreasing in K and $\lim_{K\to\infty} \tilde{\Lambda}_Z^K(\theta) = \tilde{\Lambda}_Z(\theta) \leq \Lambda_Z(\theta)$. This directly implies that $\tilde{\theta}^K \geq \eta$. Take $\theta > \eta$, so that $\Lambda_Z(\theta) = \infty$. If $\tilde{\Lambda}_Z(\theta) < \infty$ then, for all K, we have $\tilde{\Lambda}_Z^K(\theta) \leq \tilde{\Lambda}_Z(\theta) < \infty$. But, we have

$$\tilde{\Lambda}_Z^K(\theta) = \inf_n \frac{1}{n} \log \mathrm{E}[\exp(\theta \tilde{Z}_{[1,n]}^K)],$$

so that there exists n such that

$$\mathrm{E}[\exp(\theta Z_{[1,n]}(N^0)), \max(Z_1, \ldots, Z_n) \leq K] \leq \exp(n(\tilde{\Lambda}_Z^K(\theta) + 1)) \leq \exp(n(\tilde{\Lambda}_Z(\theta) + 1)),$$

but the left-hand side tends to infinity as $K \to \infty$. Hence, we have proved that, for all $\theta > \eta$, we have $\tilde{\Lambda}_Z^K(\theta) \to \infty$ as $K \to \infty$. This implies that $\tilde{\theta}^K \to \eta$ as $K \to \infty$.

4.4. Upper bound

Lemma 9. *Under previous assumptions, we have*

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathrm{P}(Z > x) \leq -\theta^*.$$

Proof. For sufficiently large L, we have, with the convention $\sum_0^{-1} = 0$ (see Proposition 2),

$$Z \leq \sup_{n\geq 0} \left(\sum_{i=-n}^{-1} \hat{s}_i(L) - \hat{\tau}_{i+1}(L) \right) + \hat{s}_0(L) =: V(L) + \hat{s}_0(L).$$

We will show that, under previous assumptions, we have

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathrm{P}(V(L) + \hat{s}_0(L) > x) \leq -\theta_L,$$

where θ_L is defined as in (10), and the lemma will follow since $\theta_L \to \theta^*$ as $L \to \infty$ (see Lemma 7).

First note that, for all $\theta \in (0, \theta_L)$, we have

$$\max\{\mathrm{E}[\exp(\theta \hat{s}_0(L))], \mathrm{E}[\exp(\theta V(L))]\} < \infty.$$

Hence, for $\theta \in (0, \theta_L)$, we have

$$\mathrm{E}[\exp(\theta(V(L) + \hat{s}_0(L)))] = \mathrm{E}[\exp(\theta V(L))]\,\mathrm{E}[\exp(\theta \hat{s}_0(L))] \leq A,$$

for some finite constant A. Hence, by Chernoff's inequality, we obtain

$$\mathrm{P}(V(L) + \hat{s}_0(L) \geq x) \leq \exp(-\theta x)\,\mathrm{E}[\exp(\theta(V(L) + \hat{s}_0(L)))] \leq A \exp(-\theta x).$$

Since the above holds for all $0 < \theta < \theta_L$, we obtain

$$\limsup_{x\to\infty} \frac{1}{x} \log \mathrm{P}(V(L) + \hat{s}_0(L) \geq x) \leq -\theta_L.$$

https://doi.org/10.1239/jap/1183667403 Published online by Cambridge University Press

Appendix A. Recursion for queues in tandem

We consider a $G/G/1/\infty \to \cdot/G/1/\infty$ tandem queue, where $\{\sigma_n^{(i)}\}$ denotes the sequence of service times in station $i = 1, 2$ and $N = \{T_n\}$ is the sequence of arrival times at the first station. For $m \le k \le n$, we denote by $D^{(i)}_{[m,n]}(k)$ the departure time of customer k from station $i = 1, 2$ when the network starts empty and is fed by $N_{[m,n]}$. With the notation introduced in Section 2, we have $X_{[m,n]}(N) = D^{(2)}_{[m,n]}(n)$. We now derive the recursion equations satisfied by the $D_{[m,n]}$s,

$$\begin{aligned}
D^{(1)}_{[m,n]}(m) &= T_m + \sigma_m^{(1)},\\
D^{(2)}_{[m,n]}(m) &= D^{(1)}_{[m,n]}(m) + \sigma_m^{(2)}\\
&= T_m + \sigma_m^{(1)} + \sigma_m^{(2)},\\
D^{(1)}_{[m,n]}(k) &= \max(D^{(1)}_{[m,n]}(k-1), T_k) + \sigma_k^{(1)},\\
D^{(2)}_{[m,n]}(k) &= \max(D^{(2)}_{[m,n]}(k-1), D^{(1)}_{[m,n]}(k)) + \sigma_k^{(2)},
\end{aligned}$$

for $m < k \le n$. From these equations, we can easily check that

$$D^{(1)}_{[m,n]}(k) = \sup_{m \le j \le k} \left\{ T_j + \sum_{i=j}^{k} \sigma_i^{(1)} \right\},$$

$$D^{(2)}_{[m,n]}(k) = \sup_{m \le j \le k} \left\{ T_j + \sup_{j \le \ell \le k} \sum_{i=j}^{\ell} \sigma_i^{(1)} + \sum_{i=\ell}^{k} \sigma_i^{(2)} \right\},$$

and (3) follows.

References

[1] Anantharam, V. (1989). How large delays build up in a GI/G/1 queue. *Queueing Systems Theory Appl.* **5,** 345–367.

[2] Baccelli, F. and Brémaud, P. (2003). *Elements of Queueing Theory*, 2nd edn. Springer, Berlin.

[3] Baccelli, F. and Foss, S. (1995). On the saturation rule for the stability of queues. *J. Appl. Prob.* **32,** 494–507.

[4] Baccelli, F. and Foss, S. (2004). Moments and tails in monotone-separable stochastic networks. *Ann. Appl. Prob.* **14,** 612–650.

[5] Baccelli, F., Foss, S. and Lelarge, M. (2005). Tails in generalized Jackson networks with subexponential service-time distributions. *J. Appl. Prob.* **42,** 513–530.

[6] Baccelli, F., Lelarge, M. and Foss, S. (2004). Asymptotics of subexponential max plus networks: the stochastic event graph case. *Queueing Systems* **46,** 75–96.

[7] Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*, 2nd edn. Springer, New York.

[8] Duffy, K., Lewis, J. T. and Sullivan, W. G. (2003). Logarithmic asymptotics for the supremum of a stochastic process. *Ann. Appl. Prob.* **13,** 430–445.

[9] Ganesh, A. (1998). Large deviations of the sojourn time for queues in series. *Ann. Operat. Res.* **79,** 3–26.

[10] Iglehart, D. L. (1972). Extreme values in the GI/G/1 queue. *Ann. Math. Statist.* **43,** 627–635.

[11] Lelarge, M. (2006). Tail asymptotics for discrete event systems. In *Proc. 1st Internat. Conf. Performance Eval. Methodol. Tools*, ACM Press, New York.

[12] Pakes, A. G. (1975). On the tails of waiting-time distributions. *J. Appl. Prob.* **12,** 555–564.