Analyzing a two-stage queueing system with many Point Process arrivals at upstream queue

Do Young Eun [a,*]

[a] *Department of Electrical and Computer Engineering, Box 7911*
North Carolina State University
Raleigh, NC 27695-7911, U.S.A.
Phone: +1 919 513 7406
E-mail: dyeun@eos.ncsu.edu

Ness B. Shroff [b,**]

[b] *School of ECE, Purdue University, West Lafayette, IN 47907-1285, U.S.A.*
Phone: +1 765 494 3471
Fax: +1 765 494 3358
E-mail: shroff@ecn.purdue.edu

We consider a two-stage queueing system where the first (upstream) queue serves many flows, of which a fixed set of flows arrive to the second (downstream) queue. We show that as the capacity and the number of flows aggregated at the upstream queue increases, the overflow probability at the downstream queue converges to that of a simplified single queue obtained by removing the upstream queue from the original two-stage queueing system. Earlier work shows such convergence for fluid traffic, by exploiting the large deviation result that the workload goes to zero almost surely, as the number of flows and capacity is scaled. However, the analysis is quite different and more difficult for the *point process* traffic considered in this paper. The reason is that for point process traffic the large deviation rate function need not be strictly positive (i.e., $I(0) = 0$), hence the workload at the upstream queue may not go to zero even though the number of flows and capacity go to infinity. The results in this paper thus make it possible to decompose the original two-stage queueing system into a simple single-stage queueing system.

---

[*] The research was conducted when this author was at Purdue University

[**] This work has been partially supported by the National Science Foundation through the Special Projects Award ANI-0099137 and the Indiana 21st Century Research and Technology Award 1220000634.

Keywords: Queueing networks, overflow probability, many-sources-asymptotic, point processes

1. Introduction

In current telecommunication networks, the link capacity (or bandwidth) at routers or switches in the network has continued to increase, thus allowing a large number of traffic flows simultaneously to traverse the network. To analyze the behavior of a queueing system with a large capacity (where many flows are multiplexed), various approaches have been used [1,3,4,9,10].

Let us first consider a queue serving N *i.i.d.* traffic flows in a FIFO manner with deterministic rate NC. For each arriving flow, $A_i(s,t)$, $s,t \in \mathbb{R}$ represents the number of customers (or packets) of type i (or flow i) that arrive in time interval $(s,t]$. The workload of the queue at time t then becomes

$$q^N(t) := \sup_{s \le t} [\sum_{i=1}^{N} A_i(s,t) - NC(t-s)], \tag{1.1}$$

assuming that the system started at $-\infty$. Under this scaling, it is well known that the behavior of the queue $q^N(t)$ is described by the following *many-sources-asymptotic* upper bound [1,4,9]:

Under appropriate conditions, we have

$$\limsup_{N\to\infty} \frac{1}{N} \log P\{q^N(t) > Nb\} \le -I(b), \quad b \ge 0, \tag{1.2}$$

where

$$I(b) := \inf_{t>0} \sup_{\theta} \ [\theta(Ct+b) - \log E\{e^{\theta A(0,t)}\}]. \tag{1.3}$$

This result has received much attention and has been widely used as an estimate of the overflow probability when the number of sources and capacity increase proportionally. However, (1.2) also reveals another aspect of the behavior of the workload $q^N(t)$. For example, suppose that $I(0)$ is positive. Then, we obtain $\sum_{N=1}^{\infty} P\{q^N(t) > 0\} \le \sum_{N=1}^{\infty} \exp(-I(0)N + o(N)) < \infty$, which implies that, by the Borel-Cantelli Lemma, $q^N(t)$ converges to zero almost surely as N goes to infinity for any *fixed* time t. Based on this observation, Wischik has shown, in a discrete time setting (for which it is always true that $I(0) > 0$), that the moment generating function of an output process converges to that of an averaged version of the arrival processes [14,15]. This work sheds some light on how each traffic

will be affected by passing through a number of switches with large capacities. However, due to the large deviation framework used in that paper, the queue dynamics in the network are described only in a log-asymptotic sense. Further, all the queues in the network are scaled in the same way (averaged by the number of sources N) to apply this result. To be specific, an output traffic flow is taken as the departure from a queue with capacity C, and with input being averaged over its *i.i.d.* copies, i.e., $\frac{1}{N}\sum_{i=1}^{N} A_i(s,t)$. Then, this output traffic flow is again averaged over its *i.i.d.* copies in order to establish the large deviations (many-sources-asymptotic) for the queue at the next stage. However, in this paper, the departure flows are taken *as is*, and they are *not independent* for any fixed N due to the interaction among different flows in the upstream queue $q^N(t)$. (See Figure 1.)

Previously, we have developed network decomposition types of results when a large number of the fluid sources are multiplexed at certain queues in the network [7,8]. More recently, again, within a fluid-model framework, the authors in [11,12] obtained the overflow asymptotics in a network of small buffers with buffer size $o(N)$ whereas the number of flows and the capacity linearly increases with N. In particular, when all the sources require the same QoS, they showed that asymptotically the admissible region corresponds to that which is obtained by assuming that flows pass through each node unchanged. *The common feature of all of these results [7,8,11,14] is that they have been obtained in a fluid framework, where the large deviation rate function is strictly positive (i.e. $I(0) > 0$).*

However, in this paper, we assume that each traffic flow is modeled as a stationary point process. This is a fundamentally different regime from the fluid model. For example, suppose that each arrival process to the queue $q^N(t)$ is Poisson with rate λ, then it is straightforward to see that the distribution of $q^N(t)$ does not vary with N. More generally, it has been recently shown [2] that for general stationary point process inputs, $q^N(t)$ converges in distribution to a queue-length random variable when the input is Poisson. Hence, for point process inputs, the queue $q^N(t)$ does not converge to zero in any sense, thereby implying that $I(0) = 0$.

This paper is organized as follows. In Section 2, we describe our problem in detail. We next provide some preliminary results and model assumptions in Section 3. In Section 4, we present our theorem with detailed proofs, and discuss several issues.

2. Problem Description

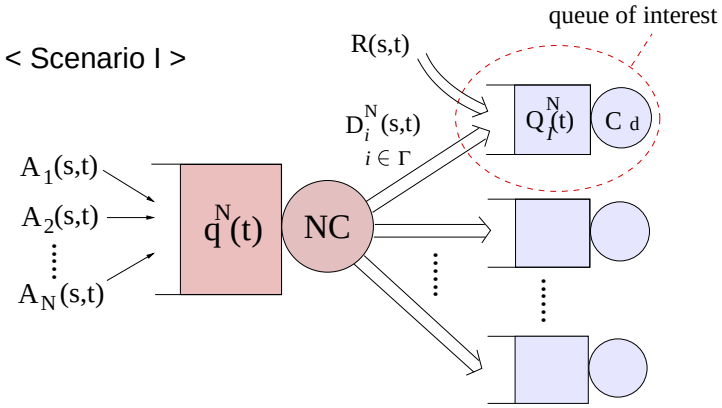

Figure 1. Queueing network: Scenario I

Consider a two-stage queueing system depicted in Figure 1. The upstream queue (with workload $q^N(t)$ at time t) represents a node (or station) that is capable of serving a large number of traffic flows in a queueing network, while each downstream queue at the second stage could be a node with only a small capacity, e.g., a node at the network periphery. In a telecommunication networking setting, we can view the upstream queue as one of the core routers in the network with large capacity, and the downstream queue as one of the edge routers with a much smaller capacity. In Figure 1, after being served at the upstream queue, the N different flows are routed to many different nodes, each of which serves only a fraction of flows. Specifically, among the N flows, a *fixed subset*[1] (non-empty and not dependent on N) of the flows i ($i \in \Gamma$) after being served at the upstream queue arrives to one of the downstream queues (with workload $Q_I^N(t)$ and capacity C_d) with an arbitrary interfering traffic $R(s,t)$, while the rest of flows are routed to other nodes or depart the system. We can thus write the steady-state workload at that downstream node as

$$Q_I^N(0) := \sup_{t \geq 0} \; [\sum_{i \in \Gamma} D_i^N(-t,0) + R(-t,0) - C_d t].$$

[1] The case where Γ scales with N could be also important in certain scenarios, for which our result does not go through. However, it is not the focus of this paper, and hence, we will not discuss it further.

We are now interested in estimating the steady-state overflow probability $P\{Q_I^N(0) > x\}$ for a given buffer level x. In order to do that, we consider a simple single-stage queueing system shown in Figure 2, a simplified version of the original two-stage queueing system in Figure 1. In Scenario II, the queue has the same interfering traffic $R(s,t)$ and the same service capacity C_d as that of Scenario I, except that the traffic arrival of interest to the queue is now $A_i(s,t)$ instead of $D_i^N(s,t)$. Specifically, we write the steady-state workload in Scenario II as

$$Q_{II}(0) := \sup_{t \geq 0} [\sum_{i \in \Gamma} A_i(-t,0) + R(-t,0) - C_d t].$$

Thus, we obtain Scenario II if we remove the upstream queue in Scenario I (the queue with large capacity). Note that $Q_{II}(0)$ does not depend on N, while $Q_I^N(0)$ does. Also note that under this scaling, the overflow probability at the downstream queue does not converge to zero as N grows in contrast to the scaling used in [14].

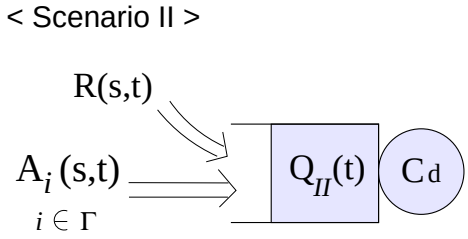

Figure 2. Scenario II: a simplified version of Scenario I

In general, the difference between the two random variables $Q_I^N(0)$ and $Q_{II}(0)$ depends on the *entire* past history of the previous queue, i.e., $q^N(t)$ for all $t \leq 0$. Thus, pointwise convergence of $q^N(t)$ to zero by itself is not sufficient to establish the convergence of $P\{Q_I^N(0) > x\}$ to $P\{Q_{II}(0) > x\}$. However, in [7], we recently showed that under $I(0) > 0$ and some technical assumptions, the overflow probability of the original system ($P\{Q_I^N(0) > x\}$) in fact converges uniformly to that of the simplified system ($P\{Q_{II}(0) > x\}$). Further, we showed that the speed of convergence is at least exponentially fast. In [7], the condition $I(0) > 0$, which guarantees the convergence of $q^N(t)$ to zero, is also shown to be satisfied for any discrete time model as well as for continuous time models under "fluid-type" of assumptions. Some examples of the "fluid" assumptions are that

(i) each arrival traffic has a finite peak rate or (ii) the sample path of $A_i(0,t)$ is smooth enough to define its derivative (See Proposition 2 in [7]).

In this paper, we can also show that $P\{Q_I^N(0) > x\}$ converges to $P\{Q_{II}(0) > x\}$ for stationary point process inputs, i.e., for non-fluid arrivals. As noted earlier, for point process inputs, the upstream queue $q^N(t)$ does not converge to zero in any sense. Nevertheless, in this paper, we show that we are still able to approximate $P\{Q_I^N(0) > x\}$ by $P\{Q_{II}(0) > x\}$ with an offset of one packet (or customer) difference (See Theorem 4.1 in the paper), which allows us to decompose the network for analysis.

3. Preliminaries

Let $A_i(s,t)$ represent the number of packets (or customers) that arrive during a time interval $(s,t]$ for flow i. We model each traffic arrival $A_i(s,t)$ as a *simple stationary point process*. A point process is said to be simple [5] when

$$P\{A\{t\} = 0 \text{ or } 1 \text{ for any } t\} = 1,$$

where $A\{t\}$ (by a little abusing the notation) denotes the number of arrivals at time t. By stationarity, we mean that $A_i(s, s+t)$ is stationary in s. Let λ be the finite intensity of the point process $A_i(0,t)$, i.e., $E\{A_i(s, s+t)\} = \lambda t$. Then, the following result is a direct consequence of a simple stationary point process from [5].

Lemma 3.1. For any simple stationary point process A_i with finite intensity λ, we have

$$P\{A_i(0,t) = 1\} = \lambda t + o(t) \quad \text{and} \quad P\{A_i(0,t) \geq 2\} = o(t). \tag{3.1}$$

Note that, by definition, $A_i(s, s+t)$ is non-decreasing in t. Throughout the paper, we assume that $A_i(s,t)$, $i = 1, 2, \ldots, N$ are independent and identically distributed.

We define

$$J(t) := \sup_{\theta > 0} \left[\theta C t - \log E\{e^{\theta A_i(0,t)}\}\right], \tag{3.2}$$

and also define $q_i(t)$ as the workload of a queue at time t with capacity C fed by a single input $A_i(s,t)$, i.e.,

$$q_i(t) := \sup_{s \le t} [A_i(s,t) - C(t-s)], \tag{3.3}$$

where $C > \lambda$. In this paper, we will impose the following assumptions on each arrival process $A_i(s,t)$.

(A1): $\limsup_{t \downarrow 0} \log E\{e^{\theta A_i(0,t)}\} = 0$ for any $\theta > 0$.

(A2): $\liminf_{t \to \infty} J(t)/\log t > 0$ where $J(t)$ is defined in (3.2).

(A3): There exists $\epsilon > 0$ such that $E\{(q_i(0))^{1+\epsilon}\} < \infty$.

Assumption (A1) is merely a technical one. In particular, it becomes trivial once $E\{e^{\theta A_i(0,t)}\} < \infty$ for some $t > 0$ from the Dominated Convergence Theorem. Assumption (A3) is quite general in that it includes almost all traffic models typically considered in the literature. For example, any long-range dependent traffic model with $\log P\{q_i(0) > x\} \sim -\alpha x^{\beta}$, where $\alpha > 0$ and $0 < \beta \le 1$, satisfies (A3). In fact, in this case, all the moments of $q_i(0)$ exist. Also, even if the workload is Pareto-distributed (i.e., having infinite variance) with parameter $1 < p \le 2$, (A3) still holds with $\epsilon = (p-1)/2 > 0$. Note that $E\{q_i(t)\}$ does not depend on i or t due to the stationarity and *i.i.d.* assumption on $A_i(s,t)$.

Assumption (A2) was first introduced by Likhanov and Mazumdar [9] to establish the many-sources-asymptotic large deviations results in the discrete time setting, and (A1) has been used to carry the proof of the result from the discrete time case over to the continuous time setting [1]. The authors in [9] showed that (A2) is quite general and even holds for on-off sources with heavy-tailed on-time distribution. Similarly, for point process models, it has been shown in [2] that (A2) is satisfied by a large class of stationary point processes. For example, (A2) holds for (i) any simple renewal process A_i whose inter-arrival distribution X with $E\{X\} = \lambda^{-1} > 0$ and (ii) any simple Poisson on-off process with the on-time and off-time distributions T_{on} and T_{off}, respectively, satisfying $E\{T_{\text{off}}\} < \infty$ and $E\{t_{\text{on}}^{1+\zeta}\} < \infty$ for some $\zeta > 0$. As a consequence, we see that Assumptions (A1)–(A3) are quite general and include most of the traffic models studied in the literature.

Lemma 3.2. Under Assumptions (A1) – (A3), we have

$$\lim_{N \to \infty} E\{q^N(t)/N\} = 0.$$

Proof. From Assumptions (A1) and (A2), note that we have the many-sources-asymptotic upper bound given by (1.2) where the rate function $I(b)$ is as in (1.3). Since $\lambda < C$, it follows that $I(b) > 0$ for any $b > 0$. Thus, for any given $\epsilon > 0$,

$$\sum_{N=1}^{\infty} P\{\frac{q^N(t)}{N} > \epsilon\} \leq \sum_{N=1}^{\infty} e^{-I(\epsilon)N+o(N)} < \infty,$$

implying that $q^N(t)/N$ converges to zero almost surely. To show the convergence in the mean, we use the following result from [6].

Theorem 3.8 in [6]: Suppose $X_n \to X$ a.s. and there are continuous functions $g, h \geq 0$ with $g(x) > 0$ for large x and $|h(x)|/g(x) \to 0$ as $|x| \to \infty$ and $E\{g(X_n)\} \leq K < \infty$ for all n. Then $E\{h(X_n)\} \to E\{h(X)\}$.

Together with Assumption (A3) and from the definition of $q_i(t)$ in (3.3), we have for some $\epsilon > 0$ and $K < \infty$,

$$\begin{aligned} E\Big\{\Big(q^N(t)/N\Big)^{1+\epsilon}\Big\} &= E\Big\{\Big(\frac{1}{N}\sup_{s\leq t}[\sum_{i=1}^{N} A_i(s,t) - NC(t-s)]\Big)^{1+\epsilon}\Big\} \\ &\leq E\Big\{\Big(\frac{1}{N}\sum_{i=1}^{N}\Big(\sup_{s\leq t}\ [A_i(s,t) - C(t-s)]\Big)\Big)^{1+\epsilon}\Big\} \\ &\leq \frac{1}{N}\sum_{i=1}^{N} E\Big\{\Big(\sup_{s\leq t}\ [A_i(s,t) - C(t-s)]\Big)^{1+\epsilon}\Big\} \\ &= E\{(q_1(t))^{1+\epsilon}\} = E\{(q_1(0))^{1+\epsilon}\} < K, \end{aligned} \tag{3.4}$$

for any N, where the second inequality follows from Jensen's inequality and the convexity of $x^{1+\epsilon}$. Hence, from Theorem 3.8 in [6] with choices of $g(x) = x^{1+\epsilon}$ and $h(x) = x$, the result follows. □

As mentioned in the introduction, however, the workload of the upstream queue, $q^N(t)$ does not converge to zero for point process inputs. Specifically, suppose that there exists $\theta_0 > 0$ such that $E\{\exp(\theta_0 A(0,t))\} < \infty$ for every $t \geq 0$. Under this assumption and under (A2), Cao and Ramanan [2] showed that for each x, $P\{q^N(t) > x\}$ converges to the tail of the workload distribution for a corresponding queue with Poisson input.

Before we proceed to our main section, we need the following result, a slight extension of (3.1) to the superposition of a finite number of arrival processes.

Lemma 3.3. Let $A_i(s,t)$, $i = 1, 2, \ldots$, be *i.i.d.* and let each of them be a stationary simple point process. Then, for any Γ with $|\Gamma| < \infty$, we have

$$P\{\sum_{i\in\Gamma} A_i(0,t) > 1\} = o(t).$$

Proof. See Appendix. □

4. Main Results

In this section, we will prove our main theorem as follows:

Theorem 4.1. Suppose that (A1) – (A3) hold for each simple stationary point process A_i, and that $(A2)$ holds for the interfering traffic $R(s,t)$. Then, for any $x \geq 1$ and any given $\delta_1, \delta_2 > 0$, we have

$$\begin{aligned} P\{Q_{II}(0) > x + \delta_1\} &\leq \liminf_{N\to\infty} P\{Q_I^N(0) > x\} \\ &\leq \limsup_{N\to\infty} P\{Q_I^N(0) > x\} \leq P\{Q_{II}(0) > x - 1 - \delta_2\}. \end{aligned}$$

As mentioned earlier, the difference between $Q_I^N(0)$ and $Q_{II}(0)$ depends on the behavior of the workload $(q_i^N(-t))$, due to flow i, over the entire past history, i.e., over $t \geq 0$. To prove our theorem, we divide the whole interval $[0,\infty)$ into $[0,T)$ and $[T,\infty)$. We deal with $q_i^N(-t)$ over each interval to show that the above convergence holds true for any fixed (independent of N), sufficiently large $T > 0$. Then, we will send T to infinity to prove the theorem.

In the subsequent section, we present a result on the behavior of the workload $(q_i^N(-t))$ for $t \geq T$. This will be used in proving our main theorem. In Section 4.2, we provide the proof of Theorem 4.1. In Section 4.3 we discuss some issues about point process inputs versus fluid-like inputs. In Section 4.4, for illustration, we derive an upper bound on the speed of convergence for the special case when each input is Poisson.

4.1. Behavior of the workload $q_i^N(-t)$ for $t \geq T$

We define $q_i^N(t)$ as the workload (number of customers or packets) corresponding to flow i in the upstream queue (with total workload $q^N(t)$; see Figure 1) at time t. From this definition, we have $q^N(t) = \sum_{i=1}^N q_i^N(t)$. Also, from the *i.i.d.* assumption on A_i $(i = 1, 2, \ldots, N)$, $q_i^N(t)$ $(i = 1, 2, \ldots, N)$ have identical

distributions (but clearly dependent on each other due to the interaction among different flows in the queue), and

$$E\{q_i^N(t)\} = \frac{E\{q^N(t)\}}{N} \tag{4.1}$$

by symmetry.

Proposition 4.2. Under Assumptions (A1) and (A3), we have for any $\delta > 0$,

$$\lim_{T\to\infty} P\Big\{ \sup_{t\geq T} \sum_{i\in\Gamma} \frac{q_i^N(-t)}{t} > \delta \Big\} = 0, \tag{4.2}$$

uniformly in N, provided that $|\Gamma| < \infty$.

Proof. Since

$$P\Big\{ \sup_{t\geq T^2} \sum_{i\in\Gamma} \frac{q_i^N(-t)}{t} > \delta \Big\} \leq P\Big\{ \sum_{i\in\Gamma} \sup_{t\geq T^2} \frac{q_i^N(-t)}{t} > \delta \Big\} \leq \sum_{i\in\Gamma} P\Big\{ \sup_{t\geq T^2} \frac{q_i^N(-t)}{t} > \frac{\delta}{|\Gamma|} \Big\},$$

without loss of generality, we only need to show that for each i, the following expression

$$P\Big\{ \sup_{t\geq T^2} \frac{q_i^N(-t)}{t} > \delta \Big\} = P\Big\{ \sup_{t\geq T} \frac{q_i^N(-t^2)}{t^2} > \delta \Big\}$$

converges uniformly (in N) to zero as T increases.

We first divide the interval $[T, \infty)$ into smaller intervals, each of which has equal length h, and then work within each interval. Specifically, let $s_n := T + nh$ and $S(n,h) = [s_{n-1}, s_n]$ where $n = 1, 2, \ldots$. Then, we have

$$\begin{aligned} \sup_{t\geq T} \frac{q_i^N(-t^2)}{t^2} &= \sup_{n\geq 1} \sup_{t\in S(n,h)} \frac{q_i^N(-t^2)}{t^2} \\ &\leq \sup_{n\geq 1} \frac{q_i^N(-s_n^2)}{s_n^2} + \sup_{n\geq 1} \sup_{t\in S(n,h)} \left(\frac{q_i^N(-t^2)}{t^2} - \frac{q_i^N(-s_n^2)}{s_n^2} \right). \end{aligned} \tag{4.3}$$

For $t \in S(n,h)$, we can write $t = s_n - u$, where $u \in [0,h]$. Thus, after simple calculations, we get

$$\begin{aligned} \left(\frac{q_i^N(-t^2)}{t^2} - \frac{q_i^N(-s_n^2)}{s_n^2} \right) &= \frac{2s_n u - u^2}{s_n^2 (s_n - u)^2} q_i^N(-s_n^2) \\ &\quad + \frac{1}{(s_n - u)^2} \Big(q_i^N(-(s_n - u)^2) - q_i^N(-s_n^2) \Big). \end{aligned}$$

Since $1/(s_n-u)^2 \le 1/(s_n-h)^2$ and $2s_nu-u^2 \le 2s_nu \le 2s_nh$ for all $u \in [0,h]$, we have

$$\sup_{t\in S(n,h)}\left(\frac{q_i^N(-t^2)}{t^2}-\frac{q_i^N(-s_n^2)}{s_n^2}\right) \le \frac{2s_nh}{s_n^2(s_n-h)^2}q_i^N(-s_n^2) + \frac{1}{(s_n-h)^2}\sup_{u\in[0,h]}\left(q_i^N(-(s_n-u)^2)-q_i^N(-s_n^2)\right). \tag{4.4}$$

Observe that for any s and any positive number t, we have

$$q_i^N(s+t) \;\le\; q_i^N(s)+A_i(s,s+t). \tag{4.5}$$

Thus,

$$\sup_{u\in[0,h]}\left(q_i^N(-(s_n-u)^2)-q_i^N(-s_n^2)\right) \le \sup_{u\in[0,h]} A_i(-s_n^2,-(s_n-u)^2) = A_i(-s_n^2,-(s_n-h)^2) \tag{4.6}$$

since $A_i(s,s+t)$ is non-decreasing in t.

Combining (4.3) – (4.6), we have

$$\begin{aligned} P\left\{\sup_{t\ge T}\frac{q_i^N(-t^2)}{t^2}>\delta\right\} \le\; & P\left\{\sup_{n\ge1}\frac{q_i^N(-s_n^2)}{s_n^2} \;>\; \frac{\delta}{3}\right\} \\ & +\; P\left\{\sup_{n\ge1}\frac{1}{(s_n-h)^2}A_i(-s_n^2,-(s_n-h)^2) \;>\; \frac{\delta}{3}\right\} \\ & +\; P\left\{\sup_{n\ge1}\frac{2s_nh}{s_n^2(s_n-h)^2}q_i^N(-s_n^2) \;>\; \frac{\delta}{3}\right\}, \end{aligned} \tag{4.7}$$

where $s_n = T+nh$. We will now show that, as T increases, each of the RHS of the above decreases to zero uniformly in N using the following two lemmas. This will complete the proof of Proposition 4.2. □

Lemma 4.3. Let $s_n = T+nh$ where $h>0$. Under (A3), we have

$$\lim_{T\to\infty} P\left\{\sup_{n\ge1}\frac{q_i^N(-s_n^2)}{s_n^2} \;>\; \frac{\delta}{3}\right\} = 0,$$

uniformly in N.

Proof. From Jensen's inequality, we have $E\{X\} \leq (E\{X^{1+\epsilon}\})^{\frac{1}{1+\epsilon}}$. Thus, from (3.4), we know that there exists $M < \infty$ such that $E\{q^N(t)/N\} < M$ for any N. Together with (4.1), this means that $E\{q_i^N(t)\} < M$ for any i, N and t. Hence,

$$P\left\{\sup_{n\geq 1}\frac{q_i^N(-s_n^2)}{s_n^2} > \frac{\delta}{3}\right\} \leq \sum_{n=1}^{\infty} P\left\{\frac{q_i^N(-s_n^2)}{s_n^2} > \frac{\delta}{3}\right\}$$
$$\leq \sum_{n=1}^{\infty}\frac{3}{\delta}\frac{M}{(T+nh)^2} \leq \frac{C_1}{T}$$

for some positive constant $C_1 < \infty$, where the second inequality follows from Markov's inequality and the fact that $s_n = T + nh$. Thus, the first term on the RHS of (4.7) goes to zero uniformly in N, as T increases. □

Lemma 4.4. Let $s_n = T + nh$ where $h > 0$. Under (A1) and (A3), the second and the third terms on the RHS of (4.7) converges to zero, uniformly in N, as T increases.

Proof. First, note that for any convex function f, using the property that $f(ax+by) \leq af(x) + bf(y)$ whenever $a + b = 1$ and $a, b \geq 0$, we have

$$f\left(\frac{A(0,t+s)}{t+s}\right) = f\left(\frac{t}{t+s}\left(\frac{A(0,t)}{t}\right) + \frac{s}{t+s}\left(\frac{A(t,t+s)}{s}\right)\right)$$
$$\leq \frac{t}{t+s}f\left(\frac{A(0,t)}{t}\right) + \frac{s}{t+s}f\left(\frac{A(t,t+s)}{s}\right).$$

Thus, by taking expectation and from the stationary increments assumption on $A_i(s,t)$, we see that the function $s(t) := E\{(A_i(0,t))^2\}/t$ is subadditive in $t > 0$ by choosing $f(x) = x^2$. Thus, $\lim_{t\to\infty} s(t)/t$ exists and is finite from Assumption (A1), which guarantees that $E\{(A_i(0,t))^2\} < \infty$ for some $t > 0$. This implies that there exists $V < \infty$ such that $E\{(A_i(0,t))^2\} \leq Vt^2$ for all sufficiently large t. In particular, we have

$$E\{(A_i(-s_n^2, -(s_n-h)^2))^2\} = E\{(A_i(0, 2s_nh - h^2))^2\}$$
$$\leq V(2s_nh - h^2)^2 \leq V(2s_nh)^2$$

for all large T (recall that $s_n = T + nh$). Hence, similarly as before, the second term on the RHS of (4.7) is dominated by

$$\sum_{n=1}^{\infty} P\left\{\frac{1}{(s_n-h)^2}A_i(-s_n^2,-(s_n-h)^2) \ > \ \frac{\delta}{3}\right\} \le \sum_{n=1}^{\infty}\frac{V}{(\delta/3)^2}\frac{(2s_nh)^2}{(s_n-h)^4}$$
$$= \sum_{n=1}^{\infty}\frac{V}{(\delta/3)^2}\frac{(2(T+nh)h)^2}{(T+(n-1)h)^4} \le \frac{C_2}{T}$$

where the first inequality follows from Markov's inequality (after taking squares in both sides) and $C_2 < \infty$ is some positive constant. Hence the second term on the RHS of (4.7) also goes to zero as T increases, and so does the third term by the same method and the fact that $E\{q_i^N(t)\} < M$. Therefore, the assertion follows. □

Note that in the proof of Proposition 4.2, we only used the conditions $E\{(A_i(0,t))^2\} < \infty$ and $E\{q_i(0)\} < \infty$, which are much weaker than (A1) and (A3), respectively. The following lemma will also be used in the proof of our theorem.

Lemma 4.5. Let $x_n \ge 0$ be a sequence converging to zero. Then, there exists a non-decreasing sequence a_n with $\lim_{n\to\infty} a_n = \infty$ such that $\lim_{n\to\infty} a_n x_n = 0$.

Proof. Since x_n converges to zero, for each integer $k > 0$, we can select N_k (N_k is non-decreasing in k and $N_k \uparrow \infty$) such that $x_n < 1/k^2$ for all $n \ge N_k$. The result then follows by setting $a_n = k$ for $N_k \le n < N_{k+1}$. □

4.2. Proof of Theorem 4.1

Proof of the lower bound. As before, let $q_i^N(t)$ denote the number of customers (packets) from flow i in the queue $q^N(t)$ at time t. Then, since $(\sup f - \sup g) \le$

$\sup(f-g)$, we have

$$\begin{aligned}\sup_{t\geq 0}[\sum_{i\in\Gamma}A_i(-t,0)+R(-t,0)-C_dt]-\sup_{t\geq 0}[\sum_{i\in\Gamma}D_i^N(-t,0)+R(-t,0)-C_dt]\\ \leq \sup_{t\geq 0}\Big(\sum_{i\in\Gamma}A_i(-t,0)-\sum_{i\in\Gamma}D_i^N(-t,0)\Big)\\ = \sup_{t\geq 0}\Big(\sum_{i\in\Gamma}q_i^N(0)-\sum_{i\in\Gamma}q_i^N(-t)\Big)\\ \leq \sum_{i\in\Gamma}q_i^N(0). \qquad (4.8)\end{aligned}$$

Thus, from an inequality $P\{X>x+\delta_1\}-P\{Y>x\}\leq P\{X-Y>\delta_1\}$, the definitions of $Q_{II}(0)$ and $Q_I^N(0)$, and from (4.8) we have

$$P\{Q_{II}(0)>x+\delta_1\}-P\{Q_I^N(0)>x\}\leq P\{\sum_{i\in\Gamma}q_i^N(0)>\delta_1\}.$$

Since (4.1) holds for any i,t and N, we have

$$P\{\sum_{i\in\Gamma}q_i^N(0)>\delta_1\}\leq\frac{|\Gamma|}{\delta_1}\frac{E\{q^N(0)\}}{N}\to 0, \qquad (4.9)$$

as N increases from Lemma 3.2. This establishes the lower bound. □

Proof of the upper bound. Let $\epsilon>0$ be given. We first divide the whole interval $[0,\infty)$ into $[0,T]$ and $[T,\infty)$ to get

$$\begin{aligned}P\Big\{\sup_{0\leq t\leq T}[\sum_{i\in\Gamma}D_i^N(-t,0)+R(-t,0)-C_dt]\ >\ x\Big\}\ \leq\ P\{Q_I^N(0)>x\}\\ \leq\ P\Big\{\sup_{0\leq t\leq T}[\sum_{i\in\Gamma}D_i^N(-t,0)+R(-t,0)-C_dt]\ >\ x\Big\}\\ +\ P\Big\{\sup_{t\geq T}[\sum_{i\in\Gamma}D_i^N(-t,0)+R(-t,0)-C_dt]\ >\ 0\Big\},\end{aligned}$$

for all $x\geq 0$ and $T>0$. Similarly,

$$\begin{aligned}P\Big\{\sup_{0\leq t\leq T}[\sum_{i\in\Gamma}A_i(-t,0)+R(-t,0)-C_dt]\ >\ x\Big\}\ \leq\ P\{Q_{II}(0)>x\}\\ \leq\ P\Big\{\sup_{0\leq t\leq T}[\sum_{i\in\Gamma}A_i(-t,0)+R(-t,0)-C_dt]\ >\ x\Big\}\\ +\ P\Big\{\sup_{t\geq T}[\sum_{i\in\Gamma}A_i(-t,0)+R(-t,0)-C_dt]\ >\ 0\Big\}.\end{aligned}$$

Thus, we have

$$
\begin{aligned}
& P\{Q_I^N(0) > x\} - P\{Q_{II}(0) > x - 1 - \delta_2\} \\
& \leq P\Big\{ \sup_{t \geq T} [\sum_{i \in \Gamma} D_i^N(-t,0) + R(-t,0) - C_d t] > 0 \Big\} \qquad (4.10) \\
& \quad + \Bigg(P\Big\{ \sup_{0 \leq t \leq T} [\sum_{i \in \Gamma} D_i^N(-t,0) + R(-t,0) - C_d t] \ > \ x \Big\} \\
& \qquad - \ P\Big\{ \sup_{0 \leq t \leq T} [\sum_{i \in \Gamma} A_i(-t,0) + R(-t,0) - C_d t] \ > \ x - 1 - \delta_2 \Big\} \Bigg). \qquad (4.11)
\end{aligned}
$$

We will use the following two lemmas whose proofs are deferred to the end of Section 4.2.

Lemma 4.6. Under the assumption of Theorem 4.1, for any given $\epsilon > 0$, we can find T_0 (independent of N) such that

$$
P\Big\{ \sup_{t \geq T} [\sum_{i \in \Gamma} D_i^N(-t,0) + R(-t,0) - C_d t] > 0 \Big\} < \epsilon, \quad \text{for all} \ \ T \geq T_0.
$$

Lemma 4.7. Under the assumption of Theorem 4.1, for any fixed $T > 0$, the expression in (4.11) converges to zero as N increases.

Hence, from Lemma 4.6 and 4.7, and from the inequality in (4.10)–(4.11), we can choose T sufficiently large such that

$$
\limsup_{N \to \infty} \left(P\{Q_I^N(0) > x\} - P\{Q_{II}(0) > x - 1 - \delta_2\} \right) < 2\epsilon.
$$

Hence, the upper bound follows by taking $\epsilon \downarrow 0$. This completes the proof of Theorem 4.1. □

From Theorem 4.1, the original overflow probability at the downstream queue ($P\{Q_I^N(0) > x\}$) can be approximated by that of a single queue ($P\{Q_{II}(0) > x\}$) when N is large. Note that the error between the upper and the lower bound becomes negligible as x increases. For instance, suppose that the distribution of $Q_{II}(0)$ satisfies

$$
\lim_{x \to \infty} \frac{1}{x^\beta} \log P\{Q_{II}(0) > x\} = -\alpha,
$$

for some $\alpha, \beta > 0$. Then, we immediately have

$$\lim_{x\to\infty} \frac{\log P\{Q_{II}(0) > x-1\}}{\log P\{Q_{II}(0) > x\}} = 1.$$

See Figure 3 for a graphical interpretation of Theorem 4.1.

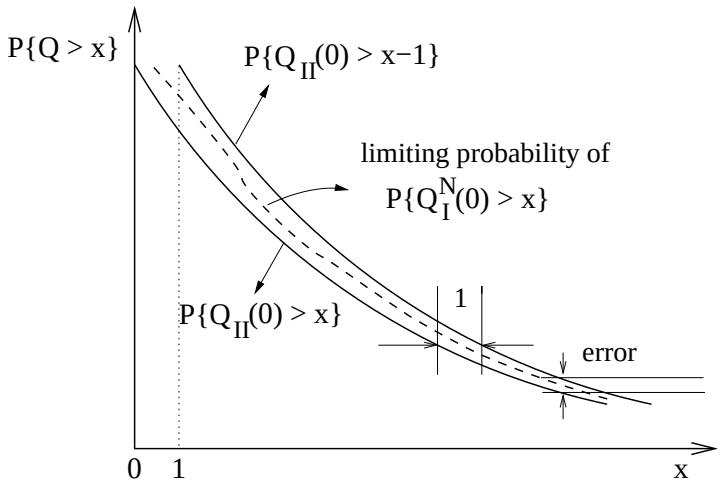

Figure 3. Illustration of Theorem 4.1

Proof of Lemma 4.6. First, we pick $\eta_1, \eta_2 > 0$ such that $C_d - \eta_1 - \eta_2$ is still larger than the mean arrival rate to the queue $Q_{II}(0)$, i.e., $\lambda|\Gamma| + \bar{r} < C_d - \eta_1 - \eta_2 < C_d - \eta_1 < C_d$, where $\lambda = E\{A(-t,0)/t\}$ and $\bar{r} = E\{R(-t,0)/t\}$ is the mean arrival rate of the interfering (or crossing) traffic. Since $D_i^N(-t,0) \le A_i(-t,0) + q_i^N(-t)$ for any t and N, we can write

$$\begin{aligned}
&P\Big\{\sup_{t\ge T}[\sum_{i\in\Gamma} D_i^N(-t,0) + R(-t,0) - C_d t] > 0\Big\} \\
&\le P\Big\{\sup_{t\ge T}[\sum_{i\in\Gamma} q_i^N(-t) + \sum_{i\in\Gamma} A_i(-t,0) + R(-t,0) - C_d t] \ > \ 0\Big\} \\
&\le P\Big\{\sup_{t\ge T}\ [\sum_{i\in\Gamma} A_i(-t,0) + R(-t,0) - (C_d - \eta_2)t] \ > \ 0\Big\} \qquad (4.12) \\
&\quad + P\Big\{\sup_{t\ge T}[\sum_{i\in\Gamma} q_i^N(-t) - \eta_2 t] \ > \ 0\Big\}. \qquad (4.13)
\end{aligned}$$

By splitting $A_i(-t,0)$ into $A_i(-t,-T)+A_i(-T,0)$ for $t \geq T$ (similarly for $R(-t,0)$), we see that the RHS of (4.12) is equal to

$$\begin{aligned}
&P\Big\{ \sup_{t\geq T} \ [\sum_{i\in\Gamma} A_i(-t,0) + R(-t,0) - (C_d-\eta_2)t] \ > \ 0\Big\} \\
&= \ P\Big\{ \sup_{t\geq T} \Big[\sum_{i\in\Gamma} A_i(-t,-T) + R(-t,-T) + \sum_{i\in\Gamma} A_i(-T,0) + R(-T,0) \\
&\quad - (C_d-\eta_1-\eta_2)(t-T) - \eta_1 t\Big] \ > \ (C_d-\eta_1-\eta_2)T\Big\} \\
&\leq \ P\Big\{ \sup_{t\geq T} \Big[\sum_{i\in\Gamma} A_i(-t,-T) + R(-t,-T) - (C_d-\eta_1-\eta_2)(t-T)\Big] > \eta_1 T\Big\} \\
&\quad + P\Big\{ \sum_{i\in\Gamma} A_i(-T,0) + R(-T,0) \ > \ (C_d-\eta_1-\eta_2)T\Big\}. \qquad (4.14)
\end{aligned}$$

From the stationary assumption on $A_i(s,s+t)$, the first term of the RHS of (4.14) is equal to

$$P\Big\{ \sup_{t\geq 0} \ [\sum_{i\in\Gamma} A_i(-t,0) + R(-t,0) - (C_d-\eta_1-\eta_2)t] \ > \ \eta_1 T\Big\}, \qquad (4.15)$$

which decreases to zero as T increases since $Q_{II}(0)$ (with service capacity C_d replaced by $C_d-\eta_1-\eta_2$) is stable. Similarly, it is not difficult to see that the second term of the RHS of (4.14) also decreases to zero as T increases. To see this, choose two positive numbers C_A and C_R such that $|\Gamma|C_A+C_R = C_d-\eta_1-\eta_2$ with $\lambda < C_A$ and $\bar{r} < C_R$. This is always possible since $\lambda|\Gamma|+\bar{r} < C_d-\eta_1-\eta_2 = C_A|\Gamma|+C_R$. Then, we have

$$\begin{aligned}
&P\Big\{ \sum_{i\in\Gamma} A_i(-T,0) + R(-T,0) \ > \ (C_d-\eta_1-\eta_2)T\Big\} \\
&\leq P\{ \sum_{i\in\Gamma} A_i(-T,0) > |\Gamma|C_A T\} + P\{R_i(-T,0) > C_R T\} \qquad (4.16) \\
&\leq |\Gamma| \exp\Big(- \sup_{\theta>0} [\theta C_A T - \log E\{e^{\theta A_i(0,T)}\}]\Big) \\
&\quad + \exp\Big(- \sup_{\theta>0} [\theta C_R T - \log E\{e^{\theta R(0,T)}\}]\Big)
\end{aligned}$$

by Markov's inequality. From Assumption (A2), we know that the first term of the above expression is bounded by

$$|\Gamma| \exp(-\alpha \log T) = |\Gamma| T^{-\alpha}$$

for some positive constant α and for all sufficiently large T. Similarly, the second term also decreases to zero as T increases. Thus, we have shown that the RHS of (4.12) goes to zero as T increases.

For (4.13), observe that

$$P\Big\{\sup_{t\geq T}[\sum_{i\in\Gamma} q_i^N(-t)-\eta_2 t] \;>\; 0\Big\} = P\Big\{\sup_{t\geq T}[\sum_{i\in\Gamma}\frac{q_i^N(-t)}{t}] \;>\; \eta_2\Big\},$$

which goes to zero (uniformly in N) as T increases from Proposition 4.2. Hence, given $\epsilon>0$, we can find T_0 (independent of N) such that (4.10) is less than ϵ for all $T>T_0$, i.e.,

$$P\Big\{\sup_{t\geq T}[\sum_{i\in\Gamma} D_i^N(-t,0)+R(-t,0)-C_d t]>0\Big\}<\epsilon, \quad \text{for all } \; T\geq T_0. \tag{4.17}$$

This completes the proof of Lemma 4.6 □

Proof of Lemma 4.7. Similarly as in (4.8), observe that

$$\begin{aligned}\sup_{0\leq t\leq T}[\sum_{i\in\Gamma} D_i^N(-t,0)+R(-t,0)-C_d t] - \sup_{0\leq t\leq T}[\sum_{i\in\Gamma} A_i(-t,0)+R(-t,0)-C_d t]\\ \leq \sup_{0\leq t\leq T}\Big(\sum_{i\in\Gamma} q_i^N(-t)-\sum_{i\in\Gamma} q_i^N(0)\Big)\\ \leq \sup_{0\leq t\leq T}\sum_{i\in\Gamma} q_i^N(-t).\end{aligned}$$

Thus, again from the inequality $P\{X>x\}-P\{Y>x-a\}\leq P\{X-Y>a\}$, (4.11) is bounded by

$$P\Big\{\sup_{0\leq t\leq T}\sum_{i\in\Gamma} q_i^N(-t)>1+\delta_2\Big\}. \tag{4.18}$$

We will show that (4.18) converges to zero as N increases for any fixed $T>0$.

We now divide the interval $[0,T]$ into smaller intervals, each of which has equal length ϵ_N. Let $S(T,\epsilon_N):=\{1,2,\ldots,\lfloor T/\epsilon_N\rfloor+1\}$. Note that

$$\begin{aligned}\sup_{0\leq t\leq T}\sum_{i\in\Gamma} q_i^N(-t) \leq \sup_{n\in S(T,\epsilon_N)}\Bigg(\sum_{i\in\Gamma} q_i^N(-n\epsilon_N)\\ + \sup_{(n-1)\epsilon_N\leq t\leq n\epsilon_N}\Big(\sum_{i\in\Gamma} q_i^N(-t)-\sum_{i\in\Gamma} q_i^N(-n\epsilon_N)\Big)\Bigg),\end{aligned}$$

and from (4.5),

$$\sup_{(n-1)\epsilon_N \le t \le n\epsilon_N} \Big(\sum_{i\in\Gamma} q_i^N(-t) - \sum_{i\in\Gamma} q_i^N(-n\epsilon_N) \Big) \le \sup_{0\le t\le \epsilon_N} \sum_{i\in\Gamma} A_i(-n\epsilon_N, -n\epsilon_N + t)$$
$$\le \sum_{i\in\Gamma} A_i(-n\epsilon_N, -(n-1)\epsilon_N).$$

Thus, we can bound (4.18) as

$$P\Big\{ \sup_{0\le t\le T} \sum_{i\in\Gamma} q_i^N(-t) > 1+\delta_2 \Big\}$$
$$\le \sum_{n\in S(T,\epsilon_N)} \left(P\Big\{ \sum_{i\in\Gamma} q_i^N(-n\epsilon_N) > \delta_2 \Big\} + P\Big\{ \sum_{i\in\Gamma} A_i(-n\epsilon_N, -(n-1)\epsilon_N) > 1 \Big\} \right)$$
$$\le (1+T/\epsilon_N) \frac{|\Gamma| E\{q^N(0)\}}{\delta_2 N} + (1+T/\epsilon_N) P\Big\{ \sum_{i\in\Gamma} A_i(0,\epsilon_N) > 1 \Big\}, \qquad (4.19)$$

where the last inequality follows from Markov's inequality and the fact that $E\{q_i^N(t)\} = E\{q^N(t)\}/N = E\{q^N(0)\}/N$ for all N and t by symmetry, and the stationarity of $A_i(s,t)$.

Since, from Lemma 3.2, $E\{q^N(0)\}/N$ decreases to zero, Lemma 4.5 asserts that there exists a non-decreasing sequence a_N ($a_N \uparrow \infty$) such that $a_N E\{q^N(0)\}/N$ decreases to zero. The first term of (4.19) now can be shown to go to zero by choosing $\epsilon_N = 1/a_N$. Since $\epsilon_N \downarrow 0$ as $N \uparrow 0$, the second term of (4.19) also decreases to zero from Lemma 3.3. Hence, we have shown that (4.11) goes to zero as N increases for any fixed T. □

4.3. Point processes vs. fluid-like processes

In the proof of Theorem 4.1, we invoke the simple point process property only through Lemma 3.3. For non-point processes (e.g., fluid-like processes), suppose that we are able to show that

$$P\{A_i(0,t) > \delta\} = o(t), \qquad (4.20)$$

for any given $\delta > 0$. We can then rewrite Theorem 4.1 to be of the following form:

Proposition 4.8. Suppose that each arrival A_i satisfies (4.20) and (A1)–(A3), and that $R(s,t)$ satisfies (A2). Then, for any $x > 0$ and any given $\delta_1, \delta_2 > 0$, we have

$$P\{Q_{II}(0) > x + \delta_1\} \leq \liminf_{N\to\infty} P\{Q_I^N(0) > x\}$$
$$\leq \limsup_{N\to\infty} P\{Q_I^N(0) > x\} \leq P\{Q_{II}(0) > x - \delta_2\},$$

i.e., $Q_I^N(0)$ converges to $Q_{II}(0)$ in distribution.

Proof. The proof is identical to that of Theorem 4.1 except that the LHS of the inequality in (4.19) now becomes

$$P\Big\{ \sup_{0\leq t\leq T} \sum_{i\in\Gamma} q_i^N(-t) > \delta_2 \Big\}.$$

This term also decreases to zero as N increases by noting that

$$P\Big\{ \sum_{i\in\Gamma} A_i(0,t) > \delta \} \leq \sum_{i\in\Gamma} P\{A_i(0,t) > \delta/|\Gamma|\} = o(t)$$

from (4.20). Hence, the result follows. □

For example, if there exists a peak rate of the input process, i.e., $A_i(0,t) \leq Pt$ for some $P < \infty$, (4.20) then follows by noting that $P\{A_i(0,t) > \delta\} \leq E\{(A_i(0,t))^2\}/\delta^2 \leq (Pt)^2/\delta^2$. In this case, as mentioned in the introduction, $I(0)$ becomes positive and, with other assumptions, $P\{Q_I^N(0) > x\}$ in fact converges to $P\{Q_I^N(0) > x\}$ uniformly in x [7]. Further, the speed of convergence is at least exponentially fast. In contrast, for point process inputs, obtaining the speed of convergence appears to be much more challenging. We are only able to obtain an upper bound on the speed of convergence for Poisson inputs (see Section 4.4). We find that this bound is quite slow. It could be that this is because the bound we obtain is conservative, or this is in fact the price we have to pay for $q^N(t)$ itself not converging to zero in the case of point processes.

Note that the one packet "offset" for the upper bound in Theorem 4.1 does not appear in the fluid case (Proposition 4.8) and can intuitively be explained as follows: For a simple point process, on its sample path basis, a packet (or customer) can arrive at any time instant. Putting it in a different way, this means that no matter how small an interval we choose, the "amount" of traffic that arrives during this interval does not always decrease to zero due to the discrete nature of point processes (for instance, see Lemma 3.1). Thus, the sample path of

$q_i^N(t)$ jumps up and down like a staircase, implying that the departure $D_i^N(s,t)$ and the arrival $A_i(s,t)$ can differ by one packet at any time instant, and so can $Q_I^N(0)$ and $Q_{II}(0)$.

Our result can also be generalized to the case of non-simple point processes including batch arrivals, provided that there exists some positive constant K with $P\{\sum_{i\in\Gamma} A_i(0,t) > K\} = o(t)$ and that $E\{q^N(t)/N\}$ decreases to zero. In this case, the offset for the upper bound in Theorem 4.1 will be K.

4.4. Speed of convergence for Poisson inputs

We pointed out that, in contrast to the case of fluid traffic arrivals, the speed of convergence in Theorem 4.1 may be quite slow depending on specific models for arrival processes. In this section, using the proof of Theorem 4.1, we provide an upper bound on the speed of convergence when each arrival is Poisson. However, as noted before, unlike in the case of fluid arrivals, this upper bound on the speed of convergence is quite slow.

We assume that each arrival A_i is a stationary Poisson process with mean rate $\lambda < C$. We also require that the interfering traffic $R(s,t)$ behave nicely in the following sense:

(A4):

(a) $E\{\sup_{t\geq 0}[R(-t,0) - C_R t]\} < \infty$, whenever $C_R > \bar{r} := E\{R(-t,0)/t\}$.

(b) There exists $H \in [0.5, 1)$ such that $\limsup_{t\to\infty} \mathrm{Var}\{R(0,t)\}/t^{2H} < \infty$

Proposition 4.9. Suppose that each A_i is a stationary Poisson process with mean $\lambda < C$ and that (A4) holds. Then, for any given $\delta_1, \delta_2 > 0$, we have

$$P\{Q_{II}(0) > x + \delta_1\} - O\left(\frac{1}{N}\right) \leq P\{Q_I^N(0) > x\} \leq P\{Q_{II}(0) > x - 1 - \delta_2\} + O\left(\frac{1}{N^\gamma}\right).$$

where $O(t)$ here means $\limsup_{t\to 0} O(t)/t < \infty$, and γ is given by

$$\gamma = \min\left\{\frac{1}{6}, \frac{2-2H}{2(3-2H)}\right\}.$$

Remark 4.10. Assumption (A4)(a) means that a single queue with capacity C_R fed by input $R(s,t)$ has a finite expectation whenever it is stable. The parameter

H in Assumption (A4)(b) is called the Hurst parameter in the literature, and used for modeling the long-range dependence of the process when $H > 1/2$.

Proof of Proposition 4.9. Observe that if A_i is a stationary Poisson process, the distribution of $q^N(t)$ in (1.1) is well-known [13], and depends only on its utilization parameter $\rho = \lambda/C$. Clearly, in this case, we have $E\{q^N(t)\} := M < \infty$ for all N and t. The lower bound then directly follows from (4.9).

For the upper bound, note first that (4.15) is bounded by $O(1/T)$ from Assumption (A4)(a). Since $\sup_{\theta>0}[\theta C_A T - \log E\{e^{\theta A_i(0,T)}\}] = \kappa T$ for some constant $\kappa > 0$ when A_i is Poisson, the first term in (4.16) is dominated by $|\Gamma|e^{-\kappa T}$. From Chebyshev's inequality and (A4)(b), we see that the second term in (4.16) is bounded by K/T^{2-2H} for some constant $K < \infty$. Similarly, from the proof of Proposition 4.2, we know that

$$P\Big\{ \sup_{t \geq T^2} \frac{q_i^N(-t)}{t} > \delta \Big\} \leq \frac{K_2}{T}$$

for large T and some constant K_2. Thus, (4.13) is less than $K_2/\sqrt{T}$. Hence, we see that (4.10) is bounded by

$$|\Gamma|e^{-\kappa T} + K/T^{2-2H} + K_2/T^{0.5}. \tag{4.21}$$

Next, we note that $P\{\sum_{i\in\Gamma} A_i(0,t) > 1\} = O(t^2)$ for sufficiently small t since the finite superposition of *i.i.d.* Poisson processes is also a Poisson process. Then, from (4.19) and the fact that $E\{q^N(t)\} = M < \infty$, we get

$$\text{RHS of (4.19)} \leq B_2 \frac{T}{\epsilon_N}\frac{1}{N} + B_3 T \epsilon_N, \tag{4.22}$$

for all sufficiently large N and ϵ_N ($\epsilon_N \downarrow 0$), where B_2 and B_3 are finite constants.

We now set $\epsilon_N = 1/N^\alpha$ and $T = N^\beta$, where

$$(\alpha, \beta) \in \mathbb{D} := \{(\alpha, \beta) : 0 < \beta < \alpha \ \text{ and } \ \alpha + \beta < 1\}.$$

With this choice of α and β, (4.21) and (4.22) clearly decrease to zero as N increases. Since $e^{-\kappa T}$ decreases much faster than $1/T^\xi$ as T increases, from (4.10)–(4.11) and (4.21)–(4.22), we can write

$$\begin{aligned} &P\{Q_I^N(0) > x\} - P\{Q_{II}(0) > x - 1 - \delta_2\} \\ &\leq \frac{B_0}{N^{\beta/2}} + \frac{B_1}{N^{\beta(2-2H)}} + \frac{B_2}{N^{1-(\alpha+\beta)}} + \frac{B_3}{N^{\alpha-\beta}} \end{aligned}$$

for all sufficiently large N, and for some constants B_i, $i = 0, \ldots, 3$. Since we can freely choose $(\alpha, \beta) \in \mathbb{D}$, it turns out that the RHS of the above is bounded by B/N^{γ} for some constant $B < \infty$ and for all sufficiently large N, where γ is given by

$$\gamma := \max_{(\alpha,\beta)\in\mathbb{D}} \Big(\min\{\frac{\beta}{2}, \beta(2-2H), 1-(\alpha+\beta), \alpha-\beta\}\Big).$$

Direct calculations yield

$$\max_{(\alpha,\beta)\in\mathbb{D}} \Big(\min\{\xi\beta, 1-(\alpha+\beta), \alpha-\beta\}\Big) = \frac{\xi}{2(\xi+1)} := y(\xi)$$

for any $\xi > 0$. Hence, we have $\gamma = \min\{y(1/2), y(2-2H)\}$. This completes the proof. □

Appendix

Proof of Lemma 3.3. Consider an event J_0 that there is no jump (arrival) during $(0, t]$. Similarly, let J_1 denote an event of one jump in the same time interval. Then, clearly from (3.1), $P\{J_0\} = 1 - \lambda t + o(t)$ and $P\{J_1\} = \lambda t + o(t)$. Hence, we have

$$\begin{aligned} P\{\sum_{i\in\Gamma} A_i(0,t) = 0\} &= P\Big\{\text{there is no jump for all } i,\ i \in \Gamma\Big\} \\ &= \big(P\{J_0\}\big)^{|\Gamma|}, \end{aligned}$$

and

$$\begin{aligned} &P\{\sum_{i\in\Gamma} A_i(0,t) = 1\} \\ &= P\Big\{\text{there is exactly one jump for some } i,\ i \in \Gamma \text{ and no jump for } j \neq i\Big\} \\ &= |\Gamma| P\{J_1\}\big(P\{J_0\}\big)^{|\Gamma|-1}. \end{aligned}$$

Combining all of the above yields

$$\begin{aligned} P\{\sum_{i\in\Gamma} A_i(0,t) > 1\} &= 1 - \sum_{k=0}^{1} P\{\sum_{i\in\Gamma} A_i(0,t) = k\} \\ &= 1 - (1-\lambda t + o(t))^{|\Gamma|} - |\Gamma|(\lambda t + o(t))(1-\lambda t + o(t))^{|\Gamma|-1} \\ &= o(t). \end{aligned}$$

□

References

[1] D. D. Botvich and N. Duffield. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems*, 20:293–320, 1995.

[2] J. Cao and K. Ramanan. A Poisson Limit for the Unfinished Work of Superposed Point Processes. *Bell Labs Tech. Report*, 2001.

[3] J. Choe and N. B. Shroff. Use of Supremum Distribution of Gaussian Processes in Queueing Analysis with Long-Range Dependence and Self-Similarity. *Stochastic Models*, 16(2), Feb. 2000.

[4] C. Courcoubetis and R. Weber. Buffer overflow asymptotics for a buffer handling many traffic sources. *Journal of Applied Probability*, 33:886–903, 1996.

[5] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes.* Springer-Verlag, 1988.

[6] R. Durrett. *Probability : Theory and Examples.* Duxbury Press, Belmont, CA, second edition, 1996.

[7] D. Y. Eun and N. B. Shroff. Network decomposition in the many-sources regime. *Advances in Applied Probability*, Sept. 2004. to appear.

[8] D. Y. Eun and N. B. Shroff. Simplification of Network Analysis in Large-Bandwidth Systems. In *Proceedings of IEEE INFOCOM*, San Francisco, CA, 2003.

[9] N. Likhanov and R. Mazumdar. Cell loss asymptotics for buffers fed with a large number of independent stationary sources. *Journal of Applied Probability*, 36(1):86–96, 1999.

[10] M. Mandjes and S. Borst. Overflow Behavior in queues with many long-tailed inputs. *Advances in Applied Probability*, 32:1150–1167, 2000.

[11] O. Ozturk, R. Mazumdar, and N. Likhanov. Many sources asymptotics for a feedforward network with small buffers. In *Proceedings of the Allerton Conference*, Monticello, IL, 2002.

[12] O. Ozturk, R. Mazumdar, and N. Likhanov. Many sources asymptotics in networks with small buffers. *submitted*, 2002.

[13] J. Roberts, U. Mocci, and J. Virtamo. *Broadband Network Teletraffic, Final Report of Action COST 242.* Springer-Verlag, New York, 1996.

[14] Damon Wischik. The output of a switch, or, effective bandwidths for networks. *Queueing Systems*, 32:383–396, 1999.

[15] Damon Wischik. Sample path large deviations for queues with many inputs. *Annals of Applied Probability*, 11:379–404, 2000.