Submitted to Advances in Applied Probability (April, 2002)

NETWORK DECOMPOSITION IN THE MANY-SOURCES REGIME

DO YOUNG EUN,* *North Carolina State University*

NESS B. SHROFF,** *Purdue University*

Abstract

We derive results that show the impact of aggregation in a queueing network. Our model consists of a two-stage queueing system where the first (upstream) queue serves many flows, of which a certain set arrive to the second (downstream) queue. The downstream queue experiences arbitrary interfering traffic. In this setup, we prove that, as the number of flows (N) being aggregated in the upstream queue increases, the overflow probability of the downstream queue $\mathbb{P}\{Q_I^N(0) > x\}$ converges uniformly in x to the overflow probability of a single queueing system $\mathbb{P}\{Q_{II}(0) > x\}$ obtained by simply removing the upstream queue in the original two-stage queueing system. We also provide the speed of convergence and show that it is at least exponentially fast. We then extend our results to non-*i.i.d.* traffic arrivals.

Keywords: aggregation; queueing network; many-sources-asymptotic; speed of convergence

AMS 2000 Subject Classification: Primary 60K25

Secondary 90B18; 60F10; 68M20

1. Introduction

In this paper, we derive results that show how traffic aggregation that typically takes place in large telecommunication networks, can help simplify network analysis. This work is motivated by the fact that the link capacity or bandwidth in telecommunication

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

** Postal address: School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907-1285, U.S.A. Tel: +1 765 494-3471, Fax: +1 765 494-3358, E-mail: shroff@ecn.purdue.edu.

This work has been partially supported by the NSF grant ANI-0099137.

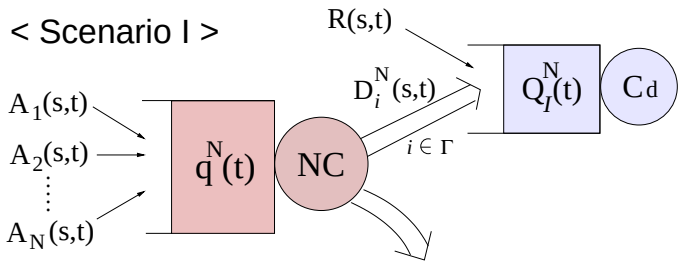

FIGURE 1: Scenario I: a two-stage queueing system

networks is becoming increasingly plentiful, thus allowing a large number of traffic flows to traverse the network.

The analysis of the workload distribution in multiple queues is a well known difficult problem, except in special cases such as Markovian queueing networks, in which product-form solutions are available [8, 9]. The difficulty in analysis primarily comes because the traffic processes lose their original statistical characteristics as they depart from the first queue in the network. Given a plethora of sophisticated techniques for analyzing a single queue, there has been some recent work that attempts to decompose the network based on large deviations techniques [14, 15, 16]. de Veciana *et. al.* [14] showed that under certain constraints, the effective bandwidth of a traffic flow, defined in terms of the exponent from the *large buffer asymptotic*, is not altered, and devised a notion of a decoupling bandwidth by which the queueing network can be decoupled or decomposed in an appropriate way. More recently, using the *many-sources-asymptotic* in discrete-time setting, Wischik [15, 16] showed that the output traffic and the input traffic after being averaged (or normalized by the number of sources) satisfies the same large deviations principle, as the number of traffic sources increases. Thus the author claims that the same set of tools using effective bandwidths (in the many-sources-asymptotic) can be used throughout the queueing network. These works shed some light on the dynamics of a network of queues. However, because of the large deviations framework, the works describe the queue dynamics in the network only in a log-asymptotic sense.

In this work, we will directly deal with the workload random variable in a queueing network. Let us consider a two-stage queueing system (Scenario I) shown in Figure 1. In this figure, the upstream queue $q^N(t)$ models a node that is capable of serving many flows. For flow i, $A_i(s,t)$ and $D_i^N(s,t)$ represent the amount of traffic arrival and

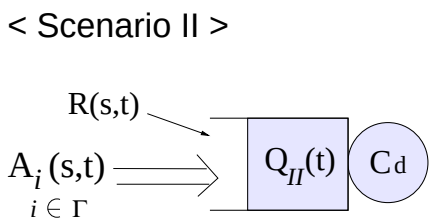

FIGURE 2: Scenario II: a simplified version of Scenario I

departure, respectively, during a time interval $[s,t)$. The server capacity of the queue is NC and $q^N(t)$ denotes the workload at time t. In Scenario I, shown in Figure 1, among the N flows, a *fixed* subset of the flows i ($i \in \Gamma$) after being served at the first (upstream) queue arrive to the downstream queue $Q_I^N(t)$ with interfering traffic $R(s,t)$, while the rest of them departs the queueing system. We are then interested in estimating the overflow probability $\mathbb{P}\{Q_I^N(0) > x\}$ 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 those of Scenario I, except that the traffic arrival of interest to the queue is now $A_i(s,t)$ instead of $D_i^N(s,t)$. Thus, we obtain Scenario II if we remove the upstream queue in Scenario I (the queue with a large number of traffic flows). Note that $Q_{II}(0)$ does not depend on N, while $Q_I^N(0)$ does.

Our main result in this paper is to show that $\mathbb{P}\{Q_I^N(0) > x\}$ converges to $\mathbb{P}\{Q_{II}(0) > x\}$ uniformly in $x > 0$, as N increases. This is much stronger than the convergence of $Q_I^N(0)$ to $Q_{II}(0)$ in distribution. By doing so, we are able to estimate the original overflow probability $\mathbb{P}\{Q_I^N(0) > x\}$ using a simple estimate of the overflow probability $\mathbb{P}\{Q_{II}(0) > x\}$, for which a myriad of techniques have been proposed in the literature. It should also be noted here that the result that $\mathbb{P}\{Q_I^N(0) > x\}$ converges to $\mathbb{P}\{Q_{II}(0) > x\}$ uniformly in x can also be proved when an arbitrary (not fixed) subset Γ of flows departs to the second queue, as long as the second queue is stable (i.e., C_d is scaled appropriately). However, the resulting convergence is trivial in the case when Γ is not fixed, and a function of N, since in this case (as will soon be evident) both probabilities

$\mathbb{P}\{Q_I^N(0) > x\}$ and $\mathbb{P}\{Q_{II}(0) > x\}$ converge to zero[†]. Thus, to avoid trivialities, we have modeled Γ, and the capacity of the downstream node C_d, to be fixed.

A motivating idea behind showing this convergence is to note that as N grows, the statistical multiplexing gain in the upstream queue also increases. For instance, the asymptotic behavior of this queue can be best described by the *many-sources-asymptotic* based on large deviations techniques [2, 4, 6, 11, 16]. It basically says that the probability that the queue $q^N(t)$ exceeds level Nb decays exponentially with the system size N. More precisely,

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}\{q^N(t) > Nb\} = -I(b), \tag{1}$$

for a certain rate function $I(b)$. Hence, as pointed out in [15, 16] under a discrete-time setting, we see from (1) that the probability that queue $q^N(t)$ is non-empty at a fixed time t goes to zero as N increases, provided that the rate function for the zero buffer level is positive, i.e., $I(0) > 0$. Accordingly, the departure traffic $D_i^N(s,t)$ will converge to the arrival traffic $A_i(s,t)$ for *fixed* s and t. However, the difference between the two random variables, $Q_I^N(0)$ and $Q_{II}(0)$, generally depends on the *entire* past history of the queue, i.e., $q^N(t)$ for all $t \leq 0$. Thus, pointwise convergence by itself is not sufficient for the required convergence of $\mathbb{P}\{Q_I^N(0) > x\}$ to hold. After proving our main result, we also derive the speed of convergence and show that it is at least exponentially fast.

This paper is organized as follows. In Section 2, we first explore the well-known many-sources-asymptotic again and develop several preliminary results that will be used in proving our main result. Then, in Section 2.2, we precisely describe our model in Scenario I shown in Figure 1 as well as required assumptions. In Section 3, we prove our main result, the uniform convergence of the overflow probability, and provide the speed of convergence. We then extend our results to non-*i.i.d.* traffic arrivals in Section 3.3. In Section 4 we discuss some issues on the implications and extensions of our results.

[†]A more meaningful result in this case would be the decay rate of the $\mathbb{P}\{Q_I^N(0) > x\}$, or any limiting result on the ratio, i.e., $\mathbb{P}\{Q_I^N(0) > x\}/\mathbb{P}\{Q_{II}(0) > x\}$. However, in this paper, we have fixed Γ and focus on the convergence of $\mathbb{P}\{Q_I^N(0) > x\}$ to a non-vanishing probability $\mathbb{P}\{Q_{II}(0) > x\}$.

2. Model and preliminaries

In the first subsection, we focus on the behavior of the queue $q^N(t)$, $(t \in T)$, in more detail and show that $\sup_{t\in[0,N^k]} q^N(-t)$ and $q^N(0)$ satisfies the same large deviations upper bound. We provide conditions under which the queue $q^N(t)$ is guaranteed to decrease to zero, namely, $I(0) > 0$. We then describe the problem that we will investigate in Section 3. Throughout the paper, we will work on both the discrete-time ($T = \mathbb{Z}$) and continuous-time domain ($T = \mathbb{R}$).

2.1. Many-sources-asymptotic revisited

Consider a queue $q^N(t)$ fed by N traffic flows $A_i(s,t)$, $i = 1, 2, \ldots, N$, $(s, t \in T)$ with capacity NC. We assume that $A_i(s,t)$, $i = 1, 2, \ldots, N$, are *i.i.d.* with stationary increments. For stability, we require $\mathbb{E}\{A_i(-t,0)\}/t := \lambda < C$. Then, assuming the system starts at $-\infty$, the steady-state workload at time t can be expressed as

$$q^N(t) := \sup_{s\le t} \big[\sum_{i=1}^{N} A_i(s,t) - CN(t-s)\big]. \tag{2}$$

Note that from the stationary increments property of $A_i(s,t)$, the distribution of $q^N(t)$ does not depend on $t \in T$. For simplicity of notation, since $A_i(s,t)$, $i = 1, 2, \ldots, N$, are *i.i.d.*, we will suppress the subscript i from $A_i(s,t)$ unless we need it explicitly. Define

$$J_t(b) := \sup_{\theta} \ \big[\theta(Ct+b) - \log \mathbb{E}\{e^{\theta A(0,t)}\}\big], \tag{3}$$

and

$$I(b) := \inf_{t>0} J_t(b) = \inf_{t>0} \sup_{\theta} \ \big[\theta(Ct+b) - \log \mathbb{E}\{e^{\theta A(0,t)}\}\big]. \tag{4}$$

We will then assume the following:

(A1) $\liminf_{t\to\infty} J_t(0)/\log t > 0$

(A2) For $T = \mathbb{R}$, $\limsup_{t\to 0} \mathbb{E}\{\ \exp(\theta \sup_{0\le u\le t} |A(0,u)|)\} = 1,\ \forall \theta > 0$

Assumption (A1) has been recently put forth by Likhanov and Mazumdar [11]. This assumption is shown to be more general than the one used in [2] that cannot be satisfied for on-off sources with heavy-tailed on-time distribution (see [11]). In [2], assumption (A2) has been shown to be sufficient to carry over the proof of the many-sources-asymptotic in the discrete-time case, to the continuous-time case. This assumption is

merely a technical one, and will be satisfied for most cases of practical interest. Recently, Mandjes [12, 13] has also shown that the many-sources-asymptotic still holds for the continuous-time case under the continuity of $I(b)$ and boundedness of traffic rate, instead of using assumption (A2). However, due to the nature of the proof in [13], the case of $b = 0$ is excluded, which is of interest in the paper. We also notice that assumption (A2) is sufficient to establish our main result on convergence in the continuous-time case. Thus we adopt assumption (A2) in this paper. While these assumptions originally have been used in order to develop large deviations results for $q^N(0)$, we show below that we can extract more (upper bound) from the stationary nature of $q^N(t)$.

Proposition 1. *Suppose that assumption (A1) holds and for the continuous-time case ($T = \mathbb{R}$), additionally assumption (A2) also holds. Then for any fixed $k \in \mathbb{N}$, we have*

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left\{ \sup_{t\in[0,N^k]} q^N(-t) > Nb \right\} \leq -I(b), \quad b \geq 0$$

where $I(b)$ is defined by (4).

Proof. ($T = \mathbb{Z}$):

For any $N > 0$, since the distribution of $q^N(t)$ does not depend on t, we have

$$\mathbb{P}\Big\{ \sup_{t\in[0,N^k]} q^N(-t) > Nb \Big\} \leq \sum_{t=0}^{N^k} \mathbb{P}\{q^N(-t) > Nb\} = (N^k + 1)\mathbb{P}\{q^N(0) > Nb\}.$$

Then since

$$\limsup_{N\to\infty} \frac{1}{N} \log(N^k + 1) = 0,$$

for any fixed $k > 0$, it remains to be shown that

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\{q^N(0) > Nb\} \leq -I(b).$$

Although this is already stated in [11], we provide a detailed proof here, in order to help us easily present the proof for the continuous-time case.

Define $W_i(t) := A_i(-t,0) - Ct$. Then, for any $\theta > 0$ and $t_0 > 0$, observe that

$$\begin{aligned}
\mathbb{P}\Big\{ \sup_{t\in\mathbb{Z}_+} \sum_{i=1}^{N} W_i(t) > Nb \Big\} &= \mathbb{P}\Big\{ \sup_{t\in\mathbb{N}} \sum_{i=1}^{N} W_i(t) > Nb \Big\} \qquad (5)\\
&\leq \sum_{t=1}^{\infty} \mathbb{P}\{ \sum_{i=1}^{N} W_i(t) > Nb \}\\
&\leq \sum_{t=1}^{\infty} \exp\big(-N\theta b + N \log \mathbb{E}\{e^{\theta W_1(t)}\}\big)\\
&\leq t_0 \max_{0<t\leq t_0} \exp\big(-N\theta b + N \log \mathbb{E}\{e^{\theta W_1(t)}\}\big)\\
&\quad + e^{-N\theta b} \sum_{t=t_0+1}^{\infty} \exp\big(N \log \mathbb{E}\{e^{\theta W_1(t)}\}\big), \qquad (6)
\end{aligned}$$

where the equality follows since $W_i(0) = 0$ and the second inequality follows from Markov's inequality. From assumption (A1), there exists $t_0 > 0$ and $\alpha > 0$ such that $J_t(0) > \alpha \log t$ for all $t > t_0$. In other words, from (3), there exists $\theta_0 > 0$ and t_0 such that $\log \mathbb{E}\{e^{\theta_0 W_1(t)}\} < -\alpha \log t$ for all $t > t_0$. Choosing such a t_0 and θ_0 in (6), we then have

$$\begin{aligned}
&\mathbb{P}\Big\{ \sup_{t\in\mathbb{N}} \sum_{i=1}^{N} W_i(t) > Nb \Big\}\\
&\quad \leq t_0 \max_{0<t\leq t_0} \exp\big(-N\theta b + N \log \mathbb{E}\{e^{\theta W_1(t)}\}\big) + \frac{e^{-\theta_0 Nb}}{N\alpha - 1} t_0^{1-N\alpha} \qquad (7)
\end{aligned}$$

for $N > 1/\alpha$. Taking logarithms, dividing by N, and taking the lim sup as $N \to \infty$ gives

$$\begin{aligned}
&\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\Big\{ \sup_{t\in\mathbb{N}} \sum_{i=1}^{N} W_i(t) > Nb \Big\}\\
&\quad \leq \max\Big\{ \max_{0<t\leq t_0} (-\theta b + \log \mathbb{E}\{e^{\theta W_1(t)}\}) \ , \ -\theta_0 b - \alpha \log t_0 \Big\}.
\end{aligned}$$

Since this holds for any $\theta > 0$ and sufficiently large t_0 and N, we take the limit $t_0 \to \infty$ and optimize with respect to θ to get

$$\begin{aligned}
\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\{q^N > Nb\} &\leq \sup_{t>0} \inf_{\theta} (-\theta b + \log \mathbb{E}\{e^{\theta W_1(t)}\})\\
&= -\inf_{t>0} \sup_{\theta} \big[\theta(Ct + b) - \log \mathbb{E}\{e^{\theta A(0,t)}\}\big]\\
&= -I(b).
\end{aligned}$$

($T = \mathbb{R}$): We will use similar arguments as used in [2] except that we are dealing with a supremum over a two-dimensional region instead. First we define

$$W_i(s,t) = A_i(-t-s,-t) - sC. \tag{8}$$

Then we have

$$\sup_{0\le t\le N^k} q^N(-t) = \sup_{0\le t\le N^k}\sup_{s\ge 0}\sum_{i=1}^{N} W_i(s,t).$$

Fix $\epsilon > 0$ and define

$$\widehat{W}_i(n,m) := \sup_{\{(n-1)\epsilon\le s\le n\epsilon\}}\ \sup_{\{(m-1)\epsilon\le t\le m\epsilon\}} W_i(s,t), \tag{9}$$

where $n \in \mathbb{N}$ and $m \in S(\epsilon) := \{1,2,\ldots,\lfloor\frac{N^k}{\epsilon}\rfloor + 1\}$.[‡]

Note that

$$\widehat{W}_i(n,m) \le W_i(n\epsilon,m\epsilon) \ + \sup_{\{(n-1)\epsilon\le s\le n\epsilon\}}\ \sup_{\{(m-1)\epsilon\le t\le m\epsilon\}} \big(W_i(s,t) - W_i(n\epsilon,m\epsilon)\big).$$

We fix $p \in (0,1)$ and apply Hölder's inequality to get

$$\begin{aligned}\log\mathbb{E}\{e^{\widehat{W}_i(n,m)}\} \ &\le\ p\log\mathbb{E}\{e^{\frac{\theta}{p}W_i(n\epsilon,m\epsilon)}\}\\ &+\ (1-p)\log\mathbb{E}\Big\{\exp\Big(\frac{\theta}{1-p}\sup_{\{(n-1)\epsilon\le s\le n\epsilon\}}\ \sup_{\{(m-1)\epsilon\le t\le m\epsilon\}}\big(W_i(s,t)-W_i(n\epsilon,m\epsilon)\big)\Big)\Big\}.\end{aligned}$$

Then, by the definition of $W_i(s,t)$ in (8), the RHS of the above is bounded by

$$\begin{aligned}&\le\quad (1-p)\log\mathbb{E}\Big\{\exp\Big(\frac{\theta}{1-p}\Big(\sup_{0\le u\le 2\epsilon}\big|A_i(-(n+m)\epsilon,-(n+m)\epsilon+u)\big| +\\ &\qquad\sup_{0\le v\le\epsilon}\big|A_i(-m\epsilon,-m\epsilon+v)\big| + C\epsilon\Big)\Big)\Big\}\\ &\le\quad \frac{1-p}{2}\log\mathbb{E}\Big\{\exp\Big(\frac{2\theta}{1-p}\sup_{0\le u\le 2\epsilon}|A_i(0,u)|\Big)\Big\}\\ &\qquad+\frac{1-p}{2}\log\mathbb{E}\Big\{\exp\Big(\frac{2\theta}{1-p}\sup_{0\le v\le\epsilon}|A_i(0,v)|\Big)\Big\} + \theta C\epsilon\\ &:=\quad \delta(\epsilon),\end{aligned}$$

where the second inequality follows from the stationarity of $A_i(s,s+t)$ in s and Jensen's inequality ($\mathbb{E}\{e^{\theta X}\} \le (\mathbb{E}\{e^{2\theta X}\})^{1/2}$). Thus, we get

$$\log\mathbb{E}\{e^{\theta\widehat{W}_i(n,m)}\} \le p\log\mathbb{E}\{e^{\frac{\theta}{p}W_i(n\epsilon)}\} + \delta(\epsilon),$$

[‡] $\lfloor x\rfloor$ is defined as the greatest integer no larger than x.

where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$ from assumption (A2). Now, observe that, from the definition of $\widehat{W}_i(n,m)$ in (9),

$$\sup_{0\le t\le N^k}\sup_{s\ge 0}\sum_{i=1}^{N} W_i(s,t) \le \sup_{m\in S(\epsilon)}\sup_{n\in\mathbb{N}}\sum_{i=1}^{N}\widehat{W}_i(n,m).$$

Note that since $W_i(s,t)$ is stationary in t, the distribution of $\widehat{W}_i(n,m)$ and $\sup_{n\in\mathbb{N}}\sum_{i=1}^{N}\widehat{W}_i(n,m)$ does not depend on m. Hence, using the union bound, we can repeat the same steps in (5) – (7) except that the RHS of them are multiplied by $\lfloor\frac{N^k}{\epsilon}\rfloor + 1$, the cardinality of the set $S(\epsilon)$. Thus, for fixed ϵ, we obtain

$$\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}\Big\{\sup_{0\le t\le N^k}\sup_{s\in\mathbb{R}_+}\sum_{i=1}^{N}W_i(s,t) > Nb\Big\} \le$$
$$-p\Big(\inf_{t>0}\sup_{\theta}\big[\theta(Ct+b) - \log\mathbb{E}\{e^{\theta A_1(0,t)}\}\big]\Big) + \delta(\epsilon).$$

Hence the result follows by taking $\epsilon \downarrow 0$ and $p \uparrow 1$.

In the proof of Proposition 1, note that N^k can be replaced by an increasing function $g(N)$ such that $\lim_{N\to\infty}\log g(N)/N = 0$. Hence for any such function $g(N)$, $\sup_{t\in[0,g(N)]} q^N(-t)$ also satisfies the same large deviation upper bound as $q^N(0)$. This kind of argument plays a pivotal role in identifying the speed of convergence in Section 3.2.

Next, we will investigate the workload random variable $q^N(0)$ itself rather than an asymptotic distribution as in Proposition 1. In particular, we will do this by evaluating the rate function $I(b)$ at $b = 0$, since this enables us to describe the random variable $q^N(0)$ in a stronger way, as will be seen later. Before we proceed, we assume the following:

(A3) For $T = \mathbb{R}$ (continuous-time case), we assume that $A(0,t)$ satisfies one of the following.

(a) For all $t > 0$, $0 \le A(0,t) \le Pt$ for some $P < \infty$.

(b) $A(0,t)$ can be represented as an integral of a stationary process $r(t)$, i.e., for all $t \ge 0$, $A(0,t) = \int_0^t r(u)du$ with $\mathbb{E}\{\sup_{0\le u\le\epsilon}|r(u)|\} < \infty$ for sufficiently small $\epsilon > 0$.

The following results will be useful.

Lemma 1. *Let X_t be a stochastic process with stationary increments. Then, for any convex function h, the function $f(t) := t \cdot \mathbb{E}\{h(X_t/t)\}$ is subadditive, i.e., $f(s+t) \leq f(s) + f(t)$ for all $s, t \geq 0$.*

Proof. Observe that

$$\begin{aligned} h\left(\frac{X_{t+s}}{t+s}\right) &= h\left(\frac{X_t + (X_{t+s} - X_t)}{t+s}\right) \\ &= h\left(\frac{t}{t+s}(\frac{X_t}{t}) + \frac{s}{t+s}(\frac{X_{t+s}-X_t}{s})\right) \\ &\leq \frac{t}{t+s}h\left(\frac{X_t}{t}\right) + \frac{s}{t+s}h\left(\frac{X_{t+s}-X_t}{s}\right). \end{aligned}$$

Thus, by taking expectations, we get the result from stationary increments property.

Lemma 2. (Corollary 2.4.5 in [5].) *Suppose that $a \leq X \leq b$ is a real-valued random variable with $\bar{x} = \mathbb{E}\{X\}$. Then, for any $\theta \in \mathbb{R}$,*

$$\mathbb{E}\{e^{\theta X}\} \leq \frac{\bar{x}-a}{b-a}e^{\theta b} + \frac{b-\bar{x}}{b-a}e^{\theta a}.$$

Proposition 2. *Suppose that assumption (A1) holds for $T = \mathbb{Z}$, and assumptions (A1), (A2) and (A3) hold for $T = \mathbb{R}$. Then, we have $I(0) > 0$ and $\lim_{N\to\infty} q^N(t) = 0$ almost surely. (Similarly, $\lim_{N\to\infty} \sup_{t\in[0,N^k]} q^N(-t) = 0$ almost surely).*

Remark 1. Assumption (A3) is indeed required to establish the almost sure convergence to zero for the continuous-time case. For example, suppose that each input is a Poisson process with rate λ. Then, clearly the aggregated input is still a Poisson process with rate $N\lambda$, and the system is scaled such that the utilization remains unchanged. Thus, we see that the workload random variable $q^N(0)$ does not converge to zero in any sense. In fact, for Poisson processes, it is easy to see that $I(0) = 0$, so we cannot expect the convergence of the workload to zero. However, the delay still converges to zero by direct application of the many-sources-asymptotic upper bound. (Note that $I(b) > 0$ if $b > 0$.)

Proof of Proposition 2. For $T = \mathbb{Z}$, it has been shown that [2, 4]

$$\begin{aligned} I(0) &:= \inf_{t>0} \sup_{\theta} \left[\theta C t - \log \mathbb{E}\{e^{\theta A(0,t)}\}\right] \\ &= \sup_{\theta} \left[\theta C - \log \mathbb{E}\{e^{\theta A(0,1)}\}\right]. \end{aligned} \tag{10}$$

From the stability condition, i.e., $C > \mathbb{E}\{A_1(0,1)\} = \lambda$, and the convexity of the function $\log \mathbb{E}\{e^{\theta A(0,1)}\}$ in θ, it immediately follows that $I(0) > 0$.

For $T = \mathbb{R}$, it is straightforward to see that both assumptions (A3).a and (A3).b satisfy (A2) by the Dominated Convergence Theorem. Thus, we can write $I(0)$ as

$$I(0) = \inf_{t>0} \sup_{\theta} \left[\theta C t - \log \mathbb{E}\{e^{\theta A_1(0,t)}\}\right].$$

Assume that $A(0,t) \leq Pt$ for some $P < \infty$. Observe that

$$\begin{aligned} I(0) &= \inf_{t>0} \sup_{\theta} \left[\theta C t - \log \mathbb{E}\{e^{\theta A(0,t)}\}\right] \\ &= \inf_{t>0} \sup_{\theta} \left[\theta C - \log \mathbb{E}\{e^{\theta \frac{A(0,t)}{t}}\}\right]. \end{aligned} \tag{11}$$

Since $0 \leq \frac{A(0,t)}{t} \leq P$ with $\mathbb{E}\{\frac{A(0,t)}{t}\} = \lambda$, from Lemma 2, we have

$$\mathbb{E}\{e^{\theta \frac{A(0,t)}{t}}\} \leq \frac{\lambda}{P} e^{\theta P} + (1 - \frac{\lambda}{P}).$$

Thus, we have

$$\sup_{\theta} \left[\theta C - \log \mathbb{E}\{e^{\theta \frac{A(0,t)}{t}}\}\right] \geq \sup_{\theta} \left[\theta C - \log(\frac{\lambda}{P} e^{\theta P} + (1 - \frac{\lambda}{P}))\right] = \sup_{\theta} h(\theta),$$

where

$$h(\theta) := \theta C - \log(\frac{\lambda}{P} e^{\theta P} + (1 - \frac{\lambda}{P})).$$

It is easy to see that h is a concave function and $h'(0) = C - \lambda > 0$. Thus, by direct calculation, we have

$$\begin{aligned} I(0) &= \sup_{\theta} \left[\theta C - \log \mathbb{E}\{e^{\theta \frac{A(0,t)}{t}}\}\right] \\ &\geq \frac{C}{P} \log(\frac{C}{\lambda}) + (1 - \frac{C}{P}) \log(\frac{P - C}{P - \lambda}) > 0, \end{aligned}$$

where $\lambda < C < P$.

Now suppose that $A(0,t) = \int_0^t r(u)du$ for a stationary process $r(t)$ satisfying assumption (A3).b. Define

$$g_\theta(t) := \mathbb{E}\{e^{\theta \frac{A(0,t)}{t}}\}.$$

From Lemma 1, the function $f(t) = t g_\theta(t)$ is subadditive since $h(x) = e^{\theta x}$ is convex. Thus, for any $t > 0$, it is easy to see that $g_\theta(\frac{t}{2^n})$ is nondecreasing in n $(n = 1, 2, \ldots)$.

From (11), we have

$$I(0) = \inf_{t>0} \sup_{\theta} \left[\theta C - \log g_\theta(t)\right] \geq \inf_{t>0} \sup_{\theta} \left[\theta C - \limsup_{n\to\infty} \log g_\theta(\frac{t}{2^n})\right]. \tag{12}$$

By the Mean Value theorem, there exists $u \in [0, t/2^n]$ such that

$$\frac{A_1(0,t/2^n)}{t/2^n} := \frac{1}{t/2^n}\int_0^{t/2^n} r(s)\ ds = r(u).$$

Thus, we get

$$\left|\frac{A_1(0,t/2^n)}{t/2^n}\right| \leq \sup_{u\in[0,t/2^n]} |r(u)|. \tag{13}$$

Since the RHS of (13) is decreasing (non-increasing) in n, and integrable for some n by our assumption (see (A3).b), we can apply Fatou's lemma to the RHS of (12) to obtain

$$\begin{aligned}\inf_{t>0} \sup_{\theta} \left[\theta C - \limsup_{n\to\infty} \log g_\theta(\frac{t}{2^n})\right] &\geq \inf_{t>0} \sup_{\theta} \left[\theta C - \log \mathbb{E}\{\ \exp(\theta \limsup_{n\to\infty} \frac{A_1(0,t/2^n)}{t/2^n})\}\right] \\ &= \sup_{\theta} \left[\theta C - \log \mathbb{E}\{e^{\theta r(0)}\}\right], \end{aligned} \tag{14}$$

where the equality follows from the integrability assumption, i.e., $\int_0^t r(u)du = A(0,t)$. Since the RHS of (14) has the same form of the RHS of (10) with $\mathbb{E}\{r(0)\} = \lambda < C$, we have $I(0) > 0$.

Finally, note that from Proposition 1, we can write, for instance, $\mathbb{P}\{q^N(t) > 0\} \leq \exp(-NI(0) + o(N))$. Thus,

$$\sum_{N=1}^{\infty} \mathbb{P}\{q^N(t) > 0\} \leq \sum_{N=1}^{\infty} \exp(-NI(0) + o(N)) < \infty.$$

Hence by the Borel-Cantelli lemma, $\lim_{N\to\infty} q^N(t) = 0$ almost surely.

2.2. Model description

Consider Figure 1 again. In this figure, for flow i, $A_i(s,t)$ and $D_i^N(s,t)$ represent the amount of traffic arrival and departure, respectively, during a time interval $[s,t)$. The server capacity of the queue is NC and $q^N(t)$ denotes the workload at time t. As before, we assume that $A_i(s,t), i = 1,2,\ldots,N$ $(s,t \in T\ )$ are *i.i.d.* with stationary increments. Later, we will weaken this *i.i.d.* assumption and obtain the same asymptotic results (see Section 3.3).

In Scenario I, among the N flows, a *fixed* subset of the flows i $(i \in \Gamma)$ after being served at the upstream queue arrive to the downstream queue $Q_I^N(t)$ with service capacity

C_d. We let $|\Gamma|$ denote the cardinality of the set Γ. We are then interested in the queueing behavior at the second (downstream) queue in Scenario I. A parameter of interest could be estimating the overflow probability $\mathbb{P}\{Q_I^N(t) > x\}$ for a given buffer level x. Here $R(s,t)$ represents other interfering traffic. We also assume that $R(s,t)$ has stationary increments with rate $\mathbb{E}\{R(0,t)\}/t = \bar{r}$. However, in our framework, this interfering traffic need not be independent from $A_i(s,t)$, i.e., we allow arbitrary dependency between $R(s,t)$ and $A_i(s,t)$ (hence $D_i^N(s,t)$). *Note that neither the service capacity C_d at the downstream node, nor the interfering traffic $R(s,t)$ is a function of N.*

Remark 2. We have seen that the queue $q^N(t)$ operates at a large deviations scale in the sense that under certain conditions, it decreases to zero almost surely as the number of aggregated traffic flows increases. However, the downstream queue is not operating in the large deviations scale. In other words, the overflow probability of the queue $\mathbb{P}\{Q_I^N(t) > x\}$ does not tend to zero as N increases and thus, in contrast to [15], the large deviations technique does not apply to the downstream queue.

We will show in Section 3 that the random variable $Q_I^N(0)$ converges to $Q_{II}(0)$ in the sense that $\left|\mathbb{P}\{Q_I^N(0) > x\} - \mathbb{P}\{Q_{II}(0) > x\}\right|$ converges to zero as N increases, *uniformly* in x. Let the system start at $-\infty$ and the queues in Scenario I be stable, i.e., the service capacity is larger than the mean arrival rate. Let $Q_I^N(0)$ and $Q_{II}(0)$ be the steady-state workload for each scenario in Figure 1 and 2. To be precise, we write

$$\begin{aligned} Q_I^N(0) &= \sup_{t\in T} \Big[\sum_{i\in\Gamma} D_i^N(-t,0) + R(-t,0) - C_d t\Big], \quad \text{and} \\ Q_{II}(0) &= \sup_{t\in T} \Big[\sum_{i\in\Gamma} A_i(-t,0) + R(-t,0) - C_d t\Big]. \end{aligned}$$

We assume that the queue in Scenario II is stable in the sense that $\mathbb{P}\{Q_{II}(0) > x\} \to 0$ as $x \to \infty$, as long as the service capacity is greater than the mean arrival rate, i.e., $C_d > \lambda|\Gamma| + r$.

Let $v_A(t) := \mathrm{Var}\{A(-t,0)\}$ and $v_R(t) := \mathrm{Var}\{R(-t,0)\}$. Then, from Lemma 1, the functions $v_A(t)/t$ and $v_R(t)/t$ are subadditive by the convexity of a function $(\cdot)^2$ and the stationary increments property. Thus, $v_A(t)/t^2$ converges to its minimum as t increases, and so does $v_R(t)/t^2$ (see Lemma 6.1.11 in [5]). If the function $v_A(t)/t^2$

converges to a positive number, i.e., $v_A(t) \sim Vt^2$, it corresponds to a pathological case that $A_i(-t,0) = X \cdot t$ for some random variable X. Hence, to avoid trivialities, we assume that $v_A(t)/t^2$ and $v_R(t)/t^2$ converge to zero as t increases.

We define $q_i(t)$ as the stationary workload with capacity C fed by single input $A_i(s,t)$, i.e.,

$$q_i(t) = \sup_{s \le t} \big[A_i(s,t) - C(t-s)\big]. \tag{15}$$

Since $A_i(s,t)$ $(i = 1, 2, \ldots, N)$ are *i.i.d.* with stationary increments, $q_i(t)$ $(i = 1, 2, \ldots, N)$ are also *i.i.d.* and stationary. Then we assume the following:

(A4) There exists $\epsilon > 0$ such that $\mathbb{E}\{|q_i(0)|^{1+\epsilon}\} < \infty$.

Assumption (A4) is technical, but necessary to prove our main theorem in Section 3. It poses certain conditions on the behavior of the arrival process $A_i(s,t)$ that are satisfied by a large class of traffic models considered in the literature. For example, any long-range dependent traffic model with $\log \mathbb{P}\{q_i(0) > x\} \sim -\alpha x^\beta$, where $\alpha > 0$ and $0 < \beta \le 1$, satisfies assumption (A4). In fact, in this case, it is easy to see that all the moments of $q_i(0)$ exist.

3. Main results

We now state and prove our main result.

Theorem 1. *Suppose that assumptions (A1) and (A4) are satisfied. Further, if $T = \mathbb{R}$, assume that (A2) and (A3) are also satisfied. Then, we have*

$$\lim_{N \to \infty} \left| \mathbb{P}\{Q_I^N(0) > x\} - \mathbb{P}\{Q_{II}(0) > x\} \right| = 0,$$

uniformly in $x > 0$.

The following subsection is devoted to the proof of Theorem 1.

3.1. Proof of Theorem 1

As mentioned in the introduction, the difference between $Q_I^N(0)$ and $Q_{II}(0)$ depends on the entire past history of the upstream queue, i.e., $q^N(-t)$ for $t \ge 0$. To prove our theorem, we divide the whole interval $[0, \infty)$ into $[0, N^k]$ and (N^k, ∞), where $k \in \mathbb{N}$ is

chosen such that $k > 1 + 1/\epsilon$ and $\epsilon > 0$ is a number that satisfies Assumption (A4). Then, from Proposition 2, $\mathbb{P}\{\sup_{t\in[0,N^k]} q^N(-t) > 0\}$ converges to zero, due to the exponential convergence in the large deviation result on $q^N(-t)$ and the stationary nature of $q^N(-t)$. Hence, the difference between $Q_I^N(0)$ and $Q_{II}(0)$, if any, will mainly come from the behavior of $q^N(-t)$ for $t \geq N^k$, and most parts in this section will be devoted to dealing with the behavior of the upstream queue over that interval.

First, note that

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

for all $x \geq 0$ and k. Similarly,

$$\begin{aligned}
\mathbb{P}\Big\{ \sup_{0\leq t\leq N^k} [\sum_{i\in\Gamma} A_i(-t,0) + R(-t,0) - C_d t] \ > \ x \Big\} \ &\leq \ \mathbb{P}\{Q_{II}(0) > x\} \\
&\leq \ \mathbb{P}\Big\{ \sup_{0\leq t\leq N^k} [\sum_{i\in\Gamma} A_i(-t,0) + R(-t,0) - C_d t] \ > \ x \Big\} \\
&\quad + \ \mathbb{P}\Big\{ \sup_{t\geq N^k} [\sum_{i\in\Gamma} A_i(-t,0) + R(-t,0) - C_d t] \ > \ 0 \Big\}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\Big| \mathbb{P}\{Q_I^N(0) > x\} - \mathbb{P}\{Q_{II}(0) > x\} \Big| \\
&\leq \ \Big| \mathbb{P}\Big\{ \sup_{0\leq t\leq N^k} [\sum_{i\in\Gamma} D_i^N(-t,0) + R(-t,0) - C_d t] \ > \ x \Big\} \\
&\qquad - \mathbb{P}\Big\{ \sup_{0\leq t\leq N^k} [\sum_{i\in\Gamma} A_i(-t,0) + R(-t,0) - C_d t] \ > \ x \Big\} \Big| \\
&\qquad + \mathbb{P}\Big\{ \sup_{t\geq N^k} [\sum_{i\in\Gamma} A_i(-t,0) + R(-t,0) - C_d t] \ > \ 0 \Big\} \\
&\qquad + \mathbb{P}\Big\{ \sup_{t\geq N^k} [\sum_{i\in\Gamma} D_i^N(-t,0) + R(-t,0) - C_d t] \ > \ 0 \Big\} \\
&:= \ \ (I) \ + \ (II) \ + \ (III).
\end{aligned}$$

We will now show that each term (I), (II), and (III) will go to zero as N increases. Let $q_i^N(t)$ denote the amount of backlog for flow i in the queue $q^N(t)$ at time t. Then,

clearly $q^N(t) = \sum_{i=1}^N q_i^N(t)$. Since $|\sup f - \sup g| \le \sup|f-g|$, we have

$$\begin{aligned}
\Big| \sup_{0\le t\le N^k} \big[\sum_{i\in\Gamma} D_i^N(-t,0) + R(-t,0) - C_d t\big] - \sup_{0\le t\le N^k} \big[\sum_{i\in\Gamma} A_i(-t,0) + R(-t,0) - C_d t\big]\Big| \\
&\le \sup_{0\le t\le N^k} \Big|\sum_{i\in\Gamma} D_i^N(-t,0) - \sum_{i\in\Gamma} A_i(-t,0)\Big| \\
&= \sup_{0\le t\le N^k} \Big|\sum_{i\in\Gamma} q_i^N(-t) - \sum_{i\in\Gamma} q_i^N(0)\Big| \\
&\le \sup_{0\le t\le N^k} \sum_{i\in\Gamma} q_i^N(-t) \\
&\le \sup_{0\le t\le N^k} q^N(-t). \qquad (16)
\end{aligned}$$

So, using the inequality $\big|\mathbb{P}\{X>x\} - \mathbb{P}\{Y>x\}\big| \le \mathbb{P}\{|X-Y|>0\}$, we can bound the term (I) to obtain

$$\begin{aligned}
(I) &\le \mathbb{P}\Big\{\sup_{0\le t\le N^k} q^N(-t) > 0\Big\} \qquad (17)\\
&\le \exp(-NI(0) + o(N)) \;\Rightarrow\; 0,
\end{aligned}$$

as $N\to\infty$ from Proposition 1 and $I(0)>0$.

Now, pick $\delta_1 > 0$ such that $C_d - \delta_1$ is still larger than the mean arrival rate to the downstream queue (with queue-length $Q_{II}(0)$), i.e., $\lambda|\Gamma| + \bar{r} < C_d - \delta_1 < C_d$, where $\lambda = \mathbb{E}\{A(-t,0)/t\}$ and $\bar{r} = \mathbb{E}\{R(-t,0)/t\}$. Then by splitting $A_i(-t,0)$ into $A_i(-t,-N^k) + A_i(-N^k,0)$ for $t \ge N^k$ (similarly for $R(-t,0)$), we have

$$\begin{aligned}
(II) &= \mathbb{P}\Big\{\sup_{t\ge N^k} \big[\sum_{i\in\Gamma} A_i(-t,0) + R(-t,0) - C_d t\big] > 0\Big\} \qquad (18)\\
&= \mathbb{P}\Big\{\sup_{t\ge N^k} \Big[\sum_{i\in\Gamma} A_i(-t,-N^k) + R(-t,-N^k) + \sum_{i\in\Gamma} A_i(-N^k,0) + R(-N^k,0) \\
&\quad - (C_d-\delta_1)(t-N^k) - \delta_1 t\Big] > (C_d-\delta_1)N^k\Big\} \\
&\le \mathbb{P}\Big\{\sup_{t\ge N^k} \Big[\sum_{i\in\Gamma} A_i(-t,-N^k) + R(-t,-N^k) - (C_d-\delta_1)(t-N^k)\Big] > \delta_1 N^k\Big\} \\
&\quad + \mathbb{P}\Big\{\sum_{i\in\Gamma} A_i(-N^k,0) + R(-N^k,0) > (C_d-\delta_1)N^k\Big\}. \qquad (19)
\end{aligned}$$

From the stationary increments assumption on $A_i(s,t)$, the first term of the RHS of (19) is equal to

$$\mathbb{P}\Big\{\sup_{t\ge 0} \big[\sum_{i\in\Gamma} A_i(-t,0) + R(-t,0) - (C_d-\delta_1)t\big] > \delta_1 N^k\Big\}, \qquad (20)$$

which decreases to zero as N increases since the downstream queue, with service capacity C_d replaced by $C_d - \delta_1$, is stable. Similarly, it is not difficult to see that the second term of the RHS of (19) also decreases to zero. To see this, first choose two positive numbers a and b such that $a + b = C_d - \delta_1$ with $\lambda|\Gamma| < a$ and $\bar{r} < b$. This is always possible since $\lambda|\Gamma| + \bar{r} < C_d - \delta_1 = a + b$. Then, we have

$$\begin{aligned}
&\mathbb{P}\Big\{ \sum_{i\in\Gamma} A_i(-N^k, 0) + R(-N^k, 0) \ > \ (C_d - \delta_1)N^k \Big\} \\
&\quad \leq \ \mathbb{P}\{ \sum_{i\in\Gamma} A_i(-N^k, 0) > aN^k \} + \mathbb{P}\{ R_i(-N^k, 0) > bN^k \} \\
&\quad \leq \ \frac{1}{(a - \lambda|\Gamma|)^2} \frac{|\Gamma| v_A(N^k)}{(N^k)^2} + \frac{1}{(b - \bar{r})^2} \frac{v_R(N^k)}{(N^k)^2} \qquad (21)
\end{aligned}$$

by Chebyshev's inequality. Since $v_A(t)/t^2$ and $v_R(t)/t^2$ decrease to zero as t increases and by our choice of a and b, we have shown that (21) converges to zero as N increases. Hence (II) also converges to zero as N increases.

We now focus on the third term (III). Pick $\delta_1, \delta_2 > 0$ such that $\lambda|\Gamma| + \bar{r} < C_d - \delta_1 - \delta_2 < C_d - \delta_1 < C_d$. Since $D_i^N(-t, 0) \leq A_i(-t, 0) + q_i^N(-t)$ for any t and N, (III) is bounded by

$$\begin{aligned}
(III) \ &\leq \ \mathbb{P}\Big\{ \sup_{t\geq N^k} [\sum_{i\in\Gamma} q_i^N(-t) + \sum_{i\in\Gamma} A_i(-t, 0) + R(-t, 0) - C_d t] \ > \ 0 \Big\} \\
&\leq \ \mathbb{P}\Big\{ \sup_{t\geq N^k} \ [\sum_{i\in\Gamma} A_i(-t, 0) + R(-t, 0) - (C_d - \delta_2)t] \ > \ 0 \Big\} \\
&\quad + \mathbb{P}\Big\{ \sup_{t\geq N^k} [\sum_{i\in\Gamma} q_i^N(-t) - \delta_2 t] \ > \ 0 \Big\}. \qquad (22)
\end{aligned}$$

Since the first term of the RHS of (22) is identical to (18) except C_d being replaced by $C_d - \delta_2$, it can also be shown to converge to zero by repeating the same steps in (18) – (21) with C_d replaced by $C_d - \delta_2$. We now have

$$\begin{aligned}
\mathbb{P}\Big\{ \sup_{t\geq N^k} [\sum_{i\in\Gamma} q_i^N(-t) - \delta_2 t] \ > \ 0 \Big\} \ &= \ \mathbb{P}\Big\{ \sup_{t\geq N^k} [\sum_{i\in\Gamma} \frac{q_i^N(-t)}{t}] \ > \ \delta_2 \Big\} \\
&\leq \ \mathbb{P}\Big\{ \sup_{t\geq N^k} \frac{q^N(-t)}{t} > \delta_2 \Big\}.
\end{aligned}$$

Thus, we only need to show that $\mathbb{P}\{ \sup_{t\geq N^k} \frac{q^N(-t)}{t} > \delta_2 \} \to 0$ as $N \uparrow \infty$ and we are done.

Lemma 3. *For any* $\delta > 0$, $\lim_{N\to\infty} \mathbb{P}\{ \sup_{t\geq N^k} q^N(-t)/t > \delta \} = 0$.

Proof of Lemma 3. Discrete-time case ($T = \mathbb{Z}$): Note that

$$\sup_{t \geq N^k} \frac{q^N(-t)}{t} \leq \sup_{t \geq N^k} \frac{q^N(-t)}{t^{\frac{k-1}{k}} N}.$$

For a given $\delta > 0$, define

$$B_{t,N} := \Big\{ \frac{q^N(-t)}{t^{\frac{k-1}{k}} N} > \delta \Big\}. \tag{23}$$

Then we see that

$$\mathbb{P}\big\{ \sup_{t \geq N^k} \frac{q^N(-t)}{t} > \delta \big\} \leq \mathbb{P}\big\{ \cup_{t=N^k}^{\infty} B_{t,N} \big\}. \tag{24}$$

Let $\epsilon > 0$ be the positive number in assumption (A4), and note that

$$\begin{aligned} q^N(-t) &:= \sup_{s \geq t} \big[\sum_{i=1}^{N} A_i(-s,-t) - CN(s-t) \big] \\ &\leq \sum_{i=1}^{N} \sup_{s \geq t} \big[A_i(-s,-t) - C(s-t) \big] \\ &= \sum_{i=1}^{N} q_i(-t), \end{aligned}$$

where $q_i(-t)$ is defined as in (15). Then, from the convexity of a function $(\cdot)^{1+\epsilon}$ and Jensen's inequality, we have

$$\left(\frac{q^N(-t)}{N} \right)^{1+\epsilon} \leq \left(\frac{1}{N} \sum_{i=1}^{N} q_i(-t) \right)^{1+\epsilon} \leq \frac{1}{N} \sum_{i=1}^{N} (q_i(-t))^{1+\epsilon}.$$

Since $q_i(-t)$ is *i.i.d.*, by taking expectation, we get

$$\mathbb{E}\big\{ \big(\frac{q^N(-t)}{N} \big)^{1+\epsilon} \big\} \leq \mathbb{E}\{ (q_i(-t))^{1+\epsilon} \} := M < \infty \tag{25}$$

from assumption (A4). By Markov's inequality, we have from (23) and (25),

$$\mathbb{P}\{B_{t,N}\} \leq \frac{M}{\delta^{1+\epsilon}} \cdot \frac{1}{t^{\frac{k-1}{k}(1+\epsilon)}} = \frac{M}{\delta^{1+\epsilon}} \cdot \frac{1}{t^p},$$

where $p := \frac{k-1}{k}(1+\epsilon) > 1$ by the choice of k. (Recall that $k > 1 + \frac{1}{\epsilon}$.) Thus,

$$\mathbb{P}\big\{ \cup_{t=N^k}^{\infty} B_{t,N} \big\} \leq \sum_{t=N^k}^{\infty} \mathbb{P}\{B_{t,N}\} \leq \sum_{t=N^k}^{\infty} \frac{M}{\delta^{1+\epsilon}} \cdot \frac{1}{t^p} \leq M_1 \cdot \frac{1}{(N^k)^{p-1}}$$

for some constant $M_1 < \infty$. Hence, from (24), we are done.

Continuous-time case ($T = \mathbb{R}$): We use similar techniques as in the proof of Proposition 1 for the continuous-time case. We first divide the interval $[N^k, \infty)$ into small intervals, each of which has equal length h, and then work within each interval. Specifically, let $s_n := N^k + nh$ and $S(n,h) = [s_{n-1}, s_n]$ where $n = 1, 2, \ldots$. Then, we have

$$\begin{aligned}\sup_{t \geq N^k} \frac{q^N(-t)}{t} &\leq \sup_{n \geq 1} \sup_{t \in S(n,h)} \frac{q^N(-t)}{t} \\ &\leq \sup_{n \geq 1} \frac{q^N(-s_n)}{s_n} + \sup_{n \geq 1} \sup_{t \in S(n,h)} \left| \frac{q^N(-t)}{t} - \frac{q^N(-s_n)}{s_n} \right|. \end{aligned} \tag{26}$$

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

$$\left| \frac{q^N(-t)}{t} - \frac{q^N(-s_n)}{s_n} \right| \leq \frac{q^N(-s_n)}{s_n} \left| \frac{u}{s_n - u} \right| + \frac{1}{s_n - u} \Big| q^N(-s_n + u) - q^N(-s_n) \Big|.$$

Since $s_{n-1} = s_n - h$ by definition and $|\frac{u}{s_n - u}| \leq 1$ for all n and $u \in [0,h]$ whenever $h \leq N^k/2$, we have for a fixed $h > 0$,

$$\begin{aligned}\sup_{t \in S(n,h)} \left| \frac{q^N(-t)}{t} - \frac{q^N(-s_n)}{s_n} \right| &\leq \frac{q^N(-s_n)}{s_n} \\ &+ \frac{1}{s_{n-1}} \sup_{u \in [0,h]} \Big| q^N(-s_n + u) - q^N(-s_n) \Big|, \end{aligned} \tag{27}$$

for sufficiently large N. Observe that for any s and any positive number t, we have

$$q^N(s) + \sum_{i=1}^N A_i(s, s+t) - CNt \leq q^N(s+t) \leq q^N(s) + \sum_{i=1}^N A_i(s, s+t).$$

Thus, we can find an upper bound on the supremum in (27) as

$$\begin{aligned}&\sup_{u \in [0,h]} \Big| q^N(-s_n + u) - q^N(-s_n) \Big| \\ &\leq \sum_{i=1}^N \left(\sup_{u \in [0,h]} \Big| A_i(-s_n, -s_n + u) \Big| + Ch \right) \\ &:= \sum_{i=1}^N J_i(N,h) \end{aligned} \tag{28}$$

where

$$J_i(N,h) := \sup_{u \in [0,h]} \Big| A_i(-s_n, -s_n + u) \Big| + Ch. \tag{29}$$

Combining (26) – (28), we have

$$\mathbb{P}\left\{\sup_{t\geq N^k}\frac{q^N(-t)}{t}>\delta\right\} \leq \mathbb{P}\left\{\sup_{n\geq 1}\frac{q^N(-s_n)}{s_n} > \frac{\delta}{3}\right\} + \mathbb{P}\left\{\sup_{n\geq 1}\frac{1}{s_{n-1}}\sum_{i=1}^{N}J_i(N,h) > \frac{\delta}{3}\right\}. \quad (30)$$

Note that

$$\frac{1}{s_n}=\frac{1}{N^k+nh}\leq\frac{1}{(N^k+nh)^{\frac{k-1}{k}}N}. \quad (31)$$

Using the same steps as in the discrete-time case and with the same choice of $p := \frac{k-1}{k}(1+\epsilon) > 1$, we can bound the first term of the RHS of (30) as

$$\begin{aligned}\mathbb{P}\left\{\sup_{n\geq 1}\frac{q^N(-s_n)}{s_n}>\frac{\delta}{3}\right\} &\leq \sum_{n=1}^{\infty}\mathbb{P}\left\{\frac{q^N(-s_n)}{s_n}>\frac{\delta}{3}\right\}\\ &\leq \sum_{n=1}^{\infty}\frac{1}{(\delta/3)^{1+\epsilon}}\frac{h}{(N^k+nh)^p}\frac{M}{h}\\ &\leq \frac{1}{(\delta/3)^{1+\epsilon}}\frac{M}{h}\int_{N^k}^{\infty}x^{-p}dx\\ &= \frac{1}{(\delta/3)^{1+\epsilon}}\frac{M}{h}\frac{1}{(p-1)(N^k)^{p-1}},\end{aligned}$$

where M is defined in (25). Thus the first term of the RHS of (30) decreases to zero as N increases.

We will show that the second term of the RHS of (30) also converges to zero as N increases. Since $A_i(s,t)$ has stationary increments, assumption (A2) gives

$$\mathbb{E}\{(J_1(N,h))^{1+\epsilon}\}=\mathbb{E}\left\{\left(\sup_{u\in[0,h]}|A_1(0,u)|+Ch\right)^{1+\epsilon}\right\}:=M_2<\infty, \quad (32)$$

by making use of the inequality $x^{1+\epsilon}\leq e^x$ for small $\epsilon>0$ to show that M_2 is finite. Note that since $A_i(s,t)$ $(i=1,2,\ldots,N)$ are *i.i.d.*, $J_i(N,h)$ are also *i.i.d.*. (See the definition of $J_i(N,h)$ in (29).) Thus, from Jensen's inequality and the convexity of a function $(\cdot)^{1+\epsilon}$, we have

$$\mathbb{E}\left\{\left(\frac{\sum_{i=1}^{N}J_i(N,h)}{N}\right)^{1+\epsilon}\right\}\leq\mathbb{E}\{(J_1(N,h))^{1+\epsilon}\}=M_2, \quad (33)$$

where M_2 is defined in (32). Thus, similarly as above, using (31) again for $s_{n-1} = N^k+(n-1)h$,

$$\mathbb{P}\left\{\sup_{n\geq 1}\frac{1}{s_{n-1}}\sum_{i=1}^{N}J_i(N,h)>\frac{\delta}{3}\right\} \leq \frac{1}{(\delta/3)^{1+\epsilon}}\frac{M_2}{h}\frac{1}{(p-1)(N^k-h)^{p-1}},$$

where $h > 0$ is a fixed number and $p > 1$. Hence the second term of the RHS of (30) also converges to zero as N increases. Thus Lemma 3 also holds for the continuous-time case. This completes the proof of Theorem 1.

Remark 3. One might intuitively expect that N i.i.d. traffic flows should get 'shaped' as they pass through the upstream queue, making the departure process $\sum_{i\in\Gamma} D_i^N(-t,0)$ behave smoother than the arrival process $\sum_{i\in\Gamma} A_i(-t,0)$. However, this is not the case. Certainly, the total departure, i.e., $\sum_{i=1}^N D_i^N(-t,0)$ would be bounded by the capacity NC, but *each* departure flow $D_i^N(-t,0)$ for $i \in \Gamma$ can be larger than its corresponding arrival $A_i(-t,0)$. To see this, note that $D_i^N(-t,0) = A_i(-t,0) + q_i^N(-t) - q_i^N(0)$. Since N flows completely share the total capacity NC, the amount of backlog for flow i in the upstream queue ($q_i^N(-t)$) can be as large as the total backlog in the queue ($q^N(-t)$), and also, $q_i^N(-t)$ for some $t > 0$ can be much larger than $q_i^N(0)$ due to the fluctuation of the queue-length and the interaction among N flows in the upstream queue.

3.2. Speed of convergence

From Theorem 1, we know that $\mathbb{P}\{Q_I^N(0) > x\}$ converges to $\mathbb{P}\{Q_{II}(0) > x\}$. In this section, we investigate its speed of convergence: How fast does it converge? To answer this question, we need the following assumptions.

(A5) There exist $H_1, H_2 \in (0,1)$ such that $v_A(t) := \mathrm{Var}\{A_i(-t,0)\} \in O(t^{2H_1})$ and $v_R(t) := \mathrm{Var}\{R(-t,0)\} \in O(t^{2H_2})$.[§]

(A6) $\mathbb{E}\{Q_{II}(0)\} < \infty$, if the queue is stable, i.e., $C_d > \lambda|\Gamma| + \bar{r}$.

Remark 4. The parameters H_1 and H_2 in assumption (A5) correspond to the so-called "Hurst parameter" in the literature. This has been widely used to model long-range dependent traffic or self-similar traffic [10, 1, 3], for which $H \in (0.5, 1)$. Note that assumption (A6) is purely technical and almost trivial in that it only requires finiteness of the expected workload when the service capacity is larger than the mean arrival rate.

[§] $f(t) \in O(g(t))$ means $\limsup_{t\to\infty} f(t)/g(t) < \infty$.

We already observed that Proposition 1 holds even if we replace N^k by any function $g(N)$ with $\log g(N)/N \downarrow 0$. In other words, $\sup_{t\in[0,g(N)]} q^N(-t)$ satisfies the same large deviations principle as $q^N(-t)$, i.e.,

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\Big\{ \sup_{t\in[0,g(N)]} q^N(-t) > Nb \Big\} \le -I(b).$$

In fact, we can go one step further. Instead of being slower than exponential, we choose $g(N)$ such that $\limsup_{N\to\infty} \log g(N)/N < I(0)$. For example, we set $g(N) = \exp(cN)$ where $0 < c < I(0)$. Then, clearly, we see that, for $t \in \mathbb{Z}$,

$$\mathbb{P}\Big\{ \sup_{0\le t\le e^{cN}} q^N(-t) > 0 \Big\} \le \Big(\lfloor \exp(cN) \rfloor + 1 \Big) \exp(-NI(0) + o(N)),$$

which gives

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\Big\{ \sup_{0\le t\le e^{cN}} q^N(-t) > 0 \Big\} \;\le\; c - I(0) \;<\; 0. \tag{34}$$

Similarly, for $t \in \mathbb{R}$, we see that from the last part of the proof of Proposition 1,

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\Big\{ \sup_{0\le t\le e^{cN}} q^N(-t) > 0 \Big\} \le \limsup_{N\to\infty} \frac{1}{N} \log \Big(\lfloor \frac{\exp(cN)}{\epsilon} \rfloor + 1 \Big) - pI(0) + \delta(\epsilon),$$

for any $p \in (0,1)$ and any $\epsilon > 0$, where $\delta(\epsilon) \downarrow 0$ as $\epsilon \downarrow 0$. Thus, (34) still holds by taking $\epsilon \downarrow 0$ and $p \uparrow 1$. These observations allow us to develop the following result on the speed of convergence in Theorem 1.

Corollary 1. *Suppose all the assumptions in Theorem 1 hold and also assumptions (A5) and (A6) hold. Then, there exists a positive constant J^* $(0 < J^* < I(0))$ such that*

$$\limsup_{N\to\infty} \frac{1}{N} \log \Big(\sup_{x\ge 0} \Big| \mathbb{P}\{Q_I^N(0) > x\} - \mathbb{P}\{Q_{II}(0) > x\} \Big| \Big) \le -J^*.$$

Proof. From the proof of Theorem 1, note that the difference of the overflow probabilities $|\mathbb{P}\{Q_I^N(0) > x\} - \mathbb{P}\{Q_{II}(0) > x\}|$ is bounded by the following terms: (17), (20), (21) and (22). Fix a number c $(0 < c < I(0))$, and let $g(N) = \exp(cN)$. We then replace N^k by $g(N)$ at its every occurrence in the proof of Theorem 1. First, the term (17) after substitution is bounded by

$$\mathbb{P}\Big\{ \sup_{0\le t\le g(N)} q^N(-t) > 0 \Big\} \tag{35}$$

Second, note that from Markov's inequality and assumption (A6), (20) is bounded by $\frac{K_1}{g(N)}$ for some constant $K_1 > 0$. Similarly, from assumption (A5), we can find positive constants K_2 and K_3 such that (21) is bounded by

$$\frac{K_2}{(g(N))^{2-2H_1}} + \frac{K_3}{(g(N))^{2-2H_2}}. \tag{36}$$

Lastly, we will deal with the term (22). Clearly, the first term of (22) is bounded by an upper bound of the form of (36). The second term of (22) then becomes

$$\mathbb{P}\Big\{ \sup_{t \geq g(N)} \frac{q^N(-t)}{t} > \delta_2 \Big\}.$$

Since $g(N) = \exp(cN) \geq N^k$ for sufficiently large N, we can write

$$\sup_{t \geq g(N)} \frac{q^N(-t)}{t} \leq \sup_{t \geq g(N)} \frac{q^N(-t)}{t^{\frac{k-1}{k}} N}.$$

Following the same steps used in the proof of Lemma 3 gives

$$\mathbb{P}\Big\{ \sup_{t \geq g(N)} \frac{q^N(-t)}{t} > \delta_2 \Big\} \leq \frac{K_3}{(g(N))^{p-1}}, \tag{37}$$

for some constant $K_3 > 0$ and $p > 1$. Hence, we can conclude from (35), (36) and (37) that, there exist positive constants K and v such that

$$\sup_{x \geq 0} \Big| \mathbb{P}\{Q_I^N(0) > x\} - \mathbb{P}\{Q_{II}(0) > x\} \Big| \leq \mathbb{P}\Big\{ \sup_{0 \leq t \leq g(N)} q^N(-t) > 0 \Big\} + \frac{K}{(g(N))^v}, \tag{38}$$

where the constant v is independent of $g(N)$. (In fact, v can be chosen as $v = \min\{2 - 2H_1, 2 - 2H_2, p - 1\}$.)

Now, observe that $g(N) = \exp(cN)$ gives

$$\limsup_{N \to \infty} \frac{1}{N} \log \Big(\frac{K}{(g(N))^v} \Big) = -vc. \tag{39}$$

Hence, from (34), (38) and (39), it follows that

$$\limsup_{N \to \infty} \frac{1}{N} \log \left(\sup_{x \geq 0} \Big| \mathbb{P}\{Q_I^N(0) > x\} - \mathbb{P}\{Q_{II}(0) > x\} \Big| \right) \leq -\min\{I(0) - c, vc\} := -J^*,$$

where J^* is positive since $0 < c < I(0)$. This completes the proof.

Remark 5. Since the number $c \in (0, I(0))$ can be chosen arbitrarily, we can optimize the constant $J^* = J^*(c) = \min\{I(0) - c, vc\}$ over $c \in (0, I(0))$. The maximum of $J^*(c)$ occurs at $c = I(0)/(v+1)$, and the constant J^* then becomes $J^* = \frac{v}{v+1} I(0)$.

Corollary 1 implies that the speed of convergence is at least exponentially fast. In other words, we can write

$$\sup_{x\geq 0}\left|\mathbb{P}\{Q_I^N(0)>x\}-\mathbb{P}\{Q_{II}(0)>x\}\right|\leq e^{-J^*N+o(N)}. \tag{40}$$

Thus, for any given target QoS ($\mathbb{P}\{Q_I^N(0) > x\}$) in the original two-stage queueing system (Scenario I), we can replace it by a simpler system in which the first node has been removed and the error is less than the RHS of (40). Note that the error term $e^{-J^*N+o(N)}$ bounds the *maximum* difference of the overflow probabilities. Thus, for a given (fixed) buffer level, the actual error will be much smaller than the uniform bound.

3.3. Extension to non-*i.i.d.* traffic arrivals

We have assumed that the arrival processes to the upstream queue in Figure 1 are *i.i.d.* In this section, we extend the result to the case of non-*i.i.d.* traffic arrivals.

Let M (M is finite) be the number of different traffic classes and $S_j(N)$ be the set of flows of class j ($j = 1, 2, \ldots, M$). As before, we scale the upstream queue such that the total number of flows and total service capacity increase proportionally. Note, however, that the number of flows for each class can be arbitrary as long as

$$\sum_{j=1}^{M}|S_j(N)|=N, \tag{41}$$

where $|S_j(N)|$ represents the number of flows for class j. We assume that $|S_j(N)|/N$ converges to a number $\gamma_k \in [0, 1]$ as N increases. We also assume that within a class, the flows are *i.i.d.*, i.e., $A_i(s,t), i \in S_j(N)$ are *i.i.d.* for any j. We however allow the flows for different classes to be possibly *heterogeneous* and *dependent*. Then, we will show that all the results in earlier sections still remain valid.

Before we proceed, we need the following lemma. Its proof is deferred to the end of this section.

Lemma 4. *Assume $A_i(s,t) \geq 0$. Then, under the same assumption of Proposition 1 for traffic flows in each class j, ($j = 1, 2, \ldots, M$), there exists a positive constant I^* such that*

$$\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}\{q^N(t)>0\}\leq -I^*. \tag{42}$$

Proposition 3. *Suppose that we require the corresponding assumptions for each traffic arrival only within a class, (we allow any heterogeneity and dependency among different classes.) and that $A_i(s,t)$, the amount of traffic arrival in $[s,t)$ is nonnegative. Then Theorem 1 and Corollary 1 remain unchanged.*

Proof of Proposition 3. We can write the workload of the upstream queue at time t as

$$q^N(t) = \sup_{s \le t} \left[\sum_{i=1}^{N} A_i(s,t) - CN(t-s) \right].$$

This is identical to (2), except that $A_i(s,t)$ are not *i.i.d.* anymore. Note that the distribution of $q^N(t)$ is still independent of t due to the stationarity of $A_i(s,t)$. In the proof of Theorem 1, (17), (25) and (33) are the only places in which the *i.i.d.* assumption plays a role.

From Lemma 4, (17) directly follows with $I(0)$ replaced by I^*. For (25), we need to show that $\mathbb{E}\{(q^N(t)/N)^{1+\epsilon}\}$ is still finite. Consider a sequence of positive numbers C_j, $(j = 1, 2, \ldots, M)$, satisfying

$$\sum_{j=1}^{M} C_j |S_j(N)| = NC. \tag{43}$$

This can be thought of as an assignment of the total capacity NC to each class such that class j exclusively consumes capacity of $C_j|S_j(N)|$. Let λ_j be the mean arrival rate of a single flow of class j. Then, we choose C_j such that the capacity to class j is proportional to the aggregate arrival rate of class j, i.e.,

$$C_j = NC \cdot \frac{\lambda_j}{\sum_{j=1}^{M} \lambda_j |S_j(N)|}, \quad j = 1, 2, \ldots, M. \tag{44}$$

Clearly, C_j in (44) satisfies (43), and from the stability condition ($\sum_{j=1}^{M} \lambda_j |S_j(N)| < NC$), we have $\lambda_j < C_j$ for each j.

Now, observe that

$$\begin{aligned} q^N(t) &= \sup_{s\le t}\ \Big[\sum_{i=1}^{N} A_i(s,t) - CN(t-s)\Big] \\ &= \sup_{s\le t}\ \Big[\sum_{j=1}^{M}\Big(\sum_{i\in S_j(N)} A_i(s,t) - C_j|S_j(N)|(t-s)\Big)\Big] \\ &\le \sum_{j=1}^{M}\sum_{i\in S_j(N)} \sup_{s\le t}\big[A_i(s,t) - C_j(t-s)\big] \\ &= \sum_{j=1}^{M}\sum_{i\in S_j(N)} \hat{q}_j(t), \end{aligned}$$

where $\hat{q}_j(t) := \sup_{s\le t}[A_i(s,t) - C_j(t-s)]$. Since $\lambda_j < C_j$, from assumption (A4), we get

$$\mathbb{E}\{(\hat{q}_j(t))^{1+\epsilon}\} = K_j < \infty$$

Thus, for each N,

$$\begin{aligned} \mathbb{E}\Big\{\Big(\frac{q^N(t)}{N}\Big)^{1+\epsilon}\Big\} &\le \mathbb{E}\Big\{\Big(\frac{1}{N}\sum_{j=1}^{M}\sum_{i\in S_j(N)}\hat{q}_j(t)\Big)^{1+\epsilon}\Big\} \\ &\le \frac{1}{N}\sum_{j=1}^{M}\sum_{i\in S_j(N)}\mathbb{E}\{(\hat{q}_j(t))^{1+\epsilon}\} \\ &\le \max\big\{K_1, K_2, \ldots, K_M\big\} \\ &< \infty, \end{aligned}$$

where the second inequality follows from the convexity of a function $(\cdot)^{1+\epsilon}$ and Jensen's inequality. Similarly, (33) is bounded by noting that

$$\mathbb{E}\left\{\Big(\frac{\sum_{i=1}^{N} J_i(N,h)}{N}\Big)^{1+\epsilon}\right\} \le \max_{i=1,2,\ldots,M}\mathbb{E}\{(J_i(N,h))^{1+\epsilon}\} < \infty.$$

Finally, it is straightforward to see that Corollary 1 follows from Lemma 4 with $0 < J^* < I^*$.

Lemma 4 lies at the heart of Proposition 3. We do not require traffic flows for different classes to be homogeneous or independent. In [2], similar attempts have been made to establish the many-sources-asymptotic under heterogeneous traffic sources. However, the authors required independence among all the sources and more stringent assumptions than (A1).

Proof of Lemma 4. Since M is finite, from (41), there exists at least one class j such that $\lim_{N\to\infty} |S_j(N)|/N = \gamma_j > 0$, otherwise (41) is violated. Without loss of generality, let class 1 be such a class, i.e.,

$$\lim_{N\to\infty} \frac{|S_1(N)|}{N} = \gamma_1 > 0. \tag{45}$$

We start from the (proportional) assignment given by the relation (44). Instead of using C_1, we take a small (unscaled) amount of service capacity from C_1 and distribute this surplus equally to all the other classes, while maintaining stability of every queue. Specifically, define a new set of assignment $\widetilde{C}_j$ such that

$$\begin{aligned} \widetilde{C}_1 &= C_1 - \epsilon, \quad \text{and} \\ \widetilde{C}_j|S_j(N)| &= C_j|S_j(N)| + \frac{\epsilon|S_1(N)|}{M-1}, \quad j \neq 1, \end{aligned} \tag{46}$$

where $\epsilon = (C_1 - \lambda_1)/2 > 0$. Then, clearly, for each $j = 1, 2, \ldots, M$, $\widetilde{C}_j$ satisfies the relation (43) and $\lambda_j < \widetilde{C}_j$.

Note that

$$\begin{aligned} q^N(t) &= \sup_{s\le t} \Big[\sum_{i=1}^{N} A_i(s,t) - CN(t-s)\Big] \\ &= \sup_{s\le t} \Big[\sum_{j=1}^{M} \Big(\sum_{i\in S_j(N)} A_i(s,t) - \widetilde{C}_j|S_j(N)|(t-s)\Big)\Big] \\ &\le \sum_{j=1}^{M} \sup_{s\le t} \Big(\sum_{i\in S_j(N)} A_i(s,t) - \widetilde{C}_j|S_j(N)|(t-s)\Big) \\ &= \sum_{j=1}^{M} \widetilde{q}_j^N(t), \end{aligned}$$

where

$$\widetilde{q}_j^N(t) := \sup_{s\le t} \Big[\sum_{i\in S_j(N)} A_i(s,t) - \widetilde{C}_j|S_j(N)|(t-s)\Big]. \tag{47}$$

Then, clearly,

$$\mathbb{P}\{q^N(t) > 0\} \le \mathbb{P}\left\{\sum_{j=1}^{M} \widetilde{q}_j^N(t) > 0\right\} \le \sum_{j=1}^{M} \mathbb{P}\{\widetilde{q}_j^N(t) > 0\}. \tag{48}$$

For class 1, from (45), we have

$$\begin{aligned}\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}\{\widetilde{q}_1^N(t)>0\} &= \limsup_{N\to\infty}\frac{|S_1(N)|}{N}\frac{1}{|S_1(N)|}\log\mathbb{P}\{\widetilde{q}_1^N(t)>0\}\\ &= \gamma_1\cdot\limsup_{N\to\infty}\frac{1}{|S_1(N)|}\log\mathbb{P}\{\widetilde{q}_1^N(t)>0\}\\ &\le -\gamma_1 I_1(0)\end{aligned}$$

for some positive constant $I_1(0)$. This is because $|S_1(N)|\uparrow\infty$ as $N\uparrow\infty$, and $\widetilde{q}_1^N(t)$ consists of $|S_1(N)|$ *i.i.d.* traffic arrivals, each of which satisfies the assumption of Proposition 1 (see (47)). In fact, for any class j with $\lim_{N\to\infty}|S_j(N)|/N>0$, we can repeat this procedure, yielding positive rate of $\gamma_j I_j(0)$. Thus, if every class satisfies $\lim_{N\to\infty}|S_j(N)|/N=\gamma_j>0$, then we see that from (48),

$$\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}\{q^N(t)>0\} \le \limsup_{N\to\infty}\frac{1}{N}\log\left(\sum_{j=1}^{M}\mathbb{P}\{\widetilde{q}_j^N(t)>0\}\right) \tag{49}$$

$$\le -\min\{\gamma_1 I_1(0),\ldots,\gamma_M I_M(0)\}. \tag{50}$$

In fact, in case that $\lim_{N\to\infty}|S_j(N)|/N>0$ for all j, we don't even need to define $\widetilde{C}_j$, and the proportional assignment C_j as in (44) would suffice. Hence, overall, there exists a positive constant I^* given by $I^*=\min\{\gamma_1 I_1(0),\ldots,\gamma_M I_M(0)\}$.

Now we will deal with a class j for which $\lim_{N\to\infty}|S_j(N)|/N=0$, if exists. Without loss of generality, let class 2 be such a class, i.e.,

$$\lim_{N\to\infty}\frac{|S_2(N)|}{N}=0. \tag{51}$$

Then, for class 2, note that there are $|S_2(N)|$ number of traffic flows being aggregated at the queue $\widetilde{q}_2^N(t)$, associated with capacity $\widetilde{C}_2|S_2(N)|$ defined by the relation (46). Since we have, from (45) and (51),

$$\lim_{N\to\infty}\frac{\widetilde{C}_2|S_2(N)|}{N}=\lim_{N\to\infty}\frac{1}{N}\left(C_2|S_2(N)|+\frac{\epsilon|S_1(N)|}{M-1}\right)=\frac{\epsilon\gamma_1}{M-1}>0,$$

there exists a positive constant $\kappa>0$ such that $\widetilde{C}_2|S_2(N)|\ge\kappa N$ for all N.

Pick a positive number $\bar{C}_2$ such that $\bar{C}_2>\lambda_2$. Then, we imagine a queue $q_2^{L,N}(t)$ with input $A_i(s,t)$, $i\in S_2(N)$ (class 2) and service capacity $\bar{C}_2$, scaled up by $L_2(N):=\lfloor\frac{\kappa N}{\bar{C}_2}\rfloor$. In other words, we write

$$q_2^{L,N}(t):=\sup_{s\le t}\Big[\sum_{L_2(N)}A_i(s,t)-\bar{C}_2L_2(N)(t-s)\Big],$$

where $\sum_{L_2(N)}$ denotes the superposition of $L_2(N)$ *i.i.d.* copies of a single traffic flow of class 2. Then, from (51) and the definition of $L_2(N)$ ($:= \lfloor \frac{\kappa N}{\bar{C}_2} \rfloor$), we have $L_2(N)/|S_2(N)| \uparrow \infty$ as $N \uparrow \infty$ and $\bar{C}_2 L_2(N) \leq \kappa N \leq \widetilde{C}_2 |S_2(N)|$. Thus, we see that the queue $q_2^{L,N}(t)$ has smaller capacity than $\widetilde{q}_2^N(t)$ and much larger number (*i.i.d.* copies) of input traffic than $\widetilde{q}_2^N(t)$, for sufficiently large N. Hence, from $A_i(s,t) \geq 0$ and the definition of $\widetilde{q}_j^N(t)$ in (47), we get

$$\begin{aligned} \widetilde{q}_2^N(t) &= \sup_{s \leq t} \Big[\sum_{i \in S_2(N)} A_i(s,t) - \widetilde{C}_2 |S_2(N)|(t-s) \Big] \\ &\leq \sup_{s \leq t} \Big[\sum_{L_2(N)} A_i(s,t) - \bar{C}_2 L_2(N)(t-s) \Big] = q_2^{L,N}(t), \end{aligned} \tag{52}$$

for sufficiently large N. Thus, from the inequality in (52), we have

$$\begin{aligned} \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\{\widetilde{q}_2^N(t) > 0\} &= \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\{q_2^{L,N}(t) > 0\} \\ &= \limsup_{N\to\infty} \frac{L_2(N)}{N} \frac{1}{L_2(N)} \log \mathbb{P}\{q_2^{L,N}(t) > 0\} \\ &\leq \frac{\kappa}{\bar{C}_2} \cdot \limsup_{N\to\infty} \frac{1}{L_2(N)} \log \mathbb{P}\{q_2^{L,N}(t) > 0\} \\ &\leq -\frac{\kappa}{\bar{C}_2} \cdot I_2(0), \end{aligned}$$

for some positive constant $I_2(0)$. Hence, by this construction, we can always find a positive constant I_j^* such that

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\{\widetilde{q}_j^N(t) > 0\} \leq -I_j^*.$$

Therefore, similarly as in (49) and (50), there exists a positive constant $I^* = \min\{I_1^*, \ldots, I_M^*\}$ satisfying (42). This completes the proof of Lemma 4.

4. Discussion

Thus far, we have analyzed a two-stage queueing system where traffic aggregation happens only at the upstream queue. In view of Theorem 1, the system behaves as if the upstream queue does not exist. In this section, we briefly discuss and raise some issues on the implications and extensions of our results.

First, as mentioned in the introduction, our results are also applicable to the case where $|\Gamma|$, the number of flows feeding the downstream queue also increases as N

increases. In fact, our proof of Theorem 1 reveals that all the results remain unchanged, as long as the capacity of the downstream queue C_d is scaled for stability as well. For example, let $\Gamma = \{1, 2, \ldots, N\}$ and there is no interfering traffic, i.e., $R(s,t) = 0$. We scale the capacity of the downstream queue as $C_d := C_d(N) = \tilde{C}N$ where $\lambda < \tilde{C} < C$. Then, note that

$$\mathbb{P}\{Q_I^N(0) > 0\} \;\leq\; \mathbb{P}\{Q_{II}(0) > 0\} + \left|\mathbb{P}\{Q_I^N(0) > 0\} - \mathbb{P}\{Q_{II}(0) > 0\}\right|. \tag{53}$$

The first term of the RHS of (53) decrease to zero exponentially fast with rate $\tilde{I}(0) > 0$, and so does the second term, but with smaller rate $\tilde{J}^* < \tilde{I}(0)$. Thus, we see that $\mathbb{P}\{Q_I^N(0) > 0\}$ also decreases to zero exponentially fast with rate $\tilde{J}^*$, i.e.,

$$\limsup_{N\to\infty} \frac{1}{N} \log\, \mathbb{P}\left\{ \sup_{t\geq 0} \left[\sum_{i=1}^{N} D_i^N(-t, 0) - \tilde{C}Nt \right] > 0 \right\} \leq -\tilde{J}^*. \tag{54}$$

Note that the set of departure traffic $D_i^N(s,t)$, $i = 1, 2, \ldots, N$ now do not satisfy the *same* large deviations principle as the arrival traffic $A_i(s,t)$. This is in contrast to the results in [15], in which the author showed that the output traffic satisfies the same large deviations principle as the input traffic in the many-sources regime. The difference is as follows: In [15], 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 our case, we emphasize that the departure traffic $D_i^N(s,t)$, $i = 1, 2, \ldots, N$, are taken *as is*, and they are *not independent* for any fixed N due to the interaction among different flows in the upstream queue $q^N(t)$. Nevertheless, the *i.i.d.* property of the exogenous traffic at the upstream queue (with other required assumptions in the paper) enables us to write an exponentially decaying relation for the downstream queue as in (54), and thus making it decrease to zero almost surely.

From a practical perspective, some concerns may arise with regard to the assumption of many *i.i.d.* traffic flows. In fact, the traffic streams accessing the output ports of a router in a network will be typically composed of independent traffic streams arriving at each input port. While it could be that this number may not be very large (although the superposed sources within a stream may be quite large), the key result in this paper, in addition to the convergence itself, is that convergence happens *at least exponentially fast.*

Hence, for any node with non-negligible queue-lengh (or with non-vanishing overflow probability), we expect that such a decomposition result will not require a large number of independent multiplexed sources. This is also verified by numerical results reported in [7].

A possible extension of our results will be a multi-stage queueing network where all the queues in each stage are capable of serving many traffic flows, except the last-stage queue (with relatively small service capacity), which is of our interest. Our results imply that we can remove every queue with large service capacity and thus simplify the whole system into a single queue with an error decaying exponentially fast, provided that each traffic flow inside the queueing network satisfies the set of assumptions in the paper. However, it may not be easy to check whether these assumptions are satisfied inside the network. Hence, we are currently investigating a queueing network scenario, in which we do not require any assumption on the traffic flows inside the queueing network. As a special case, if each traffic arrival is regulated, i.e., there exists a function $A^*(t)$ such that $A_i(s, s+t) \leq A^*(t)$ for all s, we have found that the rate of convergence J^* becomes equal to $I(0)$, and that an immediate extension to the aforementioned multi-stage queueing network is possible without any assumption on the internal traffic flows.

References

[1] Beran, J., Sherman, R., Taqqu, M. S. and Willinger, W. (1995). Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions on Communications* **43,** 1566–1579.

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

[3] Choe, J. and Shroff, N. B. (2000). Use of Supremum Distribution of Gaussian Processes in Queueing Analysis with Long-Range Dependence and Self-Similarity. *Stochastic Models* **16,** 209–231.

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

[5] Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*. Springer-Verlag. 2nd Edition.

[6] Duffield, N. G. (1996). Economies of scale in queues with sources having power-law large deviation scaling. *Journal of Applied Probability* **33,** 840–857.

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

[8] JACKSON, J. R. (1957). Networks of Waiting Lines. *Operations Research* **5,** 518–521.

[9] KINGMAN, J. F. C. (1969). Markov Population Processes. *Journal of Applied Probability* **6,** 1–18.

[10] LELAND, W. E., TAQQU, M., WILLINGER, W. AND WILSON, D. V. (1994). On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Transactions on Networking* **2,** 1–15.

[11] LIKHANOV, N. AND MAZUMDAR, R. (1999). Cell loss asymptotics for buffers fed with a large number of independent stationary sources. *Journal of Applied Probability* **36,** 86–96.

[12] MANDJES, M. AND BORST, S. (2000). Overflow Behavior in queues with many long-tailed inputs. *Advances in Applied Probability* **32,** 1150–1167.

[13] MANDJES, M. AND KIM, J. H. (2001). Large deviations for small buffers: an insensitivity result. *Queueing Systems* **37,** 349–362.

[14] VECIANA, G. D., COURCOUBETIS, C. AND WALRAND, J. (1994). Decoupling bandwidths: A decomposition approach to resource management in networks. In *Proceedings of IEEE INFOCOM.* pp. 466–473.

[15] WISCHIK, D. (1999). The output of a switch, or, effective bandwidths for networks. *Queueing Systems* **32,** 383–396.

[16] WISCHIK, D. (2000). Sample path large deviations for queues with many inputs. *Annals of Applied Probability* **11,** 379–404.