Packet Reordering in Networks with Heavy-Tailed Delays

Marc Lelarge

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Packet Reordering in Networks with Heavy-Tailed Delays

Marc Lelarge

N° 5783

December 2005

THÈME 1

ISSN 0249-6399 ISRN INRIA/RR--5783--FR+ENG

INRIA
ROCQUENCOURT

Packet Reordering in Networks with Heavy-Tailed Delays

Marc Lelarge *

Thème 1 — Réseaux et systèmes
Projet TREC

Rapport de recherche n° 5783 — December 2005 — 31 pages

Abstract: An important characteristic of any TCP connection is the sequencing of packets within that connection. Out-of sequence packets indicate that the connection suffers from loss, duplication or reordering. It is thus of interest to study the magnitude of out-of sequence packets within Internet TCP connection and to identify their causes.

More generally, in many distributed applications (e.g., remote computations, database manipulations, or data transmission over a computer network), information integrity requires that data exchanges between different nodes of a system be performed in a specific order. However, due to random delays over different paths in a system, the packets or updates may arrive at the receiver in a different order than their chronological order. In such a case, a resequencing buffer at the receiver has to store disordered packets temporarily.

We analyze both the waiting time of a packet in the resequencing buffer and the size of this resequencing queue. We derive the exact asymptotics for the large deviation of these quantities under heavy-tailed assumptions. In contrast with results obtained for light-tailed distributions, we show that there exists several "typical paths" that lead to the large deviation. We derive explicitly these different "typical paths" and give heuristic rules for an optimal balancing.

Key-words: Asymptotics, parallel queues, resequencing, heavy tail

* INRIA-ENS, ENS, 45 rue d'Ulm, 75005 Paris, France, Marc.Lelarge@ens.fr

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 — Télécopie : +33 1 39 63 53 30

Reclassement de paquets dans des réseaux avec des délais à queue lourde

Résumé : Une caractéristique importante d'une connection TCP est l'ordre des paquets au sein de cette connection. Des paquets déclassés indiquent que la connection souffre de pertes, duplications ou reclassements. Il est donc intéressant d'étudier l'ampleur du nombre de paquets déclassés au sein d'une connection TCP et d'identifier ses causes.

Plus généralement, dans de nombreuses applications distribuées (manipulations de bases de données ou communication sur un réseau informatique), l'intégrité de l'information exige que l'échange de données entre les différents nœuds soit faite dans un ordre précis. Cependant à cause des délais aléatoires sur les différents chemins du système, les paquets ou mises à jour peuvent arriver au recepteur dans un ordre différent de leur ordre chrnonologique. Dans ce cas, un tampon de reclassement doit garder temporairement, au niveau du recepteur les paquets décalssés.

Nous analysons le temps d'attente dans le tampon de reclassement ainsi que la taille de la file dans ce tampon. Nous obtenons les asymptotiques exactes pour les grandes déviations de ces quantités sous des hypothèses de distributions à queue lourde. Nous montrons qu'il existe différents "chemin typiques" menant à ces grandes déviations et les décrivons de manière explicite. Enfin nous donnons des règles heuristiques pour un équilibrage optimal.

Mots-clés : Asymptotiques, files en parallèle, reséquencement, distributions à queue lourde

1 Motivation

Reordering the out-of-order arrival packets at the destination is a common phenomenon in the Internet [6], [25]. The major cause of reordering has been found to be the parallelism in Internet components (switches) and links [6]. Reordering greatly impacts the performance of applications in the Internet. In a TCP connection, the reordering of three or more packet positions within a flow may cause fast retransmission and fast recovery multiple times resulting in a reduced TCP window and consequently in a drop in link utilization and hence in less throughput for application [24]. For delay-based real-time service in UDP (such as VoIP or video conference), the ability to restore order at the destination will likely have finite limits. The deployment of a real-time service necessitates certain reordering constraints to be met. To verify whether these QoS requirements can be satisfied, knowledge about the reordering behavior in the Internet seems desirable. In this paper, we analyze the effect of heavy-tailed delays on the resequencing delay and on the number of out-of-order arrival packets.

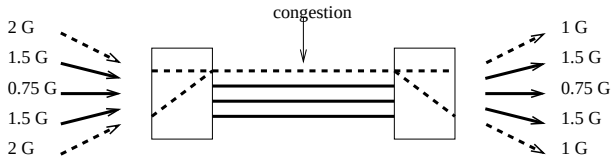

Figure 1: Without Trunking

Consider the following situation where there are four physical connections between the same pair of devices. Each link has a capacity of 2 Gbit/sec. Assume now that this system has to process the following traffic's profile (in Gbit/sec): 2, 1.5, 0.75, 1.5 and 2 from five different flows. If the system allocates to each physical connection a set of flows, it cannot proceed the whole traffic and at least one physical link is congested whereas others are underutilized. In the case of Figure 1, the effective throughput is 5.75 Gbits/sec, whereas the maximum throughput for the entire group of links is 8 Gbits/sec.

One way to solve this problem is to aggregate the four links into a single logical 8 Gbits/sec trunk group. This trunk group combines the bandwidth of all links to optimize data traffic load-sharing (see Figure 2).

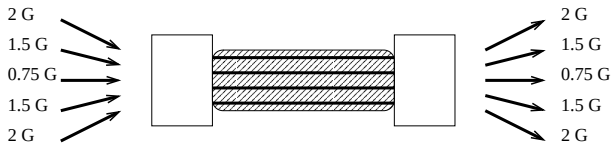

Figure 2: With Trunking

There exists several ways to load balance the traffics across parallel links and we will not discuss this issue here. We will only give some examples where the situation described above happens.

Storage area networks (SANs) provide the data communication infrastructure for storage systems consisting of switches, servers and storage systems. In this context, redundancy of physical connection is a protection against failure. A failure does not completely "break the pipe" but simply makes the pipe thinner. In the case of SANs, there exists software solutions in order to increase performance with Inter-Switch Link (ISL) Trunking see for example [17].

It allows up to four ISLs between two 2Gbit/sec switches to be logically combined in order to balance data traffic (as described in our example above). There are several advantages for this technology: it distributes heavy SAN frame traffic across all the ISLs in a trunk, it eliminates the need for data rerouting if one link of the trunk fails and it simplifies management by reducing the number of ISLs required.

In the context of communication networks, multiple links are quite frequently placed between a pair of devices in order to increase performance by "widening the pipe" (without going to a newer, more expensive technology). In this situation, establishing Point-to-Point Protocol (PPP) over each connection independently is far from an ideal solution, because it is then necessary to manually distribute the traffic over the two (or more) channels or links. A better solution would combine multiple links and use them as if they were one high-performance link. Some hardware devices actually allow this to be done at the hardware level itself; in Integrated Services Digital Network (ISDN) this technology is sometimes called bonding when done at layer one. For those hardware units that do not provide this capability, PPP makes it available in the form of the PPP Multilink Protocol (MP). This protocol was originally described in RFC 1717, and was updated in RFC 1990 [19]. MP allows PPP to bundle multiple physical links and use them like a single, high-capacity link. Once operational, it works by fragmenting whole PPP frames and sending the fragments over different physical links. For the transmission, MP first encapsulates the datagrams into a modified version of the regular PPP frame. It then takes that frame and decides how to transmit it over the multiple physical links. Typically, this is done by dividing the frame into fragments that are evenly spread out over the set of links. These are then encapsulated and sent over the physical links. At the reception, MP takes the fragments received from all physical links and reassembles them into the original PPP frame. That frame is then processed like any PPP frame, by looking at its Protocol field and passing it to the appropriate network layer protocol. The fragmenting of data in MP introduces a number of complexities that the protocol must handle. In particular, since fragments are being sent roughly concurrently, we need to identify them with a sequence number to allow reassembly.

Note that it is also possible to load balance the traffic at the layer 3 (Internet). For example it can be enable on Cisco Express Forwarding thanks to the command ip load-sharing (see [18]): "Per-packet load balancing allows the router to send data packets over successive equal-cost paths without regard to individual destination hosts or user sessions. Path utilization is good, but packets destined for a given destination host might take different paths and might arrive out of order."

Indeed, in all the examples cited above, due to random delays over different paths, the method of per-packet load balancing can mean that packets in a particular connection or flow arrive at their destination out of sequence. It can cause problems, especially for the streaming media, like video and audio. In the case of reliable protocols such as TCP, packets have to be delivered to the receiving application in the order they are transmitted at the sender. In order to deliver the arrived packet to the application in sequence, the receiver's transport layer needs to temporarily buffer out-of order packets and resequence them as more packets arrive.

In this paper, we analyze a model where mis-ordering is caused by multipath routing. Packets are generated according to a renewal process. Then they arrive at a disordering feed-forward network. A resequencing buffer follows the disordering network. We assume that the delays (which will be made precise in what follows) at the different stations of the disordering network are heavy-tailed. We will see that this stochastic assumption allows to model transfer of very large files or failure of devices or links (that are typically very long in

comparison to packet service times). In an Internet context, there are statistical evidence of the presence of heavy-tailed file size or transfer time (see [27], [10] and the references therein). We should stress that presence of heavy-tailed distribution is a strong motivation for load balancing. In particular, it allows in certain conditions to guarantee a minimal bandwidth to each class of flows [11], [9]. Moreover, the structural model proposed in [26] provides a direct link between the observed self-similarity characteristic of measured aggregate network traffic, and the strong empirical evidence in favor of heavy-tailed, infinite variance phenomena at the level of individual network connections. In our model, we are able to derive exact asymptotics of the packet resequencing delay and of the size of the resequencing queue. In view of the recent experiment results obtained in [30], our main results are the first networking model that could explain the power law observed for the reordered packet length and the long tailness of packet lag (see [29] for a simplified model). To the best of our knowledge, there are very few results on this problem under heavy-tailed assumptions. A similar framework has been studied by Jean-Marie and Gün [22], where the disordering network consists of K parallel $M/GI/1$ queues and the corresponding distribution of the resequencing delay is derived. A survey is given in [5]. More recently, Xia and Tse [28] consider a $2 - M/M/1$ queueing system and derive large deviation results for the resequencing queue size. The main results of our paper allow to get the exact asymptotics (which is much more precise than in the log scale as usually done in the large deviation literature) of the resequencing delay and of the resequencing queue size for this disordering network under heavy-tail assumptions.

In the next section, we describe our mathematical model and give a summary of our results. In Section 3 we give the exact asymptotics that we are able to derive for the resequencing delay and the resequencing queue size. We give some heuristic rules for optimal balancing. Concerning the resequencing delay, the results are derived from [12]. The cases that are not covered in this reference are proved in Section 8. For the resequencing queue size, new arguments have to be derived. They are given in Section 4 and then developed in Sections 5 and 6. Finally, concluding remarks are discussed in Section 7.

2 Mathematical Model

Consider the situation described in the introduction with two links and with cross traffic, illustrated in Figure 3. As an user of the network, we are interested by the delays of our own packets. In our model, the tagged flow will correspond to these specific packets. We assume

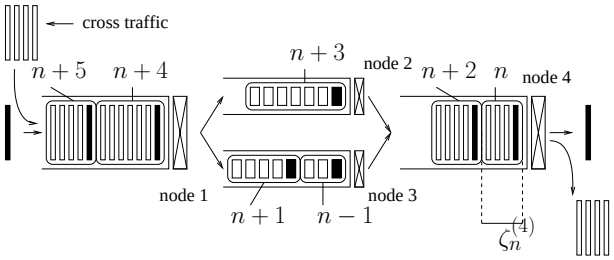

Figure 3: The case of two links.

that the service times of packets of the tagged flow are negligible compared to the queueing delays. We see that the time spent in a server is mainly due to the cross traffic. Thus in order to analyze the delay of the tagged packets, we define the virtual service times for each

tagged packet to be the amount of cross traffic arrived between two successive arrivals of the tagged customers. This (virtual) service time is denoted as $\zeta_n^{(i)}$ (for link i). These sequences will be assumed to be stochastically independent and to have heavy-tailed distributions. The resulting queueing system (with such virtual service times) is a single class FIFO queueing network.

From now on, packets will correspond to the virtual service times (i.e. there is only one packet for each tagged packet). We assume that they are ordered when they arrive in node 1. In our example, we want to model a situation where a packet that leaves the first node is randomly routed up (to node 2) or down (to node 3). In the case of Figure 3, packets with numbers n and $n+2$ are waiting in the resequencing buffer for the packet number $n-1$ (and then $n+1$). Note that in this case, packets n and $n+2$ have necessarily been routed to node 2 when they left node 1. Indeed the following remark will be of importance in what follows:

Remark 1 *Packets in the resequencing queue are always coming from the same node and waiting for a packet in the other node.*

When we will consider the resequencing queue size, we are counting the number of (virtual) packets. In order to go back to the real number of packets, we have to take into account a multiplicative factor that depends on the statistic of the cross traffic.

We now briefly give a flavor of our results in the case depicted on Figure 3:

Result 1 *Assume that the aggregate service time of node i is heavy-tailed:*

$$\mathbb{P}(\zeta_n^{(i)} > x) \sim \ell_i x^{-\alpha_i}, \text{ as } x \to \infty,$$

where $\alpha_i > 1$, $\ell_i \geq 0$. We assume that a fraction p of the packets are sent to node 2 and the remaining fraction $(1-p)$ is sent to node 3.

The asymptotics of the resequencing delay R (i.e. the time spent in node 4) is given by

$$\mathbb{P}(R > x) \sim C_R(p) x^{-\max_i(\alpha_i)-1}, \text{ as } x \to \infty,$$

where $C_R(p)$ is a function of the routing probability p that is given in an explicit form in what follows.

The asymptotics of the resequencing queue size Q_r (i.e. the number of packets in node 4) is given by

$$\mathbb{P}(Q_r > n) \sim C_Q(p) n^{-\max_i(\alpha_i)-1}, \text{ as } n \to \infty,$$

where the function $C_Q(p)$ is different from $C_R(p)$.

We will optimize delay and queue size asymptotics, by determining optimal values of p that minimize $C_R(p)$ or $C_Q(p)$ and we will see in details the implications due to the fact that $C_R(p) \neq C_Q(p)$. This optimization is important because it is well known that TCP performs poorly under significant packet reordering that is not necessarily caused by packet losses [8]. Our results can explain the power laws observed empirically by Zhou and Van Mieghem in [30] for the resequencing of UDP packets. We should stress that reordering can become a significant factor in the future Internet as result of increased parallelism. Our model is accurate for the study of other networks with radically different characteristics too. This is the case of wireless networks, in particular multi-hop mobile ad-hoc networks. In this case

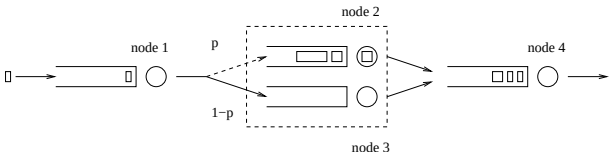

Figure 4: Resequencing problem.

routing protocols need to recompute routes often which may lead to packet reordering (due to the multiple paths as in our example).

In the sequel, we shall thus consider the model described in Figure 4. We assume that packets arrive in the first queue according to a renewal process $\{T_n\}$. We will model the routing at node 1 by a Bernoulli routing. Once packet k reaches the receiver, it leaves the system if all packets j with $j < k$ have already left the system. Otherwise it stays in the resequencing buffer, where it waits for the packets with number less than k.

One of our main tool in our analysis will be the work of Baccelli and Foss in [3] and the subsequent works done in [4] and [12]. We briefly describe how we can model the system described above thanks to the (max,plus) algebra (or Petri net). This formalism is not essential in our paper and we refer to [2] for more references. However, we stress that (max,plus) formalism is crucial for the proof of Theorem 1 below (that can be found in [12]). In view of the key role of this theorem for our analysis, we chose to show the connection between our model and (max,plus) algebra. Moreover, it allows for a rigorous presentation of our model.

Consider the standard fork and join system as depicted (with Petri net formalism) in Figure 5. In this auxiliarly model, a different kind of routing is in action. Each time a packet (say k) finishes its service $\zeta_k^{(1)}$ in node 1, there is one packet sent up and one packet sent down simultaneously. The 'up'-packet ('down'-packet) is then also the k-th packet for node 2 (for node 3 respectively). The k-th packet joins the queue of node 4 once both packets have left node 2 and 3 respectively. Each node is a standard $\cdot/G/1/\infty$ queue.

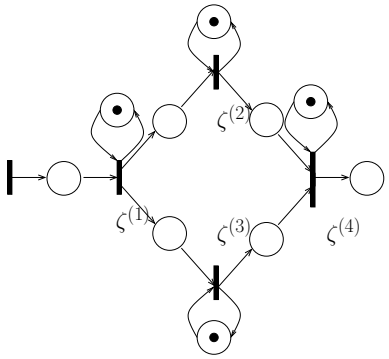

Figure 5: Fork and join model.

The (max, plus) semi-ring $\mathbb{R}_{\max}$ is the set $\mathbb{R} \cup \{-\infty\}$, equipped with max, written additively (i.e., $a \oplus b = \max(a, b)$) and the usual sum, written multiplicatively (i.e., $a \otimes b = a + b$). Let $X_n^{(i)}$

denotes the departure time of the n-th packet from node i. We have the following equations:

$$\begin{aligned} X_{n+1}^{(1)} &= (T_{n+1} \oplus X_n^{(1)}) \otimes \zeta_{n+1}^{(1)}, \\ X_{n+1}^{(2)} &= (X_{n+1}^{(1)} \oplus X_n^{(2)}) \otimes \zeta_{n+1}^{(2)}, \\ X_{n+1}^{(3)} &= (X_{n+1}^{(1)} \oplus X_n^{(3)}) \otimes \zeta_{n+1}^{(3)}, \\ X_{n+1}^{(4)} &= (X_{n+1}^{(4)} \oplus X_{n+1}^{(3)} \oplus X_n^{(4)}) \otimes \zeta_{n+1}^{(4)}. \end{aligned}$$

Note that our system is linear in the (max, plus) semi-ring $\mathbb{R}_{\max}$, which allows us to use results of [12].

In order to model the desired routing mechanism we will use the idea of clones, i.e., packets that behave like real packets except that they never require any service time: their service time is null. Suppose that the real route of packet k is up. Then at the end of its service in the first node, a clone is sent to node 3. Since $\zeta_k^{(3)} = 0$, the departure time of packet k from node 3 is $X_k^{(3)} = \max(X_k^{(1)}, X_{k-1}^{(3)})$. Similarly, if the real route of packet k is down, then a clone is sent up. In both cases the "real" packet k joins the queue of node 4 once "real" packet $k-1$ has joined it (and not before). In particular packets are ordered when they leave node 4. This shows that the resequencing delay of packet n is given by

$$R_n = \max\left\{X_n^{(4)} - X_n^{(2)}, X_n^{(4)} - X_n^{(3)}\right\}. \tag{1}$$

It corresponds to the time spent by "real" packet in the resequencing buffer. In particular, if we take $\zeta_k^{(4)} = 0$ for all k, this delay is purely due to multi-path routing.

We now give the stochastic assumptions made on the service times distribution. Here and later in the paper, for positive functions f and g, the equivalence $f(x) \sim dg(x)$ with $d > 0$ means $f(x)/g(x) \to d$ as $x \to \infty$. By convention, the equivalence $f(x) \sim dg(x)$ with $d = 0$ means $f(x)/g(x) \to 0$ as $x \to \infty$; this is written as $f(x) = o(g(x))$. Recall that a distribution function G on $\mathbb{R}_+$ is called subexponential if $\overline{G^{*2}}(x) \sim 2\overline{G}(x)$, where $\overline{G^{*2}}$ is the tail of the twofold convolution of G. Note that a subexponential distribution G is long-tailed: for any $y > 0$, we have

$$\overline{G}(x+y) \sim \overline{G}(x). \tag{2}$$

A tail distribution $\overline{G}$ is called regularly varying with index $-\alpha \leq 0$, $\overline{G} \in \mathcal{R}(-\alpha)$ if

$$\lim_{x\to\infty} \frac{\overline{G}(tx)}{\overline{G}(x)} = t^{-\alpha},\ \forall t > 0.$$

Such distributions with $\alpha > 0$ are automatically subexponential. Weibull or lognormal distributions are other examples of subexponential distributions (that are not regularly varying), see [13] for more details on subexponential distributions.

Throughout, we let F be a distribution function on $\mathbb{R}_+$ such that:

- F is subexponential, with finite first moment;
- The integrated distribution F^s of F with the tail

$$\overline{F}^s(x) := 1 - F^s(x) := \min\left\{1, \int_x^\infty \overline{F}(u)du\right\}$$

is subexponential.

For example if $\overline{F} \in \mathcal{R}(-\alpha)$ with $\alpha > 1$, then the distribution function F satisfies previous assumptions. Weibull and lognormal distributions are other examples that satisfy these assumptions.

Let $\{\sigma_n = (\sigma_n^{(1)}, \ldots \sigma_n^{(4)})\}_n$ be an i.i.d. sequence of mutually independent random variables with finite mean and such that the following equivalence holds when x tends to infinity (with $d^{(j)} \geq 0$):

$$\mathbb{P}(\sigma_1^{(j)} > x) \sim d^{(j)}\overline{F}(x),$$

for all $j = 1, \ldots, 4$ with $\sum_{j=1}^4 d^{(j)} > 0$. In particular at least one component of σ_n is heavy-tailed. Note that some components are allowed to be light-tailed in which case we take the corresponding $d^{(j)} = 0$. Moreover if two components of σ_n are heavy-tailed then, we take for $\overline{F}$ the "dominating one". For example if $\mathbb{P}(\sigma_0^{(1)} > x) \in \mathcal{R}(-\alpha_1)$ and $\mathbb{P}(\sigma_0^{(2)} > x) \in \mathcal{R}(-\alpha_2)$ with $1 < \alpha_1 < \alpha_2$ and the other components are light-tailed. Then we can take $\overline{F}(x) = \mathbb{P}(\sigma_0^{(1)} > x)$, $d^{(1)} = 1$ and $d^{(i)} = 0$ for $i = 2, 3, 4$.

Let $\{r_n\}_{n\in\mathbb{Z}}$ be a sequence of i.i.d. random variables, independent of everything else, with values in $\{2, 3\}$. We write $\mathbb{P}(r_n = 2) = 1 - \mathbb{P}(r_n = 3) =: p$, and assume that $0 < p < 1$. In order to apply our idea of clones, we define

$$\begin{aligned}
\zeta_n^{(1)} &:= \sigma_n^{(1)}, \\
\zeta_n^{(2)} &:= \sigma_n^{(2)} \mathbb{1}_{\{r_n=2\}}, \\
\zeta_n^{(3)} &:= \sigma_n^{(3)} \mathbb{1}_{\{r_n=3\}}, \\
\zeta_n^{(4)} &:= \sigma_n^{(4)}.
\end{aligned}$$

We denote $\gamma_i := \mathbb{E}[\zeta_0^{(i)}]$ and $\gamma := \max_{i=1,\ldots,3} \gamma_i$. In our context, we will always assume that the service time at the resequencing queue is null, $\sigma_n^{(4)} = 0$ for all n. But our results are still valid if it is such that $d^{(4)} = 0$ and that $\gamma_4 < \min_{i=1,\ldots,3} \gamma_i$.

We will always assume that

$$\gamma < a := \mathbb{E}[T_1 - T_0]. \tag{3}$$

We know that under this condition, the system is stable, see [2]. In particular, we can define the stationary end-to-end delay (or sojourn time in the whole network) Z and the stationary resequencing delay R (which is the stationary version of (1)).

3 Main Results

The following proposition follows form the analysis made in [12] (the first part of the theorem given by equation (4) is a slight extension of [4]),

Theorem 1 *We have as $x \to \infty$,*

$$\mathbb{P}(Z > x) \sim \left(\frac{d^{(1)}}{a - \gamma} + \frac{pd^{(2)}}{a - \gamma_2} + \frac{(1-p)d^{(3)}}{a - \gamma_3} \right) \overline{F}^s(x). \tag{4}$$

We denote

$$\overline{G}(x) = \left(\frac{pd^{(2)}}{a - \gamma_2} + \frac{(1-p)d^{(3)}}{a - \gamma_3} \right) \overline{F}^s(x). \tag{5}$$

We have

1. *if* $\gamma_1 > \max(\gamma_2, \gamma_3)$, *then*

$$\mathbb{P}(R > x) = \overline{G}(x) + o(\overline{F}^s(x)).$$

2. *if* $\gamma_3 > \max(\gamma_1, \gamma_2) := \gamma_{1\vee 2}$, *then*

$$\mathbb{P}(R > x) = \frac{d^{(1)}}{a - \gamma_3} \overline{F}^s \left(\frac{a - \gamma_{1\vee 2}}{\gamma_3 - \gamma_{1\vee 2}} x \right) + \overline{G}(x) + o(\overline{F}^s(x)).$$

3. *if* $\gamma_3 = \gamma_2 \geq \gamma_1$ *and* $\overline{F}^s \in \mathcal{R}(-\alpha)$ *with* $\alpha > 0$, *then*

$$\mathbb{P}(R > x) = \overline{G}(x) + o(\overline{F}^s(x)).$$

Remark 2 *Note that results of [4] and [12] deal with networks with a general topology that belongs to the class of stochastic event graphs. This theorem follows directly from these results except case 3 when* $\gamma_3 = \gamma_2 \geq \gamma_1$ *(see Section 8 for a proof). We can interchange indices 2 and 3 in the case 2.*

An important consequence of our result is that the asymptotics of the resequencing delay (or the end-to-end delay) does NOT change with the link speed, assuming the traffic characteristics are not altered by the technology change. The reason for this is that if c is the speed factor gained thanks to the technology, then we have to replace each $a = \mathbb{E}[T_1 - T_0]$ by ca in the formulas above. But if we want to keep the same load for the system, we will put more traffic and hence adapt the intensity of the arrival process $\{T'_n\}$ such that $\mathbb{E}[T'_1 - T'_0] = a' = a/c$ and we see that for a fixed load the asymptotics are preserved.

We give now some practical applications of ou results:

- Consider the case depicted on Figure 3, where a link of 4 Gbit/sec feeds two links of 2 Gbit/sec. In order to take into account the factor 2 between these throughputs, we assume that

$$\sigma_n^{(2)} \stackrel{d}{=} \sigma_n^{(3)} \stackrel{d}{=} 2\sigma_n^{(1)}.$$

  Hence in this case, we have $\gamma_2 = 2p\gamma_1$ and $\gamma_3 = 2(1-p)\gamma_1$. Moreover if we denote $\overline{F}(x) = \mathbb{P}(\sigma_0^{(1)} > x)$, then we have $\mathbb{P}(\sigma_0^{(2)} > x) = \mathbb{P}(\sigma_0^{(3)} > x) = \overline{F}(x/2)$. In particular if $\overline{F} \in \mathcal{R}(-\alpha)$, we have $d^{(1)} = 1$ and $d^{(2)} = d^{(3)} = 2^\alpha$. In the case $p = 1/2$, we are in case 3 and we have

$$\mathbb{P}(R > x) \sim \overline{G}(x) = \frac{2^\alpha}{a - \gamma_1} \overline{F}^s(x).$$

  In the case $p < 1/2$, we have $\gamma_2 < \gamma_1 < \gamma_3$, and we are in case 2. The resequencing delay will have a tail asymptotics of the same order but with a constant $C(p)$ that is larger than in the case $p = 1/2$. In particular $C(p) > C(1/2) = \frac{2^\alpha}{a - \gamma_1}$ for $p \neq 1/2$.

- Since node 1 represents the access to the links represented by node 2 and node 3, it may be more realistic to make different assumptions on the tail distribution of the service times in node 1. One can assume that this access is not bottleneck in which case, we can consider $d^{(1)} = 0$. Consider now a link of capacity C that feeds two links of respective capacities uC (represented by node 2) and $(1-u)C$ (represented by node 3) with $0 < u \leq 1/2$. Hence, we assume that

$$u\sigma_n^{(2)} \stackrel{d}{=} (1-u)\sigma_n^{(3)}. \tag{6}$$

Taking the expectation, we get $u\mathbb{E}[\sigma_n^{(2)}] = (1-u)\mathbb{E}[\sigma_n^{(3)}] := c$. Hence, we have

$$\gamma_2 = \frac{pc}{u},\ \gamma_3 = \frac{(1-p)c}{1-u}.$$

Note that if $\min\{c/u, c/(1-u)\} = c/(1-u) < a$, i.e. if $u < 1 - c/a$, then the system with $p = 0$ is still stable and there is no need for resequencing: we can put all the traffic on one link. We assume in particular that $1 - c/a < 1/2$, i.e. $\beta := a/c < 2$.

For the tail distributions, we assume that they are regularly varying and from (6), we have:

$$\mathbb{P}(\sigma_n^{(2)} > x) \sim u^{-\alpha}\overline{F}(x),\ \mathbb{P}(\sigma_n^{(3)} > x) \sim (1-u)^{-\alpha}\overline{F}(x).$$

For this model, the asymptotics of the resequencing delay is given by $\overline{G}(x)$ in (5). We denote by $p(u)$ the value of the routing parameter p that gives the minimal asymptotics, i.e. that minimizes the coefficient in (5). We have (with $\beta = a/c < 2$)

$$p(u) = \frac{u^{\alpha/2}\beta + (1-u)^{\alpha/2}\left(\frac{1}{1-u} - \beta\right)}{u^{\alpha/2-1} + (1-u)^{\alpha/2-1}},\ u > 1 - \frac{1}{\beta}. \tag{7}$$

Figure 6 gives the curves of $p(u)/u = \gamma_2(u)/c$ for different values of $\alpha \in \{1, \ldots, 2.5\}$.

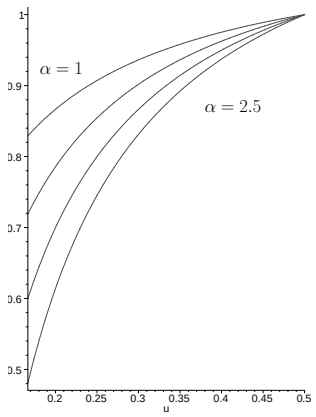

Figure 6: Optimal load in the link with small capacity for different values of α: $u \mapsto \gamma_2(u) = p(u)c/u$ represents the amount of traffic that should be sent in link with capacity u in order to optimize the resequencing delay.

We see that we have $\gamma_2(u) < c < \gamma_3(u)$ for $u < 1/2$. Hence we have to put less traffic on the link with minimal capacity and a little more on the link with maximal capacity.

The reason why this asymmetric choice is optimal is the following: in any case, as we will see the cause for a large delay is that one service time in the network is very big whereas all the others are close to their means. But if this big service time occurs in the link with small capacity, it takes a long time for this link to process it. Hence we see that we have to compare the relative efficiency of both links for processing big services in order to get the optimal solution. We define for $1-\beta(1-u) \leq p \leq \beta u$,

$$\begin{aligned} C_2(p,u) &:= \frac{p}{u^\alpha(a-\gamma_2(p,u))}, \\ C_3(p,u) &:= \frac{1-p}{(1-u)^\alpha(a-\gamma_3(p,u))}, \end{aligned}$$

Looking back to the coefficient in Equation (5), it is equal in our model to,

$$C(p,u) := C_2(p,u) + C_3(p,u) \approx \max\left(C_2(p,u), C_3(p,u)\right),$$

where the approximation is good when $p/u \to \beta$ (i.e. $\gamma_2(p,u)) \to a$) or $(1-p)/(1-u) \to \beta$ (i.e. $\gamma_3(p,u)) \to a$). The approximation above is due to the fact that for fixed u, the function $p \mapsto C_2(p,u)$ is non-decreasing whereas the function $p \mapsto C_3(p,u)$ is non-increasing. It has the following interpretation: when $p \approx 1-\beta(1-u)$ then the large delay is mainly due to node 3 (since the effect of a big service time in node 3 dominates $C(p,u)$). We put too much traffic in this link. The symmetric case is $p \approx \beta u$, which corresponds to an overloading of link 2.

The quantities $C_2(p,u)$ and $C_3(p,u)$ are measuring the "cost" of each node for a given (p,u). In view of the monotonicity of these functions and of the fact that the total cost is approximately the maximum of the costs, we see that the optimal parameter p^* is defined by the following equation $C_2(p^*,u) = C_3(p^*,u)$. It turns out that this rule gives a very good approximation. Indeed it can be more simplified while remaining accurate.

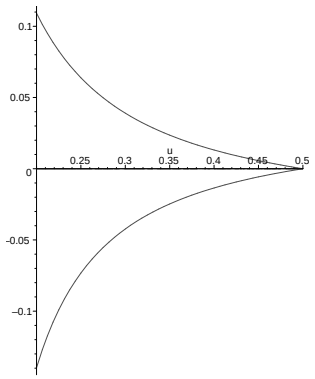

Figure 7: Heuristic rule for $1.5 < \alpha < 2.5$, the error made on the routing parameter p (in the optimization of the resequencing delay) by using the approximative rule lies between the two curves.

Define:

$$\tilde{C}_2(p,u) := \frac{1/u}{1-\rho_2(p,u)},\ \tilde{C}_3(p,u) := \frac{1/(1-u)}{1-\rho_3(p,u)}, \tag{8}$$

with $\rho_i(p,u) = \gamma_i(p,u)/a$. The solution $\tilde{p}^*$ of the equation $\tilde{C}_2(p,u) = \tilde{C}_3(p,u)$ is clearly a good approximation of p^*. Figure 7 shows the curve $(p(u) - \tilde{p}^*(u))/p(u)$ for different values of α. The error in the choice of p is less than 10% for $1.5 < \alpha < 2.5$ and decreases when $u \to 1/2$ when we use the approximative rule compare to the optimal $p(u)$ defined in (7).

The main interest of the rule defined by (8) is that we can apply it when only the average load of each link is known. Moreover this rule is quite natural. Assume that we are only given the average load of each link as a function of p (this can be done empirically). With so little information, we can only use well-known formulas for the $M/M/1$ queue and estimate that the mean sojourn time in each node is approximately given by $T_i := \mathbb{E}[\sigma^{(i)}]/(1-\rho_i)$. In particular, in order to minimize the resequencing delay, we should choose p such that $T_2 = T_3$. In view of (6), this equation is exactly the one given by the rule defined by (8)!

- Consider now the case $d^{(1)} > 0$. Note that we have to take into account a new effect:

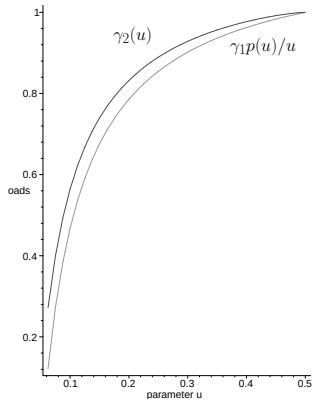

Figure 8: Optimal loads for two different models: the curve $\gamma_1 p(u)/u$ gives the optimal load of link with capacity u if node 1 (representing I/O delays) is light-tailed and the curve $\gamma_2(u)$ corresponds to the case where the node 1 is heavy-tailed.

when $\gamma_2 \neq \gamma_3$ there is an additional term that corresponds to the case where the big service time occurs in node 1. Then, there is a resequencing delay that is due to the mismatch between the throughputs of nodes 2 and 3. This effect will affect the optimal choice of p. In a model, where

$$\sigma_n^{(1)} \stackrel{d}{=} u\sigma_n^{(2)} \stackrel{d}{=} (1-u)\sigma_n^{(3)},$$

we have $\mathbb{E}[\sigma_n^{(1)}] = u\mathbb{E}[\sigma_n^{(2)}] = (1-u)\mathbb{E}[\sigma_n^{(3)}] := \gamma_1$. Figure 8 gives the curves (for $u \leq 0.5$) corresponding to

- (upper one) $\gamma_2(u)$ the optimal load of node 2 in this model, computed thanks to case 2 of Theorem 1;
- (lower one) $\gamma_1 p(u)/u$ the optimal load of node 2 in previous model.

The other parameters are fixed as follows: $\alpha = 3/2, a = 1.2, \gamma_1 = 1$. The fact that $\gamma_2(u) \geq \gamma_1 p(u)/u$ shows the impact of a big service time in node 1.

- We can consider other scenarios. For example, we may not control the traffic in the links in parallel that our own traffic is using. In the case of Figure 9, we see that service times at node 1 represent the I/O delays and are completely independent of the service times at node 2 and 3. In particular, depending on what we want to model, results will be

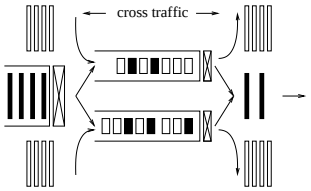

Figure 9: User's perspective

quite different. For example, if the I/O is bottleneck, then we may assume that $d^{(1)} = 1$ whereas $d^{(2)} = d^{(3)} = 0$. In this case, we see that it is quite important that $\gamma_2 = \gamma_3$, i.e. that the load is equally distributed among the different links.

Let Q_r be the resequencing queue size just after the arrival of packet number 0 to the resequencing queue when the system is in the stationary regime.

In the next sections, we will show the following theorem:

Theorem 2 *Let Q_r be the stationary size of the resequencing buffer. We denote*

$$\overline{H}(n) = \left(\frac{pd^{(2)}}{a-\gamma_2} \overline{F}^s \left(\frac{an}{1-p} \right) + \frac{(1-p)d^{(3)}}{a-\gamma_3} \overline{F}^s \left(\frac{an}{p} \right) \right).$$

Assume that $\overline{F}^s \in \mathcal{R}(-\alpha)$ with $\alpha > 0$, then we have as $n \to \infty$,

1. *if $\gamma_1 > \max(\gamma_2, \gamma_3)$ or if $\gamma_3 = \gamma_2$, then*

$$\mathbb{P}(Q_r > n) = \overline{H}(n) + o(\overline{F}^s(n)). \tag{9}$$

2. *if $\gamma_3 > \max(\gamma_1, \gamma_2) := \gamma_{1\vee 2}$, then*

$$\mathbb{P}(Q_r > n) = \frac{d^{(1)}}{a-\gamma_3} \overline{F}^s \left(\frac{a-\gamma_{1\vee 2}}{\gamma_3 - \gamma_{1\vee 2}} \frac{\gamma_3 n}{p} \right) + \overline{H}(n) + o(\overline{F}^s(n)). \tag{10}$$

Comparing both theorems shows that results are quite similar. In particular, the same remarks as the ones done after Theorem 1 can be done for the resequencing delay. In the heavy-tailed case, the "typical paths" that lead to a large resequencing delay or to a large resequencing queue size are the same. We will describe these paths more precisely in the next section and in the proof of the theorem. However, we see on Figure 10 that considering the resequencing delay or the resequencing queue size is slightly different. The upper curve represents the optimal load for the resequencing delay and the lower one for the resequencing queue size (for $\alpha = 1.5$). We see that we have to put even less traffic on the link with small capacity if we want to optimize the resequencing queue length. Considering the case when $p = p(u)$, we know that a big service time in node 2 or in node 3 will have the same impact

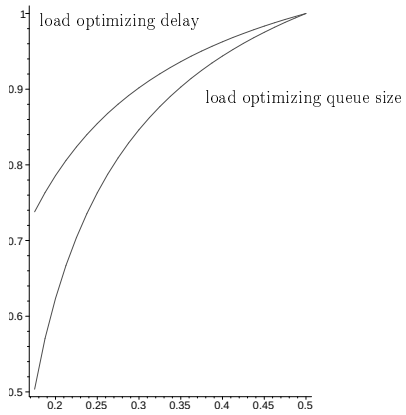

Figure 10: Comparison of the optimal loads for the resequencing delay and the resequencing queue size.

on the resequencing delay. But given a resequencing delay and the fact that $p < 1/2$, it seems clear that if the resequencing delay is due to a big service time in node 2, the resequencing queue size will be bigger than if the delay is due to a big service time in node 3. Lowering p will have two contradictory effects: it will lower the probability of a large delay due to node 2 and it will make the resequencing queue size even bigger if the delay is due to node 2. Figure 10 tells us that the first effect dominates the second one and we have to lower p.

Note that Theorem 2 can be seen as an extension of Theorem 2 in [12]. In particular, we consider here a much more complex function of the network: compare the representation of Q_r given in Section 6.1 with the function Φ introduced in Section 3 of [12]. We give here original arguments and a self-contained proof of Theorem 2 that could be adapted to prove Theorem 1 too.

4 Idea of the proof

We present now how we adapt the ideas of Baccelli and Foss [3] to our framework. We will give a generalization of the so called "single big event theorem", well known for isolated queues, to the resequencing buffer. In the $GI/GI/1$ queue, this theorem states that in the case of subexponential service times, large workloads occur on a typical event where a single large service time has taken place in a distant past, and all other service time are close to their mean. Similarly, in our resequencing problem with subexponential service times, large resequencing delays or queue sizes occur when a single large service time has taken place in one of the nodes 1, 2 or 3, and all other service times are close to their mean.

The first step consists in finding an upper bound for Q_r for which the asymptotic is easy to derive (and satisfy the "single-big-event" theorem). Thanks to Remark 1, we note that there are only two possibilities for Q_r to be positive (we take $r_0 = 2$, the case $r_0 = 3$ is symmetric):

1. when packet 0 arrives in the resequencing buffer, there are only packets coming from the (same) node 2 and waiting for packets with $r_i = 3$ and $i < 0$.

2. when packet 0 arrives in the resequencing buffer, there are only packets coming from the (other) node 3 waiting for a packet with $r_i = 2$ and $i > 0$ (with our definition of Q_r, packets that were waiting for the number 0 are not counted in Q_r).

In the first case, the total number of packets in the network at time T_0 is clearly an upper bound for Q_r. To make this definition precise, we consider the network in its stationary regime and define Q_t as the minimal integer k such that packet $-k$ has left node 4 at the arrival time of packet 0, T_0. Note that thanks to the FIFO assumption, Q_t corresponds exactly to the total number of packets in the network of Figure 4 (or in the fork-and-join model if we do not count the clones at node 2 and 3). In case 1, we have $Q_r \leq Q_t$ since any packet that is in the resequencing queue when packet number 0 joins this queue was already in the network at time T_0.

In case 2, the situation is different since all the packets present in the resequencing queue arrived after packet 0 in the network. In this case, we have clearly $Q_r \leq \sum_{i=1}^{\infty} \mathbb{1}_{\{T_i \leq Z\}} := N(Z)$, where Z is the time spent in the whole network by packet 0 and N is the counting process associated with the arrival times $\{T_n\}$.

Hence we have

$$Q_r \leq \max(Q_t, N(Z)) := U.$$

We are able to derive the tail asymptotics of U as follows. First note that we have

$$\mathbb{P}(U > x) \leq \mathbb{P}(Q_t > x) + \mathbb{P}(N(Z) > x).$$

Thanks to the distributional Little's law [16], we have $Q_t \stackrel{d}{=} N(Z)$ in distribution. Hence we have

$$\mathbb{P}(Q_t > x) \sim \mathbb{P}(N(Z) > x).$$

We are able to derive the tail asymptotics of this latter quantity thanks to Theorem 1 and to the results of [1] or [14]:

Proposition 1 *Assume that $\overline{F}^s \in \mathcal{R}(-\alpha)$ with $\alpha \in (0, \infty)$. Then we have as $n \to \infty$,*

$$\mathbb{P}(Q_t > n) \sim \mathbb{P}(N(Z) > n) \sim \mathbb{P}(Z > na),$$

where $a = \mathbb{E}[T_1 - T_0]$.

Proof. We want to apply Theorem 3.1 of [14]. Note that by definition of regular variation [7] we have

$$\begin{aligned} g(x) &= -\log \overline{F}^s(x) \\ &= \alpha \log(x) - \log L(x), \end{aligned}$$

where $L \in \mathcal{R}(0)$. Hence, for any function $d(x)$ such that $d(x) \to \infty$ as $x \to \infty$, we have,

$$g\left(x + \frac{x}{d(x)}\right) = g(x) + o(1), \quad \text{as } x \to \infty.$$

Thanks to Theorem 3.1 of [14], we have

$$\mathbb{P}(Q_t > n) = \mathbb{P}(N(Z) > n) \sim \mathbb{P}(Z > na),$$

because $\mathbb{P}(Z > x)$ is α insensitive by Theorem 1. □

This proposition allows us to define a typical event in the same spirit as in [3]. This event $T_x = A_x \cup B_x$ is such that:

1. $\mathbb{P}(A_x) \sim \mathbb{P}(Q_t > x)$ and $\mathbb{P}(B_x) \sim \mathbb{P}(N(Z) > x)$;
2. $\mathbb{P}(Q_t > x, A_x^c) = o(\mathbb{P}(Q_t > x))$ and $\mathbb{P}(N(Z) > x, B_x^c) = o(\mathbb{P}(N(Z) > x))$.

Hence the event T_x "describes" the way the rare events $\{Q_t > x\}$ and $\{N(Z) > x\}$ occur. It will be made clear in the next section. Once we defined this event, we have

$$\begin{aligned}\mathbb{P}(Q_r > x) &= \mathbb{P}(Q_r > x, T_x) + \mathbb{P}(Q_r > x, T_x^c) \\ &\leq \mathbb{P}(Q_r > x, T_x) \\ &\quad +\mathbb{P}(\max(Q_t, N(Z)) > x, T_x^c) \\ &\leq \mathbb{P}(Q_r > x, A_x) \\ &\quad +\mathbb{P}(Q_t > x, A_x^c) + \mathbb{P}(N(Z) > x, B_x^c).\end{aligned}$$

But the term $\mathbb{P}(Q_t > x, A_x^c) + \mathbb{P}(N(Z) > x, B_x^c) = o(\overline{F}^s(x))$ thanks to Proposition 1 and Theorem 1. Hence we have

$$\mathbb{P}(Q_r > x) = \mathbb{P}(Q_r > x, T_x) + o(\overline{F}^s(x)).$$

The last part of the proof (Section 6) is the computation of the quantity $\mathbb{P}(Q_r > x, A_x)$.

5 Typical Event for Q_t and $N(Z)$

In this section we first derive the typical event A_x for the total queue length Q_t.

We first consider a $GI/GI/1/\infty$ queue with mean inter-arrival times $a = \mathbb{E}[\tau_n]$ and mean service times $b = \mathbb{E}[\sigma_n]$, where $a > b$. Denote by F the distribution function of σ and assume that $\overline{F}^s$ is regularly varying. We define

$$S_n^\tau = \sum_1^n \tau_{-i},\ S_n^\sigma = \sum_1^n \sigma_{-i},\ S_n = S_n^\sigma - S_n^\tau.$$

For any sequence ϵ_n, we define the following sequence of events, $n \geq x$,

$$\begin{aligned}K_{n,x} &= \left\{\left|\frac{S_k^\tau}{k} - a\right| \leq \epsilon_k,\ \forall x \leq k \leq n\right\} \\ &\bigcap \left\{\left|\frac{S_k^\sigma}{k} - b\right| \leq \epsilon_{k+1},\ \forall x \leq k < n\right\}.\end{aligned}$$

Due to the stong law of large numbers, there exists a non-increasing sequence ϵ_n such that $\epsilon_n \to 0$ and $n\epsilon_n \to \infty$ as $n \to \infty$ and

$$\sup_{n \geq x} \mathbb{P}(K_{n,x}) \to 1 \quad \text{as } x \to \infty. \tag{11}$$

Lemma 1 *Let Q be the stationary queue length in the FIFO $GI/GI/1/\infty$ queue. Let $K_{n,x}$ be defined as above and $\eta_n = 3\epsilon_n$, let*

$$A_{n,y} = K_{n,y} \cap \{\sigma_{-n} > y + n(a - b + \eta_n)\},\ A_x = \bigcup_{n \geq x} A_{n,xb}.$$

Then, as $x \to \infty$,

$$\mathbb{P}(Q > x) \sim \mathbb{P}(Q > x, A_x) \sim \mathbb{P}(A_x) \sim \sum_{n \geq x} \mathbb{P}(A_{n,xb}). \tag{12}$$

Proof. Simple calculations using the fact that $\overline{F}^s$ is regularly varying show that, as $x \to \infty$,

$$\begin{aligned}\sum_{n\geq x} \mathbb{P}(A_{n,xb}) &= \sum_{n\geq x} \mathbb{P}(K_{n,xb})\mathbb{P}(\sigma_{-n} > xb + n(a-b+\eta_n)) \\ &\sim \sum_{n\geq 0} \overline{F}(xb + (n+x)(a-b+\eta_n)) \\ &\sim \sum_{n\geq 0} \overline{F}(xa + n(a-b)) \sim \frac{1}{a-b}\overline{F}^s(ax),\end{aligned}$$

where we used the independence between $K_{n,x}$ and σ_{-n} for the first line and (11) to get the second line.

Note that Proposition 1 is also valid for a $GI/GI/1$ queue (see [14]), hence we have $\mathbb{P}(Q > x) \sim \frac{1}{a-b}\overline{F}^s(ax)$.

Thus, if we show that the sequences $\{K_{n,y}\}$ and $\{\eta_n\}$ are such that, for all sufficiently large x, (a) the events $A_{n,xb}$ are disjoint for all $n \geq x$, (b) $A_{n,xb} \subset \{Q > x\}$ for all $n \geq x$, then

$$\mathbb{P}(Q > x) \geq \mathbb{P}(Q > x,\ A_x) = \mathbb{P}(A_x) \sim \frac{1}{a-b}\overline{F}^s(ax).$$

Hence in this case we have the equivalence (12). We now show that the sequences $\{K_{n,x}\}$ and $\{\eta_n\}$ satisfy (a) and (b).

We denote for $\ell \geq 0$, $S^\sigma_{[-n,-n+\ell]} = \sum_{i=-n}^{-n+\ell} \sigma_i$ and $Q_{[-n,0]}$ the size of the queue at time T_0+ in the system fed by packets $-n, -n+1, \ldots, 0$ only (or in other words, the system fed with the arrival process $T_{-n}, T_{-n+1}, \ldots, T_0$). Note that $S^\sigma_{[-n,-n+\ell]} \leq S^\tau_n$ implies that $Q_{[-n,0]} \leq n - \ell$ and that $S^\sigma_{[-n,-n+\ell]} \geq S^\tau_n$ implies that $Q_{[-n,0]} \geq n - \ell$. Hence we have $Q_{[-n,0]} = n - \sup\{\ell, S^\sigma_{[-n,-n+\ell]} \leq S^\tau_n\}$. For $n \geq x$, on the event $A_{n,xb}$, we have $\sigma_{-n} + (b - \epsilon_n)Q_{[-n,0]} \leq n(a+\epsilon_n)$, which implies

$$\begin{aligned}Q &\geq Q_{[-n,0]} \geq \frac{xb}{b-\epsilon_n} + \frac{n(\eta_n - 2\epsilon_n)}{b-\epsilon_n} \\ &\geq x.\end{aligned}$$

Hence we showed that (b) is satisfied. For the part (a), if $\epsilon_x \leq (a-b)/2$ then the events $A_{n,xb}$ are disjoint for $n \geq x$. Indeed, on the event $A_{n,xb}$, we have $S_n > xb$ and $S^*_{n-1} = \max_{x\leq j\leq n-1} S_j \leq \max_{0\leq j\leq n-1} j(b-a+2\epsilon_x) \leq 0$; and the events $\{S^*_{n-1} \leq 0\} \cap \{S_n > x\}$ are clearly disjoint. □

We now extend previous lemma to our more general framework. We still denote $S^\tau_n = \sum_1^n \tau_{-i}$ and we now introduce $S^{(\ell)}_n = \sum_1^n \zeta^{(\ell)}_{-i}$. We denote by N_x a function of x such that $N_x \to \infty$ and $N_x/x \to 0$ as $x \to \infty$. For any sequence ϵ_n, we define the following sequence of events, $n \geq N_x$,

$$\begin{aligned}K_{n,x} &= \left\{\left|\frac{S^\tau_k}{k} - a\right| \leq \epsilon_k,\ \forall N_x \leq k \leq n\right\} \\ &\bigcap \left\{\left|\frac{S^{(\ell)}_k}{k} - \gamma_\ell\right| \leq \epsilon_{k+1},\ \forall N_x \leq k < n,\ \forall \ell \in [1,3]\right\}.\end{aligned} \tag{13}$$

Note that $K_{n,x}$ is independent of the vector $(\zeta^{(1)}_{-n}, \zeta^{(2)}_{-n}, \zeta^{(3)}_{-n}, \zeta^{(4)}_{-n})$ and that due to the strong law of large number, we have still (11).

We define $\gamma_{[\geq 1]} = \max(\gamma_1, \gamma_2, \gamma_3)$, $\gamma_{[\geq i]} = \gamma_i$ for $i = 2, 3$.

Proposition 2 *Let Q_t be the stationary total number of packets in the network (defined in section 4). Let η_n be a sequence tending to 0 as n tends to infinity, and $K_{n,y}$ the sequence of events defined above. Let,*

$$\begin{aligned} A^{(j)}_{n,y} &= K_{n,y} \cap \{\zeta^{(j)}_{-n} > y + n(a - \gamma_{[\geq j]} + \eta_n)\}, \quad j \in [1,3] \\ A^{(j)}_{x} &= \bigcup_{n \geq x} A^{(j)}_{n,x\gamma_{[\geq j]}}, \end{aligned}$$

and $A_x = \cup_{j=1}^{3} A^{(j)}_x$. Then, as $x \to \infty$,

$$\begin{aligned} \mathbb{P}(Q_t > x) &\sim \mathbb{P}(Q_t > x, A_x) \\ &\sim \mathbb{P}(A_x) \sim \sum_j \sum_{n \geq x} \mathbb{P}(A^{(j)}_{n,x\gamma_{[\geq j]}}). \end{aligned} \tag{14}$$

Proof. Note that with the same argument as in the proof of Lemma 1, the events $\{A^{(j)}_{n,x}\}$ are disjoint for x sufficiently large. Hence we have

$$\begin{aligned} \mathbb{P}(A_x) &= \sum_j \sum_{n \geq x} \mathbb{P}(A^{(j)}_{n,x\gamma_{[\geq j]}}) \\ &= \sum_j \sum_{n \geq x} \mathbb{P}(K_{n,x\gamma_{[\geq j]}}) \cdot \\ &\quad \cdot \mathbb{P}(\zeta^{(j)}_{-n} > x\gamma_{[\geq j]} + n(a - \gamma_{[\geq j]} + \eta_n)) \\ &\sim \sum_j \tilde{d}^{(j)} \sum_{n \geq x} \overline{F}(x\gamma_{[\geq j]} + n(a - \gamma_{[\geq j]})) \\ &\sim \sum_j \frac{\tilde{d}^{(j)}}{a - \gamma_{[\geq j]}} \overline{F}^s(ax). \end{aligned}$$

Hence thanks to Theorem 1, we have $\mathbb{P}(Q_t > x) \sim \mathbb{P}(A_x)$. To prove the proposition, we have only to prove that $A_x \subset \{Q_t > x\}$, i.e. that $A^{(j)}_{n,x\gamma_{[\geq j]}} \subset \{Q_t > x\}$ for all j and $n \geq x$. We proceed as in previous proof. Take $j = 2$, we have $Q_t \geq n - k_n$ where

$$k_n = \sup \left\{ \ell; \zeta^{(2)}_{-n} + \max_{\ell \leq i \leq n-1} \left(S^{(2)}_{[-n+1,-i]} + S^{(4)}_{[-i,-l]} \right) \leq S^{\tau}_n \right\},$$

with $S^{(j)}_{[u,v]} = \sum_{i=u}^{v} \zeta^{(j)}_i$ for $u \leq v$ and $j \in [1,4]$. Note that k_n is the number of departures from the system constituted of nodes 2 and 4 without taking into account the node 3 (i.e. if service times ar null in node 3). Hence k_n is clearly an upper bound on the real number of departures from the whole network on the interval of time $(-S^{\tau}_n, 0]$. On the event $A^{(2)}_{n,x\gamma_{[\geq 2]}}$, we have $\zeta^{(2)}_{-n} + (\max\{\gamma_2, \gamma_4\} - \epsilon_n)k_n \leq n(a + \epsilon_n)$. For simplicity we omit now the terms ϵ_n or η_n since they do not play any role. We have

$$Q_t \geq n - k_n \geq n - \frac{na - x\gamma_{[\geq 2]} - n(a - \gamma_{[\geq 2]})}{\gamma_{[\geq 2]}} = x.$$

The proof is similar to show that $A^{(j)}_{n,x\gamma_{[\geq j]}} \subset \{Q_t > x\}$ for $j = 1, 3$. □

Consider a sequence of $J_{n,x}$ such that

1. $\sup_{n\geq x}\mathbb{P}(J_{n,x})\to 1$ as $x\to\infty$;
2. $J_{n,x}$ is independent of the vector $(\zeta_{-n}^{(1)},\zeta_{-n}^{(2)},\zeta_{-n}^{(3)},\zeta_{-n}^{(4)})$.

Note that

$$\mathbb{P}(K_{n,x}\cap J_{n,x})\geq\mathbb{P}(K_{n,x})+\mathbb{P}(J_{n,x}-1,$$

hence we have $\sup_{n\geq x}\mathbb{P}(K_{n,x}\cap J_{n,x})\to 1$ as $x\to\infty$. Then the following result follows directly from the proof of Proposition 2:

Corollary 1 *The result of Proposition 2 holds when we replace the events $K_{n,x}$ by the event $\tilde{K}_{n,x}=K_{n,x}\cap J_{n,x}$.*

We now define the typical event for $N(Z)$. It directly follows from the typical event for the random variable Z given in [4], Property 2 on page 85.

Proposition 3 *Let η_n be a sequence tending to 0 as n tends to infinity, and $\tilde{K}_{n,y}$ the sequence of events defined above. Let,*

$$B_x^{(j)} = \bigcup_{n\geq N_x}\tilde{A}_{n,xa}^{(j)},$$

and $B_x=\cup_{j=1}^3 B_x^{(j)}$. Then, as $x\to\infty$,

$$\begin{aligned}\mathbb{P}(N(Z)>x) &\sim \mathbb{P}(N(Z)>x,B_x)\\ &\sim \mathbb{P}(B_x)\sim\sum_j\sum_{n\geq N_x}\mathbb{P}(\tilde{A}_{n,xa}^{(j)}).\end{aligned}\tag{15}$$

Proof. It follows directly from Property 2 of [4] that

$$\mathbb{P}(Z>ax) \sim \mathbb{P}(B_x)\sim\sum_j\sum_{n\geq N_x}\mathbb{P}(\tilde{A}_{n,xa}^{(j)}),$$

and that $B_x\subset\{Z>ax\}$. Now on the event $K_{n,x}$, we have

$$N(Z)=\sum_k\mathbb{1}_{\{S_k^\tau\leq Z\}}\geq\sum_{k\geq N_x}\mathbb{1}_{\{k(a-\epsilon_k)\leq ax\}}\geq x-CN_x.$$

Hence we have

$$\begin{aligned}\mathbb{P}(N(Z)>x,B_x^c) &\leq \mathbb{P}(N(Z)\in(x-CN_x,x))\\ &= \mathbb{P}(N(Z)>x-CN_x)-\mathbb{P}(N(Z)>x)\\ &= o(\mathbb{P}(N(Z)>x)),\end{aligned}$$

by definition of N_x. □

We summarize the results of this section in the following corollary,

Corollary 2 *Consider a sequence of $J_{n,x}$ such that*

1. $\sup_{n\geq x}\mathbb{P}(J_{n,x})\to 1$ as $x\to\infty$;

2. $J_{n,x}$ is independent of the vector $(\zeta_{-n}^{(1)}, \zeta_{-n}^{(2)}, \zeta_{-n}^{(3)}, \zeta_{-n}^{(4)})$.

Let $\tilde{K}_{n,x} = K_{n,x} \cap J_{n,x}$ where $K_{n,x}$ is the event defined in (13). Let η_n be a sequence tending to 0 as n tends to infinity, and

$$\begin{aligned} \tilde{A}_{n,y}^{(j)} &= \tilde{K}_{n,y} \cap \{\zeta_{-n}^{(j)} > y + n(a - \gamma_{[\geq j]} + \eta_n)\}, \quad j \in [1,3] \\ \tilde{A}_x^{(j)} &= \bigcup_{n \geq N_x} \tilde{A}_{n,x\gamma_{[\geq j]}}^{(j)}, \end{aligned}$$

and $T_x = \cup_{j=1}^3 \tilde{A}_x^{(j)}$. Then, as $x \to \infty$,

$$\begin{aligned} \mathbb{P}(Q_r > x) &= \mathbb{P}(Q_r > x, T_x) + o(\overline{F}^s(x)) \\ &= \sum_{j=1}^{3} \sum_{n \geq N_x} \mathbb{P}\left(Q_r > x, \tilde{A}_{n,x\gamma_{[\geq j]}}^{(j)}\right) + o(\overline{F}^s(x)). \end{aligned} \tag{16}$$

Note that $A_x \cup B_x \subset T_x$, and then the corollary follows from previous propositions.

6 Computation of $\mathbb{P}(Q_r > x, T_x)$

In this section, we will see how to chose the events $J_{n,x}$ of Corollary 2 in order to make the computation of the sum (16) as easy as possible.

6.1 Representation of Q_r

We give here an explicit representation of Q_r that follows directly from the argument given in Section 4. For $n \geq 0$, let

$$D_n^i = \sup_{k \leq n, r_k = i} \left\{ T_k + \sup_{k \leq \ell \leq n} \left(S_{[k,\ell]}^1 + S_{[\ell,n]}^i \right) \right\},$$

be the departure time of packet $n \in \mathbb{Z}$ from node $i = 2, 3$. Define $N_-^2 := \max\left\{k,\ D_{-k}^3 > D_0^2\right\}$, $N_{++}^2 = \max\left\{k \geq 0,\ D_0^2 > D_k^3\right\}$ and $N_+^2 = \min\left\{j \geq 1, r_j = 2\right\}$ with the convention $\max \emptyset = -\infty$. Then we have

$$Q_r \mathbb{1}_{\{r_0=2\}} = \underbrace{\sum_{i=-N_-^2}^{0} \mathbb{1}_{\{r_i=2\}}}_{\Sigma^2} + \underbrace{\sum_{i=N_+^2}^{N_{++}^2} \mathbb{1}_{\{r_i=3\}}}_{\Delta^2},$$

with the convention $\sum_{\infty}^{0} = 0$ or $\sum_{k}^{-\infty} = 0$. The symmetric formula holds for $r_0 = 3$, hence Q_r is explicitly given. We denote by Σ^3, Δ^3 the corresponding quantities.

Note that only one sum is non-null, in particular, we have the following decomposition of the event $\{Q_r \mathbb{1}_{\{r_0=2\}} > x\} = \{\Sigma > x, r_0 = 2\} \cup \{\Delta > x, r_0 = 2\}$. Hence we see that in order to compute the sum (16), we have to compute the quantities where the dot is for Σ or Δ,

$$\sum_{n \geq N_x} \mathbb{P}\left(.^i > x, r_0 = i, \tilde{A}_{n,x\gamma_{[\geq j]}}^{(j)}\right),$$

for $j = 1, 2, 3$ and $i = 2, 3$. We will compute each of these terms separately. It is possible since if we define an event $J_{n,x}$ satisfying the conditions of Corollary 2, then the intersection of these events, will still satisfy these conditions.

6.2 The case $j = 2$ or 3

Lemma 2 *We have as $x \to \infty$,*

$$\sum_{n \geq N_x} \mathbb{P}\left(\Sigma^2 > x, r_0 = 2, \tilde{A}^{(2)}_{n,x\gamma_2}\right) = o(\overline{F}^s(x)).$$

Proof.

Note that on the event $\tilde{A}^{(2)}_{n,x\gamma_2}$, we have $r_{-n} = 2$, hence conditionally on $\tilde{A}^{(2)}_{n,x\gamma_2}$, $\{D^3_{-k}\}_{k\geq 0}$ and $\zeta^{(2)}_{-n}$ are independent. On the event $A^{(2)}_{n,x\gamma_2} \cap \{r_0 = 2\}$, we have

$$\begin{aligned} D^2_0 &\geq \zeta^{(2)}_{-n} - n(a + \epsilon_n) + n(\gamma_2 - \epsilon_n) \\ &\geq x\gamma_2. \end{aligned}$$

Moreover we have $D^3_{-k} \leq D^3_0$ for all $k \geq 0$. Since $N^2_- = \max\{k,\ D^3_{-k} > D^2_0\}$, we have

$$\{\Sigma^2 > x\} \subset \{N^2_- \neq -\infty\} \subset \{D^3_0 \geq x\gamma_2\}.$$

From this, we have the following upper bound

$$\begin{aligned} &\mathbb{P}(\Sigma^2 > x, r_0 = 2, A^{(2)}_{n,x\gamma_2}) \\ &\leq \mathbb{P}(D^3_0 \geq x\gamma_2)\mathbb{P}(\zeta^{(2)}_{-n} > x\gamma_2 + n(a - \gamma_2 + \eta_n)). \end{aligned}$$

Hence summing over n, we get

$$\begin{aligned} &\sum_{n > N_x} \mathbb{P}(\Sigma^2 > x, r_0 = 2, A^{(2)}_{n,x\gamma_2}) \\ &\leq \mathbb{P}(D^3_0 \geq x\gamma_2) \sum_{n>0} \mathbb{P}(\zeta^{(2)}_{-n} > x\gamma_2 + n(a - \gamma_2 + \eta_n)) \\ &\leq o(\overline{F}^s(x)). \end{aligned}$$

□

Lemma 3 *We have as $x \to \infty$,*

$$\begin{aligned} &\sum_{n \geq N_x} \mathbb{P}\left(\Sigma^3 > x, r_0 = 3, \tilde{A}^{(2)}_{n,x\gamma_2}\right) \\ &= \frac{p(1-p)d^{(2)}}{a - \gamma_2} \overline{F}^s\left(\frac{ax}{1-p}\right) + o(\overline{F}^s(x)). \end{aligned}$$

Proof.

We have $\Sigma^3 = \sum_{i=-N^3_-}^{0} \mathbb{1}_{\{r_i = 3\}}$ with $N^3_- = \max\{k,\ D^2_{-k} > D^3_0\}$. We will denote by z_x a function of x such that $z_x \to \infty$ and $z_x = o(x)$.

We denote $X_{-k}(i) := \left(T_i + \sup_{i\le\ell\le -k} S^1_{[i,\ell]} + S^2_{[\ell,-k]}\right)\mathbb{1}_{r_i=2}$ and define the following sequence of events,

$$\begin{aligned} J_{n,x} &= \bigcap_{k<n}\left\{\sup_{-n+1\le i\le -k} X_{-k}(i) \le z_x\right\} \\ &\cap \bigcap_{i\le -n}\left\{T_i - T_{-n} + \sup_{i\le\ell\le -n-1} S^1_{[i,\ell]} + S^2_{[\ell,-k]} \le z_x\right\} \\ &\cap \left\{\left|\frac{\sum_{i=-k}^{0}\mathbb{1}_{\{r_i=3\}}}{k} - (1-p)\right| \le \epsilon_k,\ N_x \le k \le n\right\} \\ &\cap \{D^2_{-n-1} \le z_x\}. \end{aligned}$$

We can choose a non-increasing sequence ϵ_n such that the conditions of Corollary 2 are satisfied (note in particular that $J_{n,x}$ is independent of ζ_{-n}). Once we chose the sequence ϵ, we define the function $z_x := x\epsilon_{\lfloor x\rfloor}$. For $0 \le k \le n$, we have

$$\begin{aligned} D^2_{-k} &= \sup_{i\le -k} X_{-k}(i) \\ &\le \sup_{-n+1\le i\le -k} X_{-k}(i) \\ &\quad + X_{-k}(-n) + \sup_{i\le -n-1}\left(X_{-k}(i) - X_{-k}(-n)\right) \end{aligned}$$

Moreover, we have

$$\begin{aligned} &\sup_{i\le -n-1}\left(X_{-k}(i) - X_{-k}(-n)\right) \\ &\quad\le \sup_{i\le -n} T_i - T_{-n} + \sup_{i\le\ell\le -n-1} S^1_{[i,\ell]} + S^2_{[\ell,-k]}. \end{aligned}$$

Hence on the event $J_{n,x}$, we have

$$D^2_{-k} \le 2z_x + T_{-n} + \sup_{-n\le\ell\le -k} S^1_{[i,\ell]} + S^2_{[\ell,-k]},$$

and we have clearly for $k \le n$,

$$D^2_{-k} \ge T_{-n} + \sup_{-n\le\ell\le -k} S^1_{[i,\ell]} + S^2_{[\ell,-k]}.$$

In what follows, C denotes a constant that may vary from place to place, but remains always finite and does not depend on n or x. On the event $\tilde{K}_{n,x\gamma_2}$, we have for all $k \le n$,

$$\begin{aligned} D^2_{-k} &\ge -na + \zeta^{(2)}_{-n} + (n-k)\gamma_2 - C(n\epsilon_n + z_x) \\ D^2_{-k} &\le -na + \zeta^{(2)}_{-n} + (n-k)\gamma_2 + C(n\epsilon_n + z_x). \end{aligned}$$

We will take the following convenient convention for summarizing these inequalities

$$D^2_{-k} \approx -na + \zeta^{(2)}_{-n} + (n-k)\gamma_2 \mp C(n\epsilon_n + z_x).$$

Hence on the event $\tilde{K}_{n,x\gamma_2} \cap \{r_0 = 3, D_0^3 \le z_x\}$, we have since $D^2_{-n-1} \le z_x$,

$$\begin{aligned} N^3 &= \max\{k \le n,\ D^2_{-k} > D^3_0\} \\ &\approx n - \frac{(na - \zeta^{(2)}_{-n})^+}{\gamma_2} \mp C(n\epsilon_n + z_x) \\ &\approx \min\left(n, \frac{\zeta^{(2)}_{-n} + n(\gamma_2 - a)}{\gamma_2}\right) \mp C(n\epsilon_n + z_x) \end{aligned}$$

Hence we have

$$\Sigma^3 \approx (1-p)\min\left(n, \frac{\zeta^{(2)}_{-n} + n(\gamma_2 - a)}{\gamma_2}\right) \mp C(n\epsilon_n + z_x).$$

Using the fact that $n \ge N_x$ and ϵ_n is non decreasing so that $\forall n \ge N_x$, $x\epsilon_n \le x\epsilon_{N_x} \le Cz_x$, we derive from previous equation,

$$\begin{aligned} \Sigma^3 > x \ &\Rightarrow \ \begin{cases} n > \frac{x}{1-p} - Cz_x \\ \zeta^{(2)}_{-n} > \frac{\gamma_2 x}{1-p} + n(a-\gamma_2) - C(n\epsilon_n + z_x) \end{cases} \\ \Sigma^3 > x \ &\Leftarrow \ \begin{cases} n > \frac{x}{1-p} + Cz_x \\ \zeta^{(2)}_{-n} > \frac{\gamma_2 x}{1-p} + n(a-\gamma_2) + C(n\epsilon_n + z_x) \end{cases} \end{aligned}$$

Note that we have

$$\begin{cases} n > \frac{x}{1-p} \\ \zeta^{(2)}_{-n} > \frac{\gamma_2 x}{1-p} + n(a-\gamma_2) \end{cases} \Rightarrow \begin{cases} n > N_x \\ \zeta^{(2)}_{-n} > \gamma_2 x + n(a-\gamma_2) \end{cases} \tag{17}$$

We have

$$\begin{aligned} &\mathbb{P}(\Sigma^3 > x, \tilde{A}^{(2)}_{n,\gamma_2 x}, r_0 = 3) \\ &\quad = \mathbb{P}(\Sigma^3 > x, \tilde{A}^{(2)}_{n,\gamma_2 x}, r_0 = 3, D^3_0 \le z_x) \\ &\qquad + \mathbb{P}(\Sigma^3 > x, \tilde{A}^{(2)}_{n,\gamma_2 x}, r_0 = 3, D^3_0 > z_x). \end{aligned}$$

Note that the second term is upper bounded by $\mathbb{P}(\tilde{A}^{(2)}_{n,\gamma_2 x}, r_0 = 3, D^3_0 > z_x)$ and its sum over n can be shown to be $o(\overline{F}^s(x))$ as in the proof of previous lemma. Hence we have

$$\begin{aligned} &\sum_{n>x} \mathbb{P}(\Sigma^3 > x, \tilde{A}^{(2)}_{n,\gamma_2 x}, r_0 = 3) \\ &\quad = \sum_{n>x} \mathbb{P}(\Sigma^3 > x, \tilde{A}^{(2)}_{n,\gamma_2 x}, r_0 = 3, D^3_0 \le z_x) + o(\overline{F}^s(x)) \\ &\quad = (1-p) \sum_{n > x/(1-p)} \mathbb{P}(\zeta^{(2)}_{-n} > \frac{\gamma_2 x}{1-p} + n(a-\gamma_2)) \\ &\qquad + o(\overline{F}^s(x)), \end{aligned}$$

where the last inequality follows from (17) and the fact that $\overline{F}^s$ is regularly varying.

Finally we get

$$\begin{aligned}\sum_{n>x} \mathbb{P}(\Sigma^3 > x, r_0 = 3, \tilde{A}^{(2)}_{n,x\gamma_2}) \\ = \quad \frac{(1-p)pd^{(2)}}{a-\gamma_2}\overline{F}^s\left(\frac{ax}{1-p}\right) + o(\overline{F}^s(x)).\end{aligned}$$

□

Lemma 4 *We have as* $x \to \infty$

$$\begin{aligned}\sum_{n\geq N_x} \mathbb{P}(\Delta^2 > x, r_0 = 2, \tilde{A}^{(2)}_{n,x\gamma_2}) \\ = \quad \frac{p^2 d^{(2)}}{a-\gamma_2}\overline{F}^s\left(\frac{ax}{1-p}\right) + o(\overline{F}^s(x)).\end{aligned}$$

Proof. With the same argument as above, we can define a sequence of events $J_{n,x}$ satisfying the conditions of Corollary 2 and such that for $n > N_x$, we have on the event $J_{n,x}$, for $k \geq 0$,

$$\begin{aligned} D_0^2 &\approx \left(-na + \zeta^{(2)}_{-n} + n\gamma_2\right) \mp C(n\epsilon_n + z_x), \\ D_k^3 &\approx ka \mp C(k\epsilon_k + z_x), \\ N_+^2 &\leq z_x. \end{aligned}$$

From these, we derive as in proof of Lemma 3 the following approximation

$$N^2_{++} \approx \frac{\zeta^{(2)}_{-n} - na + n\gamma_2}{a} \mp C(n\epsilon_n + z_x).$$

Hence, we have

$$\Delta^2 \approx (1-p)\frac{\zeta^{(2)}_{-n} - na + n\gamma_2}{a} \mp C(n\epsilon_n + z_x).$$

The end of the proof follows exactly the steps of the poof of Lemma 3:

$$\begin{aligned}\sum_{n\geq N_x} \mathbb{P}(\Delta^2 > x, r_0 = 3, \tilde{A}^{(2)}_{n,x\gamma_2}) \\ = \quad p\sum_{n\geq N_x} \mathbb{P}\left(\zeta^{(2)}_{-n} > \frac{ax}{1-p} + n(a-\gamma_2)\right) + o(\overline{F}^s(x)).\end{aligned}$$

□

Lemma 5 *We have as* $x \to \infty$

$$\sum_{n\geq N_x} \mathbb{P}(\Delta^3 > x, r_0 = 3, \tilde{A}^{(2)}_{n,x\gamma_2}) = o(\overline{F}^s(x)).$$

Proof. In the same way as in the proof of Lemma 2, we have

$$D^2_{N^3_+ - 1} \geq x\gamma_2,$$

hence we have

$$\{N^3_{++} \neq -\infty\} \subset \{D^3_0 \geq x\gamma_2\},$$

and the end of the argument is completely similar. □

6.3 The case $j = 1$

6.3.1 The case $\gamma_1 > \max(\gamma_2, \gamma_3)$

We only consider the case $r_0 = 2$ (the case $r_0 = 3$ is symmetric).

Lemma 6 *We have as $x \to \infty$*

$$\begin{aligned}\sum_{n \geq N_x} \mathbb{P}(\Sigma^2 > x, r_0 = 2, \tilde{A}^{(1)}_{n,x\gamma_{[\geq 1]}}) &= o(\overline{F}^s(x)),\\ \sum_{n \geq N_x} \mathbb{P}(\Delta^2 > x, r_0 = 2, \tilde{A}^{(1)}_{n,x\gamma_{[\geq 1]}}) &= o(\overline{F}^s(x)).\end{aligned}$$

Proof. We can define a sequence of events $J_{n,x}$ satisfying the conditions of Corollary 2 and such that for $n > N_x$, we have on the event $J_{n,x}$, for $k \leq n$,

$$\begin{aligned}D^3_{-k} &\approx -na + \zeta^{(1)}_{-n} + (n-k)\gamma_1 \mp C(n\epsilon_n + z_x),\\ D^2_0 &\approx -na + \zeta^{(1)}_{-n} + n\gamma_1 \mp C(n\epsilon_n + z_x).\end{aligned}$$

From these equations, we derive that $N^2_- \leq C(n\epsilon_n + z_x) \leq Cn\epsilon_n$ for $n > x$ (by definition of z_x). But we have $\Sigma^2 \leq N^2_- \leq Cn\epsilon_n$, hence $\Sigma^2 > x \Rightarrow Cn\epsilon_n > x$ and we can find a function $h(x)$ (depending only on the sequence ϵ_n such that $\Sigma^2 > x \Rightarrow Cn > h(x)$ and $h(x)/x \to \infty$. Hence we have

$$\begin{aligned}&\sum_{n>x} \mathbb{P}(\Sigma^2 > x, \tilde{A}^{(1)}_{n,x\gamma_1}, r_0 = 2)\\ &\leq \sum_{n>h(x)} \mathbb{P}(\zeta^{(1)}_{-n} > x\gamma_1 + n(a - \gamma_1 + \eta_n))\\ &\leq C\overline{F}^s(h(x)).\end{aligned}$$

For all $M > 0$, we have $h(x) \geq Mx$ for x sufficiently large, hence $\overline{F}^s(h(x)) \leq \overline{F}^s(Mx) \leq \frac{2^\alpha}{M^\alpha}\overline{F}^s(x)$ for x sufficiently large. This implies that $\overline{F}^s(h(x)) = o(\overline{F}^s(h(x)))$. The case of Δ^2 can be done following the same kind of argument. □

6.3.2 The case $\gamma_3 > \max(\gamma_1, \gamma_2) = \gamma_{1\vee 2}$

Lemma 7 *We have as $x \to \infty$*

$$\begin{aligned}&\sum_{n \geq N_x} \mathbb{P}(\Sigma^2 > x, r_0 = 2, \tilde{A}^{(1)}_{n,x\gamma_{[\geq 1]}})\\ &= \frac{pd^{(1)}}{a - \gamma_3}\overline{F}^s\left(\frac{x\gamma_3(a - \gamma_{1\vee 2})}{p(\gamma_3 - \gamma_{1\vee 2})}\right) + o(\overline{F}^s(x)).\end{aligned}$$

Proof. It is easy to construct an appropriate sequence of events $J_{n,x}$ satisfying the conditions of Corollary 2 such that, we have on the event $J_{n,x}$,

$$\begin{aligned}D^3_{-n-1} &\leq z_x\\ D^3_{-k} &\approx -na + \zeta^{(1)}_{-n} + (n-k)\gamma_3 \mp C(n\epsilon_n + z_x),\ \forall k \leq n\\ D^2_0 &\approx \max\left(-na + \zeta^{(1)}_{-n} + n\gamma_{1\vee 2}, 0\right) \mp C(n\epsilon_n + z_x).\end{aligned}$$

From these equations, we derive that

$$N_-^2 \approx \min\left(n\frac{\gamma_3-\gamma_{1\vee 2}}{\gamma_3}, \frac{\zeta_{-n}^{(1)}-n(a-\gamma_3)}{\gamma_3}\right) \mp C(n\epsilon_n + z_x),$$

and then that

$$\Sigma^2 \approx p\min\left(n\frac{\gamma_3-\gamma_{1\vee 2}}{\gamma_3}, \frac{\zeta_{-n}^{(1)}-n(a-\gamma_3)}{\gamma_3}\right) \mp C(n\epsilon_n + z_x).$$

With the same argument as above, we have

$$\begin{aligned}
&\sum_{n>x} \mathbb{P}(\Sigma^2 > x, \tilde{A}^{(1)}_{n,\gamma_3 x}, r_0 = 2) \\
&= \sum_{n>\gamma_3 x/p(\gamma_3-\gamma_{1\vee 2})} p\mathbb{P}\left(\zeta_{-n}^{(1)} > x\gamma_3/p + n(a-\gamma_3)\right) \\
&\quad + o(\overline{F}^s(x)) \\
&= \frac{pd^{(1)}}{a-\gamma_3}\overline{F}^s\left(\frac{x\gamma_3(a-\gamma_{1\vee 2})}{p(\gamma_3-\gamma_{1\vee 2})}\right) + o(\overline{F}^s(x)).
\end{aligned}$$

□

Lemma 8 *We have as* $x \to \infty$

$$\begin{aligned}
&\sum_{n\geq N_x} \mathbb{P}(\Delta^3 > x, r_0 = 3, \tilde{A}^{(1)}_{n,x\gamma_{[\geq 1]}}) \\
&= \frac{(1-p)d^{(1)}}{a-\gamma_3}\overline{F}^s\left(\frac{x\gamma_3(a-\gamma_{1\vee 2})}{p(\gamma_3-\gamma_{1\vee 2})}\right) + o(\overline{F}^s(x)).
\end{aligned}$$

Proof. We construct events $J_{n,x}$ such that

$$\begin{aligned}
D_0^3 &\approx \left(\zeta_{-n}^{(1)} - na + n\gamma_3\right)^+ \mp C(n\epsilon_n + z_x), \\
D_k^2 &\approx \max\left(\zeta_{-n}^{(1)} - na + (n+k)\gamma_3, ka\right) \mp C(n\epsilon_n + z_x), \\
N_+^3 &\leq z_x.
\end{aligned}$$

From these equations, we derive (we omitted the term: $\mp C(n\epsilon_n + z_x)$)

$$\begin{aligned}
N_{++}^3 &\approx \min\left(\frac{n(\gamma_3-\gamma_{1\vee 2})}{\gamma_{1\vee 2}}, \frac{\zeta_{-n}^{(1)} - na + n\gamma_3}{a}\right), \\
\Delta^3 &\approx p\min\left(\frac{n(\gamma_3-\gamma_{1\vee 2})}{\gamma_{1\vee 2}}, \frac{\zeta_{-n}^{(1)} - na + n\gamma_3}{a}\right).
\end{aligned}$$

Hence we have

$$\begin{aligned}
&\sum_{n\geq N_x} \mathbb{P}(\Delta^3 > x, r_0 = 3, \tilde{A}^{(1)}_{n,x\gamma_{[\geq 1]}}) \\
&= (1-p) \sum_{n\geq \frac{\gamma_{1\vee 2}x}{p(\gamma_3-\gamma_{1\vee 2})}} \mathbb{P}\left(\zeta_{-n}^{(1)} > \frac{ax}{p} + n(a-\gamma_3)\right) \\
&\quad + o(\overline{F}^s(x)),
\end{aligned}$$

from which the result follows. □

Lemma 9 *We have as* $x \to \infty$

$$\begin{aligned}\sum_{n \geq N_x} \mathbb{P}(\Sigma^3 > x, r_0 = 3, \tilde{A}^{(1)}_{n,x\gamma_{[\geq 1]}}) &= o(\overline{F}^s(x)),\\ \sum_{n \geq N_x} \mathbb{P}(\Delta^2 > x, r_0 = 2, \tilde{A}^{(1)}_{n,x\gamma_{[\geq 1]}}) &= o(\overline{F}^s(x)).\end{aligned}$$

Proof. The proof of these results is very similar to the proof of Lemma 6 and is omitted. □

7 Conclusion

We analyzed the impact of heavy-tailed delays on different paths for the resequencing delay and the resequencing queue size. We compute the exact asymptotics for a simple model of 2 queues in parallel. These asymptotics are dominated by the heaviest tail distribution of the delays among the 2 queues. We studied in details different stochastic assumptions and gave the corresponding rule-of-thumb that minimizes the resequencing delay or queue size (in an asymptotic sense). Surprisingly, we found that the values of the optimal load-balancing is not the same depending on the quantity to optimize. Moreover, in certain cases, the rule-of-thumb agrees with a natural rule when only the average load of each link is known and a $M/M/1$ model is used.

As a final remark, we should stress that our results can be extended to other network topologies. The framework of [4] and [12] covers the class of stochastic event graphs.

There are two straightforward extensions of our results:

- one is to consider K nodes in parallel with $K > 2$;
- the other is to replace each node in parallel by a more complex network belonging to the class of the stochastic event graphs.

In both cases, we see that the upper bound derived in Section 4 is still valid. Moreover Theorem 1 as stated in [12] covers these cases. Hence we can apply exactly the same methodology and we will get the exact asymptotics of the resequencing queue size.

These extensions to other load-balanced routing networks can help us to understand the theoretical bounds and be useful to the general class of multicommodity network problems. In particular, the problem of reordering packets appears in the architecture of parallel switches [21], [23]. In the context of ad-hoc networks, it is known that load-balancing packets across multiple paths increase the throughput [15]. In this context, empirical studies [20] show that the distribution of the inter-contact time between nodes in an opportunistic networking environment follows an approximate power law over a large range, which gives credit to our analysis.

8 Proof of Theorem 1 point 3

Note that this is the only case, that is not covered by [12].

We have

$$R = \max\{X^{(4)} - X^{(2)}, X^{(4)} - X^{(3)}\},$$

where $X^{(i)}$ denote the departure time of packet or clone 0 from node i in the stationary regime. We have here

$$X^{(2)} = \sup_{k\le 0}\left\{T_k + \sup_{k\le\ell\le n}\left(S^1_{[k,\ell]} + S^2_{[\ell,n]}\right)\right\}.$$

For ϵ such that $\epsilon < a-\gamma_2$, we define

$$\begin{aligned} S^{(2)}_{[\ell,n]}(\epsilon) &= \sum_{i=\ell}^{n}(\zeta^{(2)}+\epsilon)^+, \\ X^{(2)}_\epsilon &= \sup_{k\le 0}\left\{T_k + \sup_{k\le\ell\le n}\left(S^1_{[k,\ell]} + S^2_{[\ell,n]}(\epsilon)\right)\right\}. \end{aligned}$$

First consider the case $\epsilon > 0$. Note that this quantity is finite and corresponds exactly to the same system where we add ϵ at each service time in the node 2 in such a way that the system is still stable. We have obviously $X^{(2)}_\epsilon \ge X^{(2)}$, which implies

$$R_\epsilon = \max\{X^{(4)} - X^{(2)}_\epsilon, X^{(4)} - X^{(3)}\} \le R.$$

We can apply Theorem 1 to R_ϵ and we get

$$\begin{aligned} \mathbb{P}(R_\epsilon > x) &= \frac{d^{(1)}}{a-\gamma_2-\epsilon}\overline{F}^s\left(\frac{a-\gamma_3}{\epsilon}x\right) \\ &\quad + \left(\frac{pd^{(2)}}{a-\gamma_2-\epsilon} + \frac{(1-p)d^{(3)}}{a-\gamma_3}\right)\overline{F}^s(x) \\ &\quad + o(\overline{F}^s(x)). \end{aligned}$$

Hence if we assume that $\overline{F}^s \in \mathcal{R}(-\alpha)$ with $\alpha > 0$, we have

$$\begin{aligned} &\lim_{x\to\infty}\frac{\mathbb{P}(R_\epsilon > x)}{\overline{F}^s(x)} \\ &= \left(\frac{d^{(1)}}{a-\gamma_2-\epsilon}\left(\frac{\epsilon}{a-\gamma_3}\right)^\alpha + \frac{pd^{(2)}}{a-\gamma_2-\epsilon} + \frac{(1-p)d^{(3)}}{a-\gamma_3}\right). \end{aligned}$$

Since $\mathbb{P}(R > x) \ge \mathbb{P}(R_\epsilon > x)$ we have

$$\liminf_{x\to\infty}\frac{\mathbb{P}(R>x)}{\overline{F}^s(x)} \ge \frac{pd^{(2)}}{a-\gamma_2} + \frac{(1-p)d^{(3)}}{a-\gamma_3}$$

For the upper bound, one can proceed similarly with $\epsilon < 0$.

References

[1] S. Asmussen, C. Klüppelberg, and K. Sigman. Sampling at subexponential times, with queueing applications. *Stochastic Process. Appl.*, 79(2):265–286, 1999.

[2] F. Baccelli, G. Cohen, G. J. Olsder, and J.-P. Quadrat. *Synchronization and Linearity*. Wiley, 1992. Available at http://www-rocq.inria.fr/metalau/cohen/SED/book-online.html.

[3] F. Baccelli and S. Foss. Moments and tails in monotone-separable stochastic networks. *Ann. Appl. Probab.*, 14(2):612–650, 2004.

[4] F. Baccelli, M. Lelarge, and S. Foss. Asymptotics of subexponential max plus networks: the stochastic event graph case. *Queueing Syst.*, 46(1-2):75–96, 2004.

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

[6] J. C. R. Bennett, C. Partridge, and N. Shectman. Packet reordering is not pathological network behavior. *IEEE/ACM Trans. Netw.*, 7(6):789–798, 1999.

[7] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1989.

[8] E. Blanton and M. Allman. On making TCP more robust to packet reordering. *SIGCOMM Comput. Commun. Rev.*, 32(1):20–30, 2002.

[9] S. Borst, O. Boxma, and P. Jelenković. Reduced-load equivalence and induced burstiness in GPS queues with long-tailed traffic flows. *Queueing Syst.*, 43(4):273–306, 2003.

[10] M. E. Crovella. Performance evaluation with heavy tailed distributions. In *Lecture Notes in Computer Science 1786*, pages 1–9, Mar. 2000.

[11] M. E. Crovella, R. Frangioso, and M. Harchol-Balter. Connection scheduling in Web servers. In *1999 USENIX Symposium on Internet Technologies and Systems (USITS '99)*, 1999.

[12] A. B. Dieker and M. Lelarge. Tails for (max,plus) recursions under subexponentiality. *IBM report, available at: http://www.di.ens.fr/∼lelarge/.*

[13] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer Verlag, 2003.

[14] S. Foss and D. Korshunov. Sampling at a random time with a heavy-tailed distribution. *Markov Process. Related Fields*, 6(4):543–568, 2000.

[15] M. Grossglauser and D. N. C. Tse. Mobility increases the capacity of ad-hoc wireless networks. In *INFOCOM*, pages 1360–1369, 2001.

[16] R. Haji and G. F. Newell. A relation between stationary queue and waiting time distributions. *J. Appl. Probability*, 8:617–620, 1971.

[17] http://www.brocade.com.

[18] http://www.cisco.com/univercd/home/home.htm.

[19] http://www.ietf.org/rfc/rfc1990.txt.

[20] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket switched networks and human mobility in conference environments. In *WDTN '05: Proceeding of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, pages 244–251, New York, NY, USA, 2005. ACM Press.

[21] S. Iyer and N. W. McKeown. Analysis of the parallel packet switch architecture. *IEEE/ACM Trans. Netw.*, 11(2):314–324, 2003.

[22] A. Jean-Marie and L. Gün. Parallel queues with resequencing. *J. Assoc. Comput. Mach.*, 40(5):1188–1208, 1993.

[23] I. Keslassy, C. Chang, N. McKeown, and D. Lee. Optimal load-balancing. In *Infocom 2005, Miami, Florida.*, 2005.

[24] M. Laor and L. Gendel. The effect of packet reordering in a backbone link on application throughput. *IEEE Network*, September 2002.

[25] V. Paxson. Automated packet trace analysis of TCP implementations. In *SIGCOMM '97: Proceedings of the ACM SIGCOMM '97 conference on Applications, technologies, architectures, and protocols for computer communication*, pages 167–179, New York, NY, USA, 1997. ACM Press.

[26] W. Willinger, V. Paxson, R. H. Riedi, and M. S. Taqqu. Long-range dependence and data network traffic. In *Theory and applications of long-range dependence*, pages 373–407. Birkhäuser Boston, Boston, MA, 2003.

[27] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Trans. Netw.*, 5(1):71–86, 1997.

[28] Y. Xia and D. Tse. On the large deviation of resequencing queue size: 2-$M/M/1$ case. *IEEE INFOCOM Proceedings*, 2004.

[29] Y. Xia and D. N. C. Tse. Analysis on packet resequencing for reliable network protocols. In *INFOCOM*, 2003.

[30] X. Zhou and P. V. Mieghem. Reordering of IP packets in internet. *Lecture Notes in Computer Science*, 3015:237–246, Jan 2004.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399