HAL
open science

Asymptotic Tail Distribution of End-to-End Delay in Networks of Queues with Self-Similar Cross Traffic

Marc Lelarge, Zhen Liu, Cathy Xia

To cite this version:

Marc Lelarge, Zhen Liu, Cathy Xia. Asymptotic Tail Distribution of End-to-End Delay in Networks of Queues with Self-Similar Cross Traffic. [Research Report] RR-4951, INRIA. 2003. inria-00071628

HAL Id: inria-00071628

https://inria.hal.science/inria-00071628

Submitted on 23 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Asymptotic Tail Distribution of End-to-End Delay in Networks of Queues with Self-Similar Cross Traffic

Marc Lelarge — Zhen Liu — Cathy Xia

N° 4951

October 2003

THÈME 1

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--4951--FR+ENG

Asymptotic Tail Distribution of End-to-End Delay in Networks of Queues with Self-Similar Cross Traffic

Marc Lelarge * , Zhen Liu † , Cathy Xia ‡

Thème 1 —Réseaux et systèmes
Projet TREC

Rapport de recherche n° 4951 —October 2003 —27 pages

Abstract: We consider the steady state distribution of the end-to-end delay of a tagged flow in queueing networks where the queues have self-similar cross traffic. We assume that such cross traffic at each queue, say queue i, is modeled by fractional Brownian Motion (FBM) with Hurst parameter $H_i \in [1/2, 1)$, and is independent of other queues.The arrival process of the tagged flow is renewal. Two types of queueing networks are considered.

We show that the end-to-end delay of the tagged flow in a tandem queueing network, and more generally in a tree network, is completely dominated by one of the queues. The dominant queue is the one with the maximal Hurst parameter. If several queues have the same maximal Hurst parameter, then we have to compare the ratio $\frac{(1-\rho)^H}{\sigma}$ to determine the dominant queue, where ρ is the load of the queue.

In the case that the tagged flow is controlled through a window based congestion control mechanism, the end-to-end delay is still asymptotically Weibullian with the same shape parameter. We provide upper and lower bounds on the constant that determines the scale parameter of the corresponding Weibull distribution.

Key-words: fractional brownian motion, open queueing network, tandem queue, (max, plus)-network

* INRIA-ENS, ENS, 45 rue d'Ulm, 75005 Paris, France, Marc.Lelarge@ens.fr
† IBM T. J. Watson Research Center, Yorktown Heights, P.O. Box 704, NY 10598, USA, {zhenl,cathyx}@us.ibm.com
‡ IBM T. J. Watson Research Center, Yorktown Heights, P.O. Box 704, NY 10598, USA, {zhenl,cathyx}@us.ibm.com

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

Asymptotique des queues de distribution du délais de bout en bout dans des réseaux de files d'attentes avec du trafic transverse auto-similaire

Résumé : Nous considérons la distribution staionnaire du délais de bout en bout dans un réseau de files d'attentes qui reçcoivent chacune un trafic transverse auto-similaire. Nous supposons que le trafic transverse arrivant à chacune des files est modélisé par exemple pour la file i par un mouvement brownien fractionnaire avec un paramètre de Hurst $H_i \in [1/2, 1)$ et est indépendant des autres files. Le processus des arrivées dans le réseau est un processus de renouvellement. Deux types de réseaux ont été considérés.

Dans le cas de files d'attente en série et plus généralement dans tout type de réseaux ayant une structure arborescente, nous montrons que le délais de bout en bout est complètement dominé par l'une des files. La file dominante est celle qui possède le paramètre de Hurst maximal. Si plusieurs files ont le même paramètre de Hurst, le ratio $\frac{(1-\rho)^H}{\sigma}$ permet de déterminer la file dominante (ρ correspond à la charge de la file).

Dans le cas d'un réseau contrôlé par un mécanisme de contrôle de flux par fenêtre, le délais de bout en bout est encore asymptotiquement weibullien avec le même paramètre de forme. Nous donnons des bornes supérieure et inférieure sur la constante qui détermine le paramètre d'échelle.

Mots-clés : mouvement brownien fractionnaire, réseau ouvert, files d'attente en série, réseau (max,plus)

1 Introduction

Since the seminal work of Leland et al. [18], there has been tremendous research work on the characterization and modeling of the Internet traffic. A number of studies have confirmed the prevalence of the Long Range Dependence (LRD) and fractal behavior in LAN and WAN as well as in Web servers (see e.g. [23]). Self-similar (SS) scaling exists in large time scales (i.e. one second or larger), cf. [18, 25, 12]. More recently, multi-fractal phenomena were observed and analyzed, see [1, 15, 16]. At small time scale, the traffic still has mono-fractal behavior in the Internet backbone [29]. Many people have also investigated into the possible causes of these long-range-dependent and self-similar phenomena, see [23] for an overview.

These observations and analyses have motivated a number studies of the performance impact in the network. In order to account for the self-similar nature of the traffic in the Internet, Fractional Brownian Motion (FBM) input model has been introduced, mostly due to its mathematical simplicity. A detailed treatment of FBM processes can be found in [26]. Its use for traffic modeling is discussed in [21] and in [23] (and references therein).

Recall that a standard FBM process with Hurst parameter $H \in [1/2, 1)$ is a Gaussian centered process with stationary increments, continuous paths and such that

$$E[Z(s)Z(t)] = \frac{1}{2}\left(s^{2H} + t^{2H} - |s-t|^{2H}\right).$$

for all $s, t \in \mathbb{R}$.

Queues with FBM input process has received much attention in the literature. Studies [21, 13, 19, 22, 28] have focused primarily on the workload W of a *single* server queue, where $W := \sup_{t>0}(\rho t + \sigma Z_t - t)$, with mean input rate ρ, standard deviation σ, and server capacity 1. A lower bound $\mathbb{P}(W > x)$ was first obtained in [21], this lower bound has been later shown in [13] to be asymptotically exact in logarithm using large deviation principle, further extensions on deriving exact expression and stronger asymptotic estimates are developed in [22] and [19]. All these studies assert that the workload W of a single server queue is asymptotically Weibullian, namely,

$$\log \mathbb{P}(W > x) \sim -\frac{1}{2\sigma^2} x^{2(1-H)} \frac{(1-\rho)^{2H}(1-H)^{2(H-1)}}{H^{2H}}. \tag{1}$$

In this paper, we focus on the end-to-end delay in a network setting. Indeed, there are many studies on the characterization of end-to-end delay in the Internet, see e.g. [6, 20, 24]. In particular, Mukherjee [20] found that packet delay along several Internet paths could be modeled using a shifted gamma distribution whose parameters vary from path to path and on time scales of hours. There also exists work on the inference of end-to-end delay, see e.g. [7, 9], where nonparametric models are used.

To the best of our knowledge, there exist few results on the tail asymptotic of the end-to-end delay in a network setting. Under the assumptions of independent and identically distributed (i.i.d.) service times and of the existence of moment generating functions, large deviation results were derived in [27] and [17] for stochastic event graphs. In case when the service times are i.i.d. and subexponential, exact asymptotics were obtained in [4] for stochastic event graphs, where the end-to-end delay has subexponential tail distribution. In the current paper, we focus on another cause of heavytailness for the end-to-end delay, namely LRD, what has not been done in a network context.

We consider the steady state distribution of the end-to-end delay denoted by D, of a tagged flow in queueing networks where some of the queues have self-similar cross traffic. We assume that such cross traffic, say at queue i, is modeled by Fractional Brownian Motion (FBM) with Hurst parameter

$H_i \in [1/2, 1)$, and is independent of other queues. Note that when $H_i = 0.5$, we have an ordinary Brownian motion model. We assume that at least one of the queues has the Hurst parameter that is strictly greater than 0.5. The arrival process of the tagged flow is renewal. Two types of queueing networks are considered.

We show that the end-to-end delay of the tagged flow in a tandem queueing network is completely dominated by one of the queues. The dominant queue is the one with the maximal Hurst parameter. If several queues have the same maximal Hurst parameter, then we have to compare the ratio $\frac{(1-\rho)^H}{\sigma}$ to determine the dominant queue, where ρ is the load of the queue. We have then

$$\log \mathbb{P}(D > x) \sim \log \mathbb{P}(W > x),$$

where W is the steady-state workload of a single server queue with the same FBM inputs as the dominant queue, which is known [21, 13] to be asymptotically Weibullian.

We also consider the case that the tagged flow is controlled through a window-based congestion control mechanism, where at any time, the number of customers of the tagged flow is upper bounded by the window size. For any arbitrarily fixed window size, we show that the end-to-end delay is still asymptotically Weibullian with the same shape parameter. We also provide upper and lower bounds on the constant that determines the scale parameter of the corresponding Weibull distribution. These results further imply round-trip time (RTT) of TCP sessions is asymptotically Weibullian, see comments in Section 4.

All these results are shown to hold not only in tandem queueing networks, which correspond to the *unicast* flows, but also for *tree queueing networks* which model the *multicast* case. The tree queueing network is defined as a queueing network with a tree topology, where, at the completion of the service at a queueing station, a customer is forked into multiple customers, one for each downstream queueing station. Such queueing networks are also referred to as *disassembly queueing networks* in the queueing literature. In such a case, the end-to-end delay is defined as the delay for a customer to traverse all the queues. Thus, our result shows that the end-to-end delay in multicast is still asymptotically Weibullian, provided that at least one of the queues has FBM cross traffic.

In the next section, we introduce the general model for a network of queues using the (max,plus)-linear presentation, where the two types of networks of interest are described in detail. In Section 3, we focus on general tree queueing networks and show our main results on the logarithmic tail asymptotic of the end-to-end delay D in steady-state. The detailed technical proofs and necessary auxiliary results are also included, where we show how the general Borell inequality for the distribution of the supremum of a centered Gaussian process can be applied to derive logarithmic estimate for $\log \mathbb{P}(D > x)$. Our studies on networks with fixed window control are then presented in Section 4. We show that the end-to-end delay is still asymptotically Weibullian and give upper and lower bound for the constant that determines the scale parameter of the corresponding Weibull distribution. Finally, concluding remarks are discussed in Section 5.

Notation

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 $x \to \infty$. We will also use the notation $f(x) = \mathcal{O}(g(x))$ to mean $\limsup f(x)/g(x) < \infty$ and $\liminf f(x)/g(x) > 0$.

2 Model

2.1 Taking Cross Traffic into Account

Consider a network of queues with cross traffic, illustrated in figure 1. We assume that the service times of customers 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 our model, in order to analyze the delay of the tagged customers, we define the virtual service times for each tagged customer to be the amount of cross traffic arrived between two successive arrivals of the tagged customers. This (virtual) service time is denoted as v_n^i (for server i). The resulting queueing system (with such virtual service times) is a single class FIFO queueing networks. In the sequel, we shall thus consider only such FIFO queues, with the (virtual) service times to be possibly self-similar.

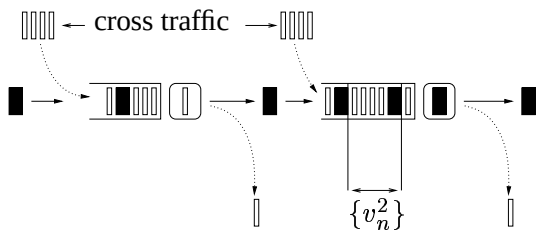

Figure 1: Queues with cross traffic

2.2 (max,plus)-Linear Systems

In our proofs, we shall use the (max,plus) algebra to describe the dynamics of the queueing networks. The reader is referred to [3] for details.

Definition 1. *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. $a \otimes b = a + b$). The zero element is denoted $\epsilon = -\infty$.*

For matrices of appropriate sizes, we define $(A \oplus B)^{(i,j)} = A^{(i,j)} \oplus B^{(i,j)} = \max(A^{(i,j)}, B^{(i,j)})$, $(A \otimes B)^{(i,j)} = \bigoplus_k A^{(i,k)} \otimes B^{(k,j)} = \max_k(A^{(i,k)} + B^{(k,j)})$.
Let s be an arbitrary fixed natural number. Assume the following to be given:

- $N = \{T_n, n \in \mathbb{Z}\}$, where $T_n \in \mathbb{R}$, the arrival time sequence;
- $\{A_n, n \in \mathbb{Z}\}$, where A_n is a $s \times s$ matrix;
- $\{B_n, n \in \mathbb{Z}\}$, where B_n, is a $s-$dimensional vector.

The associated (max,plus)-linear recurrence is that with state variable sequence $\{X_n, n \in \mathbb{Z}\}$, where X_n is a $s-$dimensional vector, which satisfies the evolution equation:

$$X_n = A_n \otimes X_{n-1} \oplus B_n \otimes T_n. \tag{2}$$

For $k < n$, we define $\Delta_{[k,n]} = \bigotimes_{j=n}^{k+1} A_j = A_n \otimes \cdots \otimes A_{k+1}$ and $\Delta_{[n,n]} = E$, the identity matrix. By definition, the stationary maximal dater is

$$\begin{aligned} D &= \lim_{n\to\infty} D_{[-n,0]} \\ &= \max_{0\le k} \left[(\Delta_{[-k,0]} \otimes B_{-k})^{(s)} + T_{-k} - T_0 \right]. \end{aligned}$$

Existence and finiteness of this limit will be addressed in section 2.4.
In the following section, we will consider two type of networks. We derive the (max,plus)-linear recurrence (2) that corresponds to each of these networks and give the interpretation of D.

2.3 Examples

2.3.1 Queues in tandem

We consider two queues in tandem, as illustrated in figure 2. Let v_n^i be the n-th service time at the i-th

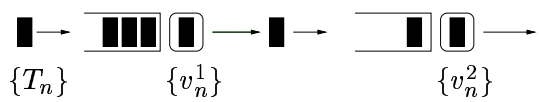

Figure 2: Queues in Tandem

server. We denote by x_n^1 (resp. x_n^2) the end of the n-th service in queue 1 (resp. 2). We have then

$$x_n^1 = (T_n \oplus x_{n-1}^1) \otimes v_n^1, \tag{3}$$
$$x_n^2 = (x_n^1 \oplus x_{n-1}^2) \otimes v_n^2. \tag{4}$$

Putting (3) in (4) gives

$$x_n^2 = (x_{n-1}^1 \oplus T_n) \otimes (v_n^1 \otimes v_n^2) \oplus x_{n-1}^2 \otimes v_n^2,$$

with

$$X_n = \begin{pmatrix} x_n^1 \\ x_n^2 \end{pmatrix}, \quad B_n = \begin{pmatrix} v_n^1 \\ v_n^1 + v_n^2 \end{pmatrix},$$
$$A_n = \begin{pmatrix} v_n^1 & \epsilon \\ v_n^1 + v_n^2 & v_n^2 \end{pmatrix},$$

so that

$$X_n = A_n \otimes X_{n-1} \oplus B_n \otimes T_n.$$

2.3.2 Tree Queueing Network

Consider a tree with nodes numbered by $1, 2, \ldots, m$ such that j is a successor of i (we write $j \in Suc(i)$) implies $i \leq j$. In particular node 1 is the root. We associate to this tree the network with m queues and in which departure process of queue i is the input process of queues $j \in Suc(i)$. Queues in tandem is a special case of tree network. Note that in the literature, such tree networks are also referred to as disassembly networks.

We take the example of a tree with three queues, as illustrated in figure 3. The end-to-end delay here is defined as the delay for a customer to traverse all the queues, which is taken care of by the dummy node (a max operator) in the end.

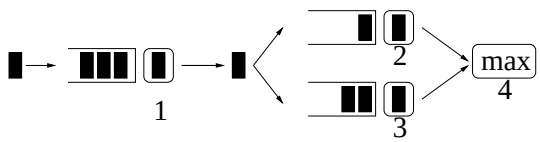

Figure 3: Tree Network

Let v_n^i be the n-th service time at the i-th server. We denote by x_n^i the end of the n-th service in server i. We have then

$$
\begin{aligned}
x_n^1 &= (T_n \oplus x_{n-1}^1) \otimes v_n^1, && (5)\\
x_n^2 &= (x_n^1 \oplus x_{n-1}^2) \otimes v_n^2, && (6)\\
x_n^3 &= (x_n^1 \oplus x_{n-1}^3) \otimes v_n^3, && (7)\\
x_n^4 &= x_n^2 \oplus x_n^3. && (8)
\end{aligned}
$$

Putting equation (5) in (6), (7), (8) we obtain the desired recursion equation with

$$
\begin{aligned}
X_n &= \begin{pmatrix} x_n^1 \\ x_n^2 \\ x_n^3 \\ x_n^4 \end{pmatrix}, \quad B_n = \begin{pmatrix} v_n^1 \\ v_n^1 + v_n^2 \\ v_n^1 + v_n^3 \\ v_n^1 + \max(v_n^2, v_n^3) \end{pmatrix},\\
A_n &= \begin{pmatrix} v_n^1 & \epsilon & \epsilon & \epsilon \\ v_n^1 + v_n^2 & v_n^2 & \epsilon & \epsilon \\ v_n^1 + v_n^3 & \epsilon & v_n^3 & \epsilon \\ v_n^1 + \max(v_n^2, v_n^3) & v_n^2 & v_n^3 & \epsilon \end{pmatrix}.
\end{aligned}
$$

2.3.3 Queueing network with fixed window control

We consider now m queues in tandem with a window-based control which does not allow more than L customers in the system. In other words, the n th customer can enter the first queue only after the $n-L$ th customer leaves the last queue in the tandem queueing network. We denote by x_n^i the end of the n-th service in queue i.

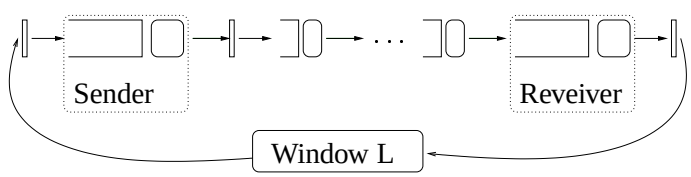

Figure 4: Tandem Queueing Network with Fixed Window Control

For the network of two queues in figure 4, we have then

$$
\begin{aligned}
x_n^1 &= (T_n \oplus x_{n-1}^1 \oplus x_{n-L}^2) \otimes v_n^1, && (9)\\
x_n^2 &= (x_n^1 \oplus x_{n-1}^2) \otimes v_n^2. && (10)
\end{aligned}
$$

From these equations, if we put (9) in (10), we obtain the desired recursion equation with

$$X_n = \begin{pmatrix} x^2_{n-L+1} \\ x^2_{n-L+2} \\ \vdots \\ \vdots \\ x^2_{n-1} \\ x^1_n \\ x^2_n \end{pmatrix}, \quad B_n = \begin{pmatrix} \epsilon \\ \epsilon \\ \vdots \\ \vdots \\ \epsilon \\ v^1_n \\ v^1_n + v^2_n \end{pmatrix},$$

$$A_n = \begin{pmatrix} \epsilon & 0 & \epsilon & \dots & \dots & \dots & \epsilon \\ \epsilon & \epsilon & 0 & \epsilon & \dots & \dots & \epsilon \\ \vdots & \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 & \epsilon & \epsilon \\ \epsilon & \epsilon & \dots & \dots & \epsilon & \epsilon & 0 \\ v^1_n & \epsilon & \dots & \dots & \epsilon & v^1_n & \epsilon \\ v^1_n + v^2_n & \epsilon & \dots & \dots & \epsilon & v^1_n + v^2_n & v^2_n \end{pmatrix}.$$

In the tree network case, the window control with size L is implemented in such a way that the n th customer can enter the first queue (root of the tree) only after all the $n - L$ th customers quit the leaf queues in the tree queueing network.

2.3.4 Interpretation of D

Consider one of the previous networks with only customers $-n, -n+1, \dots, 0$ that arrived in the network at respective times $T_{-n}, T_{-n+1}, \dots, T_0$. The time at which the system empties is exactly $X_0^{(s)}$ and thanks to the recursion equation we can calculate X_0 in this case:

$$X_0 = X_{[-n,0]} = \max_{-n \le k \le 0} (\Delta_{[k,0]} \otimes B_k) + T_k,$$

with notations of section 2.2. Hence the end-to-end delay of customer 0 is (with n customers in the network)

$$\begin{aligned} D_{[-n,0]} &:= X_0^{(s)} - T_0 \\ &= \max_{0 \le k \le n} \left[(\Delta_{[-k,0]} \otimes B_{-k})^{(s)} + T_{-k} - T_0 \right] \end{aligned}$$

Now $D = \lim_{n \to \infty} D_{[-n,0]}$ corresponds to the stationary end-to-end delay of the network if it is finite (see next section).

Remark 1. We always implicitly assume that the v^i_n are non-negative to get a dynamical interpretation of the (max,plus) equations. Nevertheless, the construction of recurrence (2) does not require any assumption on the sign of the v^i_n. We will use the notation $\{\beta^i_n\}$ instead of $\{v^i_n\}$ to make a clear difference if the β^i_n's do not have to be non-negative.

2.4 General Framework and Stochastic Assumptions

In what follows, we will consider:

- a sequence of arrival times $N = \{T_n\}_{n\in\mathbb{Z}}$ that is a renewal process: interarrival times $\{\tau_n = T_{n+1} - T_n\}$ are i.i.d. We assume moreover that $\mathbb{E}[\tau_0] = 1$ and $T_0 = 0$ (under Palm probability).
- sequences $\{\beta_n^i\}_{n\in\mathbb{Z}}, i \in \{1, \ldots, m\}$ that are constructed as follows

$$\beta_n^i = S^i(T_{n+1}) - S^i(T_n), \quad \text{with} \quad S^i(t) = \rho_i t + \sigma_i Z^i(t),$$

  where Z^i is a FBM with Hurst parameter $1/2 \leq H_i < 1$. The FBM Z^i are independent of each other and $\max\{H_i\} > 1/2$.

Remark 2. The condition on the mean of τ_0 is not restrictive, we can take any renewal process with positive intensity. Moreover, we see that our virtual service times β_n^i are not non-negative but each sequence is self-similar and long range dependent if $H_i > 1/2$.

From now on, we consider networks with the associated recursion:

$$X_n = A_n \otimes X_{n-1} \oplus B_n \otimes T_n,$$

of dimension s. This means that the matrices $\{A_n, B_n\}$ and vectors that are used in the recursion are obtained via two applications f and g such that:

$$\begin{array}{rccc} f: & \mathbb{R}^m & \to & \mathbb{M}_{(s,s)}(\mathbb{R} \cup \{\epsilon\}) \\ & \beta = (\beta^1, \ldots, \beta^m) & \mapsto & A(\beta), \end{array}$$

$$\begin{array}{rccc} g: & \mathbb{R}^m & \to & \mathbb{M}_{(s,1)}(\mathbb{R} \cup \{\epsilon\}) \\ & \beta = (\beta^1, \ldots, \beta^m) & \mapsto & B(\beta), \end{array}$$

via the formula:

$$f(\beta_n) = A_n, \quad g(\beta_n) = B_n, \quad \text{with } \beta_n = (\beta_n^1, \ldots, \beta_n^m).$$

Stability of the system:
Each sequence $\{\beta_n^i\}_{n\in\mathbb{Z}}$ is stationary and ergodic (see [11] Theorem 14.2.1), hence we have:

$$\lim_{k\to\infty} \frac{(\Delta_{[-k,0]})^{(s,i)}}{k} = \lim_{k\to\infty} \frac{(\Delta_{[-k,0]} \otimes B_{-k})^{(s)}}{k} = \gamma, \tag{11}$$

where γ is the top Lyapounov exponent of the sequence $\{A_n\}$ (see [3]). We will always assume that

$$\gamma < 1. \tag{12}$$

Under this condition we know that $D < \infty$ since $(\Delta_{[-k,0]} \otimes B_{-k})^{(s)} + T_{-k} \to -\infty$ a.s.

Lemma 1. *Consider the matrix $P = A(\rho)$, with $\rho = (\rho_1, \ldots, \rho_m)$. We denote by $\overline{\rho}$ the maximal (max, plus)-eigenvalue of P. We have:*

$$\max(\rho_i) \leq \overline{\rho} \leq \gamma.$$

3 Logarithmic Tail Asymptotic for Tree Networks

3.1 Main result

Theorem 1 (Main Result). *Let D be the stationary end-to-end delay associated to a tree network. Consider the set of queues with maximal Hurst parameter denoted H,*

$$\mathcal{S} := \arg\max\{H_i\},$$

now define the subset of dominant queues as follows

$$\mathcal{D} := \arg\min_{i\in\mathcal{S}} \left\{ \frac{(1-\rho_i)^H}{\sigma_i} \right\}.$$

We denote by W the workload of a single server queue with the same parameter as one of the queues in $\mathcal{D}$, then we have

$$\log \mathbb{P}(D > x) \sim \log \mathbb{P}(W > x). \tag{13}$$

Note that the tandem queueing network is a special case of tree networks. The result of Theorem 1 thus holds for tandem queueing network as well.

3.2 First result with deterministic arrival times

In fact, applications f and g of section 2.4 can be viewed as purely algebraic objects. Following §2.3 of [3], we can associate to each application f and g a directed graph, respectively $\mathcal{G}_A$ and $\mathcal{G}_B$. For f, the set of nodes is $\{1,\dots,s\}$ and an arc from i to j is introduced in $\mathcal{G}_A$ if $A(0)^{(j,i)} \neq \epsilon$. For g, the set of nodes is $\{0,1,\dots,s\}$ and an arc form 0 to i is introduced in $\mathcal{G}_B$ if $B(0)^{(i)} \neq \epsilon$. We denote $\mathcal{G} = \mathcal{G}_A \cup \mathcal{G}_B$. Each coefficient of A and B is a (max,plus)-expression $expr := \bigoplus_{j=1}^{d} \bigotimes_{k\in\mathcal{K}_j} \beta^k$, and we put in $\mathcal{G}$, d copies of the original arc and give to each of them a weight that is the associated set $\mathcal{K}_j$. We obtain a weighted graph $\mathcal{G}_w$. For each arc $e \in \mathcal{G}_w$, $\mathcal{W}(e)$ denotes the weight of e (i.e. a set of indices). We give here the graph corresponding to queues in tandem with window control (section 2.3.3), line style corresponds to the mark: dashed=$\{1\}$, dotted=$\{2\}$, solid=$\{1,2\}$ and dash-dot=$\emptyset$ (observe that in this case $s = L+1$).

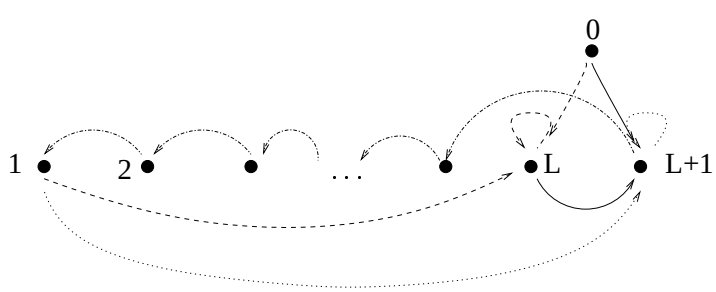

Figure 5: Graph $\mathcal{G}_w$ for Tandem Queueing Network with Fixed Window Control

We denote by Ξ the set of paths in $\mathcal{G}_w$ going from node 0 to node s. For $\xi = (e_0, e_1, \ldots, e_l) \in \Xi$, we denote:

$$\begin{aligned} |\xi|_l &= l+1, \quad \rho(\xi) = \sum_{i=0}^{l} \sum_{j \in \mathcal{W}(e_i)} \rho_j, \\ \mathcal{Z}(\xi) &= \sum_{i=0}^{l} \sum_{j \in \mathcal{W}(e_i)} \sigma_j (Z^j(i+1) - Z^j(i)), \\ |\xi|_w &= \mathbb{E}\left[\mathcal{Z}(\xi)^2\right]. \end{aligned}$$

In the special case $T_n = n$, the end-to-end delay can be expressed as:

$$\begin{aligned} D &= \max_{k \geq 0} \left[(\Delta_{[-k,0]} \otimes B_{-k})^{(s)} - k \right] \\ &\stackrel{\mathrm{d}}{=} \sup_{\xi \in \Xi} \left[\rho(\xi) + \mathcal{Z}(\xi) - (|\xi|_l - 1) \right]. \end{aligned}$$

First we rewrite the event $\{D > x\}$:

$$\begin{aligned} \{D > x\} &= \{\exists \xi \in \Xi,\ \rho(\xi) + \mathcal{Z}(\xi) - |\xi|_l + 1 > x\} \\ &= \{\sup_{\xi \in \Xi} \frac{\mathcal{Z}(\xi)}{x - 1 + |\xi|_l - \rho(\xi)} > 1\}. \end{aligned} \tag{14}$$

To consider the event $\{D > x\}$ or $\{D > x+1\}$ does not change the asymptotic. For the simplicity of notations, we consider the latter in what follows.
Based on (14), to study the tail asymptotic for D, it suffices to focus on the supremum of the centered Gaussian process $\{\frac{\mathcal{Z}(\xi)}{x + |\xi|_l - \rho(\xi)}\}_{\xi \in \Xi}$.
We claim the following logarithmic tail asymptotic for D.

Property 1. *Consider D the stationary end-to-end delay of a tree queueing network. We assume deterministic arrival times, $T_n = n$. Then as $x \to \infty$, we have*

$$\log \mathbb{P}(D > x) \quad \sim \quad -\inf_{\xi \in \Xi} \frac{(x + |\xi|_l - \rho(\xi))^2}{2|\xi|_w}. \tag{15}$$

To prove the above main result, we shall need the so-called "Borell's inequality"[2, p.43,p.47] for the supremum of a Gaussian process which we recall below.

3.3 Borell's Inequality

In what follows, we shall always assume that T has a countable dense subset and the processes we consider are always separable. We recall that a real stochastic process $\{X_t\}_{t \in T}$ is separable if there is a sequence $\{t_j\}$ of parameter values and a set Λ of probability 0 such that, if A is any closed interval and I is any open interval, the sets

$$\{X_t(\omega) \in A, t \in I \cap T\}, \quad \{X_{t_j}(\omega) \in A, t_j \in I \cap T\},$$

differ by at most a subset of Λ.
The following property can be found in [2], theorem 2.1:

Property 2. *Let $\{X_t\}_{t\in T}$ be a centered Gaussian process with sample paths bounded a.s. Let $\|X\| = \sup_{t\in T} X_t$. Then $\mathbb{E}\|X\| < \infty$, and for all $\lambda > 0$*

$$\mathbb{P}\{|\|X\| - \mathbb{E}\|X\|| > \lambda\} \leq 2\exp\left(-\frac{1}{2}\lambda^2/\sigma_T^2\right), \tag{16}$$

where $\sigma_T^2 := \sup_{t\in T} \mathbb{E}X_t^2$. In particular, for all $\lambda > \mathbb{E}\|X\|$, equation (16) may be rewritten as follows:

$$\mathbb{P}\{\|X\| > \lambda\} \leq 2\exp\left(-\frac{(\lambda - \mathbb{E}\|X\|)^2}{2\sigma_T^2}\right). \tag{17}$$

The only assumption made on the parameter space T is that T is totally bounded in the canonical metric. We recall that the canonical metric is defined as follows

$$d(s,t) := \sqrt{\mathbb{E}(X_s - X_t)^2}. \tag{18}$$

We denote by $N(\epsilon)$ the smallest number of closed d-balls of radius ϵ that cover T. T is totally bounded if the function $N(\epsilon)$ is finite for all $\epsilon > 0$.
In fact, following proof of theorem 2.1 in [2], we see that this assumption may be relaxed.
Consider a centered Gaussian process with sample paths bounded a.s. $\{X_t\}_{t\in T}$. Let $\{T_n\}_{n\geq 1}$ be an increasing sequence of subsets of T that tends to a dense subset of T containing the sequence $\{t_j\}$ of points satisfying the conditions of the separability definition. We suppose that each T_n is totally bounded in the canonical metric, and we denote $\|X\|_n = \sup_{t\in T_n} X_t$. Then for any n, thanks to property 2, we have $\mathbb{E}\|X\|_n < \infty$, and for all $\lambda > 0$

$$\mathbb{P}\{|\|X\|_n - \mathbb{E}\|X\|_n| > \lambda\} \leq 2\exp\left(-\frac{1}{2}\lambda^2/\sigma_n^2\right), \tag{19}$$

where $\sigma_n^2 := \sup_{t\in T_n} \mathbb{E}X_t^2$. Moreover, we have $\sigma_n^2 \uparrow \sigma_T^2$. We consider the case $\sigma_T^2 < \infty$ and first show that $\mathbb{E}\|X\| < \infty$ like in [2].
Suppose $\mathbb{E}\|X\| = \infty$ and choose $\lambda_0 > 0$ such that

$$e^{-\lambda_0^2/2\sigma_T^2} \leq 1/4, \quad \mathbb{P}\left[\sup_{t\in T} X_t < \lambda_0\right] \geq 3/4.$$

Such a constant exists since σ_T is finite and the random variable $\sup_{t\in T} X_t$ is finite a.s.
Now since $\mathbb{E}\|X\|_n \uparrow \mathbb{E}\|X\| = \infty$, we can find n such that $\mathbb{E}\|X\|_n > 2\lambda_0$. Borell's inequality on the space T_n then gives

$$\begin{aligned}
\frac{1}{2} \geq 2e^{-\lambda_0^2/2\sigma_T^2} &\geq 2e^{-\lambda_0^2/2\sigma_n^2} \\
&\geq \mathbb{P}[|\|X\|_n - \mathbb{E}\|X\|_n| > \lambda_0] \\
&\geq \mathbb{P}[\mathbb{E}\|X\|_n - \|X\| > \lambda_0] \\
&\geq \mathbb{P}[\lambda_0 > \|X\|] \geq 3/4.
\end{aligned}$$

This proved the required contradiction, and so $\mathbb{E}\|X\| < \infty$. Since $\|X\|_n \uparrow \|X\|$ a.s.(separability condition), we have for all $\lambda > 0$

$$\mathbb{P}\{|\|X\| - \mathbb{E}\|X\|| > \lambda\} \leq 2\exp\left(-\frac{1}{2}\lambda^2/\sigma_T^2\right). \tag{20}$$

Application 1. Consider the process $\{G_t = \frac{Z(t)}{1+t}\}_{t\in[0,\infty)}$. Since $\lim_{t\to\infty} Z(t)/t = 0$, this process is a.s. bounded. Here we take $T_n = [0, n]$, and $T = [0, \infty)$. Each T_n is totally bounded (see [19]) and $\sigma_T^2 = \sup_{t\geq 0} \mathbb{E}G_t^2 = H^{2H}(1-H)^{2(1-H)}$. Hence Borell's inequality applies for this process on the whole interval $[0, \infty)$.

Application 2. If T is countable, then Borell's inequality applies. Just take T_n finite and hence totally bounded !

3.4 Auxiliary Results

In this section, we derive some necessary auxiliary results before we prove the main results as claimed in Property 1. Define

$$\left\{ X_\xi^x = \frac{\mathcal{Z}(\xi)}{x + |\xi|_l - \rho(\xi)} \right\}_{\xi\in\Xi},$$

and

$$m_x := \mathbb{E}\left[\sup_{\xi\in\Xi} X_\xi^x\right], \quad \text{and} \quad \sigma_x^2 = \sup_{\xi\in\Xi} \mathbb{E}\left(X_\xi^x\right)^2. \tag{21}$$

The process $\{X_\xi^x\}$ is a centered Gaussian process. The stability condition (12) $\gamma < 1$ ensures that $D < \infty$ almost surely, from which the boundedness of the sample path of process $\{X_\xi^x\}$ follows.
In our context, the parameter set Ξ is countable as the countable union of the finite sets: $\Xi_n = \{\xi \in \Xi, |\xi|_l = n\}$. Hence Borell's inequality applies (see Application 2 in previous section) and if $m_x \leq 1$, we will have

$$\mathbb{P}\left(\sup_{\xi\in\Xi} X_\xi^x > 1\right) \leq 2\exp\left(-\frac{(1-m_x)^2}{2\sigma_x^2}\right). \tag{22}$$

Lemma 2. *In the case of tree networks, we have* $\lim_{x\to\infty} m_x = 0$.

Proof:
For any path $\xi = (e_0, \ldots, e_l)$, we write $t^0 = 0$ and $t^k = t^{k-1} + \sum_{i=0}^{l} \mathbf{1}_{\{k\in\mathcal{W}(e_i)\}}$, then we have

$$\begin{aligned}
\frac{|\mathcal{Z}(\xi)|}{x + |\xi|_l - \rho(\xi)} &= \frac{\left|\sum_{k=1}^m \sigma_k (Z^k(t^k) - Z^k(t^{k-1}))\right|}{x + |\xi|_l - \rho(\xi)} \\
&\leq \frac{\sum_k \sigma_k |Z^k(t^k) - Z^k(t^{k-1})|}{x + (1-\overline{\rho})|\xi|_l} \\
&\leq \sum_k \frac{\sigma_k |Z^k(t^k) - Z^k(t^{k-1}))|}{x + (1-\overline{\rho})(t^k - t^{k-1})},
\end{aligned}$$

where first inequality follows from $\forall\xi,\ \rho(\xi) \leq \overline{\rho}|\xi|_l$. Hence, we have

$$\begin{aligned}
\left|\mathbb{E}\left[\sup_{\xi\in\Xi} X_\xi^x\right]\right| &\leq \mathbb{E}\left[\sup_{\xi\in\Xi} |X_\xi^x|\right] \\
&\leq \mathbb{E}\left[\sup_{t^k} \sum_k \frac{\sigma_k |Z^k(t^k) - Z^k(t^{k-1}))|}{x + (1-\overline{\rho})(t^k - t^{k-1})}\right] \\
&\leq \sum_k \mathbb{E}\left[\sup_u \frac{\sigma_k |Z^k(u)|}{x + (1-\overline{\rho})u}\right]
\end{aligned}$$

But we know that $\lim_{u\to\infty} Z^k(u)/u = 0$, hence

$$\sup_u \frac{\sigma_k |Z^k(u)|}{x + (1-\overline{\rho})u} \to 0 \quad \text{as } x \to \infty,$$

and each term of the sum goes to zero as $x \to \infty$ by monotone convergence.
□
If X and Y are centered Gaussian random variables with respective variances σ_X^2 and σ_Y^2, we will write $X \leq_{\text{var}} Y \Leftrightarrow \sigma_X \leq \sigma_Y$.

Lemma 3. *We have*

$$X_\xi^x \leq_{var} \frac{\sum_{i=1}^m \sigma_i Z^i(|\xi|_l)}{x + |\xi|_l(1-\gamma)}.$$

Proof:
We first prove that

$$\mathcal{Z}(\xi) \leq_{\text{var}} \sum_{i=1}^m \sigma_i Z^i(|\xi|_l). \tag{23}$$

Take $t_1 < t_2 < t_3 < t_4$, we use the notation: $\Delta_1 = t_2 - t_1$, $\Delta_2 = t_3 - t_2$, $\Delta_3 = t_4 - t_3$, $\Delta = t_4 - t_1$ and $Z(\Delta_1) = Z(t_2) - Z(t_1), \ldots$ We have $Z(\Delta_3) + Z(\Delta_1) \leq_{\text{var}} Z(\Delta)$. This follows from the following inequalities with $1/2 \leq H < 1$ (recall that $a^{2H} + b^{2H} \leq (a+b)^{2H}$):

$$\begin{aligned}
&\mathbb{E}(Z(\Delta_3) + Z(\Delta_1))^2 \\
&\quad = \mathbb{E}Z(\Delta_3)^2 + \mathbb{E}Z(\Delta_1)^2 + 2\mathbb{E}Z(\Delta_3)Z(\Delta_1) \\
&\quad = \Delta_3^{2H} + \Delta_1^{2H} + \Delta^{2H} - (\Delta_1 + \Delta_2)^{2H} + \Delta_2^{2H} - (\Delta_2 + \Delta_3)^{2H} \\
&\quad \leq \Delta^{2H} + \Delta_1^{2H} + \Delta_2^{2H} + \Delta_3^{2H} - \Delta_1^{2H} - \Delta_2^{2H} - \Delta_2^{2H} - \Delta_3^{2H} \\
&\quad \leq \Delta^{2H} = \mathbb{E}Z(\Delta)^2.
\end{aligned}$$

We have then

$$\begin{aligned}
\sum_{i=0}^{|\xi|_l - 1} \mathbf{1}_{j \in \mathcal{W}(e_i)} (Z^j(i+1) - Z^j(i)) \quad &\leq_{\text{var}} \quad Z^j(|\xi|_l), \\
\text{hence,} \quad \mathcal{Z}(\xi) \quad &\leq_{\text{var}} \quad \sum_{i=1}^m \sigma_i Z^i(|\xi|_l).
\end{aligned}$$

By definition, we have $\rho(\xi) \leq |\xi|_l \overline{\rho} \leq |\xi|_l \gamma$ for any $\xi \in \Xi$, and we get $x + |\xi|_l - \rho(\xi) \geq x + |\xi|_l(1-\gamma)$. Now thanks to (23), we have

$$\frac{\mathcal{Z}(\xi)}{x + |\xi|_l - \rho(\xi)} \leq_{\text{var}} \frac{\sum_{i=1}^m \sigma_i Z^i(|\xi|_l)}{x + |\xi|_l(1-\gamma)}$$

□
From lemma 3, we derive

Lemma 4. *We have* $\sigma_x^2 := \sup_{\xi \in \Xi} \mathbb{E}(X_\xi^x)^2 \to 0$ *as* $x \to \infty$. *If we denote* $H := \max\{H_j\}$, *we have* $\sigma_x^2 = \mathcal{O}(x^{2(H-1)})$.

Proof :
Consider the process $C_t^x = \frac{\sum_{i=1}^m \sigma_i Z^i(t)}{x+t(1-\gamma)}$, by a change of variable, we have

$$C^x_{xt/(1-\gamma)} \quad = \quad \frac{\sum_{i=1}^m \sigma_i Z^i(xt/(1-\gamma))}{x+xt},$$

and the self-similarity of the FBM $Z^i(t)$ ensures that the process $\{C^x_{xt/(1-\gamma)}\}$ has the same distribution as the process

$$\frac{\sum_{i=1}^m \sigma_i (x/(1-\gamma))^{H_i} Z^i(t)}{x+xt} \quad = \quad \sum_{i=1}^m \frac{\sigma_i x^{H_i-1}}{(1-\gamma)^{H_i}} G_t^i,$$

with $G_t^i = \frac{Z^i(t)}{1+t}$. Thanks to previous lemma, we have $\mathbb{E}(X_\xi^x)^2 \leq \mathbb{E}(C^x_{|\xi|_l})^2$, hence

$$\sup_{\xi\in\Xi} \mathbb{E}(X_\xi^x)^2 \leq \sup_{t>0} \mathbb{E}(C_t^x)^2,$$

but we have

$$\sup_{t>0} \mathbb{E}(C_t^x)^2 = \sum_{i=1}^m \frac{(\sigma_i)^2 x^{2(H_i-1)}}{(1-\gamma)^{2H_i}} \sup_{t>0} \mathbb{E}(G_t^i)^2.$$

A simple calculation gives:

$$\sup_{t>0} \mathbb{E}(G_t^i)^2 = (H_i)^{2H_i}(1-H_i)^{2(1-H_i)},$$

then

$$\sigma_x^2 \leq \sum_{i=1}^m \frac{(\sigma_i)^2 x^{2(H_i-1)}}{(1-\gamma)^{2H_i}} (H_i)^{2H_i}(1-H_i)^{2(1-H_i)}.$$

Since each β_n^j is on the diagonal of the matrix A_n, we know that for any $l \geq 1$, there exists a path in Ξ such that

$$|\xi_0|_l = l, \quad \rho(\xi_0) \geq \rho_j l, \quad |\xi_0|_w \geq (\sigma_j)^2 l^{2H_j}.$$

Hence we have

$$\mathbb{E}(X_{\xi_0}^x)^2 \geq \frac{(\sigma_j)^2 l^{2H_j}}{(x+l(1-\rho_j))^2}.$$

Taking an index j, such that $H_j = H$, we have

$$\begin{aligned}
\sigma_x^2 \quad &\geq \quad \sup_{l\geq 1} \frac{(\sigma_j)^2 l^{2H}}{(x+l(1-\rho_j))^2} \\
&= \quad (\sigma_j)^2 x^{2(H-1)} \sup_{l\geq 1} \frac{(l/x)^{2H}}{(1+l/x(1-\rho_j))^2} \\
&\sim \quad (\sigma_j)^2 x^{2(H-1)} \sup_{t>0} \frac{t^{2H}}{(1+t(1-\rho_j))^2}.
\end{aligned}$$

This gives the last result. □

3.5 Proof of Property 1

First observe that the right-hand term in (15) goes to $-\infty$ when $x \to \infty$ since

$$-\inf_{\xi\in\Xi} \frac{(x+|\xi|_l-\rho(\xi))^2}{2|\xi|_w} = -\frac{1}{2\sigma_x^2},$$

and thanks to lemma 4, we know that $\sigma_x^2 \to 0$. We have to prove that $\lim_{x\to\infty} 2\sigma_x^2 \log \mathbb{P}(D>x) = -1$.
Upper bound:
Taking the logarithm of equation (22), we obtain

$$2\sigma_x^2 \log \mathbb{P}(D>x) \le 2\sigma_x^2 \log(2) - (1-m_x)^2 .$$

Thanks to lemmas 4 and 2, we have

$$\limsup_{x\to\infty} 2\sigma_x^2 \log \mathbb{P}(D>x) \le -1. \tag{24}$$

Lower bound:
We denote:

$$\overline{\Phi}(y) = \frac{1}{\sqrt{2\pi}} \int_y^\infty e^{-x^2/2} dx.$$

We have:

$$\begin{aligned}
\mathbb{P}(D>x) &= \mathbb{P}\left(\sup_{\xi\in\Xi} \rho(\xi) + \mathcal{Z}(\xi) - |\xi|_l > x\right) \\
&\ge \sup_{\xi\in\Xi} \mathbb{P}\left(\mathcal{Z}(\xi) > x + |\xi|_l - \rho(\xi)\right) \\
&= \sup_{\xi\in\Xi} \overline{\Phi}\left(\frac{x+|\xi|_l-\rho(\xi)}{\sqrt{\mathbb{E}[\mathcal{Z}(\xi)^2]}}\right) \\
&= \overline{\Phi}\left(\inf_{\xi\in\Xi} \frac{x+|\xi|_l-\rho(\xi)}{\sqrt{|\xi|_w}}\right).
\end{aligned}$$

Using the approximation $\log \overline{\Phi}(y) \sim -y^2/2$, we obtain

$$\log \mathbb{P}(D>x) \ge -\inf_{\xi\in\Xi} \frac{(x+|\xi|_l-\rho(\xi))^2}{2|\xi|_w},$$

hence

$$\liminf_{x\to\infty} 2\sigma_x^2 \log \mathbb{P}(D>x) \ge -1. \tag{25}$$

Equations (24) and (25) give the desired asymptotic for deterministic arrival times.
□

3.6 Computation of σ_x^2

3.6.1 The case of Single Server Queue

Equation (15) takes the simple from:

$$\begin{aligned}\inf_{\xi\in\Xi}\frac{(x+|\xi|_l-\rho(\xi))^2}{2|\xi|_w} &= \inf_{n\geq 1}\frac{(x+n(1-\rho))^2}{2\sigma^2 n^{2H}}\\ &= \frac{1}{2\sigma^2}x^{2(1-H)}\inf_{n\geq 1}\frac{(1+n/x(1-\rho))^2}{(n/x)^{2H}}\\ &\sim \frac{1}{2\sigma^2}x^{2(1-H)}\inf_{t>0}\frac{(1+t(1-\rho))^2}{t^{2H}}.\end{aligned}$$

The infimum is attained in $t^*_{single}=\frac{H}{(1-\rho)(1-H)}$ and we have

$$\log\mathbb{P}(W>x) \sim -\frac{1}{2\sigma^2}x^{2(1-H)}\frac{(1-\rho)^{2H}(1-H)^{2(H-1)}}{H^{2H}}. \tag{26}$$

3.6.2 The case of 2 Queues in Tandem

For $\xi=(e_0,\ldots,e_l)\in\Xi$, we define:

$$m=\sum_{i=0}^{l}\mathbf{1}_{\{1\in\mathcal{W}(e_i)\}},\quad n=\sum_{i=0}^{l}\mathbf{1}_{\{2\in\mathcal{W}(e_i)\}}.$$

Then we have

$$\mathbb{E}(X_\xi^x)^2 = \frac{(\sigma_1)^2m^{2H_1}+(\sigma_2)^2n^{2H_2}}{(x+m+n-\rho_1 m-\rho_2 n)^2}. \tag{27}$$

We first suppose that $H_1>H_2$, then we have

$$\begin{aligned}\sigma_x^2 &= x^{2(H_1-1)}\sup_{m,n}\frac{(\sigma_1)^2(\frac{m}{x})^{2H_1}+(\sigma_2)^2(\frac{n}{x})^{2H_2}x^{2(H_2-H_1)}}{(1+\frac{m+n}{x}-\rho_1\cdot\frac{m}{x}-\rho_2\cdot\frac{n}{x})^2}\\ &\sim x^{2(H_1-1)}\sup_{t>0,u\in[0,1]}\frac{(\sigma_1)^2(ut)^{2H_1}}{(1+(1-\rho(u))t)^2}\end{aligned}$$

with $\rho(u)=\rho_1 u+\rho_2(1-u)$. The supremum is attained in $u=1$ and $t^*=\frac{H_1}{(1-\rho_1)(1-H_1)}$, and we obtain

$$\sigma_x^2 \sim x^{2(H-1)}\frac{\sigma^2H^{2H}}{(1-\rho)^{2H}(1-H)^{2(H-1)}},$$

with $H:=H_1$, $\rho:=\rho_1$ and $\sigma:=\sigma_1$.
The case $H_2>H_1$ is exactly the same. Hence $H=H_2,\rho=\rho_2$ and $\sigma=\sigma_2$.

We suppose now that $H_1 = H_2 = H$, then we have

$$\begin{aligned}
\sup_{m,n} \frac{(\sigma_1)^2 m^{2H} + (\sigma_2)^2 n^{2H}}{(x+m+n-\rho_1 m - \rho_2 n)^2} &= x^{2(H-1)} \sup_{m,n} \frac{(\sigma_1)^2 (\frac{m}{x})^{2H} + (\sigma_2)^2 (\frac{n}{x})^{2H}}{(1 + \frac{m+n}{x} - \rho_1 \cdot \frac{m}{x} - \rho_2 \cdot \frac{n}{x})^2} \\
&\sim x^{2(H-1)} \sup_{t>0, u\in[0,1]} \frac{(\sigma_1)^2 (ut)^{2H} + (\sigma_2)^2 ((1-u)t)^{2H}}{(1+(1-\rho(u))t)^2} \\
&= x^{2(H-1)} \sup_{u\in[0,1]} \Big\{ \big[(\sigma_1)^2 u^{2H} + (\sigma_2)^2 (1-u)^{2H}\big] \cdot \\
&\qquad \cdot \sup_{t>0} \frac{t^{2H}}{(1+(1-\rho(u))t)^2} \Big\} \\
&= x^{2(H-1)} \sup_{u\in[0,1]} \Big\{ \big[(\sigma_1)^2 u^{2H} + (\sigma_2)^2 (1-u)^{2H}\big] \cdot \\
&\qquad \cdot \frac{H^{2H}}{(1-\rho(u))^{2H}(1-H)^{2(H-1)}} \Big\}
\end{aligned}$$

But the function $u \mapsto \frac{(\sigma_1)^2 u^{2H} + (\sigma_2)^2 (1-u)^{2H}}{(1-\rho(u))^{2H}}$ is either monotone on $[0,1]$, or non-increasing on $[0,u^*]$ and non-decreasing on $[u^*,1]$ for a certain u^*. Thus the supremum is attained either at 0 or at 1, and we have

$$\sigma_x^2 \sim x^{2(H-1)} \frac{\sigma^2 H^{2H}}{(1-\rho)^{2H}(1-H)^{2(H-1)}}$$

with

$$\frac{(1-\rho)^{2H}}{\sigma^2} = \min\left\{ \frac{(1-\rho_1)^{2H}}{(\sigma_1)^2}, \frac{(1-\rho_2)^{2H}}{(\sigma_2)^2} \right\}.$$

This gives the desired result for 2 queues in tandem.

3.6.3 General Tree Networks

First observe that previous result holds true for k queues in tandem. We have to find the supremum of the function $(x_1, x_2, \ldots, x_k) \mapsto \frac{(\sigma_1)^2 x_1^{2H} + \cdots + (\sigma_k)^2 x_k^{2H}}{(1-\rho(x_1,\ldots,x_k))^{2H}}$ under the constraints $x_i \geq 0$ and $\sum_i x_i = 1$. Suppose the supremum is attained in $(x_1^*, \ldots, x_k^*)$ with $x_i^* x_j^* > 0$ for some $i < j$. Then the study of the function

$$u \mapsto \frac{(\sigma_1)^2 (x_1^*)^{2H} + \cdots + (\sigma_i)^2 (x_i^* + u)^{2H} + \cdots + (\sigma_j)^2 (x_j^* - u)^{2H} + \cdots + (\sigma_k)^2 (x_k^*)^{2H}}{(1-\rho(x_1^*, \ldots, x_i^* + u, \ldots, x_j^* - u, \ldots, x_k^*))^{2H}}$$

is similar to previous section. Hence we see that the supremum is attained in a certain $(0, \ldots, 1, \ldots, 0)$. Then we have

$$\sigma_x^2(k) \sim x^{2(H-1)} \max_{j\in\mathcal{D}^k} \frac{(\sigma_j)^2 H^{2H}}{(1-\rho_j)^{2H}(1-H)^{2(H-1)}},$$

where H is the maximal Hurst parameter and $\mathcal{D}^k$ is defined as in Theorem 1 for the k queues.
But for a general tree network, the variance σ_x^2 is the maximum of the variance corresponding to a

path going from the root of the tree to any leaf, i.e. a network of queues in tandem, hence we have directly

$$\sigma_x^2 \sim x^{2(H-1)} \max_{k\in\mathcal{D}} \frac{(\sigma_k)^2 H^{2H}}{(1-\rho_k)^{2H}(1-H)^{2(H-1)}}.$$

Remark 3. We recall that we always assume that $H > 1/2$. Nevertheless, for now on, we never use this assumption and in fact for deterministic arrival times, Theorem 1 is still true with $H = 1/2$. This corresponds to Brownian queues and in the case of tandem queues the large deviation technics used in [17] apply and it is straightforward that Theorem 1 of [17] gives exactly the same result as our Theorem 1 for deterministic arrival times.

3.7 From deterministic times to random arrival times

In previous section, we proved by computing σ_x^2 that Theorem 1 holds true in the case of deterministic arrival times. We prove now that the result extends to random arrival times.
We denote $\Psi(\epsilon) = \sup_n(T_{-n} + n(1-\epsilon))$. There exist C and λ such that for sufficiently large x, we have

$$\mathbb{P}[\Psi(\epsilon) \geq x] \leq Ce^{-\lambda x}.$$

We have

$$\begin{aligned} D &= \sup_{\xi} \left\{\rho(\xi) + \mathcal{Z}(\xi) + T_{-|\xi|_l}\right\} \\ &\leq \sup_{\xi} \left\{\rho(\xi) + \mathcal{Z}(\xi) - |\xi|_l(1-\epsilon)\right\} + \sup_{n} \left\{T_{-n} + n(1-\epsilon)\right\} \\ &:= D^{(1-\epsilon)} + \Psi(\epsilon). \end{aligned}$$

Therefore, for any $\alpha > 0$,

$$\begin{aligned} \mathbb{P}[D > x] &\leq \mathbb{P}[D^{(1-\epsilon)} + \Psi(\epsilon) > x] \\ &= \mathbb{P}[D^{(1-\epsilon)} + \Psi(\epsilon) > x, \psi(\epsilon) < \alpha x] + \mathbb{P}[D^{(1-\epsilon)} + \Psi(\epsilon) > x, \psi(\epsilon) \geq \alpha x] \\ &\leq \mathbb{P}[\psi(\epsilon) < \alpha x]\mathbb{P}[D^{(1-\epsilon)} > (1-\alpha)x] + Ce^{-\lambda\alpha x}. \end{aligned}$$

Hence for sufficiently small ϵ, we have

$$\begin{aligned} \log \mathbb{P}[D > x] &\leq \log\left\{\mathbb{P}[\psi(\epsilon) < \alpha x]\mathbb{P}[D^{(1-\epsilon)} > (1-\alpha)x] + Ce^{-\lambda\alpha x}\right\} \\ &\sim -[(1-\alpha)x]^{2(1-H)} \frac{((1-\epsilon)-\rho)^{2H}(1-H)^{2(H-1)}}{2\sigma^2 H^{2H}}, \end{aligned}$$

because we have $H > 1/2$. Letting α go to 0, we get

$$\log \mathbb{P}[D > x] \leq -x^{2(1-H)} \frac{((1-\epsilon)-\rho)^{2H}(1-H)^{2(H-1)}}{2\sigma^2 H^{2H}}. \tag{28}$$

Now due to the strong law of large number, we can choose $G \equiv G(\epsilon)$, such that

$$\mathbb{P}(T_{-n} \geq -n(1+\epsilon) - G,\ \forall n \geq 0) \geq 1 - \epsilon.$$

We denote this event by K_ϵ. Then, on the event K_ϵ, we have

$$\begin{aligned} D &= \sup_\xi \left\{\rho(\xi) + \mathcal{Z}(\xi) - |\xi|_l(1+\epsilon) + T_{-|\xi|_l} - |\xi|_l(1+\epsilon)\right\} \\ &\geq \sup_\xi \left\{\rho(\xi) + \mathcal{Z}(\xi) - |\xi|_l(1+\epsilon)\right\} - G \\ &:= D^{(1+\epsilon)} - G. \end{aligned}$$

Then we have for sufficiently small ϵ

$$\begin{aligned} \mathbb{P}(D > x) &\geq \mathbb{P}(D > x, K_\epsilon) \\ &\geq \mathbb{P}(D^{(1+\epsilon)} > x + G)(1-\epsilon). \end{aligned}$$

Taking the logarithm of this equation gives

$$\begin{aligned} &\log \mathbb{P}(D > x) \\ &\geq -(x + G(\epsilon))^{2(1-H)} \frac{((1+\epsilon)-\rho)^{2H}(1-H)^{2(H-1)}}{2\sigma^2 H^{2H}} \\ &\sim -x^{2(1-H)} \frac{((1+\epsilon)-\rho)^{2H}(1-H)^{2(H-1)}}{2\sigma^2 H^{2H}}. \end{aligned} \tag{29}$$

Letting ϵ go to 0, we get the same lower bound (29) and upper bound (28). This concludes the proof of theorem 1.
□

4 Tree Networks with Window Control

4.1 Tree Network with Window Control

Theorem 2 (Main Result). *Let D be the stationary end-to-end delay in a tree network with window control. Let L be the size of the window and d the depth of the tree. Define*

$$\begin{aligned} H &:= \max H_i, \\ \mathcal{S} &:= \arg\max\{H_i\}, \\ \mathcal{D} &:= \arg\min_{i\in\mathcal{S}} \left\{\frac{(1-\rho_i)^H}{\sigma_i}\right\}. \end{aligned}$$

Denote by W the workload of a single server queue with the same parameter as one of the queues in $\mathcal{D}$. There exists a constant $0 < C \leq 1$ such that

$$C \leq \underline{\lim} \frac{\log \mathbb{P}(D > x)}{\log \mathbb{P}(W > x)} \leq \overline{\lim} \frac{\log \mathbb{P}(D > x)}{\log \mathbb{P}(W > x)} \leq 1, \tag{30}$$

We have the following inequalities for the constant C:

1. *if $L \geq d$ then:*

$$(1-\gamma)^2 \leq C, \tag{31}$$

where γ is the top Lyapounov exponent associated to the network defined in equation (11);

2. *if $L < d$ then:*

$$(1-\gamma)^2 \frac{(1-\gamma)^{2H}}{(1-\rho_i)^{2H}} \frac{(\sigma_i)^2}{\sum_{j\in\mathcal{S}}(\sigma_j)^2} \leq C, \tag{32}$$

for any $i \in \mathcal{S}$.

Note that such a max-plus formulation can be extended to variable window control mechanisms, such as AIMD (additive-increase-multiplicative-decrease) type window control, see [5]. In particular, the inequalities (30) still hold. Moreover the Lyapounov exponent is monotonic in the window size: given the structure of a network with window control of size L, we can associate to it the Lyapounov exponent $\gamma(L)$ and this function is non-increasing in L. So that inequalities (31) and (32) still hold, with the left hand side being replaced by the $\gamma(L_{min})$ of the minimum window size L_{min}. This implies that, with the TCP model of [5], the asympotics of the end-to-end delay presented in Theorem 2 extends to RTT of TCP sessions.

4.2 Proof of Theorem 2

We use the same notation as in section 3. Lemmas 3 and 4 of section 3.4 are demonstrated in great generality and hold true for tree networks with window control. Moreover if we can prove that $m_x \leq 1$ for sufficiently large x, we have the following lower and upper bound in great generality

$$\begin{aligned}
\liminf_{x\to\infty} 2\sigma_x^2 \log \mathbb{P}(D > x) &\geq -1, \\
\limsup_{x\to\infty} 2\sigma_x^2 \log \mathbb{P}(D > x) &\leq -(1 - \limsup_{x\to\infty} m_x)^2,
\end{aligned}$$

where

$$X_\xi^x = \frac{\mathcal{Z}(\xi)}{x + |\xi|_l - \rho(\xi)}, \quad m_x := \mathbb{E}\left[\sup_{\xi\in\Xi} X_\xi^x\right], \quad \text{and} \quad \sigma_x^2 = \sup_{\xi\in\Xi} \mathbb{E}\left(X_\xi^x\right)^2.$$

We have now to replace lemma 2 by

Lemma 5. *We have* $\limsup_{x\to\infty} m_x \leq \gamma < 1$.

Proof:
The function $x \mapsto \sup_\xi \frac{\mathcal{Z}(\xi)^+}{x+|\xi|_l-\rho(\xi)}$ is non-increasing since

$$\begin{aligned}
x \leq y \quad &\Rightarrow \quad \frac{\mathcal{Z}(\xi)^+}{x + |\xi|_l - \rho(\xi)} \geq \frac{\mathcal{Z}(\xi)^+}{y + |\xi|_l - \rho(\xi)} \\
&\Rightarrow \quad \sup_\xi \frac{\mathcal{Z}(\xi)^+}{x + |\xi|_l - \rho(\xi)} \geq \sup_\xi \frac{\mathcal{Z}(\xi)^+}{y + |\xi|_l - \rho(\xi)}.
\end{aligned}$$

Thanks to Borell's inequality, we have $\mathbb{E}\left[\sup_{\xi\in\Xi} X_\xi^1\right] < +\infty$ and by symmetry, $\mathbb{P}(\sup_\xi |X_\xi^1| > \lambda) \leq 2\mathbb{P}(\sup_\xi X_\xi^1 > \lambda)$, hence we have $\mathbb{E}\left[\sup_{\xi\in\Xi}(X_\xi^1)^+\right] \leq \mathbb{E}\left[\sup_{\xi\in\Xi}|X_\xi^1|\right] < +\infty$. Then we can use monotone convergence to derive

$$\lim_{x\to\infty} \mathbb{E}\left[\sup_{\xi\in\Xi}(X_\xi^x)^+\right] = \mathbb{E}\left[\lim_{x\to\infty}\sup_{\xi\in\Xi}(X_\xi^x)^+\right].$$

Thanks to (11), we know that for any $0 < \epsilon < 1-\gamma$, there exists a finite random variable L, such that

$$|\xi|_l \geq L \quad \Rightarrow \quad \mathcal{Z}(\xi) + \rho(\xi) \leq (\gamma+\epsilon)|\xi|_l.$$

Hence for such a path, we have

$$\begin{aligned} \frac{\mathcal{Z}(\xi)^+}{x+|\xi|_l-\rho(\xi)} &\leq \frac{(\gamma+\epsilon)|\xi|_l-\rho(\xi)}{x+|\xi|_l-\rho(\xi)} \\ &\leq \gamma+\epsilon. \end{aligned}$$

We define the random variable $M = \sup_{|\xi|_l \leq L} \mathcal{Z}(\xi)^+$. We have $M < +\infty$ a.s. and

$$\sup_{\xi\in\Xi}(X_\xi^x)^+ \quad \leq \quad \frac{M}{x} + \gamma + \epsilon.$$

Hence we have $\lim_{x\to\infty} \sup_{\xi\in\Xi}(X_\xi^x)^+ \leq \gamma$, and the result follows since $m_x \leq \mathbb{E}\left[\sup_{\xi\in\Xi}(X_\xi^x)^+\right]$. □

Hence, we have as $x \to \infty$

$$-1 \leq \underline{\lim}\, 2\sigma_x^2 \log \mathbb{P}(D > x) \leq \overline{\lim}\, 2\sigma_x^2 \log \mathbb{P}(D > x) \leq -(1-\gamma)^2. \tag{33}$$

We have now to compute the corresponding variance σ_x^2.

Remark 4. The bound of Lemma 5 is tight in the sense that there are cases for which we have

$$\lim_{x\to\infty} m_x = \gamma.$$

This result shows the limit of our approach since even if we can compute the variance σ_x^2, the technic used here can not give an exact asymptotic for the quantity $\log \mathbb{P}(D > x)$ in these particular cases. We take the example of two queues in tandem with window control of size $L = 1$. We recall the recursion equations with the notation of section 2:

$$\begin{pmatrix} x_n^1 \\ x_n^2 \end{pmatrix} = \begin{pmatrix} v_n^1 & v_n^1 \\ v_n^1 + v_n^2 & v_n^1 + v_n^2 \end{pmatrix} \begin{pmatrix} x_{n-1}^1 \\ x_{n-1}^2 \end{pmatrix} \oplus \begin{pmatrix} v_n^1 \\ v_n^1 + v_n^2 \end{pmatrix} T_n.$$

Take $\rho_1 = \sigma_1 = 0$ (service in station 1 is instantaneous) and $\rho_2 = 0$. We have

$$(\Delta_{[-k,0]} \otimes B_{-k})^{(2)} \quad \stackrel{\mathrm{d}}{=} \quad \sum_{i=0}^{k} \sigma_2 \left((Z^2(i+1) - Z^2(i))\right)^+.$$

Hence we have

$$\gamma = \sigma_2 \mathbb{E}\left[\left(Z^2(1)\right)^+\right] > 0.$$

We have $\mathbb{E}\left[\left(Z^2(1)\right)^+\right] \leq \mathbb{E}\left[1 + (Z^2(1))^2\right] = 2$, hence we can choose $\sigma_2 = 1/3$ and we have $\gamma < 1$. Now we see that for $\xi \in \Xi_n$, we have

$$X_\xi^x \geq 1/3 \frac{\sum_{i=0}^n \left(Z^2(i+1) - Z^2(i)\right)^+}{x+n},$$

hence

$$\begin{aligned}\sup_{\xi} X_{\xi}^{x} &\geq \sup_{n} 1/3 \frac{\sum_{i=0}^{n}\left(Z^2(i+1)-Z^2(i)\right)^+}{x+n} \\ &\geq \lim_{n\to\infty} 1/3 \frac{\sum_{i=0}^{n}\left(Z^2(i+1)-Z^2(i)\right)^+}{x+n} \\ &= \gamma > 0.\end{aligned}$$

In this specific case, thanks to Lemma 5, we have

$$\lim_{x\to\infty} m_x = \gamma,$$

hence we see that inequalities (33) can not be improved and lower and upper bounds will not coincide.

4.3 Computation of σ_x^2

4.3.1 The general case

With notation introduced in Theorem 2, we have

$$\log \mathbb{P}(W > x) \quad \sim \quad -\frac{1}{2\sigma^2} x^{2(1-H)} \frac{(1-\rho)^{2H}(1-H)^{2(H-1)}}{H^{2H}},$$

where $\frac{(1-\rho)^H}{\sigma} = \frac{(1-\rho_j)^H}{\sigma_j}$ for any $j \in \mathcal{S}$.
Thanks to lemma 4, we know that

$$\begin{aligned}\sigma_x^2 &\geq (\sigma_j)^2 x^{2(H-1)} \sup_{t>0} \frac{t^{2H}}{(1+t(1-\rho_j))^2} \\ &= (\sigma_j)^2 x^{2(H-1)} \frac{H^{2H}}{(1-\rho_j)^{2H}(1-H)^{2(H-1)}}\end{aligned}$$

for an index $j \in \mathcal{S}$. Hence we have

$$-\frac{1}{2\sigma_x^2}(\log \mathbb{P}(W > x))^{-1} \leq 1. \tag{34}$$

In proof of Lemma 4, we showed

$$\sigma_x^2 \leq \sum_{i=1}^{m} \frac{(\sigma_i)^2 x^{2(H_i-1)}}{(1-\gamma)^{2H_i}} (H_i)^{2H_i}(1-H_i)^{2(1-H_i)}.$$

Hence, we have

$$\lim_{x\to\infty} -\frac{1}{2\sigma_x^2}(\log \mathbb{P}(W > x))^{-1} \quad \geq \quad \frac{(1-\gamma)^{2H}\sigma^2}{(1-\rho)^{2H}\sum_{j\in\mathcal{S}}(\sigma_j)^2}. \tag{35}$$

Thanks to equations (34) and (35), we have:

$$\frac{(1-\gamma)^{2(H+1)}\sigma^2}{(1-\rho)^{2H}\sum_{j\in\mathcal{S}}(\sigma_j)^2} \leq \underline{\lim} \frac{\log \mathbb{P}(D > x)}{\log \mathbb{P}(W > x)} \leq \overline{\lim} \frac{\log \mathbb{P}(D > x)}{\log \mathbb{P}(W > x)} \leq 1.$$

Remark 5. So far, we never use the specific structure of the network in the proof of Theorem 2. In fact, we proved the following result for the asymptotic tail distribution of Z, the stationary reponse time for an open stochastic event graphs (defined in [4] or [3]):
Define

$$\begin{aligned} H &:= \max H_i, \\ \mathcal{S} &:= \arg\max\{H_i\}, \\ \mathcal{D} &:= \arg\min_{i\in\mathcal{S}} \left\{ \frac{(1-\rho_i)^H}{\sigma_i} \right\}. \end{aligned}$$

Denote by W the workload of a single server queue with the same parameter as one of the queues in $\mathcal{D}$, then we have

$$(1-\gamma)^2 \frac{(1-\gamma)^{2H}}{(1-\rho_i)^{2H}} \frac{(\sigma_i)^2}{\sum_{j\in\mathcal{S}}(\sigma_j)^2} \leq \underline{\lim} \frac{\log \mathbb{P}(Z>x)}{\log \mathbb{P}(W>x)} \leq \overline{\lim} \frac{\log \mathbb{P}(Z>x)}{\log \mathbb{P}(W>x)} \leq 1,$$

where γ is the top Lyapounov exponent associated to the network.

4.3.2 The case of $L \geq d$ in Tandem Queueing Network

We restrict the proof to the case of 2 queues with $L \geq 2$ and show that

$$\sigma_x^2 \sim x^{2(H-1)} \frac{(\sigma_j)^2 H^{2H}}{(1-\rho_j)^{2H}(1-H)^{2(H-1)}},$$

for $j \in \mathcal{D}$. The general case of tandem queueing network can be treated in the same way, though the proof will be lengthier.

It is easy to see from the structure of the graph $\mathcal{G}_w$ (see section 3.2) that for any $\xi \in \Xi$, with the same notation as in section 3.6.2 we have

$$\begin{aligned} \mathcal{Z}(\xi) &\leq_{var} \sigma_1 Z^1(m) + \sigma_2 Z^2(n), \\ \rho(\xi) &= m\rho_1 + n\rho_2, \\ |\xi|_l &\geq m+n. \end{aligned}$$

Hence we have

$$\mathbb{E}(X_\xi^x)^2 \leq \frac{(\sigma_1)^2 m^{2H_1} + (\sigma_2)^2 n^{2H_2}}{(x+m+n-\rho_1 m - \rho_2 n)^2}.$$

The right-hand term corresponds exactly to (27). Hence the upper bound follows.

4.3.3 General Tree Network with Window Control

Instead of considering only queues in tandem with window control, we can consider a general tree network with m queues and with a window size L. In this context, we see that results of section 4.2 remain true since we do not use the specific structure of the tandem case. The computation of the variance σ_x^2 is different, but Theorem 2 is true for tree networks with window control. In section 4.3.2, we derive the variance of the network of queues in tandem with window control $L \geq d$ because we can upper bound it by the variance of queues in tandem without window control. But the same result

holds in the tree case when $L \geq d$. We can upper bound the variance of the tree network with window control by the variance of the same tree network (i.e. corresponding to the same tree topology) without window control. Hence, thanks to the results of section 3.6.3, we see that one queue dominates for the computation of the variance σ_x^2 and we get the same upper and lower bound as in theorem 2 for the tail asymptotic of the end-to-end delay.

The demonstration for the case of renewal arrival times of section 3.7 is the same in the case of tandem queues with window control except we multiply the upper bound by a factor $(1-\gamma)^2$.

5 Concluding Remarks

In this paper, we studied the steady state distribution of the end-to-end delay of a tagged flow in queueing networks where the queues have self-similar cross traffic modeled by fractional Brownian noises, independent of other queues. The arrival process of the tagged flow is renewal. Two types of queueing networks are considered, namely, the tree networks without or with window control.

Our analysis is based on the (max,plus) representation and the general Borell's inequality. We have shown that the end-to-end delay of tree networks is completely dominated by one of the queues. The dominant queue is the one with maximal Hurst parameter (or the maximal ratio $\frac{(1-\rho)^H}{\sigma}$ in case of ties). Thus, the tail distribution of the end to end delay is asymptotically Weibullian. In the case of tandem queues with window control, the end-to-end delay is still asymptotically Weibullian and we give upper and lower bound for the constant that determines the scale parameter of the corresponding Weibull distribution.

This work shows that the long-range dependence of the cross traffic does have strong impact on the end-to-end delay tail asymptotic. However, this impact is only through a single dominant queue that has the strongest dependence structure, i.e. the queue with the highest Hurst parameter. The window control mechanisms à la TCP in such networks play a minimal role in the determination of the tail distribution of the end-to-end delay.

The fact the end-to-end delay has Weibull tail also provides a new means to the inference and prediction of end-to-end delays. Indeed, this provides a parametric model for their tail distribution. It also connects the Hurst parameters at the routers with the tail asymptotics of the end-to-end delays.

As future work, we plan to extend current framework to study other long-range dependent models for the cross traffic. For example, one well-known model is the $M/G/\infty$ model. Other possible extensions include considering tandem or general networks with dynamic/adaptive window control.

References

[1] P. Abry and D. Veitch, Wavelet Analysis of Long-Range Dependence Traffic, *IEEE Transactions on Information Theory*, 44(1), pp. 2-15, January, 1998.

[2] R.J. Adler, *Continuity, Extrema, and Related Topics for General Gaussian Processes*, Institute of Math. Stat., Lecture notes, 1990, Hayward, California.

[3] F. Baccelli, G. Cohen, G.J. Olsder and J.P. Quadrat, *Synchronization and Linearity*, Wiley, 1992.

[4] F. Baccelli, S. Foss, M. Lelarge, *Asymptotics of Subexponential Max Plus Networks the Stochastic Event Graph Case*, to appear in QUESTA, 2003.

[5] F. Baccelli and D. Hong, TCP is Max-Plus Linear and What It Tells Us on Its Throughput, *ACM Sigcomm 2000*, p.219-230.

[6] J. C. Bolot, Characterizing end-to-end packet delay and loss in the internet. *Journal of High Speed Networks*, 2(3):289– 298, September 1993.

[7] R. Caceres, N.G. Duffield, S.B. Moon, D. Towsley, Inferring Link-level Performance from End-to-End Multicast Measurements, *Proc. IEEE/ISOC Global Internet '99*, December 1999.

[8] A. Chaintreau, F. Baccelli and C. Diot, Impact of TCP-like Congestion Control on the Throughput of Multicast Group, *IEEE/ACM Transactions on Networking* vol.10, p.500-512, August 2002.

[9] N.G. Duffield, F. Lo Presti, *Multicast Inference of Packet Delay Variance at Interior Network Links*, *IEEE Infocom 2000*.

[10] A. Chaintreau, C. Diot, F. Baccelli, Impact of TCP-like Congestion control on the Throughput of Multicast Groups. *IEEE/ACM Transaction on Networking*. June/July 2002.

[11] I.P. Cornfeld, S.V. Fomin, Ya.G. Sinai, *Ergodic Theory*, Springer Verlag, Berlin, 1982.

[12] M. E. Crovella and A. Bestavros, Self-similarity in world wide web traffic evidence and possible causes. In *Proceedings of the ACM SIGMETRICS 96*, pages 160–169, Philadelphia, PA, May 1996.

[13] N. G. Duffield and N. O'Connell, Large deviation and overflow probabilities for the general single-server queue, with applications. *Mathematical Proceedings of the Cambridge Philosophical Society*, 118:363-375, 1995.

[14] A. Feldmann, A. C. Gilbert, and W. Willinger, Data networks as cascades: Investigating the multifractal nature of Internet WAN traffic. In *Proc. of the ACM SIGCOMM'98*, pages 25–38, Vancouver, B.C., 1998.

[15] A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, The Changing Nature of Network Traffic: Scaling Phenomena, *ACM Computer Communication Review*, vol. 28, pp. 5-29, Apr. 1998.

[16] A. Feldmann, P. Huang, A. C. Gilbert, and W. Willinger, Dynamics of ip traffic: A study of the role of variability and the impact of control. In *Proceedings of the ACM/SIGCOMM'99*, 1999.

[17] A.J. Ganesh, Large deviations of the sojourn time for queues in series, Annals of Operations Research, 1998.

[18] W. Leland, M. Taqqu, W. Willinger, D. Wilson, On the Self-Similar Nature of Ethernet Traffic, *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, pp. 1-15, February 1994.

[19] L. Massoulié, A. Simonian, *Large Buffer Asymptotics for the Queue with FBM Input*, Journal of Applied Probability, Vol. 36, Num. 3, pp. 894-906, 1999.

[20] A. Mukherjee, On the dynamics and significance of low frequency components of Internet load, *Internetworking: Research and Experience*, vol. 5, no. 4, pp. 163–205, 1994.

[21] I. Norros, *A storage model with self-similar input, Queueing Systems*, 16:387-396, 1994

[22] I. Norros, A. Simonian, D. Veitch, J. Virtamo, *A Benes formula for a buffer with fractional Brownian input, 9th ITC Specialists Seminar'95: Teletraffic Modelling and Measurement*, 1995.

[23] K. Park and W. Willinger (Eds.), *Self-Similar Network Traffic and Performance Evaluation*, John Wiley & Sons, New York (NY), 2000.

[24] V. Paxson, *End-to-End Routing Behavior in the Internet, IEEE/ACM Transactions on Networking*, Oct 1997, vol. 5, pp. 601–615.

[25] V. Paxson and S. Floyd, Wide-area Traffic: The Failure of Poisson Modeling, *IEEE/ACM Transactions on Networking*, pp.226-244, June 1995.

[26] G. Samorodnitsky and M.S. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models With Infinite Variance*, Chapman and Hal, New York (NY), 1994.

[27] F. Toomey, Large Deviations in a Class of Stochastic Discrete Event Systems, PhD thesis, University of Dublin, 1997.

[28] C.H. Xia and Z. Liu, Queueing Systems with Long-range Dependent Input Process and Subexponential Service Times. *Proceedings of the ACM SIGMETRICS Conference*, June, 2003.

[29] Z.-L. Zhang, V. J. Ribeiro, S. B. Moon, and C. Diot, Smalltime scaling behaviors of Internet backbone traffic: an empirical study, *INFOCOM 2003*.

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