Impact of Network Delay Variation on Multicast Session Performance With TCP-like Congestion Control

Augustin Chaintreau, François Baccelli, Christophe Diot

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Impact of Network Delay Variation on Multicast Session Performance With TCP-like Congestion Control

Augustin Chaintreau — François Baccelli — Christophe Diot

N° 3987

Août 2000

THÈME 1

ISSN 0249-6399 ISRN INRIA/RR--3987--FR

Impact of Network Delay Variation on Multicast Session Performance With TCP-like Congestion Control

Augustin Chaintreau [*] , François Baccelli [†] , Christophe Diot [‡]

Thème 1 — Réseaux et systèmes
Projet MCR

Rapport de recherche n° 3987 — Août 2000 — 23 pages

Abstract: We study the impact of random noise (queueing delay) on the performance of a multicast session. With a simple analytical model, we analyze the throughput degradation within a multicast (one-to-many) tree under TCP-like congestion and flow control. We use the (max,plus) formalism together with methods based on stochastic comparison (association and convex ordering) and on the theory of extremes (Lai and Robbins' notion of maximal characteristics) to prove various properties of the throughput.

We first prove that the throughput obtained from Golestani's deterministic model [1] is systematically optimistic. In presence of light tailed random noise, we show that the throughput decreases like the inverse of the logarithm of the number of receivers. We find an analytical upper and a lower bound for the throughput degradation. Within these bounds, we characterize the degradation which is obtained for various tree topologies. In particular, we observe that a class of trees commonly found in IP multicast sessions [9] (which we call umbrella trees) is significantly more sensitive to network noise than other topologies.

Key-words: Congestion control, flow control, internet, muticast, tree, TCP, throughput, (max,plus) algebra, Lyapunov exponents, stochastic ordering, convex ordering, association, maximal characteristic.

[*] ENS, 45 rue d'Ulm 75005 Paris, France {Augustin.Chaintreau@ens.fr}
[†] INRIA-ENS, 45 rue d'Ulm 75005 Paris, France {Francois.Baccelli@ens.fr}
[‡] Sprint ATL, 1 Adrian Court, Burlingame, CA 94010, USA {cdiot@sprintlabs.com}

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex (France)
Téléphone : 01 39 63 55 11 - International : +33 1 39 63 55 11 —Fax : (33) 01 39 63 53 30 - International : +33 1 39 63 53 30

Impact de la variabilité des délais sur les performances d'une session multipoint contrôlée par un mécanisme de type TCP

Résumé : Nous étudions l'impact de perturbations aléatoires (files d'attente) sur les performances d'une session multipoint. Grâce à un modèle analytique simple, nous analysons la dégradation du débit de la session dans le cas d'une source unique et lorsque le contrôle de congestion et de flux est de type TCP. Nous utilisons l'algèbre (max,plus) ainsi que des méthodes de comparaison stochastique (association et ordonnancement convexe) et des résultats de la théorie des extrêmes (caractéristique maximale de Lai et Robbins) pour analyser les propriétés asymptotiques du débit d'une telle session lorsque le nombre de récepteurs est grand.

Nous montrons d'abord que le débit obtenu par le modèle déterministe de Golestani [1] est systématiquement optimiste. En présence d'un bruit dont la distribution est à queue exponentielle, le débit de la session décroît en fait comme l'inverse du logarithme du nombre des récepteurs. Ce comportement est établi au moyen de bornes inférieure et supérieure explicites du debit. Nous étudions aussi l'évolution du débit entre ces deux bornes en fonction de la topologie de l'arbre multipoint. Nous montrons notamment qu'une classe d'arbres fréquemment rencontrés dans les sessions IP multipoint [9] (et que nous appelons arbres parapluie dans ce qui suit) est particulièrement sensible au bruit.

Mots-clés : Contrôle de congestion, contrôle de flux, internet, multipoint, arbre, TCP, débit, algèbre (max,plus), ordre stochastique, ordre convexe, association, caractéristique maximale.

1 Introduction

TCP-friendly congestion control has been advocated by the IRTF Reliable Multicast Research Group in the past [13], where a TCP-friendly flow is a flow that competes "fairly" with TCP-connections. Several recent papers focused on a TCP-friendly solution for the control of multicast [10, 11, 12]. In particular, Golestani has made fundamental observations on multicast flow and congestion control in [1] using a deterministic model.

The present paper goes a step forward from Golestani's in providing an understanding of further properties of TCP-like congestion control in a network with random noise. This step is of practical importance in that it establishes the dependence of a multicast session throughput on the number of receivers, and consequently refines observations realized with a deterministic model. Since multicast deployment will most probably be pushed by single source applications with high bandwidth requirements and a large number of receivers, it is important to check whether TCP-like congestion control does not in fact force multicast sessions to suffer very low bandwidth. Bhattacharyya et al. [16] analyze the impact of TCP-like congestion control on the throughput of a multicast session. They show that for loss based additive-increase multiplicative-decrease algorithms, there is a severe degradation of throughput for large multicast groups.

We generalize the findings in [1] and [16] by showing that even in the case of an ideal TCP control where the flow control window size is kept equal to its maximal value, there is a severe throughput degradation within a one-to-many multicast tree when the group size grows. Intuitively, the session throughput is expected to decrease when the number of receivers increases for the following two reasons:

- Due to the stochastic assumptions, when a new receiver joins, it can add a new link whose bandwidth is less than that of any of the links already present in the tree.
- Due to the fact that the congestion control mechanism is based on feedbacks stemming from all receivers, slow receivers will "slow down" the sender.

In other words, the higher the number of receivers is, the more likely one of them is slow enough to affect the global performance.

The influence of the tree topology on throughput that we establish analytically is another key contribution of this work.

We choose to model a multicast session as follows: packets are sent by a unique sender, located at the root of a tree of routers, to a set of receivers located at the leaves of this tree. This tree will be referred to as the forward tree.

The packet transmission at the sender is controlled by a "TCP-like" congestion control mechanism where each receiver sends acknowledgements back to the sender, and where the sender throughput is controlled by a classicl sliding window mechanism.

We have chosen to model a homogeneous tree: each receiver is equally distant from the source, and all the routers in the tree have the same service time distribution. This assumption allows us to design a simpler model without losing the properties we want to observe.

The model captures congestion via the queueing delay that each packet experiences in each router it passes through. In particular, the fluctuations due to the processing of packets of other (unicast or multicast) connections sharing the same interface of the router are represented by random service times for packets of the reference multicast connection. Our random service times are assumed to be independent in time and space, and light-tailed (i.e. the tail is exponential). The queuing strategy is assumed to be FIFO. Within this framework, the sender and the receivers are modeled as routers, possibly with different mean delays and different distributions.

Since we are not interested in the effect of losses (but only in the effect of an ideal flow control having reached its maximal window size), we assume that all routers have infinite buffers and consider that the network is lossless and that the window size is fixed.

All the assumptions that we make about the network (homogeneity), about transmission control (no losses, window size always equal to its maximal value) and about noise (light tails) have been carefully selected to provide an optimistic network environment. We will show that even in this favorable context there is a severe decrease of the throughput when the number of receivers increases.

To the best of our knowledge, this work is the first to address analytically the question of multicast session throughput degradation due to network noise (queueing delay), for different tree topologies, in a TCP-like (single-rate) control environment. Although we limit this first study to some simple cases, we believe that our mathematical methodology can be expanded to analyze more general cases (e.g. adaptive window, heavy tailed noise, non-homogeneous trees or windows) as discussed in the conclusion.

The paper is structured as follows: in Section 2, we build our analytic model on the (max,plus) formalism [3, 4, 5]. In Section 3, an algebraic simulator is derived from the model. Simulations show that the throughput obtained from Golestani's deterministic model [1] is systematically optimistic. We study throughput degradation for a large number of receivers and for different tree topologies. We further generalize our simulation results with the help of the (max,plus) model. In Section 4, we analyze the model using the notion of positive correlation (also called association), as well as the notion of maximal characteristics (Lai and Robbins). In the presence of a light tailed random noise, the throughput is shown to be bounded by functions that decrease like the inverse of the logarithm of the number of receivers. This qualitative result explains the general shape obtained by simulation for throughput degradation. Within these bounds, we characterize the fine structure of degradation depending on the tree topology. We analyze three different families of tree topologies. First, we analyze classical binary trees. We then consider a class of trees commonly found in IP multicast sessions, [9, 14], which we call umbrella trees. We show that this class of trees is significantly more sensitive to network noise than other topologies, and that in some cases, these topologies reach the upper bound. We finally characterize the throughput degradation curve for a class of optimal trees called "reverse-umbrella" trees. To make the paper more readable, the proofs are given in Appendix A-D.

2 (max,plus) Representation

We introduce first the (max,plus) algebra, show how it can be used to represent a network, and apply this technique to multicast packet transmission.

2.1 Introduction to the (max,plus) algebra

2.1.1 Operations in $\mathbb{R}_{\max}$

We consider the scalar algebra, namely, the set $\mathbb{R}_{\max} = \mathbb{R} \cup \{-\infty\}$, which we endow with two operations that are different from the usual ones: the max operation (denoted by $\vee$) replaces the usual addition, and addition, with the convention ($\forall a \in \mathbb{R}_{\max}, -\infty + a = -\infty$), replaces the usual multiplication.

Note that this structure has all the properties required to make a commutative semi-ring: associativity, commutativity, neutral elements[1], and distributivity ($\forall a, b, c \in \mathbb{R}_{\max}, a+(b \vee c) = (a+b) \vee (a+c)$).

2.1.2 $\mathbb{R}_{\max}$ matrices

$(\mathbb{R}_{\max}, \vee, +)$ being a semi-ring, we can construct operations on matrices as in the conventional algebra, with the addition of matrices obtained by term by term maximization, and multiplication defined by the rule :

$$(\mathbb{A}\mathbb{B})_{i,j} = \max_k (\mathbb{A}_{i,k} + \mathbb{B}_{k,j}) \tag{1}$$

[1] for $\vee$, $-\infty$ that we will denote by ε, and for $+$, 0 that we will denote by e.

We will denote by ε the matrix filled with ε everywhere, and $\mathbb{I}$ the identity matrix (e on the diagonal and ε everywhere else).

Norm Let $\|.\|$ denote the matrix norm:

$$\|\mathbb{A}\| = \max_{i,j}(\mathbb{A}_{i,j})$$

2.2 (max,plus) Representation of a Network

We illustrate the (max,plus) modeling of a network via a simple example : a point-to-point end-to-end connection through routers $R_i, 1 \leq i \leq L$, with window flow control with a fixed size window W (this model was introduced in [6]). The sender is modeled as the first router, and the receiver as the last router. The multicast model we will present in the next section is an extension of this network model.

Each router is represented as a FIFO queue with an infinite buffer[2] and a service time for each packet of the connection. The service time includes the waiting time due to the processing of packets of other connections present in the routers. Assuming that the multicast session reaches a steady state, it is natural to make the assumption that the service times in all routers are identically distributed. We will also assume service time independence for the sake of simplicity, i.e. service times on different routers in the network are independent and the sequence of service times on a router is made of independent and identically distributed random variables.

We denote by $s_m^{(i)}$ the service time of the m-th packet on router i, and $x_m^{(i)}$ the time when router i has completed the processing and forwarding of packet m.

- Router $R_i, i > 1$ starts processing packet m as soon as it has finished processing packet $m-1$, and the upstream router has forwarded packet m. Once a router R_i starts processing packet m, it takes $s_m^{(i)}$ to complete processing. $s_m^{(i)}$ actually includes the processing time of all the packets of the other connections interleaved between packet $m-1$ and packet m. So we have for $i > 1$:

$$x_m^{(i)} = (x_{m-1}^{(i)} \vee x_m^{(i-1)}) + s_m^{(i)}.$$

- The sender R_1 sends packet m as soon as it has finished with packet $m-1$, provided that the window control allows packet m to be sent. So that we have :

$$x_m^{(1)} = (x_{m-1}^{(1)} \vee x_{m-W}^{(L)}) + s_m^{(1)}.$$

Let X_m be the vector of dimension L with entries $(x_m^{(i)})_{1 \leq i \leq L}$, and Y_m be the block-vector of dimension LxW with blocks $X_m, X_{m-1}, \ldots, X_{m-W+1}$. We can capture the dynamics of the network by a $(\max, +)$ linear recurrence

$$\begin{cases} Y_0 = \text{ vector with all its coordinates equal to } e \\ Y_m = \mathbb{P}_m Y_{m-1} \text{ for } m > 0 \end{cases} \tag{2}$$

where the matrix $\mathbb{P}$ consists of the different square blocks (each of dimension L) as indicated below:

$$\mathbb{P}_m = \begin{pmatrix} \mathbb{S}_m & \varepsilon & \ldots & \varepsilon & \mathbb{W}_m \\ \mathbb{I} & \varepsilon & \ldots & \varepsilon & \varepsilon \\ \varepsilon & \mathbb{I} & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \varepsilon & \varepsilon \\ \varepsilon & \ldots & \varepsilon & \mathbb{I} & \varepsilon \end{pmatrix}$$

[2]Note that buffers being of infinite size, no loss occurs. As a result, the window size is assumed to reach its maximal value and to remain constant. The effect of congestion is expressed by the variations of service times in routers.

- $\mathbb{W}_m$ represents the window control mechanism. In this case we have $(\mathbb{W}_m)_{i,j}$ equal to ε if $j \neq L$ and to $s_m^{(1)} + \cdots + s_m^{(i)}$ for $i = 1, \ldots, L$ and $j = L$.
- $\mathbb{S}_m$ represents the forwarding mechanism in the network, and $(\mathbb{S}_m)_{i,j}$ is generally given by the maximum over all paths leading from i to j of the sum of service times for packet m on the path from router j to router i (including both i and j).

Note that if service times are independent and identically distributed, then the matrices $(\mathbb{P}_m)_m$ are also independent and identically distributed; we can apply Corollary 1 shown in appendix A, which implies the existence of $\lim_{m\to\infty} \frac{\|Y_m\|}{m} = \gamma$ in $\mathbb{L}_1$ and with probability 1. γ is called the *Lyapounov exponent* of this sequence of matrices. Since $\|Y_m\|$ represents the epoch between the times when packet m started at the sender and arrived at the receiver, γ is therefore the *inverse* of the asymptotic throughput, which can be defined as the long term averaging of the instantaneous throughput at which packets are sent by the source *(i.e. average throughput since the beginning of the session).*

In the following section, we extend this model to multicast.

2.3 Representation of Multicast Flow Control

2.3.1 Multicast extension of the network model

In our multicast model, a single source broadcasts packets over a unidirectional homogeneous tree to N receivers. Homogeneous means that the path from the sender to each receiver is statistically the same for all receivers; i.e. there is the same number of routers and the service time distribution is the same for all routers of level l). Note that homogeneity allows for quite complex tree structures, where nodes may have different degrees depending on their level in the tree.

Acknowledgements are forwarded through a backward tree which is a mirror version of the forward tree up to the source. The backward tree is chosen to be functionally independent of the forward tree for the sake of simplicity. Similarly, the way back from a receiver to the sender is statistically the same for all receivers. The implosion problem (see [1]) is taken care of via the aggregation of acknowledgements in each router of the backward tree.

Packet transmission is controlled by a unique fixed size window (instead of various receiver based windows). This is a major difference with Golestani's model. But, given his conclusion that each receiver should choose a window size proportional to its distance from the source, it makes sense to consider a unique window in the case of a homogeneous tree.

This model is illustrated in Figure 1, in the case of a binary tree.

We denote the window size by W , the number of receivers by N, the number of routers in the network by L, and the depth of the (forward) tree by D. For all receivers i, we denote by $s_m^{(f(1,i))}, \ldots, s_m^{(f(2D,i))}$ the service times of packet m in the routers[3] from the source to itself via receiver i. Let

$$S_m^{(i)} = s_m^{(f(1,i))} + \ldots + s_m^{(f(2D,i))} \tag{3}$$

denote the round trip time of packet m on the route that contains receiver i.

2.3.2 (max,plus) Representation of the Model

The network model described above can be written in a way similar to that of Section 2.2. Let X_m be the $\mathbb{R}_{\max}$ vector of dimension L where entry i is the departure time of packet m from router i. The first entry corresponds to the first router of the network (the source), and the last entry corresponds to the last router at the end of the feedback tree, just before the control loop. Let Y_m be the block vector of dimension LxW built on top of $(X_m)_{m\in\mathbb{N}}$ and which captures the history of X_m in the same

[3] f(d,i) is the index of the d-th router on the path that leads from the source to itself via receiver i (cf figure 1)

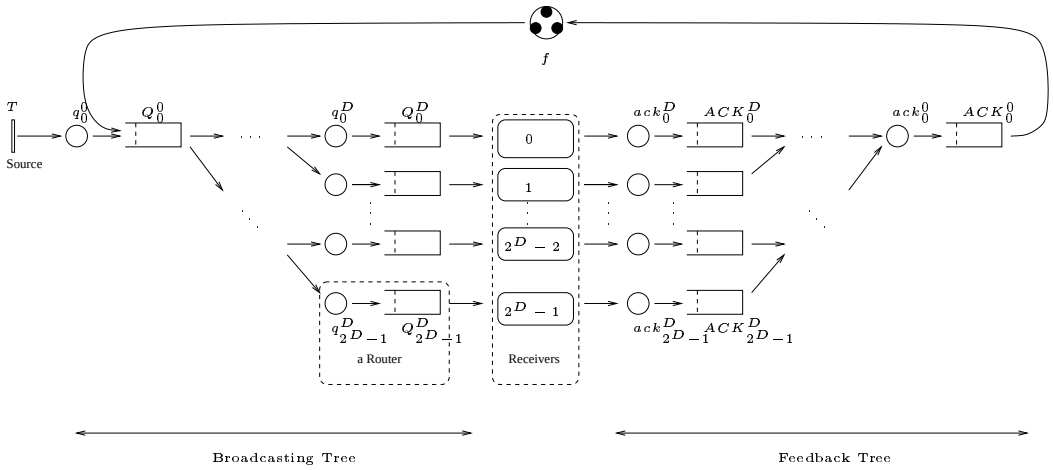

Figure 1: The "broadcast and feedback" graph

way as above. We have the same (max,plus) linear system for Y_m as in Equation (2), though with different matrices.

$\mathbb{P}_m$ has exactly the same shape as before. The block $\mathbb{W}_m$ is the same as before and the block $\mathbb{S}_m$ has the same interpretation. $\mathbb{S}_{i,j}$ is the sum of the different service times along the path from i to j in the tree if this path exists (ε if it does not).

The assumptions that we make about service times and the block structure of $(\mathbb{P}_m)_{m\in\mathbb{N}}$ allow us to deduce from Corollary 1 the existence of the Lyapounov exponent, which is given by $\lim_{m\to\infty} \frac{\|Y_m\|}{m}$ and represents the inverse of the average throughput.

3 Algebraic Simulation

The algebraic simulator described below does the same job as a discrete event simulator. Its advantages are twofold: (a) since this type of simulation consists in products of matrices and vectors, it is of low complexity, which is important when the number of receivers becomes large, and (b) the same formalism is used in the simulation and in the analytical sections.

3.1 Description of the Simulator

As explained in the previous section, the average throughput is obtained through the Lyapounov exponent, which can itself be obtained accurately by simulation. In practice, we can estimate γ by $\|Y_M\|/M$ for a large enough value of M. The algebraic simulator picks different random values for service times in the routers, then builds the matrix $\mathbb{P}$, and multiplies the current value of Y by $\mathbb{P}$. After M steps, we have Y_M, and hence a reasonable approximation of γ (if M is chosen properly). Preliminary convergence studies that we made reveal that $M = 400$ steps is sufficient to analyze the throughput behavior. As far as the simulation is concerned, there is a trade-off between accuracy and computational complexity: the estimation of the Lyapunov exponent requires the simulation of a large number of packets (or equivalently the computation of the product of a large number of matrices), whereas that simulation of multicast groups with a large number of receivers requires the manipulation of large matrices. In order to simulate large multicast groups, we had to accept moderately accurate estimates (i.e. rather large confidence intervals) for the Lyapunov exponents. This choice results in rather non-smooth shapes for most of the simulation curves produced below. However, this is sufficient to estimate the general shape and the relative ordering between the curves in question, as we see in the next section.

In the remaining of this paper, we discuss simulation and analytical results in term of average throughput rather than of in term of Lyapounov exponent.

3.1.1 Modeling the Topology

Let us first consider a complete binary tree with height D and with total number of leaves 2^D).

To study the throughput dependency on the size of the group, we need to vary the number of receivers of a multicast session. For every binary tree of size 2^D, we consider a set of N "active" receivers which is a subset of the leaves of the complete binary tree ($N \leq 2^D$). For this we simply set the service times to be equal to zero in all the routers that do not forward packets to an active receiver. So we can use the general equations for the complete binary tree with these special values of the service times to analyze the sub binary tree corresponding to this subset of N leaves.

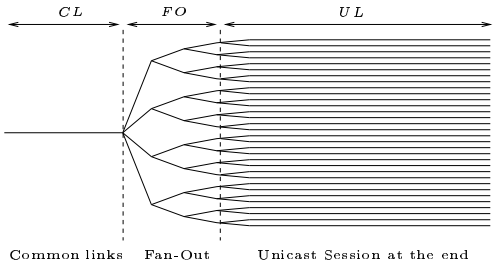

Figure 2: Trees generic topology

The above technique allows us to study any kind of tree topology that can be represented by the generic topology described in Figure 2. These topologies consist of three parts :

- A first set of CL links which is common to all receivers.
- A Fan-Out whose total depth is FO. The first step of this fan-out is k-ary (degree[4] k for the first node of the fan out), all the other fan-outs are binary (degree 2 everywhere else).
- A unicast transmission of depth UL (unicast in the sense that there is no replication of the packet in this part of the tree, and no link shared by different receivers).

Using this parametric representation of tree topologies, we simulate three types of trees represented in Figure 3. In addition to complete binary trees we consider:

- Umbrella Trees: these trees end by a long unicast transmission after a short fan-out (large value of UL). The limiting case is one path from the source to all receivers. It characterizes multicast trees where receivers share few links only. This kind of topology is often found in IP Multicast sessions [9, 14].
- Reverse Umbrella Trees: packets are forwarded first along a long common path, and then a short fan out ends the transmission (large value of CL). Intuitively this kind of topology is optimal, as receivers' behaviors differ only by few links.

These categories will be more precisely defined and analytically studied in Section 4.3.

[4]The method to emulate a k-ary fan out in binary trees consists in starting the binary expansion before the real fork and in using appropriate values for service time ensuring that this represents the desired fork.

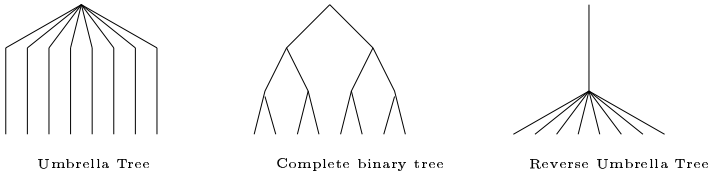

Figure 3: Fundamental types of tree topologies

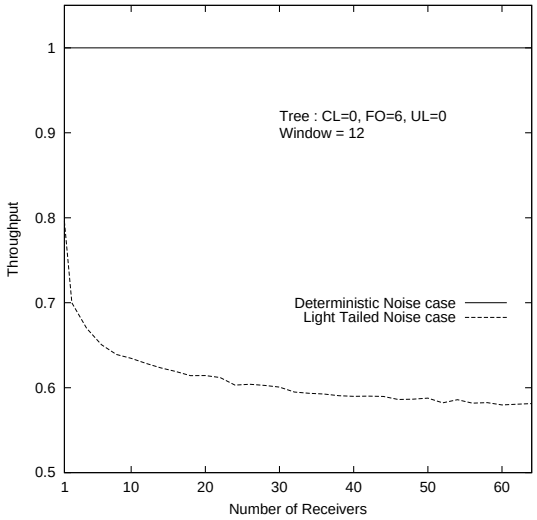

Figure 4: Average throughput vs. number of receivers in a binary tree with light noise.

3.2 Average throughput vs. Number of Receivers

This section analyzes the evolution of the average throughput with the number of receivers in a multicast group.

We start the simulation by taking one active receiver in a binary tree, and by computing the associated (max,plus) linear recurrence on Y in order to estimate γ. Then we pick another receiver in the tree, add it to the current tree and run the same type of simulation again (which gives the value of γ for two receivers). Then we progressively fill in the tree with more and more receivers.

We simulate different ways of filling in the tree. "Best Filling" consists in starting from receiver 0 (numbers refer to Figure 1) and taking at each step the "next" receiver in the order suggested by the numbering. We also consider "Random Filling", where each new receiver to join is chosen randomly.

Simulation results are shown in Figures 4 to 10. The service times in the routers follow an exponential distribution with the same parameter for each router in the network, so that the homogeneity condition is satisfied.

Note that in homogeneous trees with deterministic service times, throughput does not dependent on the number of receivers, as each receiver has the same round trip time and behaves synchronously with other receivers in the multicast group. Each plot below includes the deterministic value of the throughput as found by Golestani.

Figure 4 is obtained by simulating a complete binary tree of length 6 ($CL = 0, FO = 6, UL = 0$), with a window of size 12. Each router of the tree has an exponential service time with average 1. The feedback tree has a null service time on all routers.

The first important observation is that the average throughput decreases like the inverse of the logarithm of the number of participants.

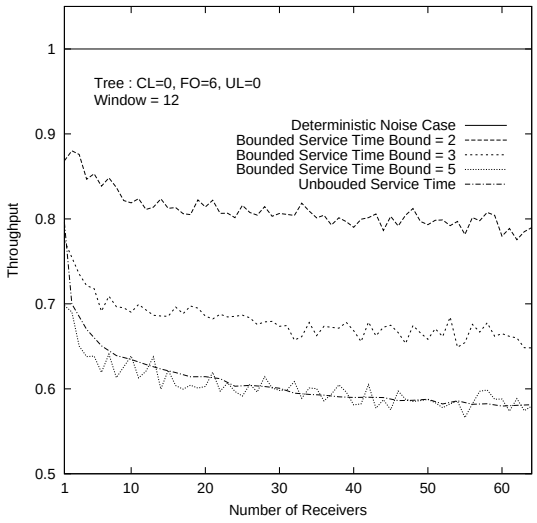

Figure 5: Average throughput in the presence of bounded and unbounded light noises.

The significant throughput drop between 1 and 2 receivers can be explained by the homogeneous nature of the tree which implies that the second receiver joins the tree with a path of length $CL + FO + UL$. Then the throughput keeps decreasing significantly until we reach 20 receivers. Between 40 and 60 receivers, the throughput degradation stabilizes at around 40% of the deterministic case. If we refer to the throughput observed with 1 receiver, the maximum degradation is limited to around 25%.

The same kind of throughput degradation has also been observed in [16] under different assumptions.

3.2.1 Influence of the noise on the throughput

Before further investigating the shape of the throughput degradation, we have to study how this degradation varies with network noise assumption.

Figure 5 shows throughput as a function of the number of receivers for service times belonging to the class of (bounded support) truncated exponential distribution functions. Since the mean values are not preserved by truncation of a given exponential density, the relative positions of the curves is not particularly meaningful.

So, the most interesting remark to be made bears on the shape of the curves. We observe the very same logarithmic decrease as in the bounded case. This is particularly clear when looking at the case where the truncation threshold is large (i.e. equal to 5), which leads to throughputs that are quite close to the unbound case. Thus, we can conclude from these curves that the shape of the degradation is not bound to the exponential assumption per se. Our conjecture is rather that all distribution functions with a tail bounded from above by a negative exponential function lead to such a decrease provided variance is large enough. For heavier tails, (e.g. Pareto tails) preliminary results suggest that the growth of the Lyapunov exponent is polynomial. So, the bounded support and the light tail cases are qualitatively the same, at least when variance is not too small, and this generic case seems to be the most favorable one when compared to heavier Pareto type tails.

3.2.2 Analysis of various network parameters

In order to understand the impact of network parameters on the throughput degradation, we vary network parameters. Figure 6 shows how the throughput decreases for various tree depth values. Trees being homogeneous, tree depth influences the throughput by the fact that each receiver joins the tree with a path whose length is the maximum tree depth (i.e. $CL + FO + UL$). The deeper the tree,

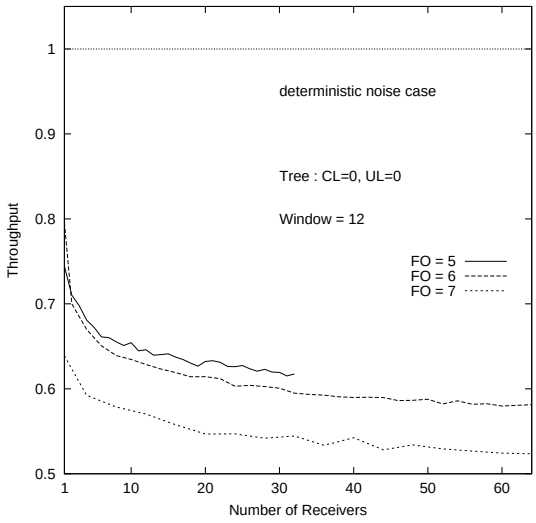

Figure 6: Throughput for different tree depths.

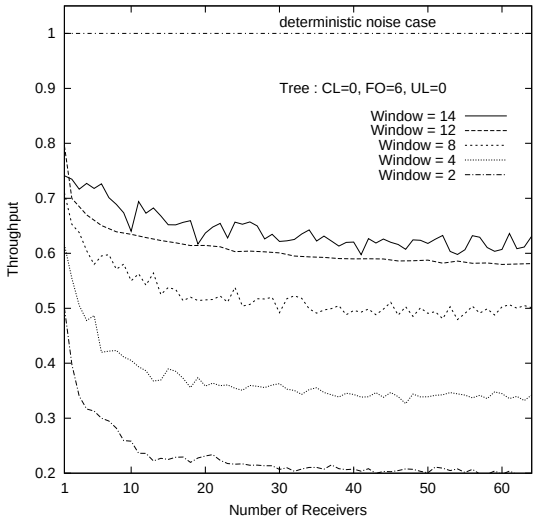

Figure 7: Throughput with varying window size.

the faster the throughput decreases. We have chosen a default tree depth of 6 because it allows us to simulate a sufficiently large number of receivers with still a good throughput accuracy.

In Figure 7, we vary the window size. As expected, we observe that the window size significantly affects the throughput. We have chosen 12 as a default window size for our experiments because it is a credible value, that it is close to the optimal throughput observed, and that it does not increase too much the complexity of the simulations.

We finally check, Figure 8, the impact of the filling algorithm on throughput degradation. As expected, randomly filled trees suffer a more severe throughput decrease than best filling trees. This difference is easy to explain. In the random filling approach, adding a new receiver generally adds more network links than in the best filling approach, where a new receiver systematically adds the minimum possible number of links.

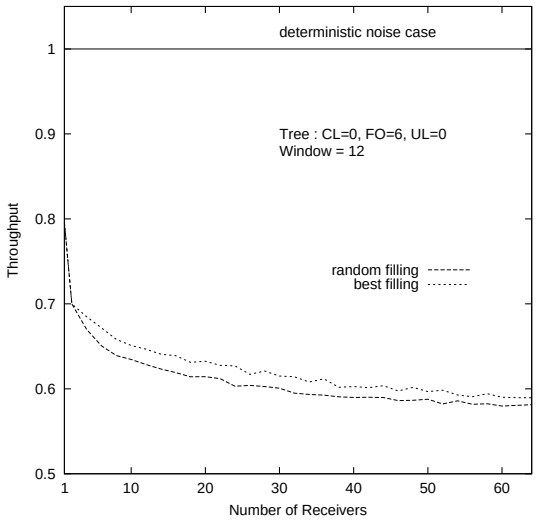

Figure 8: Throughput with different tree construction approaches.

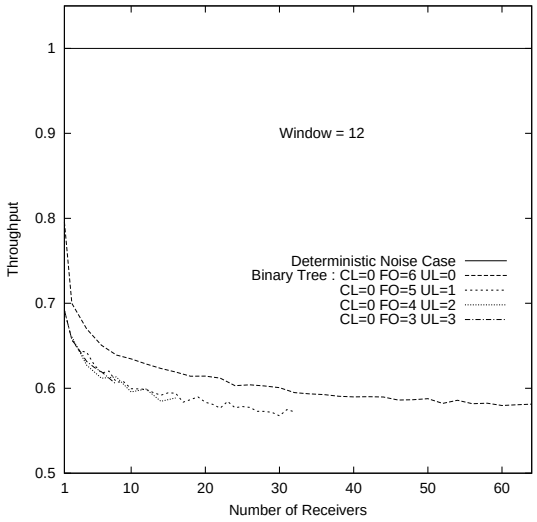

Figure 9: Throughput in the case of umbrella trees.

3.3 Analysis of the Tree Topology

We now simulate the tree topologies described in Section 3.1.1 with window size equal to 12 and with a random filling technique. The total depth of the tree is always equal to 6. We only vary the value of CL, FO, and UL with the sum being 6. In a deterministic model, all trees of same length would have the same performance. Figure 9 plots various umbrella trees and Figure 10 plots various reverse-umbrella trees. In both cases, the binary tree case is given as a reference.

First, varying the topology of the tree significantly influences the throughput (up to 20% between the best reverse-umbrella tree and the worst umbrella tree). The second observation is that reverse-umbrella trees perform systematically better than binary trees, themselves performing better than umbrella trees.

We also observe that the closer the fan-out is to the receivers, the higher the minimum throughput. Thus, trees where receivers share fewer links are much more sensitive to the number of receivers in the group. The throughput of an umbrella tree has already decreased to 60% of the deterministic

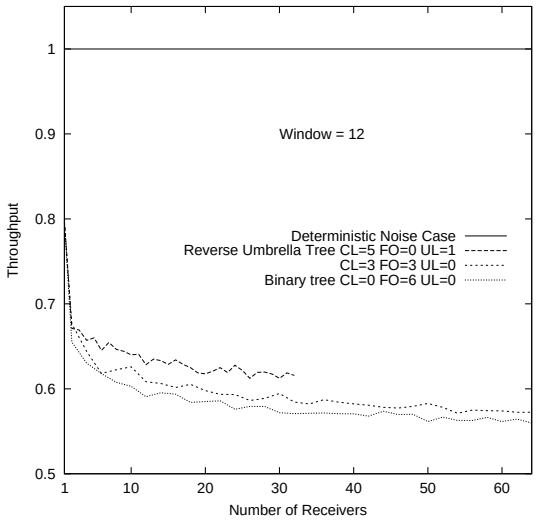

Figure 10: Throughput in the case of reverse-umbrella trees.

throughput for 3 receivers, while a reverse-umbrella tree reaches the same throughput with 10 receivers. This is an important observation as the current Internet topology favors umbrella trees [9, 14].

4 Mathematical Analysis

In this section we give mathematical arguments substantiating the average throughput degradation in $\ln(N)$ that was observed in the simulations. We propose a mathematical framework to evaluate and compare the throughput in a predictable way depending on the tree topology and on the number of receivers.

4.1 Mathematical Tools

In order to have more means to characterize the throughput in our model, we use a few basic notions of stochastic comparison *convex ordering* which will help us to compare the deterministic case and the random case; the notion of *association of random variables* which will help us to express the correlation between different receivers; lastly the notion of *maximal characteristics* which comes from the theory of extremes, and which will allow us to explicitly define bounds on the performances.

4.1.1 Comparison with the deterministic case

Proposition 1 *Let* $\mathbb{A}$ *and* $\mathbb{B}$ *be two random* $\mathbb{R}_{\max}$ *matrices. Then for all* i *and* j *:* $(\mathbb{E}[\mathbb{A}\mathbb{B}])_{i,j} \geq (\mathbb{E}[\mathbb{A}]\mathbb{E}[\mathbb{B}])_{i,j}$ *which implies* $\|\mathbb{E}[\mathbb{A}\mathbb{B}])\| \geq \|\mathbb{E}[\mathbb{A}]\mathbb{E}[\mathbb{B}]\|$

Proof We just need to verify this formula for the two basic operations. This is clear for the addition; concerning the max operation, for all random variables X and Y with value in $\mathbb{R}_{\max}$, we have (by a direct convexity argument):

$$\mathbb{E}[\max(X, Y)] \geq \max(\mathbb{E}[X], \mathbb{E}[Y])$$

□

Let now $\bar{\mathbb{P}}_m$ be the matrix describing the same network as above, where each random service time has been replaced by its mean value. We have $\bar{\mathbb{P}}_m = \mathbb{E}[\bar{\mathbb{P}}_m] \leq \mathbb{E}[\mathbb{P}_m]$ (because $\mathbb{P}$ is made of max and sums of service times) and so

$$\begin{aligned} \bar{Y}_m &= \bar{\mathbb{P}}_m \dots \bar{\mathbb{P}}_1 Y_0 \leq \mathbb{E}[\mathbb{P}_m] \dots \mathbb{E}[\mathbb{P}_1] Y_0 \\ &\leq \mathbb{E}[\mathbb{P}_m \dots \mathbb{P}_1 Y_0] = \mathbb{E}[Y_m]. \end{aligned}$$

This implies $\bar{\gamma} \leq \gamma$, proving that the deterministic model always leads to a better throughput indeed. Hence, Golestani's deterministic model is actually proven to give the best possible throughput within the range of all stochastic models of the same class and with the same means.

4.2 Bounding the Throughput Degradation

We now analyze the way throughput decreases when new receivers join the group.

4.2.1 Upper Bound

Theorem 1 *Consider a network with exponential service times in routers; then γ is bounded from above by a function that can be expanded as*

$$RTT \ln(N)(1 + o(1)) \text{ for } N \to +\infty \tag{4}$$

where N is the number of receivers, RTT is the average round trip time of a receiver ($RTT = \mathbb{E}[S_1^{(1)}]$, where $S_1^{(1)}$ is defined in (3)).

Proof is given in Appendix A.

□

This upper bound can be reached, this is the case when the window size is $W = 1$ and when we have an umbrella tree made of maximally disjoint branches.

4.2.2 Lower Bound

Theorem 2 *Consider a network with exponential noise in routers, and assume that receivers are distinct (i.e. the last link for every receiver is different); then γ is bounded from below by a function that can be expanded as*

$$\frac{\mathbb{E}[s]}{W} \ln(N)(1 + o(1)) \text{ for } N \to +\infty, \tag{5}$$

where N is the number of receivers, W the window, and s a typical service time in a router [5].

Proof is given in Appendix A.

□

It is possible to have a better lower bound for γ under some additional assumptions on the tree as we see in the next subsection. Again this bound is reached by the reverse umbrella tree with maximal common path.

[5] In the homogeneity hypothesis that we have made, $\mathbb{E}[s]$ is the average service time of the last router before receivers, that is by assumption the same for all receivers.

4.3 Tree Topology Dependence : Notion of Aggregation

We have been able to compute a lower and an upper bound on γ which both grow in proportion to $\ln(N)$. We now give a finer grain classification within these bounds, based on tree topologies. We show that, based on the tree topology, it is possible to define a partial order on trees. We consider trees with N receivers, under the assumptions of homogeneity and independence described earlier.

4.3.1 Definition of a partial order on tree topologies

We still assume that the service times on the backward tree are all equal to zero.

Consider two receivers i and j, with paths from the source to every of these two receivers. For every $l = 1, \dots D$, the service times $s_m^{(f(i,l))}$ and $s_m^{(f(j,l))}$ are either:

- The same variable - when receivers i and j share their l-th link.
- Two independent random variables with the same distribution - when paths from the source to i and j are different at a given depth, and for all the following links in the tree.

Definition: Aggregation. Given the above statements, we can define the aggregation of receivers i and j, that we write $a(i,j)$, by

$$a(i,j) = \max\{l = 1 \dots D | s_m^{(f(i,l))} \equiv s_m^{(f(j,l))}\} \tag{6}$$

The round trip times $S_m^{(i)}$ and $S_m^{(j)}$ for receivers i and j can be described as a sum of service times along the transmission path. The aggregation is exactly the number of common service times in the sum between i and j; all others terms of these sums are independent.

Aggregation indeed quantifies the correlation beetween two receivers: receivers with large aggregation appear to have similar performances. It is possible to show that a given aggregation characterizes a tree [6].

Definition: Aggregation order. We will say that a tree T is less aggregated than another tree T' if their aggregation functions are such that $a \leq a'$ (i.e. aggregation beetween any receiver and another one is higher is tree a' than in a). This is of course a partial order relation on trees. This order is compatible with the performance of the tree as shown in the next result.

Theorem 3 *If two trees T and T' are such that $a \leq a'$, then $\gamma_{a'} \leq \gamma_a$.*

Proof is given for a particular case in Appendix D, with a brief description of the general proof.

□

This theorem gives us another proof for the upper and lower bounds of Section 4.2. The upper tree[7] (given by $a(i,j) = D-1$) provides the lower bound, the lower tree (given by $a(i,j) = 0$) provides the upper bound.

4.3.2 Characterization of Umbrella Trees

Definition: Umbrella Tree of Class l. A tree (given by its aggregation a) is said to be an umbrella tree of class l if we have $a \leq D - l$. It represents a tree that finishes for every receiver by a unicast connection of length at least l.

Umbrella trees typically correspond to a worst case situation, as receivers share few links. We have observed via simulation that for this type of trees, the throughput seems to degrade more when the number of receivers increases. In an umbrella tree, the lower bound for γ can be reached.

[6] A consequence from the study of the equivalence relation $i \approx_l^a i' \Leftrightarrow a(i,i') \geq l$ is that we can build the tree using only the aggregation function.

[7] when making the assumptions that the receivers are distinct.

Theorem 4 *Under the foregoing assumptions, the Lyapounov exponent for an umbrella tree of class l is bounded from below by a function that can be expanded as:*

$$\frac{\mathbb{E}[s]}{W} R_l(1 + o(1)) \text{ for } N \to +\infty, \tag{7}$$

where s is the sum of service times of routers[8] $D - l + 1, \ldots, D$, and R_l is the function defined in Appendix C (Corollary 3).

Proof According to Theorem 3, we just need to verify this formula for the tree ($a = l$). The following formula established in the proof of the lower bound holds:

$$\gamma \geq \frac{\mathbb{E}[\max_{i=1\ldots N}(S_1^{(i)})]}{W} \tag{8}$$

Let us look at the performance of the tree ($\forall i, j,\ a(i,j) = l$). We have, for all i,

$$S_1^{(i)} \geq s_1^{(f(i,D-l+1))} + \ldots + s_1^{(f(i,D))}.$$

We can then apply Corollary 3 to this sum of random variables that has a Gamma distribution.

□

As we can see, in spite of all the optimistic assumption we have made, an umbrella tree systematically results in a severe degradation of the throughput. The results observed in the algebraic simulation, as well as the intuition we have on tree topology impact, are confirmed and generalized by analytical results. Aggregation seems to be a key concept, as it gives us a parametric representation of the tree topology which allows direct performance comparison. An important result of this study is that umbrella trees, that are frequently encountered in the Internet [9] suffer severe throughput degradation even in the case of light tailed fluctuations in queueing delay.

Notice that Theorem 3 shows that the link established on umbrella trees between aggregation and throughput is actually valid for more general trees.

5 Conclusion

In this paper, we study the impact of random noise (i.e. queueing delay) on the performance of a one-to-many multicast session in the presence of a "TCP-like" congestion control mechanism. With a simple analytical model, we analyze the degradation of throughput when the size of the multicast group increases. In addition, we study the impact of tree topology on the throughput of the multicast session.

In presence of a light tailed random noise, we show that the throughput decreases logarithmically when the number of receivers increases. We analytically find an upper and a lower bound for the throughput. Within these bounds, we characterize the degradation depending on the tree topology. The throughput decreases very significantly as the number of receivers increases from 1 to 20. With 40 receivers, the throughput is only 70% of what it would be with a single receiver. Then the throughput seems to stabilize around 60 receivers. In particular, we have identified a class of trees commonly found in IP multicast sessions [9, 14] as a worst case of throughput degradation. This observation is quantified by simulation and then explained analytically.

This work analytically proves that TCP-like congestion control might be harmfull with reliable multicast transmission. Consequently, applications may prefer multi-rate control mechanisms to single

[8] if we make the same assumption of a network where every router has the same service time, we then have $l \times \mathbb{E}[s_1^{(f(1,1))}]$, which can be interpreted has $\frac{l}{D} RTT$.

rate reliable multicast transmission. This results extends to unreliable applications that have difficulties in managing a 30% drop in average throughput when the number of receivers increases.

Multi-rate (layered) control mechanisms are best suited to multicast sessions with a large number of receivers. Rubenstein et al. [11] has shown that multi-rate control can preserve TCP fairness with regard to TCP flows sharing the same congested node, while not penalizing all receivers in case of localized congestion. Subcasting (single group with filtering in nodes) is another area of investigation.

Another major contribution of our work is a new analytical framework that can be used to study various problem related to flow and congestion contol (in multicast and unicast environements).

In future works, we will extend and generalize our analytical framework. Extension to adaptive window size is possible based on a generalization of the (max,plus) representation of TCP Tahoe and Reno known for unicast [15].

In this framework, we will also study various sub-grouping approaches and try to define new classes of congestion control mechanism that might be applicable to unicast transmission as well. We will also analyze how TCP-like congestion control affect shared trees.

6 Appendices

A Proof of the Main (max,plus) Representation Results

A.1 Product of a Large Number of Matrices, Lyapounov Exponents

We start from the following result shown in [3].

Theorem 5 *let $(\mathbb{A}_n)_{n\in\mathbb{N}}$ be a sequence of random square matrices in $\mathbb{R}_{\max}$ independent with same law and with coefficients with finite expectation; then we have :*

$$\lim_{m\to\infty} \frac{\|\mathbb{A}_m \mathbb{A}_{m-1} \dots \mathbb{A}_1\|}{m} = \gamma \tag{9}$$

in expectation and with probability 1, where γ is a constant called the (max,plus) Lyapounov exponent of this sequence of matrices.

In the article, we will make use of the following corollary where the assumptions are those of the above theorem; the dimension of the matrices is LW; $Y(n)$ is a vector of dimension LW which satisfies the (max,plus) linear recurrence :

$$Y_m = \mathbb{P}_m Y_{m-1} \text{ for } m > 0. \tag{10}$$

We assume in addition that Y_m is the block vector $(X_m, X_{m-1}, \dots, X_{m-W+1})$, where each block is of dimension L, that Y_0 has all its coordinates equal to 0 and that the sequence X_m is coordinatewise non decreasing. All these properties are satisfied in our models.

Corollary 1 *Under the foregoing assumptions,*

$$\lim_{m\to\infty} \frac{\|X_m\|}{m} = \gamma \tag{11}$$

in expectation and with probability 1, where γ is the Lyapounov exponent of the sequence $\{\mathbb{P}_n\}$.

Proof $\|X_m\| = \|y_m\| = \|\mathbb{A}_m \mathbb{A}_{m-1} \dots \mathbb{A}_1\|$

□

A.2 Proof of the Bounds on the Throughput

Theorem 6 (Upper Bound) *consider a network with exponential service times in routers. Then γ is bounded from above by a function that can be expanded as*

$$RTT \ln(N)(1 + o(1)) \text{ for } N \to +\infty, \tag{12}$$

where N is the number of receivers and RTT is the average round trip time of a receiver ($RTT = \mathbb{E}[S_1^{(1)}]$).

Proof We have

$$\frac{\|Y_m\|}{m} = \frac{\|\mathbb{P}_m \dots \mathbb{P}_1 Y_0\|}{m} \leq \frac{\|\mathbb{P}_m\| + \dots + \|\mathbb{P}_1\|}{m}$$

The strong law of large numbers gives us

$$\gamma \leq \lim_{m \to +\infty} \frac{\|\mathbb{P}_m\| + \ldots + \|\mathbb{P}_1\|}{m} = \mathbb{E}[\|\mathbb{P}_1\|]$$

with probability 1.

We will now use the interpretation we have on the elements of the matrix $\mathbb{P}_1$. The largest element in $\mathbb{P}_1$ is the sum of the service times of packet n along a path from the source to the last router, that is $\|\mathbb{P}_1\| = \max_{i=1\dots N}(S_1^{(i)})$.

The random variables $S_1^{(i)}$, $i = 1, \dots, N$ are *associated* (see Proposition 2 in the appendix), so that Proposition 3 in the appendix tells us

$$\gamma \leq \mathbb{E}[\|\mathbb{P}_1\|] = \mathbb{E}[\max_{i=1\dots N}(S_1^{(i)})] \leq \mathbb{E}[\max_{i=1\dots N}(\tilde{S}_1^{(i)})],$$

where the random variables are an independent version of the variables $S_1^{(i)}$ i.e. the random variables $\tilde{S}_1^{(i)}$ are independent and for all i, $S_1^{(i)}$ and $\tilde{S}_1^{(i)}$ have the same law.

Using the homogeneity assumption we have made in 2.3, we have:

$$\begin{aligned} \max_i \tilde{S}_1^{(i)} &= \max_i(\tilde{s}_1^{(f(1,i))} + \dots + \tilde{s}_1^{(f(2D,i))}) \\ &\leq \max_i \tilde{s}_1^{(f(1,i))} + \dots + \max_i \tilde{s}_1^{(f(2D,i))}, \end{aligned}$$

where $f(d, i)$ is the index of the d-th router on the path from the source to receiver i and then complete aggregation.

For every max we can apply Corollary 2 of appendix C, so that we have sum of $2D$ functions that can be expanded as $\ln(N)(1 + o(1))$ multiplied by $\mathbb{E}[s_1^{(f(i,1))}], \dots, \mathbb{E}[s_1^{(f(2D,i))}]$, respectively, so that the sum can be expanded in the same way with a multiplicative constant equal to $\mathbb{E}[s_1^{(f(1,i))})] + \ldots + \mathbb{E}[s_D^{(f(2D,i))})] = \mathbb{E}[S_1^{(i)}]$.

□

Theorem 7 (Lower Bound) *Consider a network with exponential service times in routers, and assume that receivers are distincts (ie. the last link for every receiver is different); then γ is bounded from below by a function that can be expanded as*

$$\frac{\mathbb{E}[s]}{W} \ln(N)(1 + o(1)) \text{ for } N \to +\infty, \tag{13}$$

where N is the number of receivers, W the window, and s a typical service time in a router [9]

[9] Due to the homogeneity hypothesis, this is the same distribution for all receivers.

Proof Let us consider the packets $W, 2W, 3W, \ldots$. Since window has a fixed size, packet kW cannot start until the packet $(k-1)W$ has arrived in all receivers (which is the definition of $\|Y_{(k-1)W}\|$). Since we then need to forward packet kW from the sender to all receivers to reach time $\|Y_{kW}\|$), we have :

$$\|Y_{mW}\| \geq \|Y_{(m-1)W}\| + \max_{i=1\ldots N} S_{mW}^{(i)},$$

where $S_m^{(i)}$ is the sum of service times for packet m, along the path from source to itself that passes through receiver i.

So that we have, using $\gamma = \lim_{m\to+\infty} \frac{\|Y_{mW}\|}{mW}$,

$$\gamma \geq \lim_{m\to+\infty} \frac{1}{mW}[\max_{i=1\ldots N}(S_W^{(i)}) + \ldots + \max_{i=1\ldots N}(S_{mW}^{(i)})]$$

Now, as $(\max_i S_{mW}^{(i)})_{m\in\mathbb{N}}$ are i.i.d random variables, the law of large number gives us the inequality :

$$\gamma \geq \frac{\mathbb{E}[\max_{i=1\ldots N}(S_1^{(i)})]}{W}. \tag{14}$$

If RTTs were independent for all receivers, we would be able to conclude immediately that there is an asymptotic behavior in $\ln(N)$. But this is not true as the RTTs of two receivers are made of a first common term which is the sum of the service times of the common routers they use from the source and of a second term, which is independent for each receiver.

Now for each receiver, there is at least a link that belongs only to the path from the source (this is indeed the last link). These links are supposed to have independent service times with the same exponential law s so that, applying Corollary 2, Equation (14) leads to the relation of the theorem.

□

B Association

Definition and construction property: A set of real random variables is said to be "associated" if for every $n \in \mathbb{N}$, and for every subset of cardinality n $X_1, \ldots, X_n$ of random variables in this set, and for all functions f and g increasing in each of its variables, we have

$$\begin{array}{c}\mathbb{E}[f(X_1,\ldots,X_n)g(X_1,\ldots,X_n)] \geq \\ \mathbb{E}[f(X_1,\ldots,X_n)]\mathbb{E}[g(X_1,\ldots,X_n)].\end{array} \tag{15}$$

The three following properties hold, that help us to build associated sets of random variables:

Proposition 2 *1. A set containing a unique random variable is associated;*

2. The union of two independent associated set is associated;

3. If $\{X_1, \ldots, X_m\}$ is associated, and ϕ is increasing in every composant, then the set $\{\phi(X_1, \ldots, X_m), X_1, \ldots, X_m\}$ is associated.

For a proof of this result and of the next two propositions, see [8].

Proposition 3 *let $X_1, \ldots, X_n$ be n associated random variables, and $\tilde{X}_1, \ldots, \tilde{X}_n$ an independent version of it (i.e. $(\tilde{X}_i)_i$ are independent and for all i, X_i and $\tilde{X}_i$ have same law); then we have*

$$P(\max_i X_i \leq t) \geq P(\max_i \tilde{X}_i \leq t) \text{ and} \tag{16}$$

$$\mathbb{E}[\max_i(X_i)] \leq \mathbb{E}[\max_i(\tilde{X}_i)] \tag{17}$$

Association and stochastic order The stochastic order (that we will note $\bar{\leq}$) is defined by

$$X \bar{\leq} Y \text{ if } F_X \geq F_Y \text{ where } F_X(x) = P(X \leq x) \tag{18}$$

We will further in the article need some property of the stochastic order.

Proposition 4 *Assume that the random variables X_i are independent, and that the random variables Y_i are independent too. Then*

1. *Stochastic order is preserved by* max *operation: if for all i, $X_i \bar{\leq} Y_i$, then $\max_i X_i \bar{\leq} \max_i Y_i$.*
2. *Stochastic order is preserved by sum : if for all i, $X_i \bar{\leq} Y_i$, then $\sum_i X_i \bar{\leq} \sum_i Y_i$.*
3. *If the random variables $(X_i)_i$ are associated, we have $\max_i X_i \bar{\leq} \max_i \tilde{X}_i$*

C Maximal Characteristics

We need an analytical tool, giving us the behavior of $\max_i^N X_i$ as a function of N and of the law of X_i, when the X_i's are independent and identically distributed. The maximal characteristics theory of Lai Robbins (described in [7]) provides such results, with a few assumptions on the law of X_i (verified by the exponential case and the Gamma law case).

Theorem 8 (Lai and Robbins maximal characteristics) *Let $(\tilde{X}_m)_{m\in\mathbb{N}}$ be a sequence of $\mathbb{R}_+$-valued i.i.d. random variables. Assume that their common distribution function F satisfies :*

$$\begin{array}{l} (\forall x \geq 0,\ F(x) < 1) \\ (\forall c > 1,\ \lim_{x\to+\infty} \frac{1-F(cx)}{1-F(x)} = 0) \end{array} \tag{19}$$

Let $m_N \hat{=} \inf\{x \geq 0, 1 - F(x) \leq 1/N\}$, then we have

$$\mathbb{E}[\max_{i=1\ldots N} \tilde{X}_i] = m_N(1 + o(1)),\ \textit{for } N \to +\infty. \tag{20}$$

Here are two corollaries (the first of which is immediate):

Corollary 2 (exponential case) *Let $(\tilde{X}_m)_{m\in\mathbb{N}}$ be a sequence of i.i.d exponential r.v.'s with parameter λ, then*

$$\mathbb{E}[\max_{i=1\ldots N} \tilde{X}_i] = \mathbb{E}[\tilde{X}_1] \ln(N)(1 + o(1)),\ \textit{for } N \to +\infty. \tag{21}$$

Corollary 3 (Gamma case) *Let $(\tilde{X}_m)_{m\in\mathbb{N}}$ be a sequence of i.i.d Gamma r.v.'s with parameter (λ, u) where $\lambda > 0$ and $u > 1$, then for $N \to +\infty$,*

$$\begin{array}{rl} \mathbb{E}[\max_{i=1\ldots N} \tilde{X}_i] & = R_u(N)(1 + o(1)) \\ & = \frac{1}{\lambda} \ln(N)(1 + o(1)) \end{array} \tag{22}$$

for $R_u(N)$ function defined by being the solution of the equation

$$X^{u-1} \exp(-\lambda X) = \frac{1}{N} \frac{\Gamma(u)}{\lambda^u}. \tag{23}$$

Proof The distribution function is:

$$F : x \to 1 - \frac{\lambda^u}{\Gamma(u)} x^{u-1} \exp(-\lambda x)$$

F verifies the conditions (19), so that we can apply the previous result, and the formula for m_N gives the definition of R. The fact that $R(N) \sim \frac{1}{\lambda} \ln(N)$ is immediate from (23) when taking the logarithm on both sides.

□

D Proof of Theorem 3

For the sake of simplicity, we will assume that the service times on the backward tree are all equal to 0.

D.1 Preliminary

The r.v.'s $s_m^{(f(i,l))}$ and $s_m^{(f(i',l'))}$ are independent if $l \neq l'$; for $l = l'$, these can be the same random value or two independent random values with the same law depending on the aggregation. All service times of the same depth have same law, that we call s_l for depth l.

Lemma 1 *Let T and T' be two trees with same service times at all levels $(s_1, \ldots, s_D)$ and such that their aggregations satisfy $a \leq a'$; then*

$$\mathbb{E}[S(a)] \geq \mathbb{E}[S(a')],$$

where $S(.)$ is the longest path in the tree :

$$S(a) = \max_{i=1,\ldots,N} (\sum_{l=1}^{D} s_1^{(f(i,l))}).$$

Before proving this lemma, we will need a few definitions and a preliminary lemma. For an aggregation a, let $\approx^a$ be the binary relation on the set of receivers, defined by

$$i \approx^a i' \Leftrightarrow a(i,i') \geq 1.$$

This is an equivalence relation, creating a partition of $\{1, \ldots, N\}$ in $J(a)$ equivalence classes $(\mathcal{C}_j^a)_{j=1\ldots J(a)}$. Each of these classes is made of receivers that share the same first link, so that we can consider for every class $\mathcal{C}_j^a$, the sub-tree a_j obtained by taking the restriction of the tree a to these receivers, and by starting after this common link. So the aggregation of a_j is equal to the restriction of the aggregation $a-1$ to the receivers of $\mathcal{C}_j^a$. The definition of $\approx$ gives us the following result:

Lemma 2 *If a and a' are two aggregations such that $a \leq a'$, then $\approx^a \subseteq \approx^{a'}$, so that for all j, there exists j' such that $\mathcal{C}_j^a \subseteq \mathcal{C}_{j'}^{a'}$.*

From the definition of $S(a)$

$$\begin{aligned} S(a) &= \max_{j=1,\ldots,J(a)} (s_1^{f(1,i_j)} + \max_{i \in \mathcal{C}_j^a} \sum_{l=2}^{D} s_1^{f(l,i)}) \\ &= \max_{j=1,\ldots,J(a)} (s_1^{f(1,i_j)} + S(a_j)) \end{aligned} \tag{24}$$

where i_j is some receiver in $\mathcal{C}_j^a$. Note that the random variables $(s_1^{f(1,i_j)})_j$ are independent, as well as the random variables $(S(a_j))_j$ (because they are made of service times taken in sub-trees that do not share any link). As a consequence of Lemma 2, if $a \leq a'$, we can write (for i_j defined as above)

$$\begin{aligned} S(a') &= \max_{j=1,\ldots,J(a)} (t_1^{f(1,i_j)} + \max_{i \in \mathcal{C}_j^a} \sum_{l=2}^{D} t_1^{f(l,i)}) \\ &= \max_{j=1,\ldots,J(a)} (t_1^{f(1,i_j)} + R(a'_j)) \end{aligned} \tag{25}$$

where $t_1^{f(l,i)}$ denotes the service times in a'; the random variables $R(a'_j)$, $j = 1, \ldots, J(a)$, are associated rather than being independent; similarly the random variables $t_1^{f(1,i_j)}$, $j = 1, \ldots, J(a)$ are associated rather than being independent.

Proof of Lemma 1 Let us show by induction on D that $S(a') \leq_{\text{st}} S(a)$ so that we will have the relation we want on expected values. The result is known for $D = 0$, let us make the induction assumption that $D \geq 1$ and that we have the result shown for $D - 1$.

For all j, the sub-trees a_j and a'_j have the same receivers (the ones in class $\mathcal{C}_j^a$), and the value of their aggregation is $a - 1$, $a' - 1$, respectively, so that we have the same comparison (i.e. $a_j \leq a'_j$), and by induction assumption we have $S(a'_j) \leq_{\text{st}} S(a_j)$.

So, to conclude the proof we just need the following classical result:

Lemma 3 *For $(X_i)_i$ and $(Y_i)_i$ independent, $(\tilde{X}_i)_i$ and $(Z_i)_i$ independent, if $(\tilde{X}_i)_i$ is an independent version of $(X_i)_i$, and if for all i $Y_i \leq_{\text{st}} Z_i$, then*

$$\max_i (X_i + Y_i) \leq_{\text{st}} \max_i (\tilde{X}_i + Z_i).$$

D.2 Proof of the theorem

We do not prove here the result with complete generality. A fundamental example is the case of a window of size 1. In this case the window control mechanism does not allow that two packets go through the tree at the same time. So that the Y_m is just an independent sum of functions like S(a).

If we have comparison of two aggregations a and a', the lemma shown in the first part of the proof tells us that we have stochastic comparison of $S(a)$ and $S(a')$. As independent sums preserve stochastic order, we can compare Y_m for a and for a', and show that the Lyapounov exponents verify $\gamma_{a'} \leq \gamma_a$.

Let us now show the main ideas of the general case. Consider two trees T and T' with the same aggregation but on pair i, j where $a(i, j) = a'(i, j) - 1$. So, in tree T, for some l, the service times $s_m^{f(l,i)}$ and $s_m^{f(l,j)}$ are independent for all m, whereas they are pathwise the same in tree T', all other service times being the same in both trees. If we can prove that $\gamma_a \geq \gamma_b$, then some induction allows one to conclude the proof of the general case.

By expanding the products of random matrices, we can represent $||y_n(a)||$ as the maximum of the sum of service times on certain paths; the set of paths can be split into three disjoint subsets: the set of those paths containing $s_n^{f(l,j)}$; the set of paths containing $s_n^{f(l,i)}$, and the set containing none of them. Using the association calculus, and some conditioning, we prove that the maximum over these sets of paths is larger for the $\leq_{\text{st}}$ order than the maximum over the same sets but when replacing these two independent r.v.'s by the same realization. An induction step in n similar to that of the case $W = 1$ allows one to conclude the proof.

References

[1] S. J. Golestani, K. K. Sabnani. "Fundamental Observations on Multicast Congestion Control in the Internet". Proceeding of IEEE ICNP '99. Ottawa. October 1999.

[2] F. Baccelli, D. Kofman and J. L. Rougier. "Self Organizing Hierarchical Multicast Trees and their Optimization". Proceedings of IEEE Infocom 1999. New York. March 1999.

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

[4] J.-Y. Le Boudec, P. Thiran and S. Giordano. "A short tutorial on Network Calculus I : fundamental bounds in communication networks". Proceedings of ISCAS'2000. Geneva. May 2000.

[5] J.-Y. Le Boudec, P. Thiran and S. Giordano. "A short tutorial on Network Calculus II : min-plus system theory applied to communication networks" Proceedings of ISCAS'2000. Geneva. May 2000.

[6] F. Baccelli and T. Bonald. "Window flow control in FIFO networks with cross-traffic". QUESTA, **32**, Special Issue on Stochastic Stability, pp. 195–231. 1999.

[7] T.L. Lai and H. Robbins. "A class of dependent random variables and their maxima". Z. Wahrscheinlichkeitsch. 1978.

[8] R. E. Barlow and F. Proschan. "Statistical Theory of Reliability and Life Testing". Holt, Rinehart and Winston. New York. 1975.

[9] R. C. Chalmers and K. C. Almeroth. "Validating the Multicast Mystique". Submitted to IEEE Infocom 2001.

[10] L. Rizzo, L. Vicisano and J. Crowcroft. "TCP-like congestion control for layered multicast data transfer". Proceedings of IEEE Infocom 98. San Francisco. March 1998.

[11] D. Rubenstein, J. Kurose and D. Towsley. "The Impact of Multicast Layering on Network Fairness". Proceedings of ACM SIGCOMM 1999. Boston. August 1999.

[12] M. Handley and S. Floyd. "Strawman Congestion Control Specifications". IRTF RMRG report (available on the RMRG web site). December 1998.

[13] IRTF Reliable Multicast research group (RMRG). URL: www.east.isi.edu/RMRG/

[14] I. Stoica, T. S. Eugene Ng and H. Zhang. "REUNITE: A Recursive Unicast Approach to Multicast". Proceedings of IEEE INFOCOM 2000. Tel-Aviv. March 2000.

[15] F. Baccelli and D. Hong. "TCP is $(\max, +)$ Linear". Proceeding of ACM Sigcomm 2000. Stockholm. August 2000.

[16] S. Bhattacharyya, D. Towsley and J. Kurose. "The Loss Path Multiplicity Problem in Multicast Congestion Control" Proceedings of IEEE Infocom 1999. New-York. March 1999.

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

Unité de recherche INRIA Lorraine : Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - B.P. 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 - B.P. 93 - 06902 Sophia Antipolis Cedex (France)

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