

Impact of TCP-Like Congestion Control on the Throughput of Multicast Groups

Augustin Chaintreau, François Baccelli, and Christophe Diot

Abstract—We study the impact of random queueing delays stemming from traffic variability 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 to prove various properties of the throughput. We first prove that the throughput predicted by a deterministic model is systematically optimistic. In the presence of light-tailed random delays, we show that the throughput decreases according to the inverse of the logarithm of the number of receivers. We find analytically an 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 is significantly more sensitive to traffic variability than other topologies.

Index Terms—Classification of tree topology, (max,+) linearity, multicast congestion control, stochastic comparison.

I. INTRODUCTION

TCP-FRIENDLY congestion control has been advocated by the Internet Research Task Force (IRTF) Reliable Multicast Research Group (RMRG),¹ where a TCP-friendly flow is a flow that competes “fairly” with TCP connections. Several recent papers have focused on a TCP-friendly solution for the control of multicast [10]–[12]. In particular, Golestani and Sabnani [1] have made some fundamental observations on multicast flow and congestion control using a deterministic model.

This paper goes a step forward from [1] in providing an understanding of additional properties of TCP-like congestion control in a network with random delays due to cross-traffic variation. This step is of practical importance in that it establishes the dependence of a multicast session’s throughput on the number of receivers, and consequently refines observations learned from 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.* [15] analyzed the impact of TCP-like congestion control on the throughput of a multicast session. They showed that for loss-based additive-increase–multiplicative-decrease (AIMD) algorithms, there is a severe degradation of throughput for large multicast groups.

We extend the findings in [1] and [15] by showing that even in the case of an ideal TCP control when the flow control window size is kept equal to its maximal value and when no losses occur, there remains a severe throughput degradation within a one-to-many multicast tree when the group size grows. Intuitively, in this ideal case, the session throughput is still 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 may 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 information stemming from all receivers, slow receivers will “slow down” the sender.

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

We have chosen to model a multicast session as follows. Packets are sent by a unique sender located at the root of a set of routers organized as a tree to a set of receivers located at the leaves of this tree. This tree is referred to as the *forward tree*.

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

The model captures congestion via the queueing delay that each packet experiences in each router it passes through. More precisely, the fluctuations due to the processing of packets of other (unicast or multicast) connections sharing the same router interface are represented by random service times for packets of the reference multicast connection. These random service times are also called *aggregated service times* (see Section II-B and Fig. 1); we will assume them to be independent in time and space, and light-tailed (i.e., the tail decreases faster than a negative exponential function). The queueing strategy is assumed to be first-in–first-out (FIFO). Within this framework, the sender and the receivers are modeled as routers, possibly with different mean delays and different distributions.

For reasons that have already been explained (i.e., 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

Manuscript received April 3, 2001; revised November 20, 2001; approved by IEEE TRANSACTIONS ON NETWORKING Editor G. Pacifici.

A. Chaintreau was with Sprint Advanced Technology Laboratory, Burlingame, CA 94010 USA. He is now with INRIA and Ecole Normale Supérieure, 75005 Paris, France (e-mail: augustin.chaintreau@ens.fr).

F. Baccelli is with INRIA and Ecole Normale Supérieure, 75005 Paris, France (e-mail: francois.baccelli@ens.fr).

C. Diot is with Sprint Advanced Technology Laboratory, Burlingame, CA 94010 USA (e-mail: cdiot@sprintlabs.com).

Publisher Item Identifier 10.1109/TNET.2002.801420.

¹[Online]. Available: <http://www.east.isi.edu/RMRG/>.

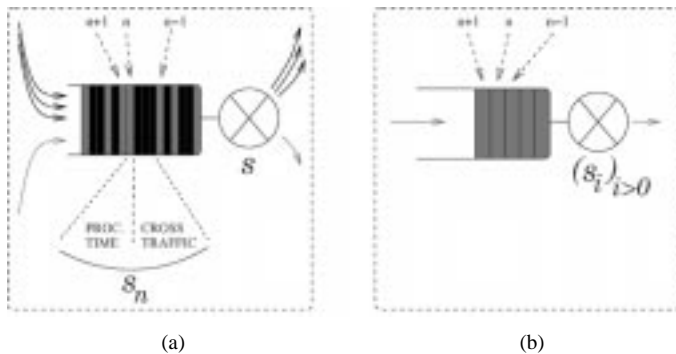


Fig. 1. Representation of a router. (a) A router. (b) Our model.

assume that all routers have infinite buffers and consider that the network is lossless and that the window size is fixed.

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

All the assumptions that we make about the network (tree homogeneity), about transmission control (no losses, window size always equal to its maximal value), and about delays (light-tailed) have been carefully selected to provide an optimistic network environment. We show that even in this favorable context there is a severe decrease of the throughput when the number of receivers increases. Note that, as will be better understood further in the paper, the homogeneous case is also the one where the size of the group has the strongest impact; this point is discussed in Section V.

To the best of our knowledge, this work is the first to address analytically the question of multicast session throughput degradation due to network queueing delays, for different tree topologies, in a TCP-like congestion control environment. Although we limited this first study to some simple cases, we believe that our mathematical methodology can be expanded to analyze more general cases either with adaptive AIMD-type window evolution, or with other situations such as heavy-tailed delays or nonhomogeneous trees, etc., as discussed in Section V.

The paper is structured as follows. In Section II, we build our analytical model on the (max,plus) formalism [3]. Note that this algebraic framework is also the basis of network calculus [4], [5]. However, the stochastic analysis and the bounds that are derived in this paper are quite different in nature from the worst-case analysis of network calculus. In Section III, we derive an algebraic simulator² from our analytical model. Simulations show that the throughput obtained from the deterministic model in [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 IV, we analyze the model using the notion of positive correlation (also called association), as well as the notion of maximal characteristics [7]. The throughput is shown to be upper and lower bounded by functions that decrease according to the inverse of the log-

²Our simulator is based on algebraic operations as found in the model, as opposed to a discrete event simulator such as NS (see Section III-A).

arithm of the number of receivers. This qualitative result primarily stems from the light-tail assumption on delays and holds for quite general (nondegenerate) tree topologies. It 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], [13] which we call *umbrella trees*. We show that this class of trees is significantly more sensitive to variability than other topologies, and that in some cases, these topologies reach the lower bound. We finally characterize the throughput degradation curve for a class of optimal trees called *reverse-umbrella trees*.

II. (MAX,PLUS) REPRESENTATION

We first introduce the (max,plus) algebra, show how it can be used to represent a connection through a network on a simple example, and apply this representation to multicast.

A. Introduction to the (max,plus) Algebra

We first 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 \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, identity elements,³ and distributivity ($\forall a, b, c \in \mathbb{R}_{\max}, a + (b \vee c) = (a + b) \vee (a + c)$).

Since $(\mathbb{R}_{\max}, \vee, +)$ is a semi-ring, we can construct matrix operations as in the conventional algebra, with the addition of matrices obtained via term-by-term maximization, and multiplication defined by the rule $(\mathbb{A}\mathbb{B})_{i,j} = \max_k (\mathbb{A}_{i,k} + \mathbb{B}_{k,j})$. We 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 $\|\cdot\|$ denote the matrix norm $\|\mathbb{A}\| = \max_{i,j} (\mathbb{A}_{i,j})$.

Product of a large number of matrices, Lyapunov exponent: We start from the following result shown in [3].

Theorem 1: Let $(\mathbb{A}_n)_{n \geq 0}$ be a sequence of random square matrices in \mathbb{R}_{\max} independent with the same law and with coefficients that are either ε with probability 1, or with finite expectation; then we have

$$\lim_{m \rightarrow \infty} \frac{\|\mathbb{A}_m \mathbb{A}_{m-1} \cdots \mathbb{A}_1\|}{m} = \gamma \quad (1)$$

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

In the following, we will make use of the following corollary where the assumptions are those of the above theorem; the dimension of the matrices is K ; Y_0 is a vector of dimension K

³Operation \vee acts as a new addition, whose zero is the element $-\infty$ (denoted by ε), and operation $+$ acts as a new multiplication, with the element 0 acting as a new unit (which we denote by e).

with its first $L \leq K$ entries equal to e and all others equal to ε . $(Y_m)_{m \geq 0}$ is defined by the (max,plus) linear recurrence

$$Y_m = \mathbb{P}_m Y_{m-1}, \quad \text{for } m > 0. \quad (2)$$

Corollary 1: If for all m , $\|\mathbb{P}_m\|$ is reached by an entry of \mathbb{P}_m within the first L lines and columns, then we have

$$\lim_{m \rightarrow \infty} \frac{\|Y_m\|}{m} = \gamma \quad (3)$$

in expectation and with probability 1, where γ is the Lyapunov exponent of the sequence $(\mathbb{P}_m)_{m \geq 0}$.

Proof: The property of $(\mathbb{P}_m)_{m \geq 0}$ is stable under multiplication of the matrix, so that $\mathbb{P}_m \mathbb{P}_{m-1} \cdots \mathbb{P}_1$ has this property. Therefore, it is easy to verify that the largest element of $\mathbb{P}_m, \mathbb{P}_{m-1}, \dots, \mathbb{P}_1$ is equal to the largest element of $\mathbb{P}_m, \mathbb{P}_{m-1}, \dots, \mathbb{P}_1 Y_0$, so we have $\|Y_m\| = \|\mathbb{P}_m, \mathbb{P}_{m-1}, \dots, \mathbb{P}_1\|$ and the conclusion follows from the theorem. ■

B. (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 L routers (numbered 1 to L) where the window flow control has a fixed-size window W . (This model and its multicast extension were introduced in [6].) The sender is incorporated into the first router, and the receiver into the last router. The multicast model we will present in Section II-C is a simple extension of this preliminary example.

Routers in the network receive packets belonging to both the reference connection and other connections sharing the same interface. Each router is represented in our model as a FIFO queue with an infinite buffer⁴ containing only packets of the reference connection. Each one is given a random service time that is also called *aggregated service time* because it includes the processing time of cross traffic packets. As shown in Fig. 1, where packets belonging to other connections are colored in black, the aggregated service time used in the queue of our model contains the processing time of the reference packet and the time taken by packets intervening between two packets of the reference connection.

Assuming that the connection under consideration stabilizes, it is natural to make the assumption that the aggregated service times of our router are identically distributed for the different packets of the connection. We also assume aggregated service time independence for the sake of simplicity, i.e., aggregated service times for different routers in the network are independent, and the sequence of aggregated service times on a router is made up of independent and identically distributed random variables. This assumption is critical for the type of degradation that is established in this paper; however, it can be significantly weakened for many other aspects like the representation of the network via products of random matrices and the subsequent characterization of throughput.

⁴Note that since the buffers are of infinite size, no loss occurs. As a result, the window size is assumed to reach its maximal value and to remain constant. Nevertheless, congestion (i.e., large aggregated service time in routers) plays a key role in our model.

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

- Router $i > 1$ starts processing packet m as soon as it has finished processing packet $m-1$, and the upstream router has forwarded packet m . After it has started processing this packet, $s_m^{(i)}$ units of time are still required to process it. This processing time actually includes the processing time of all the packets of the other connections interleaved between packet $n-1$ and packet n of the reference connection. So we have for $i > 1$

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

- The sender (considered as router $i = 1$) sends packet m as soon as it has finished with packet $m-1$, provided that the window control allows packet m to be sent. (This is the meaning behind the assumption that the source is saturated, namely, it always has packets to send). We now have

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

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 LW with blocks $X_m, X_{m-1}, \dots, X_{m-W+1}$. We can capture the dynamics of the network by a (max,plus) linear recurrence

$$\begin{cases} Y_0 = & \text{the vector with all its coordinates equal to } e \\ Y_m = & \mathbb{P}_m Y_{m-1} \quad \text{for } m > 0 \end{cases} \quad (4)$$

where the matrix \mathbb{P}_m has the following block structure (each block is a square block of dimension L):

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

$$\mathbb{S}_m = \begin{pmatrix} s_m^{(1)} & \varepsilon & \varepsilon & \cdots & \varepsilon \\ s_m^{(1)} + s_m^{(2)} & s_m^{(2)} & \varepsilon & \cdots & \varepsilon \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ s_m^{(1)} + \cdots + s_m^{(L)} & \cdots & \cdots & \cdots & s_m^{(L)} \end{pmatrix}.$$

- \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 $i = 1, \dots, L$ and $j = L$.
- \mathbb{S}_m represents the forwarding mechanism in the network, and $(\mathbb{S}_m)_{i,j}$ is more generally given by the maximum over all paths leading from i to j of the sum of aggregated service times for packet m on the path from router j to router i (including both i and j).

Note that if aggregated service times are independent and identically distributed, then the matrices $(\mathbb{P}_m)_{m \geq 0}$ are also independent and identically distributed; we can then apply Corollary 1, which gives the existence of $\lim_{m \rightarrow \infty} (\|Y_m\|/m) = \gamma$ both in

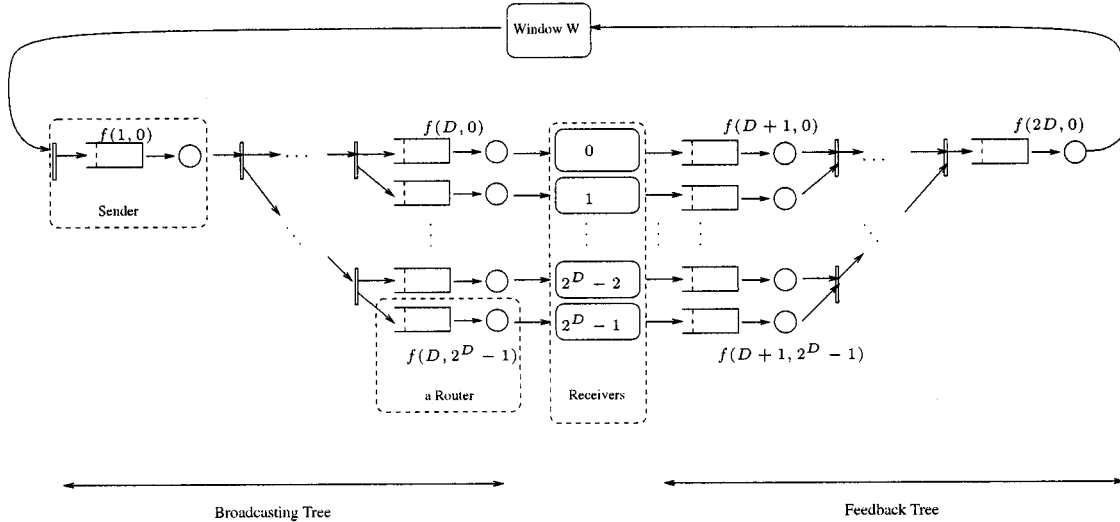


Fig. 2. “Forward and backward” graph.

expectation and with probability 1. γ is called the *Lyapunov exponent* of this sequence of matrices. Since $\|Y_m\|$ represents the epoch when packet m has arrived at its destination, the Lyapunov exponent represents the mean delay between the reception times of two successive packets and $1/\gamma$ is, therefore, the *average throughput*, i.e., the total amount of data transmitted since the beginning of the session divided by the duration of the session. In the following section, we extend this model to multicast.

C. Representation of Multicast Flow Control

A single source broadcasts packets over a unidirectional tree to N receivers. Each node in the forward tree simultaneously duplicates and forwards each packet on the downstream branches. Acknowledgments (acks) are forwarded back to the source through a feedback tree that is a mirror version of the forward tree. We assume that the feedback tree is functionally independent of the forward tree. In the feedback tree, the ack of a packet only arrives in router i when the latest ack of the same packet has been sent by the routers upstream. It is then transmitted by router i after some queuing delay. The aggregation of acknowledgments in each router of the feedback tree allows one to take care of the implosion problem (see [1]). The case of a binary tree is shown in Fig. 2.

The flow control is enforced by the sender; it is again based on a sliding window mechanism, of constant size W . The sender only sends packet $n + W$ when the ack of packet n has been received from the final router of the feedback tree.

We have chosen to model homogeneous trees only. Homogeneity means that the path from the sender to each receiver is statistically the same for all receivers, i.e., there are the same number of routers, and the aggregated service time distribution is the same for all routers at the same level.⁵

We still denote by L the total number of routers in the network, and D the depth of the forward tree. For receiver i , we let $f(1, i)$, $f(2, i)$, \dots , $f(D, i)$ denote the different routers on

the path from the sender to this receiver in the forward tree, and let $f(D + 1, i)$, $f(D + 2, i)$, \dots , $f(2D, i)$ denote the different routers in the feedback tree transmitting acknowledgments from receiver i back to the sender. By definition, the *path* of receiver i is the sequence $f(1, i)$, $f(2, i)$, \dots , $f(2D, i)$. For router $f(d, i)$, we denote by $s_n^{(f(d, i))}$ the aggregated service time of the n th packet on this router.

With this notation

$$S_m^{(i)} = s_m^{(f(1, i))} + \dots + s_m^{(f(2D, i))} \quad (5)$$

is the (minimal) round-trip time (RTT) of packet m on the path that contains receiver i .

Homogeneity is an important difference from the assumptions in Golestani’s model [1]. Given his conclusion that receivers should have a window size proportional to their distance from the source, it makes sense to consider a homogeneous tree with a single window.

Note that homogeneity allows for quite complex tree structures. Homogeneous trees are complex enough to illustrate the properties we want to stress. They also make the comparisons between different topologies easier.

The network model described earlier can be written in a way similar to that of Section II-B. 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 router of level 0 in the tree (the source), and the last entry corresponds to the router of the highest level ($2D$), at the end of the feedback tree, which can be seen as that of the final aggregation. Let Y_m be the block vector of dimension LW built on top of $(X_m)_{m \geq 0}$ and which captures the history of X_m in the same way as above. We have the same (max,plus) linear system for Y_m as in (4), though with different matrices.

\mathbb{P}_m has the same block structure as before. The block \mathbb{W}_m is defined as follows: if $f(d, i)$ is router l , then $(\mathbb{W}_m)_{l, L} = \max_{j \in \mathcal{R}(l)} s_m^{(f(1, j))} + \dots + s_m^{(f(d, j))}$, where $\mathcal{R}(l)$ is the set of receivers whose path contains router l , and all other rows are ε .

The block \mathbb{S}_m is again the maximum over all paths from l' to l of the sums of the different aggregated service times (of order

⁵Two routers in the graph have the same level if they are the same distance from the sender.

m) along the path (with the maximum over an empty set equal to ε by convention).

The sequence $(\mathbb{P}_m)_{m \geq 0}$ is again independent identically distributed (i.i.d.), which allows us to deduce from Corollary 1 the existence of the Lyapunov exponent, which is given by $\lim_{m \rightarrow \infty} (\|Y_m\|/m)$ and that represents the mean delay between the reception of two successive packets in receivers, or, equivalently, the inverse of the averaged throughput of the connection.

III. SIMULATION RESULTS

The simulator described below is based on products of matrices and vectors; the product by one more random matrix provides the exact emulation of the transmission of one more packet through the whole network under the control assumptions previously described. This simulator is equivalent to a discrete event simulator. Its advantages with respect to discrete event simulation are: 1) its algebraic nature makes it of lower complexity, which is important when the number of receivers becomes large, and 2) the same formalism is used in the simulation and in the analytical studies.

A. Description of the Simulator

We can compute the Lyapunov exponent which is the inverse of the average throughput for the connection, and which can be obtained from the simulator as the almost sure limit $\gamma = \lim_{m \rightarrow \infty} (\|Y_m\|/m)$. In practice, we can estimate γ by $\|Y_M\|/M$ for a large enough value of M . The numerical simulator samples different random variables for aggregated 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).

As far as the simulation is concerned, there is a natural computational tradeoff between the accuracy of the estimation of the Lyapunov exponent and the simulation of multicast groups with a large number of receivers. The accuracy of the throughput estimator requires the simulation of a large number of packets, or, equivalently, the computation of the product of a large number of matrices, whereas a large group implies 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. In most of the simulation runs, we took $M = 400$. This choice results in rather nonsmooth shapes for most of the simulation curves produced below. However, as we see in the next section, this is sufficient to estimate the general shape and the relative ordering between the curves.

1) *Modeling the Topology*: In addition to the homogeneity assumption that we stressed in Section II, we further assume that all aggregated service times in the feedback tree are zero (i.e., as soon as all copies of packet m have reached the receivers, the sender instantaneously receives the acknowledgment of packet m). It corresponds to an optimistic scenario. Our model can easily take into account the case where acknowledgments

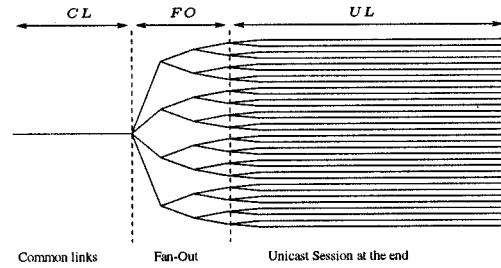


Fig. 3. Trees generic topology.

suffer random delays in the network before being received by the sender; nevertheless, we do not consider this in our paper.

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

We need first to vary the number of receivers of a multicast session to make it possible to study how the throughput varies with the size of the group. For every binary tree of size 2^D , we consider a set of N “active” receivers that is a subset of the leaves of the complete binary (forward) tree ($N \leq 2^D$). For this, we simply set the aggregated 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 aggregated service times to analyze the sub-binary tree corresponding to this subset of N leaves.

The tree topologies we study are represented in Fig. 3. 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 (degree⁶ 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 duplication 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 Fig. 4. In addition to complete binary trees, we consider:

- *Umbrella trees*. These trees end with a long unicast transmission after a short fan-out (large value of UL). The limiting case is that with one independent path from the source to each receiver. It is characteristic of a multicast tree where the receivers share only a few links. This kind of topology is identified in [9] and [13] as being often found in Mbone sessions.
- *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 the receivers’ behavior differs only by a few links.

These categories will be more precisely defined and analytically studied in Section IV-C.

⁶To emulate any topology from a binary tree, one can set aggregated service time deterministically null in some of the links; this is equivalent to not taking these links into account.

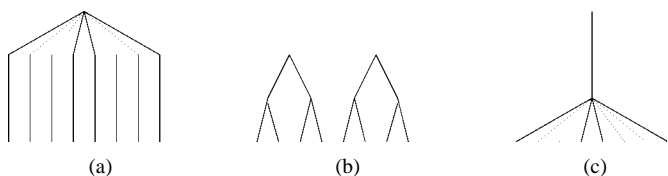


Fig. 4. Fundamental types of tree topologies observed. (a) Umbrella tree. (b) Complete binary tree. (c) Reverse umbrella tree.

B. Average Throughput Versus Number of Receivers

This section focuses on the degradation of performance (i.e., increase of the Lyapunov exponent) when the number of receivers increases in the 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 compute the same simulation (which gives the value of γ for two receivers). Then we progressively fill the tree with more and more receivers.

We simulate different ways of filling in the tree. “Best filling” consists in starting from receiver 1 (numbers refer to Fig. 2) 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. In what follows, the default option is random filling.

Simulation results are shown in Figs. 5–12. The aggregated service times in the routers follow an exponential law with the same parameter (λ) for each router in the network, so that the homogeneity condition is satisfied. In each simulation, λ is chosen in such a way that the sum of the aggregated service times along a path from the sender to any receiver has a mean value equal to 1 ($\lambda = D$).

Note that for homogeneous trees with deterministic aggregated service times, there is no dependency of the throughput on the number of receivers, since each receiver has the same RTT and behaves synchronously with other receivers in the multicast group. For each plot, we have represented the value of the throughput in the deterministic case as found by Golestani [1], which is equal to 1 for this choice of λ .

Fig. 6 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 aggregated service time with mean value 1. The feedback tree has a null aggregated service time on all routers.

The first important observation is that the average throughput decreases according to the inverse of the logarithm of the number of receivers. To verify that the throughput decreases logarithmically with the number of receivers, we plot in Figs. 6–12 the number of receivers on a logarithmic scale (x axis) and, on the vertical axis, the Lyapunov exponent instead of the throughput. Above two receivers, each curve is close to a straight line, which indeed shows the logarithmic nature of the throughput decrease. This is completely different from what is obtained with a deterministic approach which seems to give a pretty optimistic evaluation of the throughput (represented by the horizontal line). This observation will be verified analytically in the next section.

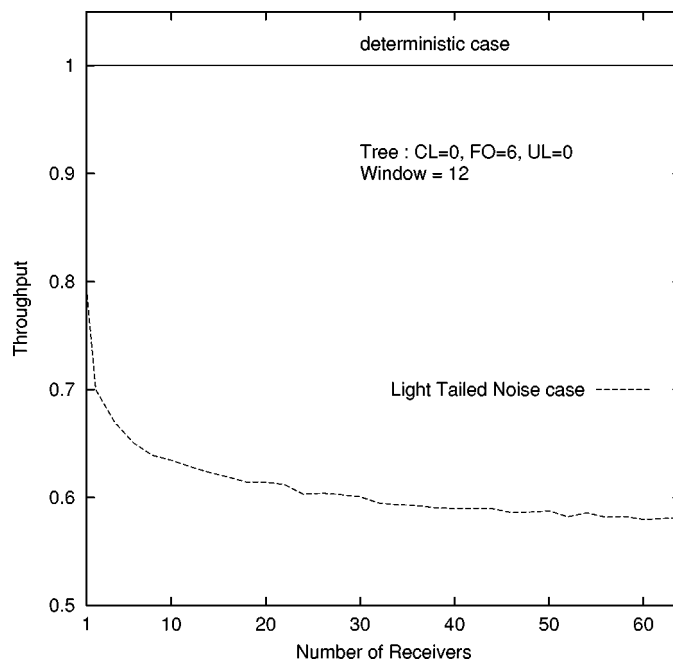


Fig. 5. Average throughput versus number of receivers in a binary tree with light-tailed delays.

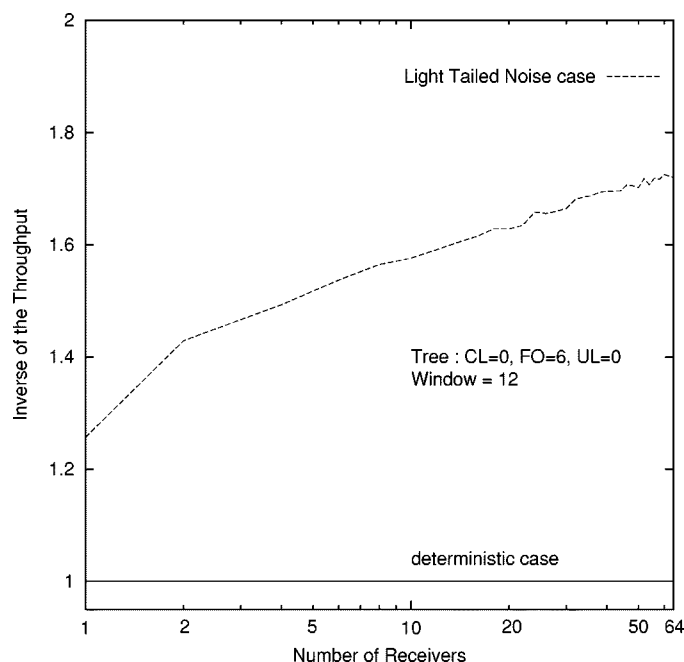


Fig. 6. Same experiment as Fig. 5: inverse of the throughput versus number of receivers in logscale.

The important throughput drop between one and two receivers can be explained by the homogeneous nature of the tree that leads the second receiver to join the tree with a path of length $CL + FO + UL$. Then the throughput keeps decreasing significantly until there are 20 participants. Between 40 and 60 receivers, the throughput stabilizes around 50% of the deterministic case.

Another important remark is that even when there is only one receiver, the throughput obtained by the stochastic model is sig-

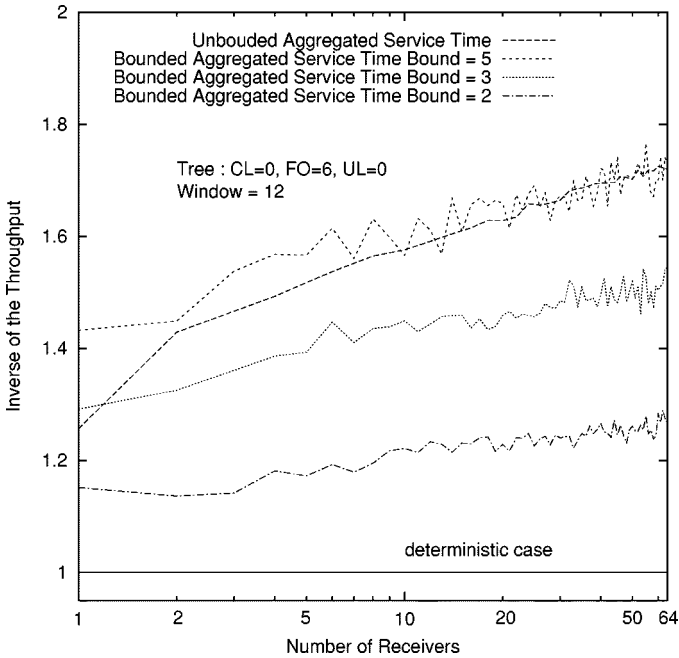


Fig. 7. Simulation for bounded and unbounded aggregated service time.

nificantly less than that of the deterministic model. For a theoretical explanation based on convex ordering, see Section IV-B-1.

The same kind of throughput degradation has also been observed in [15], where it was a consequence of the overestimation of the loss probability by TCP running over a tree instead of a route. In [15], the nature of the throughput degradation resulted from the binary assumption on the tree topology. Here, the reasons for degradation are quite different, as the window of TCP is supposed to be fixed and no loss occurs. As we shall see, a similar degradation is observed for all topologies, as soon as variability of the cross traffic leads to light-tailed aggregated service time in the network.

1) *Influence of the Stochastic Assumptions:* Before further investigating the shape of the throughput degradation, we have to verify that the shape of the degradation is not a direct consequence of the nature of the network delays.

Fig. 7 gives throughput as a function of the number of receivers for aggregated 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 are not particularly meaningful.

The most interesting remark to be made comes from 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 a throughput that is quite close to the unbounded case. Thus, we can conclude from these curves that the shape of the degradation is not bound to the exponential assumption. Our observation is rather that all distribution functions with a tail bounded from above by a negative exponential function lead to such a decrease for some finite range provided variance is nonzero. For heavier tails (e.g., Pareto tails), preliminary results seem to suggest that the growth of the Lyapunov exponent is polynomial. So, the bounded support and the light-tail cases are

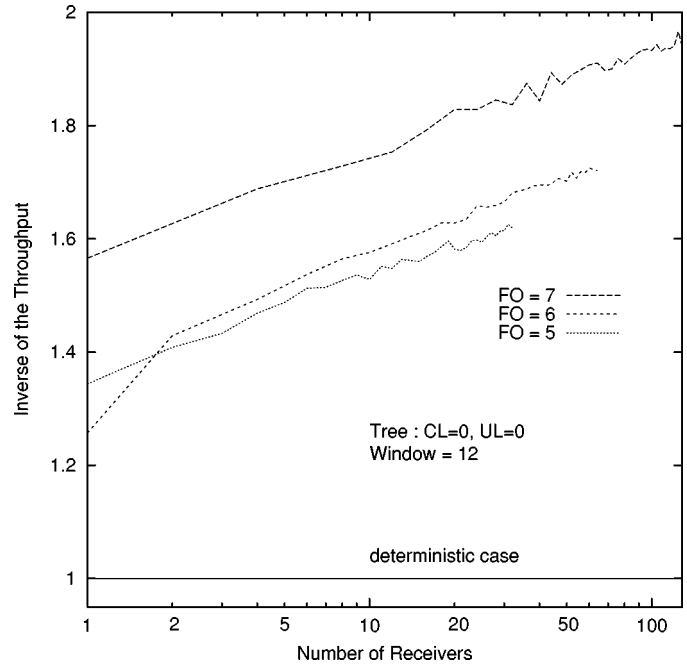


Fig. 8. Simulation for various tree depths.

qualitatively the same, at least when variance is not too small, and this generic case seems to be the most favorable when compared with heavier Pareto-type tails.

In the following simulations, we use unbounded aggregated service times.

2) *Analysis of Various Network Parameters:* In order to understand the impact of network parameters on the throughput degradation, we have varied network parameters. Fig. 8 shows how the throughput decreases for various tree depth values. When the trees are homogeneous and aggregated service times all have the same distribution, the 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, the faster the throughput decrease. We have chosen a default size of 6, which allows us to simulate a sufficiently large number of receivers.

Fig. 9 focuses on the influence of the window size. We have chosen 12 as a default value; this value is sensible and it keeps simulation times low enough.

Finally, we checked, as shown in Fig. 10, the influence 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 minimal possible number of links.

C. Analysis of the Tree Topology

We now simulate the tree topologies described in Section III-A-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 the same

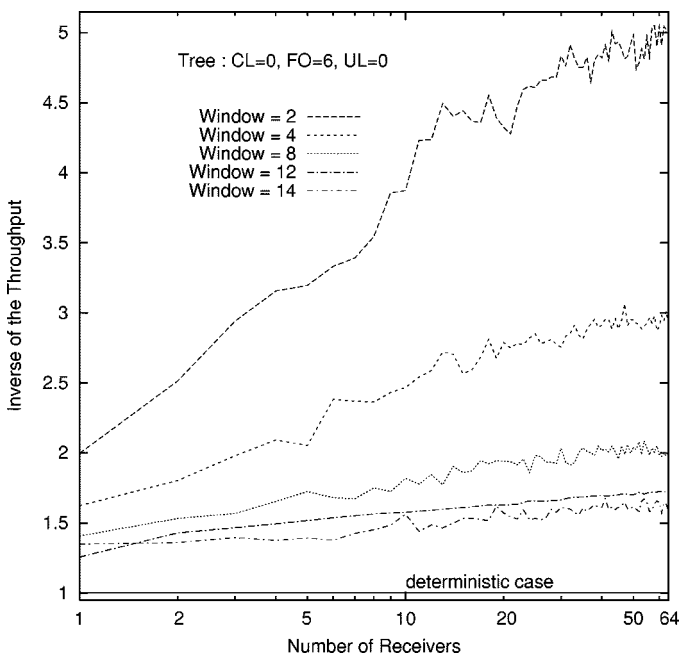


Fig. 9. Simulation for various window sizes.

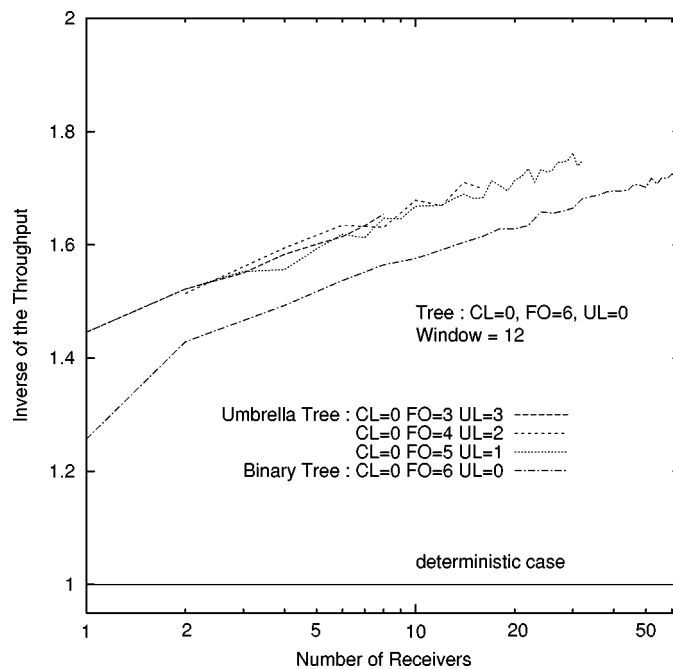


Fig. 11. Simulation in the case of umbrella trees.

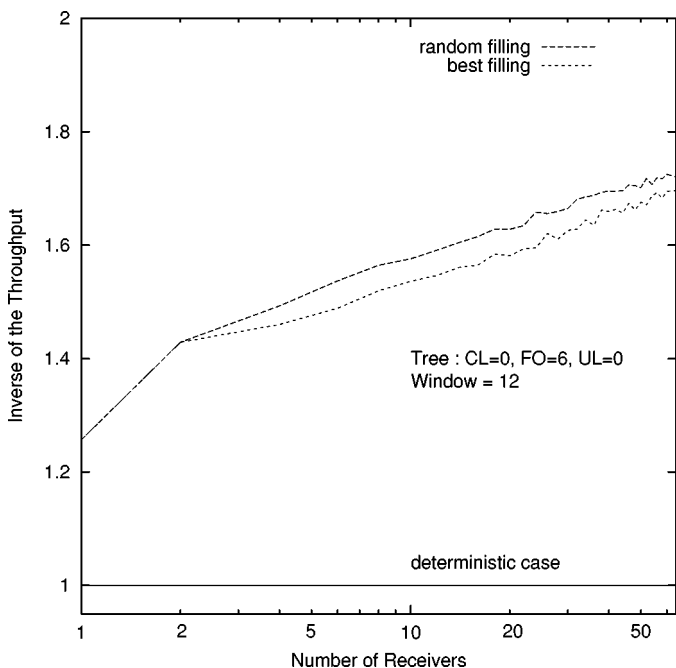


Fig. 10. Simulation for different tree construction approaches.

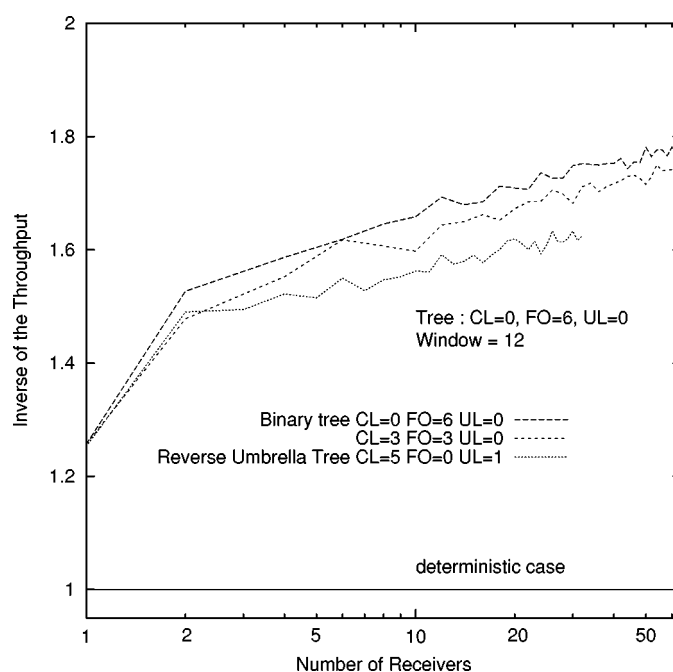


Fig. 12. Simulation in the case of reverse umbrella trees.

length would have the same performance (depending only on the RTT). Fig. 11 plots various umbrella trees and Fig. 12 plots various reverse-umbrella trees. In both case, the binary tree case is given as a reference.

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

We also observe that the closer the fan-out from the receivers, the higher the throughput. Thus, trees where receivers share few links are much more sensitive to the number of receivers in the

group. The throughput of an umbrella tree has already decreased to 78%, compared with the case with one receiver, for a group made of seven receivers, while a reverse-umbrella tree reaches the same throughput for 19 receivers. A throughput decrease of 81% (still compared with the case with one receiver) is reached with an umbrella tree for three receivers; with a binary tree, the same rate is obtained for seven receivers and with a reverse umbrella tree for ten receivers. For a group with three receivers, the throughput degradation (still compared with the case with one receiver) is 36% higher for an umbrella tree than for a binary tree. This observation is very important, as the current Internet

topology seems to favor umbrella trees. Note that such trees were also shown to be optimal in terms of network resource consumption (for more detail, see [2]).

IV. MATHEMATICAL ANALYSIS OF THROUGHPUT

In this section, our goal is first to give mathematical arguments substantiating the growth of the Lyapunov exponent in $\ln(N)$ that was observed in the simulations. Second, we give mathematical arguments explaining why and how certain trees or certain situations should compare in a predictable way.

In order 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, and the notion of *maximal characteristics*, which comes from the theory of extremes, and which will allow us to get explicit bounds on the performances.

A. Mathematical Tools

1) Association and Stochastic Order:

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, \dots, X_n of random variables in this set, and for all functions f and g increasing in each of its variables, we have

$$\mathbb{E}[f(X_1, \dots, X_n)g(X_1, \dots, X_n)] \geq \mathbb{E}[f(X_1, \dots, X_n)]\mathbb{E}[g(X_1, \dots, X_n)]. \quad (6)$$

The three following properties hold, helping us to build associated sets of random variables.

Proposition 1:

- 1) A set containing a unique random variable is associated.
- 2) The union of two independent associated sets is associated.
- 3) If $\{X_1, \dots, X_m\}$ is associated, and ϕ is increasing in every component, then the set $\{\phi(X_1, \dots, X_m), X_1, \dots, X_m\}$ is associated.

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

Proposition 2: Let X_1, \dots, X_n be n associated random variables, and $\tilde{X}_1, \dots, \tilde{X}_n$ an independent version of it (i.e., $(\tilde{X}_i)_{i=1..n}$ are independent and, for all i , X_i and \tilde{X}_i have same law); then we have

$$\text{for all } u, P\left(\max_i X_i \leq u\right) \geq P\left(\max_i \tilde{X}_i \leq u\right) \quad (7)$$

and

$$\mathbb{E}\left[\max_i(X_i)\right] \leq \mathbb{E}\left[\max_i(\tilde{X}_i)\right]. \quad (8)$$

Association and stochastic order: The stochastic order (that we note \leq_{st}) is defined by

$$X \leq_{st} Y, \quad \text{if } F_X \geq F_Y \text{ where } F_X(u) = P(X \leq u). \quad (9)$$

In particular, we can write, due to the last proposition, that if the random variables $(X_i)_i$ are associated, then $\max_i X_i \leq_{st} \max_i \tilde{X}_i$.

Later in this paper, we will use the following properties of the stochastic order.

Proposition 3: Assume that the random variables $(X_i)_{i=1..n}$ are independent, and that the random variables $(Y_i)_{i=1..n}$ are also independent. Then:

- 1) Stochastic order is preserved by max operation: if for all i , $X_i \leq_{st} Y_i$, then $\max_i X_i \leq_{st} \max_i Y_i$.
- 2) Stochastic order is preserved by sum: if for all i , $X_i \leq_{st} Y_i$, then $\sum_i X_i \leq_{st} \sum_i Y_i$.

2) **Maximal Characteristics:** We need an analytical tool that will give us the behavior of $\max_i^N X_i$ as a function of N and of the law of variables $(X_i)_{i \geq 0}$, when they are independent and identically distributed. The maximal characteristics theory of Lai and Robbins [7] provides such results, with a few assumptions on the law of X_i (satisfied by the exponential case and the Gamma law case).

Theorem 2 (Lai and Robbins Maximal Characteristics): Let $(\tilde{X}_m)_{m \geq 0}$ be a sequence of \mathbb{R}_+ -valued i.i.d. random variables. Assume that their common cumulative distribution function F satisfies

$$\begin{aligned} & (\forall u \geq 0, \quad F(u) < 1) \\ & \left(\forall c > 1, \quad \lim_{u \rightarrow +\infty} \frac{1 - F(cu)}{1 - F(u)} = 0 \right). \end{aligned} \quad (10)$$

Let $u_N \triangleq \inf\{u \geq 0 \mid 1 - F(u) \leq 1/N\}$, then we have

$$\mathbb{E}\left[\max_{i=1..N} \tilde{X}_i\right] = u_N(1 + o(1)), \quad \text{for } N \rightarrow +\infty. \quad (11)$$

We deduce two corollaries from this theorem:

Corollary 2 (Exponential Case): Let $(\tilde{X}_m)_{m \geq 0}$ be a sequence of i.i.d. exponential r.v.s with parameter λ ; then

$$\mathbb{E}\left[\max_{i=1..N} \tilde{X}_i\right] = \mathbb{E}[\tilde{X}_1] \ln(N)(1 + o(1)), \quad \text{for } N \rightarrow +\infty. \quad (12)$$

Corollary 3 (Gamma Law Case): Let $(\tilde{X}_m)_{m \geq 0}$ be a sequence of i.i.d. Gamma random variables with parameter (λ, l) where $\lambda > 0$ and l is an integer larger than or equal to 1; then for $N \rightarrow +\infty$

$$\mathbb{E}\left[\max_{i=1..N} \tilde{X}_i\right] = R_{l,\lambda}(N)(1 + o(1)) = \frac{1}{\lambda} \ln(N)(1 + o(1)) \quad (13)$$

where $R_{l,\lambda}(N)$ is the unique solution (in u) of

$$\exp(-\lambda u) \left(\sum_{k=0}^{l-1} \frac{\lambda^k u^k}{k!} \right) = \frac{1}{N} \quad (14)$$

located in the interval $(0, 1)$.

Proof: The distribution function is

$$F: u \rightarrow 1 - \left(\sum_{k=0}^{l-1} \frac{\lambda^k u^k}{k!} \right) \exp(-\lambda u).$$

F verifies the conditions in (10), so that we can apply the previous result, and the formula for u_N gives the definition of $R_{l,\lambda}$. The fact that $R_{l,\lambda}(N) \sim (1/\lambda) \ln(N)$ for $N \rightarrow +\infty$ is immediate from (14) by taking the logarithm on both sides. \square

B. Bounding Throughput Degradation

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

1) Comparison With the Deterministic Case:

Proposition 4: 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 only 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 $\mathbb{E}[\max(X, Y)] \geq \max(\mathbb{E}[X], \mathbb{E}[Y])$ by a direct convexity argument. \square

Let us interpret this result for the $(\max, +)$ system we define in our model. We denote by $\bar{\mathbb{P}}_m$ the matrix describing the same network as above, but this time with each random aggregated service time replaced by its mean value. In particular, $\bar{\mathbb{P}}_m$ is no longer a random variable, but a deterministic matrix. We denote the sequence $(\bar{Y}_m)_{m \geq 0}$ given by the deterministic recurrence equation $\bar{Y}_m = \bar{\mathbb{P}}_m \bar{Y}_{m-1}$, and \bar{Y}_0 is given as before by a vector whose entries are all $e = 0$. This is the sequence $(Y_m)_{m \geq 0}$ when every delay in the network is assumed to be deterministic. Defining $\bar{\gamma}$ as being the Lyapunov exponent given by this sequence, $1/\bar{\gamma}$ represents the throughput of the session as given by a deterministic model.

We have $\bar{\mathbb{P}}_m = \mathbb{E}[\bar{\mathbb{P}}_m] \leq \mathbb{E}[\mathbb{P}_m]$ and using Proposition 4, we find

$$\begin{aligned} \bar{Y}_m &= \bar{\mathbb{P}}_m \cdots \bar{\mathbb{P}}_1 Y_0 \leq \mathbb{E}[\mathbb{P}_m] \cdots \mathbb{E}[\mathbb{P}_1] Y_0 \\ &\leq \mathbb{E}[\mathbb{P}_m \cdots \mathbb{P}_1 Y_0] = \mathbb{E}[Y_m] \end{aligned} \quad (15)$$

which implies $\bar{\gamma} \leq \gamma$. Hence, Golestani's deterministic model [1] 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.

2) Upper Bound:

Theorem 3: Consider a network with exponential aggregated service times in routers; then γ is bounded from above by a function that can be expanded as

$$\text{RTT} \ln(N)(1 + o(1)) \quad \text{for } N \rightarrow +\infty \quad (16)$$

where N is the number of receivers and RTT is the average minimal round-trip time of a receiver [$\text{RTT} = \mathbb{E}[S_1^{(1)}]$], where $S_1^{(1)}$ is defined in (5).

Proof: We have

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

The strong law of large numbers shows that with probability 1

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

We now use the interpretation we have on the elements of \mathbb{P}_1 . The largest element in \mathbb{P}_1 is the maximum of the sums of the aggregated service times of packet 1 along paths 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 (Property 3 of Proposition 1), so that

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

where 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 (this is Proposition 2).

Using the homogeneity assumption (Section II-C), we have

$$\begin{aligned} \mathbb{E} \left[\max_i \tilde{S}_1^{(i)} \right] &= \mathbb{E} \left[\max_i (s_1^{f(1,i)} + \cdots + s_1^{f(D,i)}) \right] \\ &\leq \mathbb{E} \left[\max_i s_1^{f(1,i)} \right] + \cdots + \mathbb{E} \left[\max_i s_1^{f(D,i)} \right]. \end{aligned}$$

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

This upper bound can be reached; this is the case when the window size is $W = 1$ and when we have an umbrella tree [i.e., a tree made of many independent branches, which corresponds to the least sharing in the tree (see Section IV-C)].

3) Lower Bound:

Theorem 4: Consider a network with exponential aggregated service time 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

$$\frac{\mathbb{E}[s^{(D)}]}{W} \ln(N)(1 + o(1)) \quad \text{for } N \rightarrow +\infty \quad (17)$$

where N is the number of receivers, W is the window, and $s^{(D)}$ a typical aggregated service time in a router of level D (the last routers before receivers), which is by assumption the same for all receivers.

Proof: Let us consider the packets $W, 2W, 3W, \dots$. Since the window has a fixed size, packet kW cannot start until the acknowledgments of packet $(k-1)W$ have arrived from 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

$$\|Y_{kW}\| \geq \|Y_{(k-1)W}\| + \max_{i=1 \dots N} S_{kW}^{(i)}.$$

Then, using $\gamma = \lim_{k \rightarrow +\infty} (\|Y_{kW}\|/kW)$, we have

$$\gamma \geq \frac{1}{W} \lim_{m \rightarrow +\infty} \frac{\max_{i=1 \dots N} (S_W^{(i)}) + \cdots + \max_{i=1 \dots N} (S_{mW}^{(i)})}{m}.$$

Now, as $(\max_i S_{mW}^{(i)})_{m \geq 0}$ are i.i.d. random variables, the law of large numbers gives us the inequality

$$\gamma \geq \frac{\mathbb{E} \left[\max_{i=1 \dots N} (S_1^{(i)}) \right]}{W}. \quad (18)$$

If the 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 aggregated

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 one link that belongs only to the path from the source (for example, the last link). These links are supposed to have independent aggregated service times with the same exponential law $s^{(D)}$ so that, applying Corollary 2, (18) leads to the relation of the theorem. \square

It is possible to have a better lower bound for γ under some additional assumptions on the tree, as we will see in Section IV-C. Again, this bound is reached by a category of trees, such as the reverse umbrella tree with all links shared except the last one.

Note that the lower bound of Theorem 4 holds as soon as the edge routers before the receivers satisfy the conditions that we have assumed on random service times, regardless of the statistical behavior of the other routers. Since it is natural to associate backbone routers with internal nodes of the tree and access routers with edge routers before receivers, this means that the logarithmic shape of throughput degradation will hold as soon as access routers alone suffer from the fluctuations in question.

C. Tree Topology Dependence

We have been able to bound γ , both from above and from below, by $\ln(N)$ functions. We now give a finer grain classification of tree topologies within these two bounds. The performance inside the interval defined by the bound found above depends essentially on the nature of the tree. We show that it is possible to create a partial order on the trees, allowing one to achieve throughputs that range within the whole interval defined by the lower and the upper bound. In this section, we consider trees with N receivers, under the assumptions of homogeneity and independence that we described earlier.

1) *Sharing Function, Partial Order:* In what follows, we will assume that the delays in the feedback 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 aggregated 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 variables with the same law, 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: Sharing function. Due to the last statement, we can define the sharing value of receivers i and j , which we denote as $a(i, j)$, by

$$a(i, j) = \max \left\{ l = 1 \dots D \mid s_m^{(f(i,l))} \equiv s_m^{(f(j,l))} \right\}. \quad (19)$$

The sharing value is exactly the number of common terms in the two sums $S_m^{(i)}$ and $S_m^{(j)}$; all others terms of these sums are independent.

The sharing function indeed measures the correlation between receivers in the tree: receivers with large sharing value

appear to have similar performances. It is possible to show that a given sharing function characterizes a tree.⁷

Definition: Sharing order. We say that a tree T is less shared than another tree T' if their sharing functions a and a' are such that $a \leq a'$. 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 5: If T and T' are such that $a \leq a'$, then $\gamma_{a'} \leq \gamma_a$.

The proof is omitted here; for a simple case see [16, Appendix].

This theorem gives us another proof for the upper and lower bounds of Section IV-B. The upper tree, when making the assumptions that the receivers are distinct, given by $a(i, j) = D - 1$, provides the lower bound, and the lower tree, given by $a(i, j) = 0$, provides the upper bound.

2) *Umbrella Trees:* **Definition: Umbrella Tree of Class l .** A tree (given by its sharing function 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 of class l with large l typically correspond to a worst-case situation, as receivers share few links. We have observed via simulation that for these types of trees, the throughput seems to degrade more severely when the number of receivers increases. In an umbrella tree, the lower bound for γ can be reached.

Theorem 6: Consider a network with exponential aggregated service time in routers, where all aggregated service times are exponential random variables of the same parameter λ . Then the Lyapunov exponent of an umbrella tree of class l is bounded from below by a function that can be expanded as

$$\frac{1}{W} R_{l,\lambda}(N)(1 + o(1)) \quad \text{for } N \rightarrow +\infty \quad (20)$$

where $R_{l,\lambda}(N)$ is the function defined by (14).

Proof: According to Theorem 5, we simply need to verify this formula for the tree ($a = l$). The formula established in the proof of the lower bound, namely

$$\gamma \geq \frac{\mathbb{E} \left[\max_{i=1 \dots N} (S_1^{(i)}) \right]}{W} \quad (21)$$

holds. Let us look at the performance of the tree ($\forall i, j, a(i, j) = l$). For all i , we have $S_1^{(i)} \geq s_1^{(f(i, D-l+1))} + \dots + s_1^{(f(i, D))}$. We can then apply Corollary 3 to this sum of random variables that has a Gamma distribution of parameter (λ, l) . \square

It is immediate to check that for all $l > 1$, the upper bound on the throughput based on $R_{l,\lambda}$ (14) improves on that of Theorem 4 (namely, it is strictly smaller). However, as shown in the proof of Corollary 3, we have the equivalence $R_{l,\lambda}(N) \sim \ln(N)$ as N tends to ∞ , so that these bounds are asymptotically equivalent. At first glance, one may then think that there is no

⁷A consequence from the study of the equivalence relation $i \approx_i^a i' \Leftrightarrow a(i, i') \geq l$ is that we can build the tree using only the sharing function.

real improvement. Numerical evidence shows that for the range of the number of receivers considered in this paper, this improved bound is always much sharper than the previous one. For instance, for the umbrella tree ($CL = 0$, $FO = 3$, $UL = 3$) of Fig. 11, and for eight receivers, the new upper bound (0.95) is much closer to the throughput provided by simulation (0.61) than the upper bound of Theorem 4 (2.89).

As we can see, in spite of all the optimistic assumptions we made, an umbrella tree systematically results in a severe degradation of the throughput. The results observed in the numerical simulation, as well as the intuition we had on tree topology impact, are confirmed and generalized by analytical results. The sharing function 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 the fact that umbrella trees, which are frequently encountered in Internet architectures [9], suffer severe throughput degradation even in the case of light-tailed fluctuation of the delay.

V. CONCLUSION

In this paper, we study the impact of randomness (i.e., variation of queueing delay due to cross traffic) 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 the tree topology on the throughput of the multicast session.

In the presence of light-tailed random delays, we show that the throughput decreases logarithmically when the number of receivers increases. We find analytically an upper and a lower bound for the throughput. Within these bounds, we characterize the degradation depending on the tree topology. We identify a class of trees commonly found in IP multicast sessions [9], [13] as a worst case of throughput degradation. This observation is quantified by simulation and then explained analytically.

This work proves analytically that TCP-like congestion control can be harmful with reliable multicast transmission through homogeneous trees, even for light-tailed delays. Note that the degradation of performance with regard to the size of the group is strongest in this homogeneous case. The case with heterogeneous trees could in principle be analyzed along the same lines. Since it is well known that in such heterogeneous cases, performance is driven by slow receivers, the simplest approximation could consist of applying our analysis to the subset of slow receivers.

Consequently, applications may prefer multirate control mechanisms to single-rate reliable multicast transmission. Multirate (layered) control mechanisms are best suited to multicast sessions with a large number of receivers. Rubenstein *et al.* [11] have shown that multirate 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 contribution of our work is a new analytical framework which can be used to study various problems related to flow and congestion control in multicast and unicast environments.

In future works, we will extend and generalize our analytical framework. Extension to an adaptive window size is possible based on a generalization of the (max,plus) representation of TCP Tahoe and Reno known for unicast [14].

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

ACKNOWLEDGMENT

The authors would like to thank K. Papagiannaki, B. Lyles, and S. Bhattacharya for their comments and suggestions which helped improve the quality of this paper.

REFERENCES

- [1] S. J. Golestani and K. K. Sabnani, "Fundamental observations on multicast congestion control in the Internet," in *Proc. IEEE INFOCOM*, vol. 2, Mar. 1999, pp. 990–1000.
- [2] F. Baccelli, D. Kofman, and J. L. Rougier, "Self-organizing hierarchical multicast trees and their optimization," in *Proc. IEEE INFOCOM*, vol. 3, Mar. 1999, pp. 1081–1099.
- [3] F. Baccelli, G. Cohen, G. J. Olsder, and J. P. Quadrat, *Synchronization and Linearity*. New York: Wiley, 1993.
- [4] C. S. Chang, *Performance Guarantees in Communication Networks*. New York: Springer Verlag, 2000.
- [5] J.-Y. Le Boudec and P. Thiran, *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet*. New York: Springer-Verlag, 2002.
- [6] F. Baccelli and T. Bonald, "Window flow control in FIFO networks with cross-traffic," INRIA, Paris, France, Res. Rep. 3434, May 1998.
- [7] T. Lai and L. Robbins, "A class of dependent random variables and their maxima," *Z. Wahrsch.*, vol. 42, pp. 89–111, 1978.
- [8] R. E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing*. New York: Holt, Rinehart and Winston, 1975.
- [9] R. C. Chalmers and K. C. Almeroth, "Modeling the branching characteristics and efficiency gains of global multicast trees," in *Proc. IEEE INFOCOM*, vol. 1, Apr. 2001, pp. 449–458.
- [10] L. Rizzo, L. Vicisano, and J. Crowcroft, "TCP-like congestion control for layered multicast data transfer," in *Proc. IEEE INFOCOM*, vol. 3, Mar. 1998, pp. 996–1003.
- [11] D. Rubenstein, J. Kurose, and D. Towsley, "The impact of multicast layering on network fairness," in *Proc. ACM SIGCOMM*, Aug. 1999, pp. 27–38.
- [12] M. Handley and S. Floyd. (1998, Dec.) Strawman congestion control specifications. Internet Research Task Force (IRTF) Reliable Multicast Research Group (RMRG). [Online]. Available: <http://www.aciri.org/mjh/rmcc.ps.gz>
- [13] I. Stoica, T. S. E. Ng, and H. Zhang, "REUNITE: A recursive unicast approach to multicast," in *Proc. IEEE INFOCOM*, vol. 3, Mar. 2000, pp. 1644–1653.
- [14] F. Baccelli and D. Hong, "TCP is max,+ linear," in *Proc. ACM SIGCOMM*, Aug. 2000, pp. 219–230.
- [15] S. Bhattacharya, D. Towsley, and J. Kurose, "The loss path multiplicity problem in multicast congestion control," in *Proc. IEEE INFOCOM*, vol. 2, Mar. 1999, pp. 856–863.
- [16] A. Chaintreau, F. Baccelli, and C. Diot, "An analytical framework for the analysis of multicast congestion control," INRIA, Paris, France, Res. Rep. 3987, Sept. 2000.



Augustin Chaintreau graduated in mathematics and computer science in June 1999 from the Ecole Normale Supérieure de Paris (ENS), Paris, France, where he is currently working toward the Ph.D. degree.

He was with the IP research group of Sprint Advanced Technology Laboratory as a Research Assistant for six months, during which time he conducted the work presented here. He is currently working in the ENS-INRIA Research Group, Paris, France, on the modeling of a TCP-controlled network, evaluating its performance as a result of the

interaction of all connections sharing network resources.



François Baccelli received the Doctorat d'Etat from Paris-Sud University, Paris, France, in 1983.

He has been the head of the Mistral research group of INRIA, Sophia Antipolis, France, from its creation to 1999. He was a partner in several European projects including IMSE (Esprit 2) and ALAPEDES (TMR), and was the coordinator of the BRA Qmips project. He is currently INRIA Directeur de Recherche in the Computer Science Department of Ecole Normale Supérieure in Paris, where he started the TREC performance evaluation

group in 1999. He is also a part-time Professor in the Applied Mathematics Department of the Ecole Polytechnique. He coauthored *Elements of Queuing Theory: Palm-Martingale Calculus and Stochastic Recurrences* (New York: Springer Verlag, 1994) with P. Bremaud and *Synchronization and Linearity* (New York: Wiley, 1993), a book on an algebraic approach for discrete event dynamical systems, with G. Cohen, G. J. Olsder, and J. P. Quadrat. His research interests are in the modelling and performance evaluation of computer and communication systems. He is currently working on the modelling of TCP and on the analysis of CDMA coverage and capacity.



Christophe Diot received the Ph.D. degree in computer science from INP, Grenoble, France, in 1991.

From 1993 to 1998, he was a Research Scientist with INRIA Sophia Antipolis, France, working on new Internet architecture and protocols. He moved to Sprint Advanced Technology Laboratory, Burlingame, CA, in October 1998, to take the lead of the IP research group. His current interest is in the passive monitoring of the Sprint IP backbone in order to study IP traffic characteristics and to design new analytical models and traffic engineering

solutions for pure packet networks.

Dr. Diot is a Member of the Association for Computing Machinery and serves as an Associate Editor for the IEEE/ACM TRANSACTIONS ON NETWORKING.