Analysis and optimization of hierarchical reliable transport protocols

A. Busson * J-L Rougier D. Kofman
[abusson, rougier, kofman] @enst.fr
Ecole Nationale Supérieure des Télécommunications
46 rue Barrault, 75013 Paris, France.

Several schemes have been proposed that address the problem of scalable reliable multicast. The Tree-based Acknowledgement protocols family achieve scalable reliable transport in one-to-many communications based on a hierarchical approach. It has been recently selected as a standard-track possible protocol architecture by IETF. In these protocols, dedicated servers are used to perform error recovery and statistical aggregations. Nevertheless a number of dimensioning issues remain open, such as the number of required servers, and the optimal number of hierarchical levels. In order to evaluate and dimension such protocols, we consider an analytical model using point processes, where points are scattered in the plane in order to represent receiver locations in the Internet, and a loss model which takes correlated losses into account. We define cost functions based on the number of retransmissions in the control tree for a packet sent by the source. We proposed random geometric model allowing us to find explicit expressions for the cost functions, w.r.t. to a limited set of parameters : number of receivers and number of repair servers. We are then able to deduce optimal control tree structures (optimal number of repair servers), and generic dimensioning rules.

1. Introduction

IP multicast allows the efficient transmismission of data from one or several sources to multiple receivers, but in an unreliable manner. However, several multicast applications require reliability. Unicast reliable protocols such as TCP are not adapted to one-to-many communications because of scalability problems.

One of the main issue when designing reliable multipoint protocols is the feedback implosion problem due to receiver's acknowledgements. In a multicast session the number of participants can be very important, in this case the source cannot deal with a window state for every destinations. An approach based on NACK messages that are only sent when a loss is detected by a receiver can be used. However, even in this case, losses occuring near the source can lead to a flooding of the source with receiver NACKs.

There is a variety of ways to solve this scalability problem. Certain mechanisms are more suitable for specific applications than others. Protocols such as SRM [1] use slotting and damping to avoid feedback implosion: when a packet is missing, the receiver triggers

*This work is supported by the french government funded project GEORGES (http://www.telecom.gouv.fr/rnrt/projets/pgeorges.htm)

a random timer. If it has not received a NACK from other receivers corresponding to its own state before the timer expires, it multicasts a NACK to the whole group. Missing packets are repaired with the same algorithm by any receiver. SRM is easy to implement but has severe scalability problems. It is well suited for small multicast groups.

Protocols such as PGM [2] require specific algorithms implemented in routers of the multicast tree. When a packet is missing, a receiver sends a NACK to the upstream router which implements this protocol; the router multicasts a NACK confirmation to the subnet, thus avoiding possible consecutive NACKs. The process is repeated until the source; then, a repair packet is sent from the source only to the router ports which have received a NACK. PGM is very efficient and scalable, but requires router-assistance which makes it difficult to deploy.

TRee-based ACKnowledgement protocols (TRACK) such as RMTP II [3] use a hierarchy of dedicated servers which assures scalability to process error recovery. TRACK as described in Section 2 has been recently selected as a standard-track possible architecture by the IETF. A framework has been defined in [4] and [5] to achieve a reliable and efficient transfer in real time for one-to-many communication. But a number of dimensionning issues remain open, such as the number of required servers, and the optimal number of hierarchical levels.

This paper focuses on performance analysis and optimization of TRACK protocols. We determine dimensioning rules (w.r.t. the number of Repair Heads) for various optimization criteria. A stochastic geometric approach, inspired by [7], [8] and [6], is used to model the spatial organization of the network. In the cited paper, it has been shown that such an approach is well adapted to the study of multicast protocols. Stochastic geometry allows us to compute functionals of a macroscopic network model, and to find the protocol general behavior w.r.t. a minimal number of structural parameters, unlike classical evaluation models which have to consider a lot of detailed parameters.

A family of cost functions depending on the number of retransmissions for a typical packet sent by the source can be explicitely expressed, from which optimal server hierarchical configurations are found.

This paper begins with the presentation of TRACK in Section 2. The loss model, the mathematical model and the cost functions are presented in Section 3. Results and optimal configurations are outlined in Section 4. Finally, conclusions are given in Section 5.

2. TRACK

In this Section, we introduce the TRee-based ACKnowlegdement protocol (TRACK) as described in [4] and [5]. We only present the control tree structure and associated mechanisms to repair losses; congestion control, tree configuration, tree maintenance and security will not be considered in this paper.

TRACK is designed to reliably send data from a single sender to a group of receivers.

The multicast group is divided into subgroups which represent sets of receivers in a local region. For each subgroup, a node (called Repair Head, RH) is defined as being responsible for error recovery and receiver state aggregation. Repair Heads of different subgroups (children) are bound to a RH of a superior level (parent). This binding is repeated until the source is reached. In this sense, the source is the top Repair Head. Such a hierarchical structure allows us to achieve scalability, as the sender and any RH are only responsible for a controlled number of receivers.

2.1. Control tree structure

The control tree is the communication path used to join the hierarchical set of Repair Heads. The data channel is the channel used to multicast packets from the source to receivers and RHs (for the data channel, there is no difference between a receiver and a RH). The data channel can be very different from the control tree. In the control tree, the sender and each RH have a multicast local control channel to their children, used to perform local retransmissions and to send maintenance packets. A multicast address (different from the session address) is used to communicate in the local control channel, thus the retransmission scope of a RH is limited to the set of its children. In figure 1(a), a control tree with three levels is shown.

2.2. Loss repair

During the considered multicast session, the sender multicasts data on the data channel to the set of RHs and receivers. A receiver must unicast regular ACKs or exceptional NACKs to its corresponding RH to indicate which packets are missing.

An ACK is a bitmap indicating which packets have been received and which packets need to be retransmitted; it also provides congestion control statistics. When a missed packet is detected, the RH (or the sender at the higher level) multicasts the packet on the local control channel. In order to avoid feedback implosion, an ACK is periodically sent by a receiver (or RH), and children send this ACK at different times in order to keep the constant ratio of control traffic seen by the RH (or the source).

In addition to using ACKs, TRACK provides mechanisms for NACKs. Unlike an ACK, a NACK is sent only when a packet is lost. For NACK, the solution is, as usual, timer based. When a packet is detected missing, the receiver (resp. the RH) triggers a random timer. If either a packet or an explicit suppression message is detected for the missing packet, the timer is cancelled. Otherwise, the receiver (or the RH) generates a NACK, and its parent immediately multicasts the repair on the local control channel.

There are other solutions to reduce the number of retransmissions and their scope, such as FEC (Forward Error Correction) and Dynamically Scoped Retransmission (DSR), but the use of these mechanisms are out of scope of this paper.

3. Model

First, we introduce the model used to represent how loss probabilities are distributed when a packet is sent by the source towards the set of receivers and RHs in the data channel. This loss model also represents the distribution of losses in the control tree for repair packets between the source and its children (RHs bound to the source). Then, in the mathematical model Subsection, we present several point processes in the plane which describe the location of receivers and RHs of the multicast session.

3.1. Loss model

Our loss model is an extension of the model proposed in [9]. In their paper, Yajnik *et al.* analyzed the loss behavior of several multicast hosts located in US and Europe connected via the MBone.

Yajnik *et al.* show that a distribution of losses modelised by a modified star topology (figure 1(b)) is close to the original distribution. This model introduces two loss probabilities, p and q; where p denotes the probability of losing a packet near the source (and thus of losing a packet for all receivers). If the loss near the source does not occur, the event of losing a packet for a receiver (which probability denoted by q) is assumed independent of events of losing a packet for the others receivers.

In this paper, we use the modified star topology introduced above : a packet sent by the source is lost before it reaches a domain (e.g. AS [2], routing domain) with a given probability $p(x)$, where x denote the location of the domain. If this loss occurs, this packet would be missing for all receivers of the domain.

Once in the domain, the packet reaches any receiver independently of the others with probability (1-q). Consequently, a receiver in a given domain located at x receives a packet from the source with probability $(1-p(x))(1-q)$.

In the control tree, the loss model for retransmitted packets sent by source to its children (RH bound to the source) is the same as the one described above. Within a domains, the probability of losing a retransmitted packet in the local control tree between any RH and its children will be q (and will be independent between children).

3.2. Geometrical model

In this section, we introduce our mathematical model that uses point processes in the plane. Points which model the set of receivers and RHs of a multicast session are randomly scattered in the plane. We will look at these points on a finite window, which will be a ball of radius R_w centered at the origin. Due to the hierarchical structure of the Internet, where receivers are concentrated in local regions (AS), we have chosen to model the set of receivers by clusters. The clusters represent domains where receivers are located. This clustering, along with the loss model presented in the previous Section, allows us to take spatial loss correlation into account.

We model the given domain by a so-called Cluster point process (see [10]). In such a process, each point (called germ) of a given point process π generates a set of points (N_x for a point of π located at x) which for stability reasons must be finite. The union of these clusters is the Cluster point process N, which can be written :

$$N = \sum_{x \in \pi} N_x.$$

Logically, the RHs are distributed in the same cluster as the receivers, and thus each point of π also generates a cluster representing the set of RHs. Indeed, multicast infrastucture is in the same cluster as multicast members. For each cluster, we make a partitionning of the area where receivers are distributed, in order to allocate distinct domain of retransmission[3] for each RH.

[2] AS:Autonomous System

[3] domain where all receivers are bound to a given RH

In our paper, we choose the Neyman-Scott process which is a particular case of Cluster processes, that has nice mathematical properties (see [11]). The underlying point process is a stationary Poisson point process π of intensity λ_p. In this case, the family of clusters $(N_x)_{x\in\mathbb{R}^2}$ applied to points of π are themselves i.i.d. [4] point processes and have identical distribution around their generating point. Under these assumptions, the process is shown as stationary. We note that points of π are not themselves points of the final process. The loss model is particularly suitable with these point processes.

A cluster will represent a domain, and the distance between the source and the domain (used to compute $p(x)$) will be the distance which separates the cluster germ from the origin. Inside a domain, the probability of losing a packet is always q. Figure 1(c) represents a realization of a Neyman-Scott process with the underlying Poisson point process and generating balls (in grey) where clusters are distributed. Figure 1(d) represents a realization of the final point process.

3.3. Example of Neyman-Scott process

For the sake of clarity, we shall start the presentation of this example with two hierarchical levels (RH located in the clusters and the source node as the top RH), the extension for three levels is very simple and explained in Section 4.

For each germ x of π, the set of receivers of the cluster centered at x (noted $N_{0,x}$, where 0 is the level of receivers) is a Poisson point process of intensity λ_c scattered in a ball centered at x and with a random area noted S_x.

For the sake of simplicity, for all RHs, we allocate the same size of retransmission domain D (see [12] for more details). In this way, the family of random variables describing the number of children bound to the RHs are i.i.d. Thus, the area of balls must be multiples of D. Moreover, the value of S_x for a point of π located at x determine the number of RHs in this cluster.

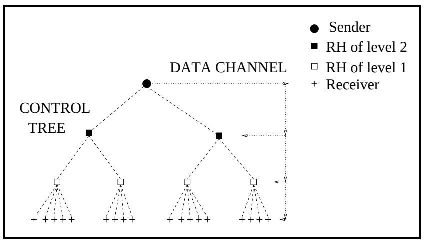
The figure 2(a) represents two clusters and their respective retransmission domain.

3.4. Cost function

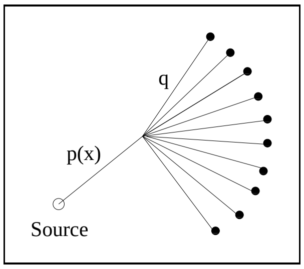
Our model allows us to define a variety of cost functions in order to evaluate hierarchical multicast transport protocols. We focus on the number of retransmissions sent at each level of the control tree for any packet sent by the source. This number of retransmissions is computed with regards to the loss model applied to a set of point processes representing network nodes. Any weighted additive function of these number of retransmissions at different hierarchical levels can be considered as cost function.

In this paper, we study two particular cost functions. The first one, denoted by $\mathbb{C}_1$ is described as follows: for a typical packet, each retransmission from a parent increases the cost by the number of children.
The second cost function ($\mathbb{C}_2$) has been chosen in order to take into account the fact that retransmissions within a given domain (cluster) are probably less expensive than retransmissions occuring at a higher level. Since, at a lower level, the distances between the parent and its children of a local retransmission tree are shorter. Thus, costs are weighted by the mean distance between parents and their children nodes.

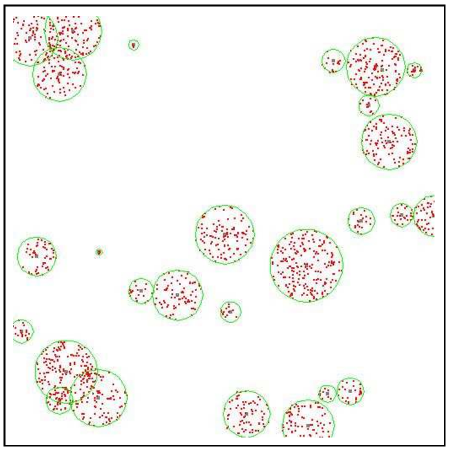
[4] i.i.d.: Independant and Identically Distributed



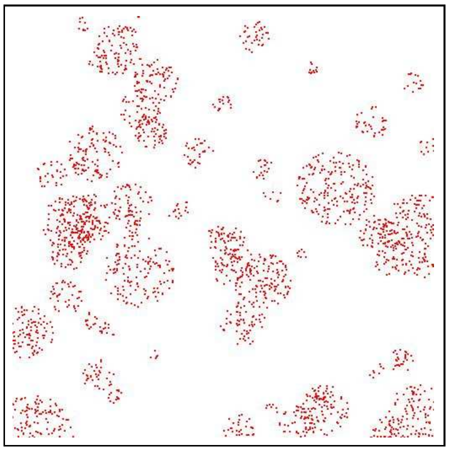
(a) Control tree



(b)
Modified star topology



(c) Neyman-Scott process 1



(d) Neyman-Scott process 2

a	*0.1*	0.6025
	0.01	0.085
	0.001	0.0088
	0.0001	0.00088

(e) probability of losing a packet

b					
		a			
		0.1	*0.01*	*0.001*	*0.0001*
q	*0.1*	0.013	0.043	0.075	0.080
	0.01	0.013	0.030	0.049	0.054
	0.001	0.013	0.013	0.018	0.022
	0.0001	0.013	0.013	0.013	0.013

(f) The constant b

Figure 1. The control tree, the modified star topology, the loss probability and the constant b

If we note by $\mathbb{E}[\tau_i]$ the mean number of retransmissions from all RHs of level i required in order for all children to have received a given packet, we have in a control tree with H levels,

$$\mathbb{C}_1 = \sum_{i=1}^{H} m_i \mathbb{E}[\tau_i] \text{ and } \mathbb{C}_2 = \sum_{i=1}^{H} l_i m_i \mathbb{E}[\tau_i]$$

where m_i is the mean number of active points in a retransmission domain (domain where children of a RH are located) of level i, and l_i is the mean distance between a RH of level i and its children.

The observation window radius R_w and the germ intensity λ_p can be chosen in order to represent the Internet topology (number of clusters representing the number of autonomous systems and the window diameter corresponding to the maximum number of AS crossed during a transmission). Similarly, the loss probability $p(x)$ can be chosen as a function of the number of AS crossed (most losses in the Internet being localized at exchange points and rarely within a domain). If we consider the loss probability at a AS border as a constant a, $p(x)$ can be expressed as a function of this parameter as shown in [12]. Table 1(e) gives the relation between a and the mean loss probability (average w.r.t. x).

4. Results

We have computed explicit formulae for the cost functions allowing us to find the optimal parameters in terms of number of RHS. But due to space constraints, formulae and the computation details are not presented in this paper, they could be found in [12]. In this Section, we give the numerical results representing the optimal tree configuration.

4.1. Results for the first cost function

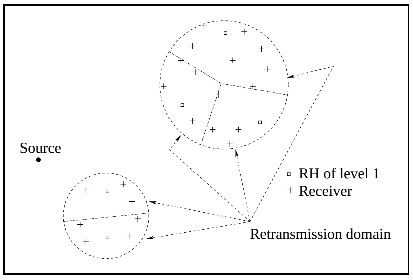
In this Subsection, we present the numerical results for the model described in 3.3 and cost functions $\mathbb{C}_1$. It can be remarked that when the number of RH increases in a given level, the mean number of children bound to these RH decreases, thus a retransmission reaches less receivers and the probability that one child has a packet missing is less important. Therefore, the cost should decrease at this level. On the other hand, the cost function at the superior level increases when the number of RH bound to a RH of superior level increases. Chosen cost functions have thus a global minimum with regards to the mean number of RH for any level. Therefore, we can find a global minimal value which corresponds to the optimal repair tree configuration w.r.t. the cost function. In this Section, we give this optimal value for $H = 2$ and for $H = 3$ with respect to λ_c which is here the mean number of receivers per cluster.

Calculations with two levels

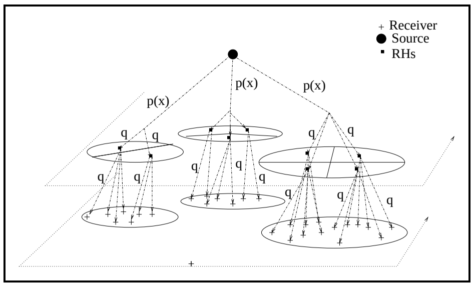
In this Subsection, we compute the optimal tree configurations in the case where we have two levels of hierarchy as shown in Figure 2(b).

In figure 2(c), we give the optimal number of RH with regard to the mean number of receivers per cluster (λ_c).

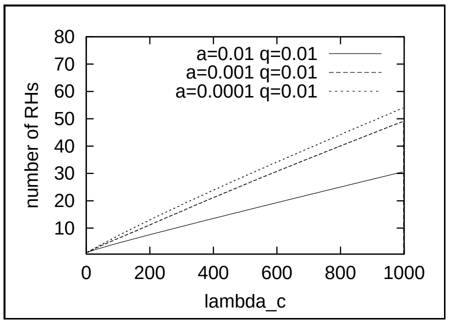
We see that the optimal value k of the mean number of RHs is close to a straight line



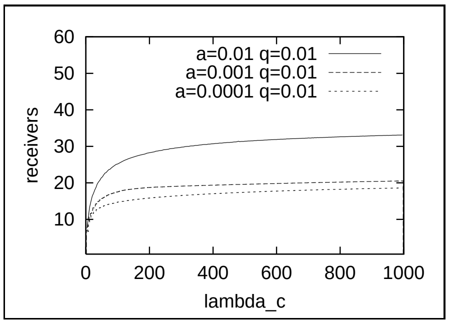
(a) Retransmission domain



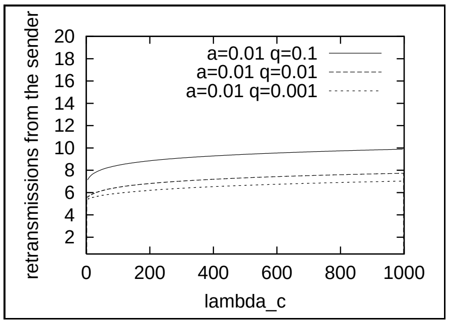
(b) Repair tree



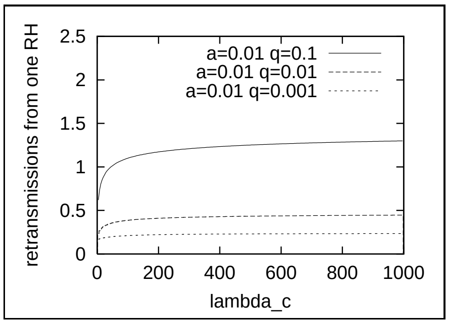
(c) optimal number of RH



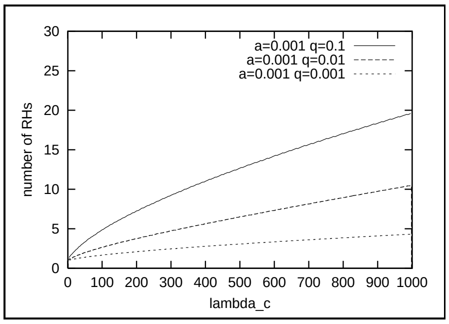
(d) mean number of receiver per RH



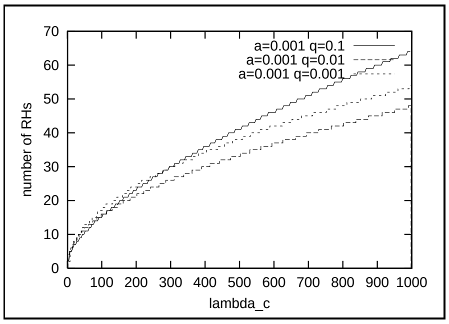
(e) mean number of retransmissions between the sender and the set of RHs



(f) mean number of retransmissions between one RH and its children



(g) optimal number of RH of level 2



(h) optimal number of RH of level 1 per RH of level 2

Figure 2. Optimal parameters w.r.t. the first cost function

w.r.t. λ_c, it can thus be expressed approximately as $k = b\lambda_c$, where b is a function of a and q, given in Table 1(f). Logically, the constant b increases when q increases. Indeed, the number of retransmissions being more important, we need more RHs. Inversely, b decreases when a increases (except when q is very small), because the loss probability a affects only the top level; so that the cost increases just between the source and RHs.

In figure 2(d), we have computed the optimal mean number of receivers per RH w.r.t. explicit formulae of the cost function. We note that, as the optimal number of RH, we can give engineering rules to easy approximate these values. Indeed, if we assume that k is linear w.r.t. λ_c, and if we denote by X the optimal number of receivers per RH, we can approximate the exact optimal value (represented in 2(d)) by :

$$X = -\lambda_c \ln\left(1 - \frac{1}{b\lambda_c}\right).$$

We observe that

$$X < \frac{1}{b} \text{ and } \lim_{\lambda_c \to \infty} X = \frac{1}{b}$$

and thus if we take $\frac{1}{b}$ receivers per RH when $\lambda_c >> 1$, we are always close to the optimal configuration.

In figure 2(e) and 2(f), we show the number of retransmissions from the source and from a typical RH. The number of retransmissions from the sender is very important (from 7 for $\lambda_c = 4$ until 10 for $\lambda_c = 1000$ when a and q are big). This number is very high, because our example corresponds to an extreme case, where all the autonomous systems of the Internet (10000) participate in the multicast session, and we have assumed that the number of RH per cluster is at least one (active or not). Moreover, the weight of the cost function favours retransmissions from the sender rather than from RHs because a retransmission from all RHs is weighted by the total number of receivers (equal to approximately 10^7 when $\lambda_c = 1000$).

Whereas, between a RH and its children the number of retransmissions is quite small (from 0.5 for $\lambda_c = 4$ to 1.3 for $\lambda_c = 1000$ in the worst case).

Finally, we can observe that the number of retransmissions varies insignificantly w.r.t. λ_c, which can be easily explained by the nature of probabilities involved in this model : since when the number of receivers (resp. RH) is multiplied by 10, their number of retransmissions is incremented only by one, for a loss of 0.1 between the parents and the set of receivers (resp. RH).

Calculations with three levels

In the following, we consider three levels of hierarchy, i.e. the source, two levels of RHs and the set of receivers. For each RH of level two (just under the source), we associate a constant number k of RHs (of level 1), the domain of retransmissions of a RH of level 1 being also constant (equal to $\frac{D}{k}$).

As in the previous Section, we compute the optimal number of RHs of level 1 and 2 w.r.t. λ_c.

For $\lambda_c >> 1$, the required number of RHs of level 2 (source children) and 1 (parent receivers) is proportional to λ_c . When λ_c increases, the optimal value of the cost function is reached for a constant proportion of children per RH of level 1 and 2. Therefore, except for transmissions from the source, the mean number of retransmissions between a RH and its children tends to be constant.

Unfortunatly, the total number of RH at the optimal is very important, 40 RHs per cluster (2.5 RHs of level 2 and 16 of level 1) when the number of receivers in a cluster is 100 with $a = 10^{-3}$ and $q = 10^{-2}$. We compared the cases $H = 2$ and $H = 3$, at the optimal configuration : the cost with three levels is much less important than with two levels. And for a given total number of RHs in the control tree, the cost is still inferior for $H = 3$. In figure 3(d) we show the cost function with $H = 2$ and 3, the cost for $H = 3$ is almost half the cost at the optimal configuration of $H = 2$.

Moreover, the cost function ($H = 3$) is slowly decreasing w.r.t. the total number of RHs. Thus, a small number of RHs can be chosen, leading to a cost function very close to the optimum. By example, in figure 3(d), we can take 8 RHs (2 RHs of level 2, and 3 RHs per RH of level 2). This result is very important as each RH has surely a cost[5].

4.2. Results for the second cost function

In this part, the cost of a retransmission depends on the mean distance between a RH (or the sender) and its children as describe in 3.4. Thus, a retransmission between the sender and a RH is more expensive than for the function $\mathbb{C}_1$.

Calculations with two levels

When a and q vary from 10^{-2} to 10^{-4}, the required number of RHs tends to a line of rate 0.013 with λ_c. The line is reached more or less quickly in function of a and q. As expected, this number is much lower than in $\mathbb{C}_1$. Although the number of local retransmissions between a RH and its children is higher (figure 3(b)), the number of retransmissions from the source does not really decrease. In figure 3(a), we show the cost function for $H = 2$ and $H = 3$ for a given number of RH and for $\lambda_c = 100$. As for $\mathbb{C}_1$, for the same total number of RHs, the cost function is higher for $H = 2$.

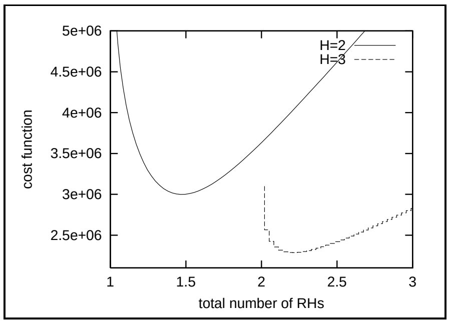
Calculations with three levels

For $H = 3$, very few RHs are located in the control tree, the mean number of RHs of level 2 reaches 1.5 in the worst case, and we have just 2 RH of level 1 per RH of superior level for all λ_c. Therefore, the mean number of retransmissions between the source and RHs of inferior levels is constant w.r.t. λ_c. Note that the number of local retransmissions between a RH and its set of receivers increases, as shown in figure 3(f).

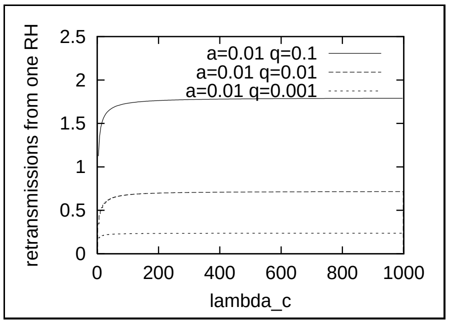
5. Conclusion

We have considered an analytical approach based on stochastic geometry to give rules for the dimensioning of TRACK protocols. We have proposed a model using point processes representing receivers in the Internet. Based on this model, we have computed explicit formulae for the mean number of retransmissions for each level of hierarchy of the control tree. A family of cost functions has been defined, and two particular functions have been studied. Exact optimal configurations of the control tree have been given w.r.t. the mean number of receivers per AS and to parameters of the loss model.

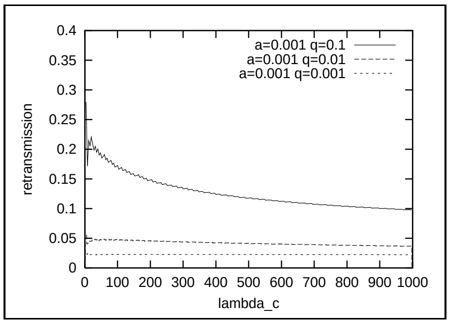
[5]By the way, we could consider cost functions which takes the cost of each RH used into account.



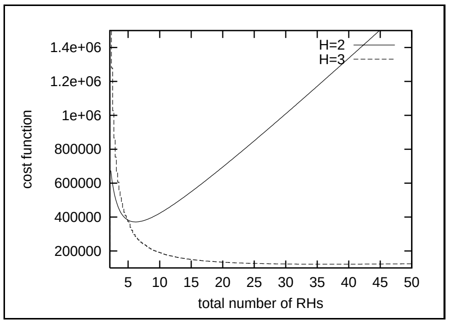
(a) Cost function for a given number of RHs with $\lambda_c = 100$



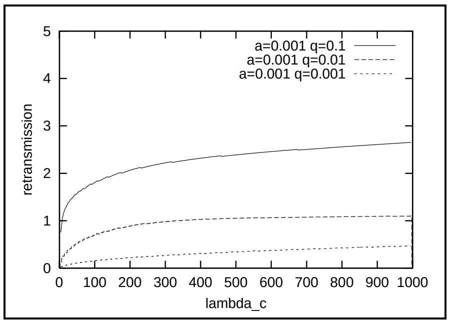
(b) number of receiver per RH



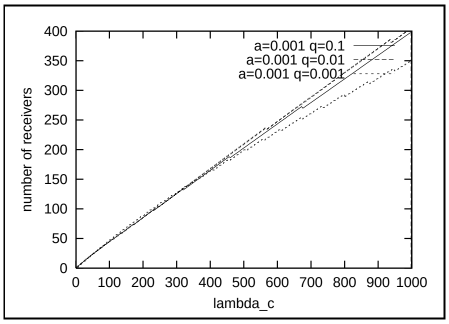
(c) number of retransmissions from a RH of level 1



(d) cost function with a=0.001 q=0.01



(e) number of retransmissions from a RH of level 1



(f) number of receiver per RH

Figure 3. optimal parameters w.r.t. the second cost function

In a control tree with two hierarchical levels, we have calculated engineering rules to compute easily the optimal number of required Repair Heads for these two cost functions. In the first cost function, each retransmission from a parent increments the cost by the number of its children. The second is counterbalanced by the mean distance between a Repair Head and its children. The first one favours the location of RHs at the top of the tree, unlike the second which locates RHs at the bottom of the tree and favours local retransmission. In these two cases, we have shown that the cost function is lower for $H = 3$ than for $H = 2$, for an equal total number of RHs in the control tree. We have also shown that the optimal structure needs a fairly large number of RHs , but a limited number of RHs can lead to good results as the cost function is very flat around the optimum.

We are currently working on new cost functions that penalize useless transmissions, i.e. the number of retransmissions received by participants which already had received the packet. We are also working on a model which can consider different hierarchical architectures (i.e. a RH responsible for several AS).

REFERENCES

1. S. Floyd *et al. A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing.* In proceedings of SIGCOMM'95, ACM, October 1995.
2. T. Speakman *et al. PGM Reliable Transport Protocol Specification.* IETF working document.
3. B.Whetten, G. Taskale, *Overview of the Reliable Multicast Transport Protocol II (RMTP-II).* IEEE Networking, Special Issue on Multicast, February 2000.
4. B. Whetten *et al. TRACK Architecture, A Scalable Real-Time Reliable Multicast Protocol.* IETF working document.
5. B. Whetten *et al. Reliable Multicast Transport Building Block for TRACK.* IETF working document.
6. F.Baccelli, D.Kofman, J.L.Rougier.*Self-Organizing Hierarchical Multicast Trees and their Optimization.* IEEE Infocom'99, New-York (USA), March 1999.
7. F.Baccelli, S.Zuyev. *Poisson-Voronoi Spanning Trees with applications to the Optimization of Communication Networks.* INRIA Research Report No. 3040. Nov.1996.
8. F. Baccelli, M. Klein, M. Lebourges and S. Zuyev, *Stochastic Geometry and Architecture of Communication Networks*, Telecommunication Systems, 7, pp. 209-227, 1997.
9. M. Yajnik, J. Kurose, and D. Towsley. *Packet Loss Correlation in the MBone Multicast Network.* IEEE Global Internet Conference 1996.
10. D.Stoyan, W.Kendall, J.Mecke. *Stochastic Geometry and its Applications.* 2nd Edition, J.Wiley & Sons Chichester, 1995.
11. D.Daley, D.Vere-Jones. *An Introduction to the Theory of Point Processes.* Springer Series in Statistics, Springer-Verlag New-York, 1988.
12. A. Busson, J-L Rougier, D. Kofman. *ENST-Research report* 2001 *D* 009, available on http://www.enst.fr/~abusson