Design and Evaluation of A New and Effective Fairness Scheme for Multicasting in Internet-Scale Distributed Systems

Yijun Lu and Hong Jiang
Department of Computer Science and Engineering
University of Nebraska-Lincoln
{yijlu, jiang}@cse.unl.edu

Abstract

This paper argues that simply applying a multiple-tree scheme does not provide sufficient fairness for applications in an Internet-scale distributed system, in terms of performance. Motivated from the observation of the tax and donation systems in our society, we believe that a better way to define fairness, for performance's sake, is to factor in nodes' proportional contributions because it provides the opportunity to support many simultaneous multicasting sessions. This paper then presents a protocol, called FairOM (Fair Overlay Multicast), to enforce proportional contribution among peers in an Internet-scale distributed system.

1. Introduction

In Internet-scale distributed systems, such as the Grid and very large-scale heterogeneous clusters, reliable and efficient data dissemination plays a very important role. While multiple-tree-based multicasting, in which each peer is supposed to contribute its bandwidth exactly once by being an interior node in one tree, has been explored to address the fairness requirement [1], we argue that a better way to define fairness is to enforce that peers' contributions are proportional to their total available outgoing bandwidths. This is analogous to taxation or donation. In taxation or donation, it is desirable for people to give the same percentage of their available capital as contribution to the society (here, we assume all the people are in the same tax bracket).

Performance-wise, enforcing proportional contribution provides an environment to support multiple simultaneous multicasting sessions that may not otherwise be achieved by simply asking every peer to contribute *arbitrarily*. Consider the following example in which peers *A* and *B* are both going to multicast a movie and each multicast will span all the peers in the network. Suppose that *A* builds its multicast forest first and one peer, *C*, is assigned to contribute 90% of its outgoing bandwidth to it. Then when *B* tries to establish its multicast forest, chances are that *C* just does not have enough bandwidth to support it because it has contributed too much to the first multicast session. In this case, the construction of forest for *B* becomes either infeasible or, barely feasible by using up all of C's outgoing bandwidth and making *C* a hot-spot/bottleneck. In this case, if we instead let each peer contribute roughly the same percentage of its outgoing bandwidth, say 20%, then *C* has a chance to support the two simultaneous multicasting sessions.

In this paper, we present a protocol called FairOM (Fair Overly Multicast) to enforce proportional contributions in Internet-scale distributed systems with the assumption that all peers play by the rules.

2. Design of FairOM

Basic idea: The basic idea of FairOM is to build a multicast forest in two phases. In the first phase, the peers join the multicast group and establish the neighborhood by a pair-wise neighborhood establishment procedure and use this neighborhood information to build an initial multicast forest that may not be complete. In the second phase, a peer contacts the source to ask for any missing stripes and finally makes the forest complete.

Establishment of neighborhood: After joining the multicast group, a new peer will eventually establish its neighbor list by running a periodical neighborhood establishment procedure.

Staged spare capacity group: It is a key data structure in FairOM to enforce proportional contribution. Suppose the spare capacity group has five stages, where each stage represents a percentage range of the capacity (e.g., stage 1 represents [0%, 20%], stage 2 (20, -40%], etc), then the source will put each of the registered peers into an appropriate stage. To

illustrate this concept, we consider a simple example as illustrated in Figure 1.

In Figure 1, suppose peer *A* has a total outgoing bandwidth of 20 (i.e., it can forward 20 stripes of data) and has already contributed 3 units of the total, then its current contribution is 15% (3/20). Because *A*'s contribution is less or equal to 20%, it is put into stage 1. *B* is put into stage 2 because its contribution is in the range (20%, 40%]. Follow the same criteria, *C* and *D* are put in stage 1 and 5, respectively.

Initial forest construction: The purpose of the initial forest construction is by no means to build a complete forest, instead, it servers as a good start and provides a skeleton on which the second phase can improve. The source first sends all the stripes out and they are forwarded to different neighbors to achieve path diversity. For each peer that receives a stripe, it forwards the stripe to as many neighbors as it can within the predefined quota. In this process, if a peer receives multiple transmissions of the same stripe, it picks one and rejects others. At this stage, let us assume that a peer picks the parent that notices it first. Then a multicast relationship between a parent and a child has been established (the parent picks the child and the child accepts it).

Making the forest complete: After the forest building process starts, each peer checks with those peers that treat it as a neighbor. If all peers it contacts have already gotten some stripes and did not choose it as a child in the initial forest construction, it will seek help from the source. Moreover, a peer contacts the source anyway if a deadline has passed.

In the message it sends to the source, the peer indicates the number of times it has requested for spare capacity and the number starts with 1 (the first time). When the source receives the message (with number 1), it only looks for parents for this peer in the first stage of the spare capacity group by randomly picking one eligible parent which has this stripe.

Then it calculates what the new contribution for the parent would be. If the new contribution ratio is beyond the quota limit of this stage (20% for the first stage), the parent's record is moved from the current stage to the next higher one (stage 2 in this case).

If a parent is found, the parent will receive the adoption request of a potential child from the source, the parent will then send a request to the potential child which needs to be adopted. Thus the peer with missing stripes can get what it wants.

If the source cannot find a parent in this stage, the peer with missing stripes waits a predefined period of time before it starts the next round of request again. By following this protocol, the source relaxes the quota

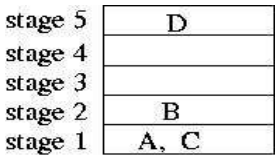

Fig. 1. Layout of the staged spare capacity group.

gradually and finally builds a complete forest in which every peer is in all trees.

Considering delay information: In the initial forest construction process, each peer sends its delay information along with the message it sent to its neighbors. When a peer receives multiple transmissions of the same stripe, it picks the one with the smallest delay and drops others. Because the dropping process is based on delay, it will not create cycles.

3. Evaluations

We measure the effectiveness of enforcing proportional contribution by *StdR*, the standard deviation of the peers' contribution. In this simulation, we run three configurations with numbers of stripes of 2, 4 and 8. In all the simulations, the algorithm satisfies the requirement to build a complete forest and satisfy all peers' bandwidth constraint. Then the mean value and *StdR* are calculated and summarized in Table 1. This result also shows that FairOM performs very well when we change the number of stripes from 2 to 8.

Table 1. Mean and Std of contribution ratios

Statistics	FairOM(2)	FairOM(4)	FairOM(8)
Mean	0.131	0.257	0.521
StdR	0.047	0.090	0.106

4. Conclusions

This paper presents the design and evaluation of FairOM, an overlay multicasting scheme for Internet-scale distributed systems. Through a two-phase forest construction process, FairOM enforces proportional contribution among peers. Simulation results show that FairOM achieves the design goal of enforcing proportional contribution.

References

[1] M. Castro, P. Druschel, A.-M. Kermarrec, A. Nandi, A. Rowstron, and A. Singh. Splitstream: High-bandwidth multicast in cooperative environment. In *Proc. of the SOSP*, Bolton Landing, New York, USA, October 2003.