Towards Deployable Large Scale End-point-based Multicast Streaming

György Dán, Viktória Fodor, Ilias Chatzidrossos and Gunnar Karlsson
Department of Signals, Sensors and Systems
KTH, Royal Institute of Technology, Stockholm, Sweden
E-mail: {gyuri,vfodor,iliasc,gk}@s3.kth.se

INTRODUCTION

Peer-to-peer overlays have proved to be an efficient means for off-line content distribution between a large number of cooperative nodes. There are several overlays coexisting in the Internet and serve the needs of millions of users. Peer-to-peer overlays are also used to provide lookup services in commercial applications, such as Skype.

But the peer-to-peer paradigm has not been that successful in all areas. The delivery of streaming media over end-point overlays has received much attention recently ([1], [2] and references therein), but such systems are still not widely used. They have found some application in the live streaming of events to a moderate number of viewers, in the order of hundreds (e.g. End System Multicast). Systems suitable for a larger population have not been developed or deployed yet.

The advantages of end-point-based multicast compared to content delivery networks are diverse. Although current commercial content delivery networks are capable of supporting many simultaneous streams, end-node-based multicast could considerably decrease the cost of large scale streaming, while being resilient to sudden surges in the client population, such as flash crowds.

In an end-point-based multicast distribution system end-points are organized or organize themselves into an application layer overlay and distribute the data among themselves. The main advantages are that such a system is easy to deploy and it reduces the load of the content provider, since the distribution cost in terms of bandwidth and processing power is shared by the nodes of the overlay.

Since the success of such schemes depends on the behavior of the participating nodes, several issues have to be dealt with, such as the effects of group dynamics, stability of the system or the incentives for nodes to collaborate. Despite of these issue there are two key requirements that a successful end-point-based multicast system should fulfill. These two requirements make end-point based streaming different from off-line content distribution.

One key requirement is robustness. The paths used for data distribution have to be updated as nodes depart from the overlay, but the updates have to be done so that the nodes that stay in the overlay suffer the least possible quality degradation. Similarly, the overlay has to be robust to data losses. The loss of packets between two adjacent nodes in the overlay should not affect nodes that receive data from the nodes connected with the erroneous path. The overlay should as well be robust to the sudden increase of the number nodes, and should possibly support a large number of spectators.

The other key requirement is the efficiency of the data distribution. Each node should receive only one copy of a piece of data. Data that reach a node more than once use bandwidth resources unnecessarily. At the same time, the delay and delay jitter that individual packets experience should be bounded.

In this work we evaluate the performance of peer-to-peer streaming architectures with mathematical models and simulations. We consider the effects of node dynamics and packet loss on the quality of the received information. We find a trade-off between the architectures' capability to recover lost information and the stability of the overlay in the presence of node departures. Based on these findings we propose a generalized architecture that inherits the good properties of the earlier solutions. We evaluate how the arhitectures perform under various delay constraints and address the problem of dynamic control of data transmission.

END-POINT BASED STREAMING SYSTEMS

Many different architectures have been proposed to solve the problem of end-point-based multicast, but only a few of them are suitable for large scale distribution. The architectures can be split into two groups, mesh based and tree based. Mesh based overlays are robust to node failures, but they do not scale well (e.g. End System Multicast [3]). Single tree based overlays offer scalability, but are more sensitive to node departures and data loss (e.g. NICE [4]). Overlays based on multiple distribution trees [1], [2] combine the robustness of mesh based systems and the efficiency and scalability of tree based systems.

In the following we describe an overlay, which generalizes the overlays presented and evaluated in [1], [2]. The overlay consists of a root node (the source of the streaming media content) and N peer nodes. Peer nodes are organized in t distribution trees, either by a distributed protocol [2] or a central entity like in [1]. We say that a node is in layer i in tree t, if the node is i hops away from the root node in tree t. The nodes are members of all t trees, and in each tree they have a different parent node from which they receive data, if it is possible. Child nodes of the root node can have the same

[1]This work has been supported in part by E-NEXT.

parent (i.e. the root) in more than one tree. Each node has up to t child nodes to which it forwards data. We denote the number of children of the root node in each tree by m, and we call it the multiplicity of the root node. We assume that nodes do not contribute more bandwidth towards their children as they use to download from their parents, so that the multiplicity of the peer nodes is one.

The root uses block based FEC, e.g. Reed-Solomon codes, so that nodes can recover from packet losses due to network congestion and node departures. To every k packets of information c packets of redundant information are added resulting in a block length of $n = k + c$. If a source would like to increase the ratio of redundancy while maintaining its bitrate unchanged, then it has to decrease its source rate. We denote this FEC scheme by FEC(n,k). Using this FEC scheme one can implement UXP, PET or the MDC scheme considered in [1]. In case of losses the lost packets can be reconstructed as long as no more than c packets are lost out of n packets. Since all nodes of the overlay may reconstruct lost packets and can distribute them to their children, it may be possible to propagate information arbitrarily far away from the root node. The root sends every t^{th} packet to its children in a given tree. Peer nodes forward data packets in at most d $(1 \leq d \leq t)$ distribution trees. Peer nodes relay the packets upon reception to their respective child nodes in the tree corresponding to the particular packets, if they have to forward data in the given tree. Once a node received at least k packets of a block of n packets it recovers the remaining c packets and sends them to the child nodes in the corresponding distribution trees, if it has to forward data in the given trees. A packet received from the parent node after it has been decoded based on other packets in the block will be discarded.

This overlay is a generalization of the overlays considered in [1]. By setting $d = t$ we get the minimum breadth tree [1], and by setting $d = 1$ we get the minimum depth tree evaluated in [1], [2]. If $d = 1$, then each node forwards packets in one tree only, and it has to be a leaf node in all other trees. If $d = t$, then each node can be an interior node in all trees, and by preference it should be located in the same or nearly the same layer in all trees to keep the time between the arrival of the packets in the different trees low.

PERFORMANCE EVALUATION

Previous analysis of the end-point-based multicast overlays described above was limited to simulations. Most of the work considered the $d = 1$ case for two reasons. First, allowing the nodes to forward packets in one tree only results in the lowest number of layers. Second, it was assumed that such a tree is more resilient to node departures, because the average number of children of the nodes is lower than for $d = t$. Simulation results shown in [1] showed that the overlay with $d = 1$ performs better than the overlay with $d = t$. Simulations shown in [2] aimed at evaluating the stability and the efficiency of the overlay with $d = 1$. The authors of both papers concluded that the overlay is sufficiently stable and efficient for the parameters considered in the simulations. But it is not clear from the results, under which circumstances the system remains stable, and how efficient it is under different network conditions, such as link failures and different rates of node departures. The lack of general results describing the behavior of these overlays led us to developing analytical models that could help to understand the behavior of the overlays, and could lead to the design of overlays with better properties.

We developed an analytical model for the overlay with $d = t$ in the presence of link failures [5]. Using the theory of discrete dynamic systems we showed that as long as the stationary packet loss probability between two adjacent nodes is below a certain threshold, such an overlay can deliver data arbitrarily far away from the root node with an arbitrarily high probability. The reception probability drops to zero if the threshold is exceeded. As an example, Fig. 1 shows the packet reception probability in layer 1000 of an overlay as a function of the loss probability between peer nodes for different FEC schemes. The model can be used to study the

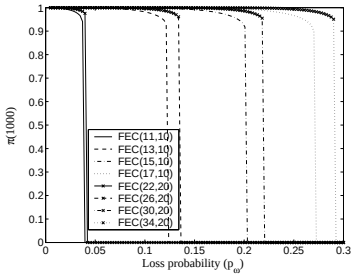

Fig. 1. Packet reception probability vs. loss probability for various FEC schemes.

effects of node dynamics as well and is an easy way to predict the performance of the overlay. Although the overlay with $d = t$ is robust to failures, its feasibility is limited. The depth of the tree is $O(N)$, and if N is large, nodes far away from the root experience large delays and possibly delay jitters.

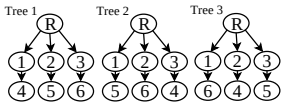

Fig. 2. Multicast tree structure for $t = 3$, $m = 3$, $N = 6$ and $d = t$.

Our results for the overlay with $d = 1$ show similarities with those for the overlay with $d = t$. The analytical model we developed shows that the packet reception probability can be made arbitrarily high under the assumption that the reception of a packet in a tree is independent of the reception of the packets in the rest of the trees. It is clear that the independence assumption does not necessarily hold, but simulations ran with an event driven simulator show that the possible correlations do not have a significant effect on the performance of the overlay. Due to that the distance of the nodes from the root node in different trees differs significantly, FEC reconstruction

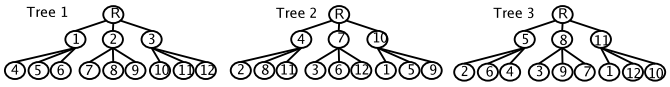

Fig. 3. Multicast tree structure for $t = 3$, $m = 3$, $N = 12$ and $d = 1$.

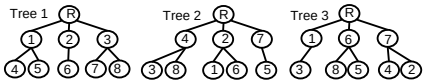

Fig. 4. Multicast tree structure for $t = 3$, $m = 3$, $N = 8$ and $d = 2$.

can introduce considerable delay. Hence, we expect that the delay tolerance of the end-nodes influences the performance of the overlay. It is an open question how the overlay for data distribution can be changed if one increases the delay budget of the end-points.

Analyzing the overlay we show that the it can not be constructed if $m < t-1$, that is, when the root has less children per tree than the number of trees minus one. Hence, such a scheme is not suitable for multicast streaming from a low bandwidth node (e.g. home broadcasting). For $m \geq t-1$ the trees can be constructed, but in the presence of node departures there is a high probability of that a node will not be able to find a parent, as the number of active nodes per tree can become unbalanced. The tree management algorithm has to reallocate nodes between the trees to resolve such an issue. Reallocating nodes is not feasible in a large overlay and should be avoided. In the overlay with $d = t$ such a phenomenon can not occur.

We can combine the good properties of the two overlays by allowing d to be between 1 and t. In such a tree, each node can be an interior node in at most d trees and is a leaf node in at least $t-d$ trees. The number of layers in the overlay for $d < t$ is $O(logN)$ like for $d = 1$, so that the trees are shallow similar to the overlay with $d = 1$. At the same time, the trees are more balanced in terms of active nodes per tree, since each node can forward data in $d \geq 2$ trees. This increases the probability of that a node finds a parent despite of the node departures. The mathematical models we developed and extensive simulations show that the packet possession probability behaves the same way as for $d = 1$, while the probability of that nodes do not find a parent is orders of magnitud lower.

The probability of that a node finds a parent is influenced by the tree management scheme as well, whether it is centralized or distributed. A central entity can construct trees with the lowest possible depth and with the highest average number of children per node, a distributed algorithm is unlikely to achieve this goal. Hence, the generalized overlay is more suitable for use in conjunction with a distributed tree management scheme. We develop quantitative measures to compare the performance of the overlays and perform extensive simulations to validate the measures.

The right choice of FEC parameters is crucial for these overlays. In a dynamic environment setting the parameters has to be done based on feedback from the peer nodes. As our mathematical models have shown it, setting the ratio of redundancy too low results in a high penalty. For this reason we are devising a feedback scheme, which avoids the underestimation of the packet losses. The scheme has to introduce low overhead, should be scalable and has to react quickly to changes, but should avoid oscillations. One prospective candidate is the filter based approach [6]. In a filter based feedback scheme, individual nodes send updates to their parents if the difference between the last reported value and the actual value of the measure of interest (the packet loss probability in our case) is larger than an adjustable threshold. Nodes aggregate and filter updates arriving from their child nodes to avoid flooding of the network.

An alternative to setting the FEC parameters based on a feedback scheme is to use incremental redundancy and let individual nodes decide how much redundancy they want to receive depending on the packet loss statistics they experience. In such a scheme not all nodes have to receive the data in all trees. Nodes with high bandwith can leverage the extra redundancy to achieve better quality. At the same time they can reconstruct more lost packets and can improve the reception for other nodes. It is not clear yet how the presence of trees where not all nodes subscribe would affect the available capacity in the overlay. Whether such a scheme performs better than the one based on feedback is not clear and it will subject of our future work to compare these two solutions.

CONCLUSION

Our work aims at designing a robust and efficient overlay architecture to be used for large scale end-point-based multicast streaming. We developed mathematical models to understand the behavior of existing overlays. Based on our knowledge on those overlays we devised an overlay, which is the generalization of two existing overlays, and inherits their robustness and efficiency. The mathematical models show that the overlay can distribute data arbitrarily far away from the root node if enough redundancy is used. We will design a feedback scheme to set the amount of redundancy in a scalable way. We expect the result of our work to be an overlay for end-point-based multimedia streaming suitable for large scale data distribution in practice.

REFERENCES

[1] V. N. Padmanabhan, H.J. Wang, and P.A Chou, "Resilient peer-to-peer streaming," in *Proc. of IEEE ICNP*, 2003, pp. 16–27.

[2] K. Sripanidkulchai, A. Ganjam, B. Maggs, and H. Zhang, "The feasibility of supporting large-scale live streaming applications with dynamic application end-points," in *Proc. of ACM SIGCOMM*, 2004, pp. 107–120.

[3] Y. Chu, S.G. Rao, S. Seshan, and H. Zhang, "A case for end system multicast," *IEEE J. Select. Areas Commun.*, vol. 20, no. 8, 2002.

[4] S. Banerjee, B. Bhattacharjee, and C. Kommareddy, "Scalable application layer multicast," in *Proc. of ACM SIGCOMM*, 2002.

[5] Gy. Dán, V. Fodor, and G. Karlsson, "On the asymptotic behavior of end-point-based multimedia streaming," in *Proc. of International Zürich Seminar on Communication*, 2006.

[6] C. Olston, J. Jiang, and J. Widom, "Adaptive filters for continuous queries over distributed data streams," in *Proc. of ACM SIGMOD*, 2003, pp. 563–574.