Approaches toward Maintaining Bi-connectivity for Resilience in Overlaid Multicasting

Ashutosh Singh and Yatindra Nath Singh

Department of Electrical Engineering,
Indian Institute of Technology,
Kanpur-208016, Uttar Pradesh, India
{singhash, ynsingh}@iitk.ac.in

Abstract

In Application layer multicast (ALM) (also called Overlay Multicast), multicast-related functionalities are moved to end-hosts. The key advantages, overlays offers, are flexibility, adaptability and ease of deployment [1]. Application layer multicast builds a peer-to-peer (P2P) overlay topology consisting of end-to-end unicast connections between end-hosts. End users self organize themselves into logical overlay networks for efficient data delivery. The major concern in designing ALM protocol is how to build and maintain a topology, to route data efficiently and reliably. We propose here a two-fold dynamic overlay tree construction and maintenance scheme in which a mesh-like topology is first built, and on top of it, a single or multiple data delivery tree(s) are built. While forming the mesh, it is ensured that two node disjoint paths are maintained between every possible pair of nodes. The overlay tree is built such that every host gets data feeds via two different paths. A fully dynamic algorithm is run over the overlay topology to insert new links to maintain biconnectivity while deleting the redundant links.

Key words: Overlay network, Reliability, node disjoint paths, Biconnectivity

1 INTRODUCTION

The Internet has seen an unprecedented growth due to the success of one-to-one applications such as reliable file transfer and electronic mail. The Internet’s unicast-only infrastructure does not provide efficient support for multicast applications viz. Internet-TV, Video conferencing, Live Lecture Delivery Systems (LDS) and software distribution over Internet where multiple copies of a message need to be transported to multiple recipients at different locations. IP multicast [2], at network layer defines an efficient way for multicast whereby the sources transmit only one copy of the data and the appropriate network nodes make duplicate copies for each receiver. However IP multicast has not been widely deployed due to the following limitations [3]. IP multicast requires routers to maintain per-group state which leads to serious scaling constraints. IP multicast calls for changes at infrastructure level (Multicast enabled routers are required) in the internet which was primarily designed for unicast applications. IP multicast is based on best-effort data delivery, and hence cannot support QoS. This is not desirable in applications such as multi-party gaming and multi-party conferencing. Finally IP multicast does not provide solution for group management, multicast address allocation and support for network management. As a result, most of the networks have not enabled multicast or provide it as a value added service. So far multicast has been limited to ‘island’ of network domains under single administrative control or in local area networks.

Application layer multicast (ALM) also known as Overlay Multicast is an attractive alternative solution, where multicast-related functionalities are moved to end-hosts. The key advantages, overlays offer, are flexibility, adaptability and ease of deployment. Application layer multicast builds a peer-to-peer (P2P) overlay topology consisting of end-to-end unicast connections between end-hosts. Multicast topology is maintained in this overlay network. End users can constantly self organize the logical overlay networks for efficient data delivery. The shifting of multicast support from routers to end systems has the potential to address most problems associated with IP multicast. Since the multicast connections are based on end hosts, there is no need of multicast enabled routers.

However ALM incurs a performance penalty over IP Multicast. Links near the end users carry redundant copy of data transmission and also the delay to some of the end users is more than in case of IP multicast. The major concern in ALM is how to route data along the topology efficiently and reliably. To enhance the reliability, we propose here a two-fold dynamic overlay tree construction and maintenance scheme in which a mesh-like topology is built first. A new host connects to two already connected hosts in the overlay network. This results in at least two node disjoint paths between any pair of nodes. Once the mesh is formed, on top of that a single or multiple data delivery tree(s) are built, using appropriate distributed algorithm. The overlay trees are built such that every host has potential to get data feed from two different paths.

To fulfill this, data is also transmitted over some extraneous links which are not the part of the tree. For example in figure 1, node 4 is getting data from one path S-1-2-3-4, where S is the source, but data is also being
transmitted over the links S-5 and 5-4, so that if current path get disturbed, data can be obtained from the path S-5-4.

![Diagram of double feed to node 4 from two node disjoint paths](image)

Figure 1 Double feed to node 4 from two node disjoint paths

A fully dynamic algorithm is run in the trees to insert new links maintaining biconnectivity and also to delete the redundant links. The next section gives a brief survey of related work. Section III describes proposed approaches toward maintaining the biconnectivity. Finally we conclude in section IV.

II RELATED WORK

In this section, we describe the proposed reliability schemes in major ALM protocols. In NARADA protocol [4], Chu et. al. proposed a mesh based topology design for small group size. Data delivery trees are constructed entirely from the overlay links present in the mesh. Shortest path spanning trees are constructed per source by running distance vector algorithms on top of the mesh. Each member periodically generates a refresh message with monotonically increasing sequence number and periodically exchanges its knowledge of group membership with its neighbors in the mesh. Mesh partition is detected, when members on one side of the partition stop receiving sequence number updates from members on the other side. Each of such members is probed to determine if it is dead, else a link is added to it. Each of the members on one side may attempt to add new links to some partitioned member on other side, this situation is probabilistically resolved such that in spite of several members simultaneously attempting to repair partition; only a small number of new links are added.

To improve the mesh quality, dynamic addition and dropping of links is also done. Each member periodically probes some random member that is not a neighbor, and evaluates the utility of adding a link to this member. Link is added if expected utility gain exceeds a given threshold. To drop a link, members periodically compute the consensus cost of its link to every neighbor. The consensus cost of a link \((i, j)\) is defined as \(\max(\text{cost}_i, \text{cost}_j)\), where \(\text{cost}_i\) is the number of members for which \(i\) uses \(j\) as next hop for forwarding packets, and \(\text{cost}_j\) is the number of members for which \(j\) uses \(i\) as next hop for forwarding packets. The link with lowest consensus cost is dropped if the consensus cost falls below a certain threshold.

In PRM enhanced NICE protocol [5], Banerjee et. al. introduced multicast data recovery scheme called Probabilistic Resilient Multicast (PRM) with two components. A proactive component called randomized forwarding in which each overlay node chooses a constant number of other overlay nodes uniformly at random and forward data to each of them with a low probability and a reactive component called triggered NAKs to handle data losses due to link errors and network congestion.

III OUR APPROACHES TOWARD BICONNECTIVITY FOR RESILIENCE

The work presented in this paper suggests approaches toward maintaining bi-connectivity in data distribution topology in an application layer multicast networks.

The basic concept behind these approaches lies in the principle that if many nodes form a ring then between any node pair inside the ring, two node disjoint path exist. Further if two such rings (or two bi-connected components) share a common link or more than one connected links then those bi-connected components are again bi-connected.

2.1 Approach 1

In the First approach, an algorithm is suggested to construct a bi-connected Mesh over which a distribution tree can be formed.

Adding newly arriving Node in the Network

Each new node connects to two already existing nodes in the network. In the beginning, when first node comes, there is no network and new node is the first node in the network (figure 1 a). The second node joins the first node. The third node connects to the first and second node forming a triangle (figure 1. c). Fourth node connects to any two of the three nodes. The fifth arriving node sees nodes 1 and 4 as having least degree hence it connects to first and fourth node. Mesh formation sequence is shown in figure 2 (a) to (e).

The rule followed is that each arriving node connects to two already existing nodes with least degrees. In case two or more than two nodes have the same degree, any two can be used for the connection.
Fig. 2 (a) to (e) Mesh Formation as nodes join to the network gradually

Continuing this way, a bi-connected mesh can be formed with any number of nodes. An example mesh with 12 nodes is shown in figure 3.

Fig. 3 A bi-connected mesh with 12 nodes

**Analysis of the Mesh so formed**

In the mesh so formed, the average degree of nodes in the mesh, as defined below increases slightly with the number of nodes and reaches to a constant value of 4 as the number of nodes increases to more than 100 (figure 4). The average degree of nodes in a network with N number of nodes can be obtained as below.

\[
\text{Average degree of network} = \frac{\sum_{n=1}^{N} \text{degree of node } n}{N}
\]

The average number of hops travelled for any source destination node pair in the mesh is calculated and plotted in figure 5. It is observed that the number of hops increase linearly with the number of nodes.

Fig. 4 Average node degree versus network size

Fig. 5 Average number of hops, averaged over all possible node pairs

**Deleting redundant Links from the Network**

For each existing links in the network the decision is taken periodically whether to maintain that link or to delete it. If the nodes connected by a link have two node disjoint paths even when the link is removed, that link is declared to be redundant and can be deleted from the mesh if no traffic is passing through it.

**Steps taken in case of node failure**

In case any node fails in the network, the attempt is made to create a ring among the first hop connected nodes to that failed node. For example, in figure 6, if node 7 fails, links 5-7, 6-7, 8-7 and 9-7 are removed and a new link 5-9 is created so that all affected nodes form a ring and thus maintaining bi-connectivity.
Advantages and limitations to approach 1

In the mesh so formed, maximum degree that a node can have is limited to 4. But the maximum number of hops between any pair of nodes increases linearly as the number of nodes increase.

2.2 Approach 2 and 3

In direct-tree protocols, biconnectivity can be achieved in two different ways as described below.

Connected leaf nodes approach

Biconnectivity is achieved by connecting all the leaf nodes in pair. By connecting leaf nodes this way, any node pair in the tree gets bi-connectivity, as rings are formed through leaf nodes. Here we have taken an example of binary tree (shown in figure 7 a) having 4-level hierarchy. Once the leaf nodes are connected (shown in figure 7 b), rings are formed between any pair of nodes.

The number of additional links required to achieve biconnectivity in this approach is (N-1)/2; where N is the number of nodes in the network.

Child-Grandparent approach

Biconnectivity in this approach is achieved by connecting all the nodes to their grand parent, which provides an alternate path to get the feed in case their parents fail. Wherever it is not possible to connect a node to its grandparent, that node is connected to its siblings. Again an example of binary tree (shown in figure 7 a) having 4-level hierarchy is considered. Figure 8 a shows the original tree along with the new added links, while figure 7 b shows only the new added links.

In the child-grandparent approach the number of additional links required is (N-2); where N is number of nodes in the network.

The plot in figure 9 shows that the child-grandparent approach uses almost double the links to achieve Biconnectivity as compared to leaf node approach.
References

[1] S. Fahmy and M. Kwon, “Characterizing Overlay multicasting Networks and their costs”, IEEE/ACM Transaction on Networking, Vol. 15, No. 2, April 2007.

[2] J. F. Kurose and K. W. Ross, “Computer Networking: A top-down approach featuring the Internet, 2nd Edition, Addison Wesley, 2002.

[3] C. Diot et.al., “Deployment issues for the I P Multicast service and architecture”, IEEE Network, Vol. 14, Issue 1, January 2000.

[4] Yang-hua Chu et. al, “A case for end system multicast,” IEEE Journal on Selected Areas in Communications, vol. 20, no. 8, pp. 1456–1471, October 2002.

[5] S. Banerjee, et. al, “Resilient Multicast using Overlays”, IEEE/ACM transaction on networking, vol. 14, no. 2, April 2006.