FatNemo: Multisource Multicast Overlay Fat-Trees

Stefan Birrer Fabián E. Bustamante Dong Lu Peter A. Dinda Yi Qiao

High-bandwidth multisource multicast among widely distributed nodes is critical for a wide range of important applications including audio and video conferencing, multi-party games and content distribution. Multicast decouples the size of the receiver set from the amount of state kept at any single node and potentially avoids redundant communication in the network.

The limited deployment of IP Multicast [10], [11] has led to considerable interest in alternate approaches implemented at the application layer that rely only on end systems [9], [12], [2], [15], [8]. In an end system multicast approach, participating peers organize themselves into an overlay topology for data delivery. Each edge in the topology corresponds to a unicast path between two peers in the underlying Internet. All multicast-related functionality is implemented at the peers, instead of at the routers of the underlying network, and the goal of the multicast protocol is to construct and maintain an efficient overlay for data transmission.

Among the proposed end system multicast protocols, tree-based systems have proven to be highly scalable and efficient in terms of physical link stress, state and control overhead, and end-to-end latency [12], [2], [7]. However, normal tree structures have inherent problems in terms of resilience and bandwidth capacity. Trees are highly dependent on the reliability of non-leaf nodes. Resilience is especially relevant to the application-layer approach, as overlays are composed of autonomous, unpredictable end systems. The high degree of transiency of end systems is one of the main challenges for these architectures [3]. Furthermore, trees are likely to be bandwidth constrained since bandwidth availability monotonically decreases as one ascends from the leaves. The bandwidth limitations of normal tree structures are particularly problematic for multisource, bandwidth-intensive applications. For a set of randomly placed sources in a tree, higher-level paths (those near the root) will become the bottleneck under high load, and tend to dominate delivery latencies. Once these links become heavily loaded or overloaded, packets will be buffered or dropped.

We have addressed the resilience issue of tree-based systems by exploiting the alternative paths introduced through *co-leaders* [4]. In this work we address the bandwidth constraints of conventional trees by importing Leiserson's *fat-trees* from parallel computing into overlay networks. Paraphrasing Leiserson, a fat-tree is similar to a real tree in that its branches become thicker as one moves away from the leaves [13]. By increasing the number of links closer to the root, a fat-tree can overcome the "root bottleneck" likely to be found by multisource multicast applications relying on conventional trees. The adoption of a fat-tree approach for overlay multicast *(i)* lowers the forwarding responsibility of the participating nodes, thus increasing system scalability to match the demands of high-bandwidth, multisource multicast applications [6], [14], [16]; *(ii)* reduces the height of the forwarding tree, hence significantly shortening delivery latencies; and *(iii)* improves the system's robustness to node transiency by increasing path diversity in the overlay [1], [6], [14].

We introduce the design, implementation and performance evaluation of *FatNemo* [5], a new application-layer multicast protocol that builds on this idea, and report on a detailed comparative evaluation.

REFERENCES

[1] D. Andersen, H. Balakrishnan, F. Kaashoek, and R. Morris. Resilient overlay networks. In *Proc. of the 18th ACM SOSP*, October 2001.

[2] S. Banerjee, B. Bhattacharjee, and C. Kommareddy. Scalable application layer multicast. In *Proc. of ACM SIGCOMM*, August 2002.

[3] M. Bawa, H. Deshpande, and H. Garcia-Molina. Transience of peers & streaming media. In *Proc. of HotNets-I*, October 2002.

[4] S. Birrer and F. E. Bustamante. Resilient peer-to-peer multicast without the cost. In *Proc. of MMCN*, January 2005.

[5] S. Birrer, D. Lu, F. E. Bustamante, Y. Qiao, and P. Dinda. FatNemo: Building a resilient multi-source multicast fat-tree. In *Proc. of IWCW*, October 2004.

[6] M. Castro, P. Druschel, A.-M. Kermarrec, A. Nandi, A. Rowstron, and A. Singh. Splitstream: High-bandwidth multicast in cooperative environments. In *Proc. of the 19th ACM SOSP*, October 2003.

[7] M. Castro, M. B. Jones, A.-M. Kermarrec, A. Rowstron, M. Theimer, H. Wang, and A. Wolman. An evaluation of scalable application-level multicast built using peer-to-peer overlays. In *Proc. of IEEE INFOCOM*, March 2003.

[8] M. Castro, A. Rowstron, A.-M. Kermarrec, and P. Druschel. SCRIBE: A large-scale and decentralised application-level multicast infrastructure. *IEEE Journal on Selected Areas in Communication*, 20(8):1489–1499, October 2002.

[9] Y.-H. Chu, S. G. Rao, S. Seshan, and H. Zhang. A case for end system multicast. *IEEE Journal on Selected Areas in Communication*, 20(8):1456–1471, October 2002.

[10] S. E. Deering. Multicast routing in internetworks and extended LANs. In *Proc. of ACM SIGCOMM*, August 1988.

[11] C. Diot, B. N. Levine, B. Lyles, H. Kassem, and D. Balensiefen. Deployment issues for the IP multicast service and architecture. *IEEE Network*, 14(1):78–88, January/February 2000.

[12] J. Jannotti, D. K. Gifford, K. L. Johnson, M. F. Kaashoek, and J. W. O'Toole Jr. Overcast: Reliable multicasting with and overlay network. In *Proc. of the 4th USENIX OSDI*, October 2000.

[13] C. E. Leiserson. Fat-trees: Universal networks for hardware-efficient supercomputing. *IEEE Transactions on Computers*, 34(10):892–901, October 1985.

[14] V. N. Padmanabhan, H. J. Wang, P. A. Chou, and K. Sripanidkulchai. Distributing streaming media content using cooperative networking. In *Proc. of NOSSDAV*, May 2002.

[15] S. Ratnasamy, M. Handley, R. Karp, and S. Shenker. Application-level multicast using content-addressable networks. In *Proc. of NGC*, November 2001.

[16] A. Young, J. Chen, Z. Ma, A. Krishnamurthy, L. Peterson, and R. Y. Wang. Overlay mesh construction using interleaved spanning trees. In *Proc. of IEEE INFOCOM*, March 2004.

The authors are with the Department of Computer Science, Northwestern University, Evanston, IL 60201, USA. Emails: {sbirrer, fabianb, donglu, pdinda, yqiao}@cs.northwestern.edu