

One-to-many TCP overlay

Zhen Liu

IBM

Thomas J. Watson Research Center

Yorktown Heights, NY 10598, USA

Abstract

This work addresses two key issues in reliable multicast overlay networks: End-to-end reliability in the presence of node failure/overflow, and throughput scalability in the presence of random perturbations. A decentralized architecture is proposed for taking care of these two issues. This architecture uses distinct point to point TCP connections between adjacent pairs of end-systems, together with a window based back-pressure control, which links adjacent pairs of TCP connections via an application-layer mechanism. Each end-system maintains forwarding buffers and a back-up buffer storing copies of the forwarded packets. The forwarding buffers are used to enforce the back-pressure mechanism whereas the back-up buffer is dedicated to re-establishing connectivity in case of end-system failure. This architecture, that we propose to call *the One-to-Many TCP Overlay*, is a natural extension of TCP to the one-to-many case, in that it adapts the rate of the group communication to local congestion in a decentralized way via the window back-pressure mechanism.

Using theoretical investigations, experimentations in the Internet, and large network simulations, we show that this architecture provides end-to-end reliability and can tolerate multiple simultaneous node failures, provided the backup buffers are sized appropriately. We also show that under random perturbations caused by cross traffic described in the paper, the throughput of this reliable group communication is always larger than a positive constant, that does not depend on the group size. This scalability result, which is of independent interest, contrasts with known results about the non-scalability of IP-supported multicast for reliable group communication.

*Joint work with François Baccelli (francois.bacelliens.fr),
Augustin Chaintreau (augustin.chaintreauens.fr), and
Anton Riabov (riabovus.ibm.com).*