第44卷 第2期 西 安 交 通 大 学 学 报 Vol. 44 №2
2010年2月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Feb. 2010

一种支持大规模应用层组播的混合拓扑覆盖网络

姜山[1,2]，郑庆华[1,2]，刘晓[1,2] 董巍[1,2]
(1.西安交通大学计算机科学与技术系，710049，西安；2.西安交通大学智能网络与网络安全教育部重点实验室，710049，西安)

摘要：提出一种支持大规模应用层组播的混合拓扑覆盖网络——H-ALM.将传输覆盖网络分为主干层和分支层，主干层采用树状结构，分支层采用网状结构.H-ALM克服了单一树状结构传输稳定性差和单一网状结构传输延迟大的不足，同时综合考虑节点规模、节点的带宽能力、节点累计转发和接收数据量等实时的状态和网络条件，从而进一步优化了混合式传输覆盖网络.实验表明，H-ALM比树状方法更稳定，比网状方法延迟更小.

关键词：应用层组播；混合拓扑；覆盖网络；稳定性；传输延迟
中图分类号：TP393 **文献标志码**：A **文章编号**：0253-987X(2010)02-0011-04

A Hybrid-Topology Overlay Network for Supporting Large-Scale Application-Level Multicast

JIANG Shan[1,2], ZHENG Qinghua[1,2], LIU Xiao[1,2], DONG Wei[1,2]
(1. Department of Computer Science and Technology, Xi′an Jiaotong University, Xi′an 710049, China; 2. Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi′an Jiaotong University, Xi′an 710049, China)

Abstract: A hybrid-topology overlay network for large-scale application-level multicast—H-ALM is proposed. The overlay network is divided into two tiers by H-ALM. One is the backbone layer that has a tree-based topology; and the other is the branch layer that has a mesh-based topology. It is known that the stability of the pure tree-based topology is not good enough for distributing streaming data; and stream delivery over the pure mesh-based topology suffers from high delay yet. The H-ALM overcomes these shortcomings by combining the merits of the tree-based topology and the mesh-based topology, and optimizes the two-tier hybrid overlay network by considering the real-time status and network conditions such as the scale of the overlay, bandwidth capacity of nodes, the cumulative data traffic that each node relayed and received and so on. Therefore, the performance of stability and communication delay of distributing streaming data are improved. Experimental results show that the H-ALM achieves better stability than the tree-based overlay and lower transmission delay than the mesh-based overlay.

Keywords: application-level multicast; hybrid-topology; overlay network; stability; transmission delay

近年来，基于P2P网络的应用层组播技术运用于流媒体直播已成为研究热点，如何有效地保证大规模应用层组播的实时性和稳定性是流媒体直播系统面临的一个重要难题.目前，应用层组播方法分为

收稿日期：2009-08-11. 作者简介：姜山(1979—)，男，博士生；郑庆华(联系人)，男，教授，博士生导师. 基金项目：国家杰出青年基金资助项目(60825202)；国家自然科学基金资助项目(60633020，60803079)；国家科技支撑计划资助项目(2006BAJ07B06)；国家高技术研究发展计划资助项目(2008AA01Z131)；新世纪优秀人才支持计划资助项目(NECT-08-0433).

两类：基于树状拓扑结构的，如 MOTOR[1]、DCMD[2]、ZigZag[3] 等；基于网状拓扑结构的，如 PRO[4]、DONet[5]、Gridmedia[6] 等. 前者传输延迟小，但稳定性差，对节点的动态性、接入网络的异构性和网络带宽的抖动适应能力差[7-8]，后者传输稳定性好，能够很好地适应异构网络环境和节点的动态性，但传输延迟大[7, 9]. 针对现有方法的不足，本文提出一种支持大规模应用层组播的混合拓扑覆盖网络(H-ALM)，将传输覆盖网络分为主干层和分支层，主干层采用树状结构，分支层采用网状结构，同时可根据用户节点规模、节点的带宽能力、已有的出度、转发与接收数据量的大小等实时状态和网络条件，对主干层和分支层覆盖网络进行优化.

1 混合拓扑结构

H-ALM 的主干层节点由专门的数据转发代理节点或性能较好的用户节点组成，并称之为数据传输单元 (Data Transmitting Unit，DTU). 主干层采用树状拓扑结构传输数据，且 DTU 充当分支层源节点. H-ALM 的分支层节点由一般的用户节点组成，构成基于 Gossip 协议的网状组播组. 在 DTU 部署不足的情况下，H-ALM 允许挑选能力强、稳定性好的用户节点充当 DTU 来加入主干层. H-ALM 的整体结构如图 1 所示.

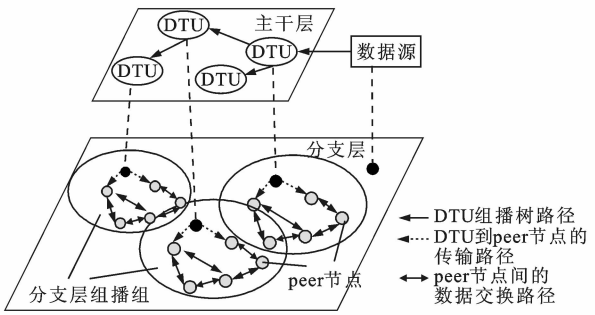


图 1 H-ALM 的整体结构(两层混合式拓扑结构)

2 主干层组播树的构建

2.1 节点加入组播树

在搭建主干层组播树时，如果仅考虑延迟优先，会使主干层节点占据上层节点的出度，那么分支层节点就只能加载在下层叶子节点之中，这样使得延迟性能整体下降. 因此，本文对主干层节点的出度进行了约束：约束连接主干层叶子节点的出度，以使用户节点能够连接到主干层节点；约束连接分支层节点的出度，以预留资源，为后续增加主干层组播树的层数所用. 主干层组播树节点出度的约束公式为

$$\left.\begin{aligned} D_{\mathrm{p}} &= \max(n, D_{\mathrm{m}}) D_{\mathrm{m}} / (n + D_{\mathrm{m}}) \\ D_{\mathrm{b}} &= D_{\mathrm{m}} - D_{\mathrm{p}} \end{aligned}\right\} \tag{1}$$

式中：D_{p} 表示连接分支层节点的出度；D_{m} 表示组播树中的节点最大出度，即组播树节点最多可支持的转发路数；D_{b} 表示连接主干层节点的出度.

新节点 i 在加入主干层组播树时，应从根节点 r 开始逐层扫描叶子节点，并将节点 r 作为候选父节点. 在每一层的扫描过程中，首先计算节点 r 通过候选父节点到达节点 i 的延迟 T_i，然后计算节点 r 通过候选父节点的孩子节点到达节点 i 的最小延迟 $\min T_i$，并将延迟最小的节点记为 $\min c$. 如果 $\min T_i \geqslant T_i$，且候选父节点满足 D_{b} 的约束，则节点 i 加入组播树，否则将 $\min c$ 替换为候选父节点，继续下一层叶子节点的比较，直到节点 i 加入组播树或候选父节点再没有叶子节点为止.

2.2 主干层与分支层的约束关系

在主干层节点(或 DTU)部署不足的情况下，H-ALM允许将性能高、稳定性好的用户节点提升到主干层，以减小网状组播传输延迟. 在图论中，平均传输延迟可近似地抽象为平均距离，所以本文用 P2P 网络的平均距离来确定组播树层数的增加情况，以 $1/(k+1)$ 作为连接的度量值，并记 k 为节点的平均邻居节点数. 根据文献[10]，则有 $\langle h^n \rangle = \langle k(k-1)\cdots(h-n+1) \rangle$，所以 $\langle h \rangle = \langle k \rangle$，$\langle h^2 \rangle = \langle k(k-1) \rangle$. 因此，平均距离

$$l = \frac{\ln N - 2\langle \ln k \rangle + \ln(\langle k^2 \rangle - \langle k \rangle) - \gamma}{\ln[\langle k \rangle - 1]} + \frac{1}{2} \tag{2}$$

当主干层组播树的层数小于 l 时，需要将用户节点提升为主干层的节点，此时增加组播树的层数，即在底层叶子节点的邻居节点中选择优先度 π 最大的节点作为新一层的叶子节点并加入到主干层的组播树之中. 优先度 π 是衡量节点综合能力的参数，它综合考虑了节点带宽和实时状态等因素. π 越大，节点的带宽就越宽，能力就越强. 考虑到网络的环境和节点的动态变化，处于主干层的用户节点需要定期收集与自己直接相连的邻居节点的 π，如果发现邻居节点的 π 比自己的大，则进行位置交换，这样做可将更合适的分支层节点提升到主干层.

2.3 节点退出组播树

节点退出组播树分正常退出和非正常退出. 节点 i 的正常退出步骤为：①节点 i 与父节点断开连接；②节点 i 的子节点向其父节点请求连接，并选择

往返时间(Round-Trip Time，RTT)最小的子节点替代节点 i，其余的子节点执行以其父节点为根节点的加入算法，快速重构组播树；③与其相连接的分支层节点执行分支层节点加入算法，并与原有的上层邻居节点断开链接；④在服务中心列表中删除节点. 节点的异常退出可通过父节点与孩子节点之间的心跳连接来确定，若父节点失效，服务中心会通知孩子节点重新加入组播树；若孩子节点失效，在父节点的孩子列表中删除该节点，并通知服务中心更新组播树拓扑.

3　分支层组播网的构建

Gossip 是一种面向 P2P 系统的通信协议，具有良好的可扩展性. 在 Gossip 协议中，消息以节点接力的方式进行传递，节点先将消息发送给周围的一组节点，接收节点在收到消息之后根据需要进行转发. H-ALM 采用 Gossip 协议在节点之间交换信息，以构建无结构的分支层网状组播组，同时综合了各方面的因素，为用户节点 i 提供一个合理的连接优先度 π_i，对组播网的搭建进行优化. π_i 考虑了节点的上行带宽能力、接收数据量、转发数据量和已有的邻居节点数等多种因素，即

$$\pi_i = \frac{B_i C_i}{K_i R_i} \Big/ \Big(\sum_{j \in s(i)} \frac{B_j C_j}{R_j K_j} + \frac{B_i C_i}{K_i R_i} \Big) \qquad (3)$$

式中：B_i 表示已经加入组播的节点 i 的上行带宽；C_i 表示节点 i 累积的转发数据量；K_j 表示接收节点 i 数据的节点数；R_j 表示节点 i 累积的接收数据量；$s(i)$表示向节点 i 发送数据的节点集合.

3.1　节点加入组播组

新用户节点 x 在加入组播组时，首先要与服务中心取得联系，并优先选择存在剩余度的主干层节点建立连接. 为了避免节点退出时造成与其连接的所有分支层节点脱离整个系统而成为孤岛，节点 x 还需要与主干层节点的父节点的分支层建立连接. 如果不存在剩余度的主干层节点，需加入具有 RTT 最小的主干层节点的分支层. 主干层节点将节点 x 的请求信息发送给与其相连的分支层节点，将生存时间(Time to Live，TTL)设为 n，每经过一个节点，TTL 减 1，直到 TTL 为 0. 此时，如果一个节点 x 不在节点 i 的邻居节点列表中，节点 i 能够接收新的邻居节点，节点 i 的优先度 $\pi_i \geqslant 1/(k+1)$，则节点 x 可加入到邻居节点列表中；否则，TTL 加 1，重复上述过程.

3.2　节点退出组播组

节点退出组播组分为正常退出和非正常退出. 节点 i 正常退出时会通知所有的邻居节点、服务中心，完成组播组的局部重构. 考虑到网状组播组中每个节点都至少与 2 个节点保持连接，所以当节点 i 退出时，若节点 i 的邻居节点按照一定的概率决定是否连接新节点. 节点的非正常退出依据的是邻居节点之间的心跳连接，若节点 i 在最长有效时间内没有收到邻居节点 x 的心跳连接，便可认为节点 x 失效，此时该节点 i 通知服务中心进行局部组播组重构.

4　实验及分析

仿真实验采用了 NS-2 平台，并通过 GT-ITM 提供的 Transit-Stub(TS)模型随机产生包含3 200 个节点的分层网络拓扑作为实验的底层网络. 骨干网内的链路带宽为 1 000 Mb/s，局部区域内链路带宽为 100 Mb/s，骨干网内每条链路的代价值为 200～400 之间的一个随机整数，局部区域内每条链路的代价值为 20～40 之间的一个随机整数. 用 4 种节点的接入带宽模拟网络的异构性，其中上行带宽分别为 5 Mb/s、1 Mb/s、384 kb/s、128 kb/s，下行带宽分别为 10 Mb/s、3 Mb/s、1.5 Mb/s、768 kb/s，端到端之间的带宽最大值为 2 Mb/s，最小值为 50 kb/s，模拟所用的通信量是一个流速为 300 kb/s 的数据流，每个数据包的大小为 1 250 B.

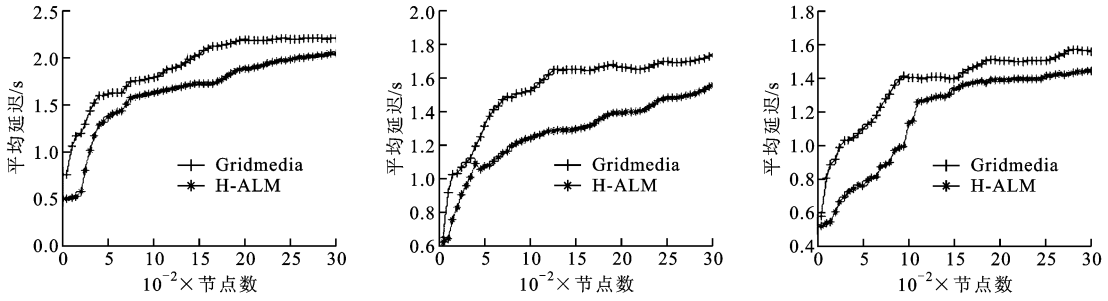
4.1　传输延迟实验

以传输延迟作为评价指标，并与 Gridmedia 方法进行对比. Gridmedia 是由清华大学研发的网络直播软件[6]，其采用 Gossip 协议构建了无结构的覆盖网络，每个节点可以独立地选择自己的伙伴节点，同时利用了推拉结合的传输策略.

传输延迟实验结果如图 2 所示，从图中可以看出，H-ALM 的平均传输延迟优于纯网状组播的 Gridmedia 方法. 这是因为，H-ALM 的主干层采用了树状拓扑，数据分发具有简易性和快速性，从整体上降低了传输延迟. 从该实验结果中还可以看出：随着组播规模的增大，数据包所经中转节点的跳数也在增加，因此平均传输延迟逐渐增大；随着邻居节点数的增加，平均传输延迟逐渐减小.

4.2　稳定性分析及实验

4.2.1　稳定性分析　稳定性是指覆盖网络对节点动态性的抗干扰能力. 在 P2P 覆盖网络中，节点具有动态特性，节点的动态退出会造成数据分发路径局部重建，这使得部分节点丢失数据. 丢包率可以反映覆盖网络的稳定性. 与树状覆盖网络相比，网状覆盖网络能够更好地应对节点的动态变化，保证



(a)邻居节点数为4 (b)邻居节点数为8 (c)邻居节点数为12

图2 平均传输延迟对比

传输的稳定性,所以H-ALM分支层采用网状结构.

主干层采用树形结构可以减小延迟,但在一定程度上却降低了网络的稳定性.主干层节点是专门的数据转发代理节点,这类节点处理能力较强,非常稳定,虽然存在用户节点充当主干层节点的情况,但只占整个主干层节点集合的一小部分,所以它们破坏覆盖网络稳定性的概率比较小.综合来看,H-ALM的稳定性较树状覆盖网络有所提高,接近于网状覆盖网络.

4.2.2 H-ALM稳定性实验 通过对比节点动态特性对H-ALM和广度优先生成树(BFT)方法的丢包率的影响,可验证H-ALM方法的稳定性,其中BFT是一种简单、有效的树状覆盖网络生成算法.随机选择一批节点并通过动态地退出系统来模拟节点的动态特性,设10 s选样10个节点退出,同时测量退出节点对系统的影响.初始的组播节点数为100,邻居节点和孩子节点数均为3,节点的带宽相同,实验结果如图3所示.从图中可以看出,BFT方法受节点动态变化的影响较大,且与随机选择的节点所处位置相关.初始阶段,节点数较多,此时随机选择孩子节点或低层节点的概率较大,而后期由于节点数减少,随机选择上层节点的概率较大,所以受影响的节点数增多,丢包率也随之增大.H-ALM方法在分支层使用了网状结构,稳定性较树状结构有很大提高.实验初期,H-ALM方法的丢包率较小,而后期由于总的节点数减少,虽然退出节点的影响逐渐增大,丢包率有所上升,但较之BFT方法仍然具有较好的稳定性.

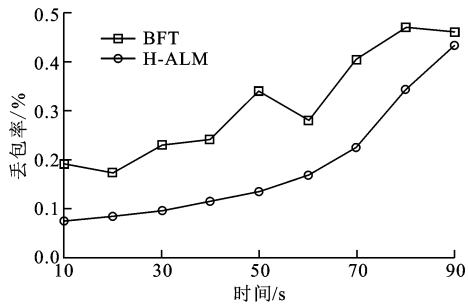


图3 节点退出时的丢包率

5 结 论

利用H-ALM可提高大规模应用层组播的稳定性和实时性.H-ALM采用了两层混合拓扑结构,优点在于:主干层采用树状结构,可以降低流媒体数据的整体传输延迟;分支层采用网状结构,可以增强系统对节点动态特性的适应能力,提高数据传输的稳定性.同时,H-ALM利用了用户节点规模、节点的带宽能力、转发和接收数据量等实时状态和网络条件,对混合式覆盖网络进行了优化,从而提高了数据传输的稳定性,减小了传输延迟.

参考文献:

[1] HOSSEINI M, GEORGANAS N D. End system multicast routing for multi-party videoconferencing applications [J]. Computer Communications, 2006, 29(11): 2046-2065.

[2] CAO Jia, LU Shiwen. A minimum delay spanning tree algorithm for the application-layer multicast [J]. Journal of Software, 2005, 16(10):1767-1773.

[3] TRAN D A, HUA K A, ZIGZAG T D. An efficient peer-to-peer scheme for media streaming[C]// Proceedings of IEEE INFOCOM. Piscataway, NJ, USA: IEEE, 2003: 1283-1292.

[4] REJAIE R, STAFFORD S. A framework for architecting peer-to-peer receiver-driven overlays[C]// Network and Operating Systems Support for Digital Audio and Video. New York, USA: ACM, 2004: 42-47.

ZHANG Xiaoxia, TANG Lixin. An ACO & SS algo-

rithm for traveling salesman problem [J]. Control and Decision, 2008, 23(7):762-766.

[5] ZHANG Xinyan, LIU Jiangchuan, LI Bo, et al. CoolStreaming/DONet: a data-driven overlay network for live media streaming [C]// Proceedings of IEEE INFOCOM. Piscataway, NJ, USA: IEEE, 2005: 2102-2111.

[6] LUO Jianguang, ZHANG Meng, ZHAO Li, et al. A large-scale live video streaming system based on P2P networks [J]. Journal of Software, 2007, 18(2): 391-399.

[7] MAGHAREI N, REJAIE R, GUO Y. Mesh or multiple-tree: comparative study of P2P live streaming services [C]// Proceedings of IEEE INFOCOM. Piscataway, NJ, USA: IEEE, 2007: 1424-1432.

[8] HEI Xiaojun, CHAO Liang, LIANG Jian, et al. A measurement study of a large-scale P2P IPTV system [J]. IEEE Trans on Multimedia, 2006, 9(8):1672-1687.

[9] LIU Yong, GUO Yang, LIANG Chao. A survey on peer-to-peer video streaming systems [J]. Peer-to-Peer Networking and Applications, 2008, 1(1): 18-28.

[10] FRONCZAK A, FRONCZAK P, JANUSZ A. Holyst: average path length in random networks [J]. Physical Review E: Statistical, Nonlinear, and Soft Matter Physics, 2004, 70: 1-7.

（编辑　苗凌）