普适环境下带宽自适应的组播策略

左 克, 王怀民, 史殿习, 胡东敏
(国防科学技术大学计算机学院, 湖南长沙 410073)

摘 要: 提出一种以树状拓扑为主干, Mesh 状拓扑为辅助的组播策略. 首先定义严格单源树状拓扑, 并根据普适环境下结点的自治性行为提出放松单源树状拓扑. 其次分析树状拓扑动态性带来的带宽瓶颈问题. 通过 Mesh 状辅助拓扑和结点带宽使用情况动态调整结点状态, 实现了自适应的带宽使用策略, 保证了较高的使用率和可用性. 分析模拟结果表明, 与现有的方法相比, 该策略能够取得较好的组播性能.

关键词: 应用层组播; 带宽自适应; 普适计算
中图分类号: TP393.03 文献标识码: A 文章编号: 0372-2112 (2009) 4A-007-05

An Bandwidth Adaptive Multicast in Pervasive Computing Environment

ZUO Ke, WANG Huai-min, SHI Dian-xi, HU Dong-min
(*Department of Computer Science, University of Defense Technology, Changsha, Hunan 410073, China*)

Abstract: This paper presents an application-level multicast using a tree-bone overlay and an auxiliary mesh. We firstly define the strictly simple resource tree-based overlay, and then propose the relaxed simple resource tree-based overlay according to the observation of self-contained performance of peers in pervasive environment. Next we analyze the churn of tree-based overlay and the resulting issue of bandwidth bottleneck. With an auxiliary mesh and state transitions based on the dynamic size of bandwidth, we design an adaptive bandwidth strategy which enables the multicast better bandwidth utilization and usability. Experimental results show that the application-level multicast achieves a good performance.

Key words: application-level multicast; adaptive bandwidth; pervasive computing

1 引言

普适计算资源的自治性、多样性、动态性等自然特性, 给普适环境中的组播问题带来了巨大挑战[1~3]. 基于 P2P(Peer-to-Peer) 拓扑的应用层组播技术由于其简单性和易用性, 目前在 Internet 上已得到大量的应用[4~7]. 根据拓扑结构的不同, 应用层组播可以分为树状拓扑和 Mesh 状拓扑两大类. Nazanin 等人在文献[2]的研究表明: (1) P2P 拓扑固有的动态性, 会导致树状拓扑中结点的可用带宽变化频繁, 当结点入度带宽之和小于组播比特率时, 随着被丢弃数据包数量的增加, 结点的数据接收质量不断降低, 直至无法满足基本的组播要求. 同时动态性还可能引起结点加入死锁问题, 限制了树状拓扑的生长, 降低了组播系统的可扩展性; (2) 尽管 Mesh 状拓扑在动态环境下为优化使用组播结点的出度带宽提供了充分条件, 但当系统中两个结点之间不存在可交换的内容, 那么他们之间的链接就没有被使用, 其带宽使用率近似为零, 因而导致内容瓶颈问题, 降低了组播系统整体的可用性.

近年来为了进一步提高组播系统的带宽使用率和可用性, 研究者们提出了网络编码技术[8], 其主要思想是根据组播系统的整体带宽大小, 将数据源编码成相互独立的段. 每段数据按照一定的拓扑结构从源结点逐步分发到参与组播的全体结点. 结点可以根据接收到的不同数据段反向解码出数据源发送的原始内容, 解码质量的好坏取决于结点接收的数据段数多少.

面对普适计算环境, 上述研究的方法缺乏对自治性、多样性、动态性等典型自然特性的考虑, 因而在实际应用时组播系统的带宽使用率和可用性较低, 对普适环境下组播系统的设计指导作用较为有限. 本文提出一种以树状拓扑为主干, Mesh 状拓扑辅助且带宽自适应的组播策略. 首先提出严格单源树状拓扑, 并根据普适环境下结点的行为自治性提出了放松单源树状拓扑. 其次分析了树状拓扑动态性带来的带宽瓶颈问题. 通过 Mesh 状辅助拓扑和结点带宽使用情况动态调整结点状态, 实现了自适应的带宽使用策略, 使得结点带宽具有较高的使用率和可用性. 分析模拟结果表明, 与现有的方法相比, 当采用相同的组播编码时, 本文提出的策略能够取得较好的组播性能.

收稿日期: 2008-11-22; 修回日期: 2008-12-15
基金项目: 国家高技术研究发展计划(863 计划) (No. 2006AA01Z198, No. 2006AA10Z237)

2 相关工作

如何借助不同的 P2P 拓扑结构进行高效而低开销的数据分发, 一直是求解组播问题的研究重点. 目前, 国内外在这方面有很多的研究成果. 从拓扑结构来看, 数据分发模型可以分为树状拓扑和 Mesh 状拓扑.

树状拓扑来源于对终端系统组播技术的扩展, 其主要思想是将参与组播的终端(以下称为结点) 依从属关系和树构造算法生成多树模型, 不同结点依据自身带宽情况选择加入一棵或多棵树中. 同时采用多重描述编码(Multiple Description Coded) 技术对源数据流进行编码, 生成多条 MDC 子流, 并设计相应的调度机制. 之后每条 MDC 子流选择某棵子树, 采用自顶向下 Push 方式分配子流. 由于 MDC 不同的子流之间满足相互独立关系, 且每条子流都能保证一定的数据接收质量, 因此结点接收到的子流个数越多, 则接收的数据质量越好. 树状拓扑性能的优劣取决于树构造算法, 算法的目标是生成满足负载平衡、结构稳定且平均高度较小的多树模型. 树状拓扑进行动态维护时, 新加入的结点查询模型中 Bootstrap 结点以获取多树模型中多个准父结点信息, 并依据数据接收质量要求与一个或多个准父结点建立父子关系, 为满足多树模型负载平衡的要求, 当不同子树中多个准父结点同时满足条件时, 新结点趋向选择具有结点总数最少的准父结点建立关系. 同时为尽量减小组播树的高度, 当某棵子树中多个准父结点同时满足条件时, 新结点趋向选择高度位置较低的准父结点建立关系; 结点退出时, 若存在以该结点为根的子树, 则子树中的所有结点同时退出, 等待一个周期时间长度后, 若该结点重新加入组播系统, 则所有结点依旧保持原子树结构一同加入, 若该结点无法加入, 则所有结点撤销原子树结构, 自由加入组播系统. 文献[2] 指出树状拓扑能有效利用各结点共享带宽以满足接收结点对接收质量的要求, 具有较好的数据分发性能; 但树状拓扑结构动态维护开销较大, 当系统振荡性(Churn) 加强时组播性能明显降低, 甚至出现结点加入死锁问题. 同时树状拓扑不适用于结点的链接稳定性、带宽大小等资源存在较大异构的应用环境.

Mesh 状拓扑结构松散灵活, 参与结点按照距离远近关系随机组成 Mesh 网络. 结构中每个结点维护一张关系路由表, 分别存储数据输入邻居结点和输出邻居结点的信息. 采用 swarming 策略分发数据, 该策略综合使用了基于 Push 的内容报告机制和基于 Pull 的内容请求. 每个结点周期性向其输出结点报告自身获取的新数据信息, 并向其输入结点请求期望得到的数据信息. 输入结点将收到的请求信息汇总, 并依据特定的数据包调度算法和请求的时间顺序分发数据. Mesh 状拓扑性能优劣取决于数据包调度算法以及 Mesh 网络的结点连接度. Mesh 状拓扑动态维护时, 新加入结点查询拓扑中 Bootstrap 结点以获取 Mesh 网络中一组可作为其输入结点的信息; 若结点退出 Mesh 网络, 向相关邻居结点发布离开信息并更新各自路由表. 文献[2] 指出由于 Mesh 网状模型结点间没有固定的组织关系, 结构灵活, 所以动态维护开销较小, 能有效地减小振荡性对于数据分发的影响, 可灵活应用于多种网络环境.

3 带宽使用分析

3.1 严格单源树状拓扑

定义 **1**(相交性)

假设 T 表示树状拓扑, V 表示 T 的所有结点集, v 表示 T 中的任意结点, 变量 i, j, k 属于自然数集合. 若 $\forall v \in \sum_{i=1}^{k} V_i$ 是某棵树 T_j 的内部结点(非叶结点), 且同时 v 也是其他 $k-1$ 棵树 $T_1, \cdots T_{j-1}, T_{j+1}, \cdots, T_k$ 中的叶结点, 则称树集合 $T_1, T_2, \cdots, T_k$ 满足相交性.

下面采用递归方式给出严格单源树状拓扑(Strictly Simple Resource Tree-based Overlay) 定义, 如图 1 所示.

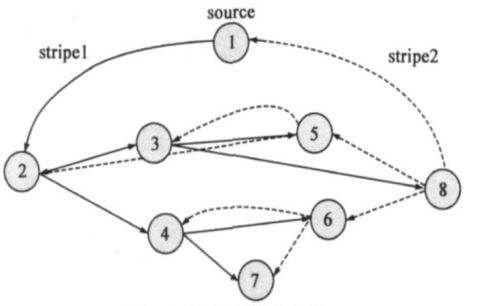

图1 严格单源树状拓扑

定义 **2** 严格单源树状拓扑是一个或多个结点的有限集合 T_o, 其中:

(1) $\exists$ s 源结点, 称作 T_o 的根;

(2) 除根 s 外, 其他结点被分为 k 个满足相交性.

定义的集合 $\{T_1, T_2, \cdots, T_k | k \geqslant 0\}$, 且每个集合都是一棵严格单源树状拓扑. 树 $T_1, T_2, \cdots, T_k$ 称作树 T_o 的子树, 且它们的根都是 s 的儿子.

定理 **1** 除源结点外, 采用网络编码的严格单源树状拓扑中的所有结点都具有最优组播质量.

证明 若严格单源树状拓扑 T_0 中存在结点 v, 且 v 的编码质量没有达到最优, 则其收到的数据段数 $m < k$, 即 v 被分配到 m 个满足相交性定义的集合中, 而另外 $k-m$ 个集合不满足相交性定义, 从而推出 T_0 不是严格单源树状拓扑.

3.2 放松单源树状拓扑

定义 **3** 放松单源树状拓扑(Relaxed Simple Source Tree-based Multicast Overlay) 不满足相交性定理的单源树

状拓扑为放松单源树状拓扑.

在普适计算环境下产生放松单源树状拓扑的原因有很多, 除了结点间的连接可靠性受网络条件限制外, 还与普适环境资源的动态性有关. 图 2 给出了连接失效时, 图 1 可能产生的一个放松单源树状拓扑. 设放松单源树状拓扑规模为 N, 网络编码将组播内容编码为 k 段. 结点 $i \in N$ 的期望接收编码大小为 $I_i (0< I_i \leqslant k)$, 已有的组播编码段数为 T_i, 最大子结点数为 C_i, 带宽最大容量 $2k$, 且平均分给入度和出度. 我们称 I_i 为结点 i 的期望入度, C_i 为期望出度. 文献[3]指出, 假如拓扑中所有结点总期望入度小于或等于总期望出度, 且存在结点 i, 其期望出度大于期望入度, 则 i 满足:

$$\forall i: C_i > I_i \Rightarrow I_i + T_i = k \qquad (1)$$

说明 i 可以接收所有 k 段组播编码, 组播质量能达到最优. 反之, 若 i 接收的组播编码段总数 T_i 小于 k 且存在可用带宽, 为了达到最优组播性能, i 通常会设法连接其他 $k-(T_i+I_i)$ 棵子树. 同时为节省入度带宽, i 通常会放弃与已有的 (T_i+I_i) 棵子树的连接, 导致入度带宽闲置. 这是普适环境下结点自治性导致的行为——即在带宽有限的环境中, 结点会考虑优化入度带宽使用, 以确保自身的组播质量尽可能达到最好, 而不会考虑优化出度带宽使用, 来满足其他结点组播质量的要求.

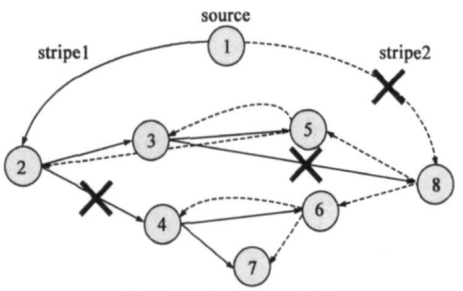

图2　放松单源树状拓扑

定义 **4**(带宽瓶颈)　在放松单源树状拓扑中, 若由于自治性导致结点主动放弃已有连接而出现闲置带宽, 称此时的结点具有带宽瓶颈.

4　带宽自适应组播策略

为了适应结点带宽动态变化的同时, 有效的降低带宽瓶颈问题的发生, 保持系统有较好的带宽使用率和组播质量, 我们设计了一个混合拓扑组播策略 Together, 以放松单源树状拓扑为主干, 在定义 Spare Capacity Set(SCS) 集合的基础上设计了 Mesh 状辅助拓扑和带宽自适应的结点状态迁移策略.

4.1　Spare Capacity Set

定义 **5**　Spare Capacity Set(SCS) 对于规模为 N 放松单源树状拓扑, SCS 由满足下面任一条件的结点组成:

(1) $\forall i \in N, I_i < k$

(2) $\forall i \in N, C_i < k$

(3)(1)、(2) 同时满足, 即 $\forall i \in N, I_i + C_i < 2k$

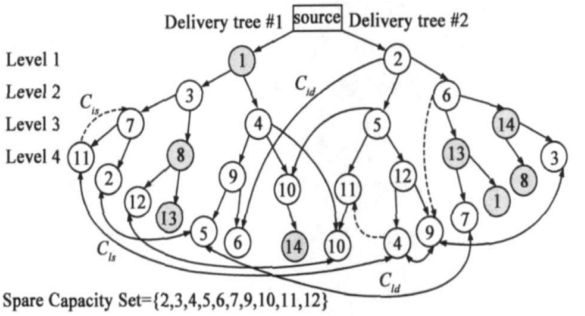

图3　放松单源树状拓扑主干

图 3 描述了一棵放松单源树状拓扑主干的情况. 图 4 描述了一个根据图 3 的 SCS 集合组成的辅助 Mesh 状拓扑.

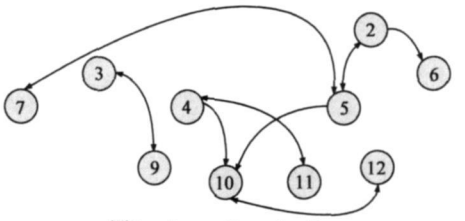

图4　Spare Capacity Set

4.2　Mesh 状辅助拓扑

当放松单源树状拓扑中结点满足定义 5, 属于 SCS 集合时, 我们选择 SCS 集合中的结点构造 Mesh 状辅助拓扑以增加这个系统的组播吞吐率和带宽使用率. 组播系统中的结点除了维护路由表和邻居表之外, 还维护一个 Mesh 状辅助拓扑的 Membership Cache(MC). MC 中保存着该结点在 Mesh 状辅助拓扑中邻居结点的相关信息. 结点间通过类似 Gossip 方式获取各自需要的编码段, 在不重新加入放松单源树状拓扑的情况下, 有效地解决带宽瓶颈问题结点的闲置带宽, 增加了整体带宽的使用率.

4.3　带宽自适应策略

在我们设计的组播策略中, 结点的状态分别在 Unknown、Pending、Tree-bone 和 Hybrid 进行迁移. 首先, 新结点 i 出现时, 由于其位于整个组播系统之外, 默认初始状态为 Unknown; 若 i 和系统中某个结点取得连接且带宽满足 $I_i = T_i = 0$, 则进入 Pending 状态, 等待合适连接(如负载平衡条件下深度较小或结点数较少的子树)以叶结点的形式加入放松单源树状拓扑主干; 当满足条件的连接存在且带宽满足 $I_i + T_i < k$, 则随机选择连接并启动 Bootstrapping 过程, 加入树状拓扑主干; 由于系统中连接的动态变化导致结点带宽的使用率稳定性较低, 如果树状拓扑中结点的带宽满足 SCS 集合的条件, 即 $(I_i < k) \cup (C_i < k)$, 则加入以 SCS 集合中结点构成的 Mesh 状辅助网络, 以保证系统动态变化的情况下, 结点根据带宽使用率调整连接类型(树状拓扑中的连接和 Mesh 状辅助网络中的连接), 保证带宽较高的使用率,

即 $I_i + T_i \cong k$. 最后, 如果结点与组播系统的连接失效, 则产生 Leaving 过程, 结点离开系统, 状态重置为 Unknown. 图 5 分别描述了结点状态迁移过程(实线箭头)与带宽自适应策略(虚线箭头).

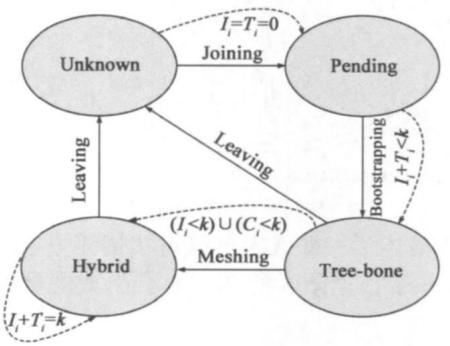

图5 结点状态迁移和带宽自适应策略

5 模拟测试

我们在 OverSim 中实现了混合组播策略 Together, 并对 Together 的带宽使用、平均组播质量和动态维护开销等特性进行评估, 将其与 SplitStream、CoolStream 和简单混合模型 mTreebone 等进行比较. 表 1 给出了模拟测试参数设置. 图 6 给出了三组测试的结果.

表1 模拟测试参数设置

结点总数	200	MDC/ stream	20
结点度数	8	时间间隔/ 数据包	4s
连接平均延时	[5ms, 25ms]	bwd	80kpbs

5.1 带宽使用

为了测试带宽使用情况, 假设任意结点 i 的带宽、结点连接度和多重描述编码率三者满足下面的关系:

$$BW_i = \tau * deg_i * bw_d (\tau > 0) \quad (2)$$

模拟参数的取值大小反应结点 i 的带宽与结点连调度、多重描述编码率之间的关系: (1) 当 $0 < \tau < 1$ 时, 结点的带宽较小, 连接度中存在不满足 bw_d 带宽要求的连接, 称这时结点带宽超负荷; (2) 当 $\tau = 1$ 时, 每个连接度都满足 bw_d 的带宽要求, 称这时结点带宽满负荷; (3) 当 $\tau > 1$ 时, 连接度中存在除满足 bw_d 要求外, 还有带宽空闲的连接, 称这时结点的带宽未满负荷; 通过调节 τ 的取值范围, 可以有效的反映出每个系统不同的带宽使用、组播质量以及动态维护的性能.

模拟实验结果如图 6(a)所示. 从图中可看出, 树状模型的带宽使用率受 τ 取值变化的影响较大, 存在一个明显的性能拐点(τ 取值接近 1 时, 实验结果为 1.2, 带宽使用率接近 90%), 而取其他值时带宽使用率较低. 分析原因发现, 当 τ 取值较小时, 结点带宽有限, 为尽量满足自身接收的数据质量, 结点往往将可用的带宽全投入到数据输入连接中, 而忽略此举对数据输出

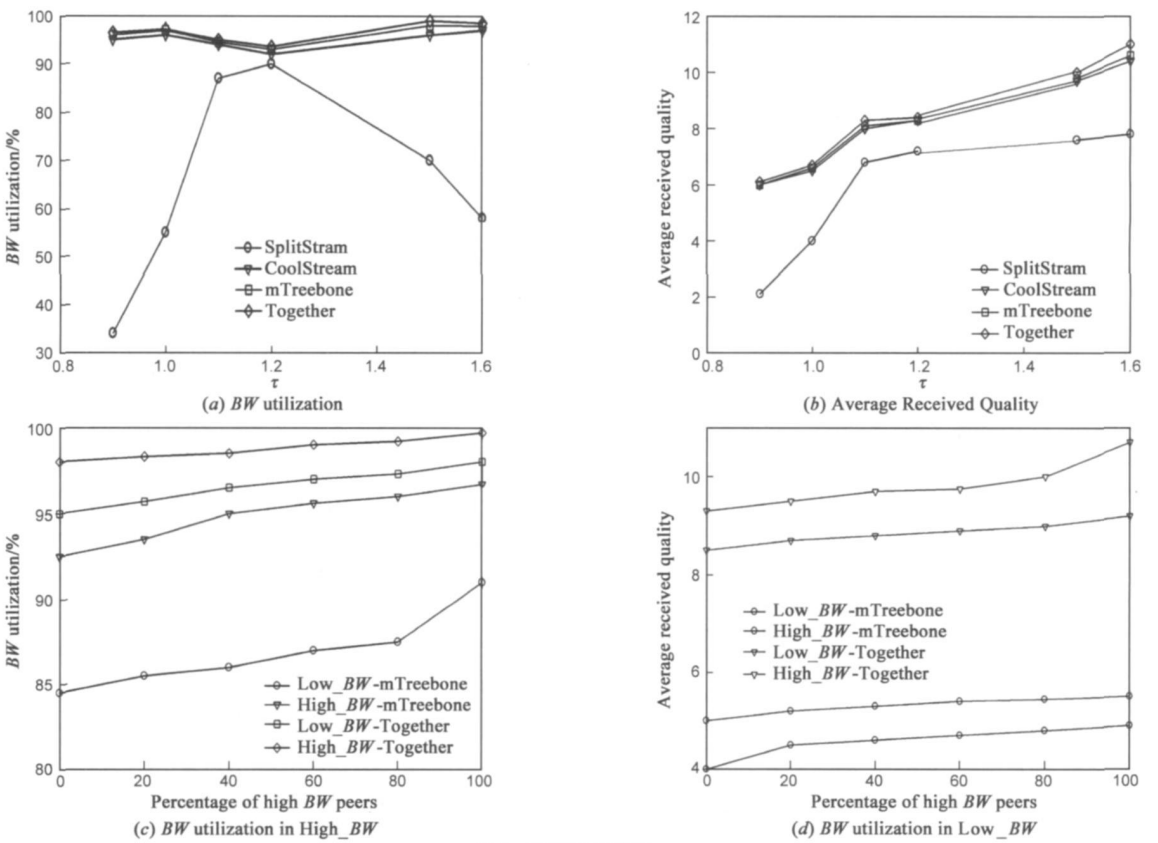

图6 模拟测试结果

连接的影响. 树状模型中多个结点此举的连锁反应导致系统整体带宽使用率较低; 当 τ 值较大时, 结点的输入输出边中存在带宽闲置的连接, 同时, 树状模型中越来越多的结点没有新数据可供分发, 从而导致模型整体带宽使用率的逐渐降低. 相比于树状模型, Mesh 状、mTreebone 和 Together 的带宽使用率基本上不受 τ 取值变化的影响, 保持在 90% 以上, 具有较高的使用率. 这一现象的主要原因是, 在 Mesh 状模型、mTreebone 和 Together 模型中, 结点间的连接不局限于任何既定的层次结构, 而是根据当前组播情况的需求, 动态的调整不同结点间的连接, 尽最大可能保证结点的可用带宽都投入到组播中, 因而模型的总体带宽使用率优于树状模型.

5.2　平均数据分发质量

首先给出平均数据分发质量(*ARQ*)的定义: 结点在一段时间内收到的平均多重描述编码个数.

试验结果如图 6(*b*)所示. 从图中可以看出, 网状模型、mTreebone 以及 Together 模型的平局数据分发质量与 τ 值基本成正比关系, 随着 τ 值的增加 *ARQ* 逐渐增常, 甚至能够超过结点带宽满负荷时 $ARQ = deg_i * bw_d = 8bw_d(\tau = 1)$. 相比之下, 当 $\tau \leqslant 1$ 时树状模型的 *ARQ* 较小. 随着 τ 值的增长, 树状模型的 *ARQ* 也逐渐增长, 但无限趋近于结点带宽满负荷时的值.

5.3　带宽自适应

为进一步测试带宽动态变化情况下 Together 和简单混合模型 mTreebone 的带宽使用率以及组播质量, 将结点分为带宽较大和带宽较小两组, 分别为 980Kbps、490Kbps. 同时不论结点带宽大小, 为保证每条连接的可用带宽大小近似, 分别指定两组中每个结点的连接度为 10、5, 使得两组结点的带宽/度数比值相同. 模拟测试结果如图 6(*c*)、图 6(*d*) 所示. 从图中可以看出, 两组结点的带宽使用率都随着系统中较大带宽结点数目的增加而上升, 且 Together 模型相比 mTreebone 模型带宽使用率要好. 分析原因, 相比以结点加入系统的时间长度为依据构造混合模型的 mTreebone, Together 模型以结点带宽使用的优化为目标, 结点间是否有连接除了可以在树状结构中存在定义外, 还可以根据具体的带宽情况灵活的加入 Mesh 状结构中, 使得系统中达到带宽满负荷状态的结点总数比 mTreebone 模型多, 因而系统整体的带宽使用率优于 mTreebone 模型. 对于平均组播质量, 有相似的分析和结论.

6　结束语

针对传统树状拓扑和 Mesh 状拓扑的不足, 本文提出了一种带宽自适应的混合组播策略. 首先提出严格单源树状拓扑, 并根据普适环境下结点的自治性行为提出了放松单源树状拓扑. 其次分析了普式环境下资源动态性带来的带宽瓶颈问题. 通过 Mesh 状辅助拓扑和自适应的带宽使用策略, 保证了较高的带宽可用性、使用率以及组播质量.

参考文献:

[1] J Peltotalo, J Harju, et al. Peer-to-Peer streaming technology survey[A]. In Proc of the 7th International Conference on Networking[C]. Cancun, Mexico: IEEE Press, 2008. 342- 350.

[2] N Magharei, R Rejaie, et al. Mesh or multiple-tree: a comparative study of live P2P streaming approaches[A]. In Proc of the IEEE INFOCOM[C]. Anchorage, Alaska, USA: IEEE Press, 2007. 1424- 1432.

[3] J Liu, S G Rao, et al. Opportunities and challenges of peer-to-peer internet video broadcast[J]. In Proc of The IEEE, Special Issue on Recent Advances in Distributed Multimedia Communications, 2008, 96(1): 11- 24.

[4] M Castro, P Druschel, AM Kermarrec, et al. Scribe: a large-scale and decentralized application-level multicast infrastructure [J]. The IEEE Journal on Selected Areas in Communications, 2002, 20(8): 98- 104.

[5] M Castro, P Druschel, et al. SplitStream: High-bandwidth multicast in cooperative environments[A]. In Proc of the ACM Symposium on Operating Systems Principles[C]. Bolton Landing, NY, USA: ACM Press, 2003. 298- 313.

[6] Shaoliang Peng, Shanshan Li, et al. SenCast: scalable multicast in wireless sensor networks[A]. In Proc of the IEEE International Parallel and Distributed Processing Symposium[C]. Rome, Italy: IEEE Press, 2008. 1- 9.

[7] J Venkataraman, P Francis. Chunkyspread: multitree unstructured peer-to-peer multicast[A]. In Proc of the 5th International Workshop on Peer-to-Peer Systems[C]. Santa, Barbara, CA, USA: IEEE Press, 2006. 2- 11.

[8] 刘亚杰, 张鹤颖, 窦文华, 陈俊峰. P2P 分层流媒体中数据分配算法[J]. 软件学报, 2006, 17(2): 325- 332.
Liu Yajie, Zhang Heying, Dou Wenhua, Chen Junfeng. Data allocation algorithms in layered P2P streaming[J]. Journal of Software, 2006, 17(2): 325- 332. (in Chinese)

作者简介:

左　克　男, 1978 年 11 月生于湖南长沙, 国防科技大学计算机学院博士生, 主要研究领域为普适计算、对等网络、资源搜索算法. E-mail: icekezuo@gmail.com

王怀民　男, 1962 年 4 月生于江苏省南京市, 获博士学位, 教授, 博士生导师, CCF 高级会员, 主要研究领域为分布计算中间件、软件 Agent、网络与信息安全.

史殿习　男, 1966 年 4 月生于山东省龙口市, 获博士学位, 副研究员. 主要研究领域为分布式计算.

胡东敏　女, 1979 年 7 月生于河北省沧州市, 国防科学技术大学计算机学院博士生. 主要研究领域为软件工程、中间件技术、软构件技术.