高技术通讯 2009 年 第 19 卷 第 5 期:487 ~ 493

doi:10.3772/j.issn.1002-0470.2009.05.009

一种具有较好用户体验的 P2P 因特网视频广播系统①

贺 磊② 郭云飞 张伟丽 刘文波 马海龙

(国家数字交换系统工程技术研究中心 郑州 450002)

摘 要 针对当前流行的 P2P 因特网视频广播系统频道切换慢、源到端时延长等问题,进行了连接节点管理算法和数据段调度算法的研究,提出了一种采用基于多树和网的方案 MTreeTV,降低了时延并能适应自治节点的高波动。该方法充分利用了结构化 P2P 覆盖网的优点,能够扩展到非常大的规模,网络效率和健壮性较好。理论分析和仿真表明,MTreeTV 可以提供较短的频道切换和源端时延(小于 9s)并具有很高的播放连续性,而且控制开销较小(小于 2%)。此外还研究了缓冲区大小、连接节点数量、节点带宽和节目速率等关键参数对 P2P 因特网视频广播系统性能的影响。

关键词 因特网视频广播,结构化 P2P 覆盖网,临近性,多树,用户体验

0 引 言

随着宽带接入的广泛部署,多媒体业务正在用户中间变得更加流行,并且在因特网中占据了越来越大的数据量。自从 20 世纪 90 年代初以来,学术界和工业界对利用网络组播(IP Multicast)[1]来支持视频广播已进行了大量研究。但由于网络组播在可扩展性、对高层功能的支持及可部署性等方面的问题,导致基于网络组播的方案未能推广。基于内容分发网络(content delivery network,CDN)[2]的方案也由于对带宽的要求过于昂贵而只能服务于少量客户。当很多网络服务提供商开始提供网络电视(IPTV)服务,并使用包交换来传递高质量视频时,就需要一种有效、一致的因特网视频广播支持方式。本文根据当前流行的 P2P 因特网视频广播系统的实际问题,提出了一种具有较好用户体验的因特网视频广播系统 MTreeTV,理论分析和仿真试验表明,该系统在时延和播放连续性上有较大优势。

1 相关研究描述

对等网络(peer-to-peer, P2P)技术对从文件下载到网络电话(VOIP)等大量应用都非常重要,但是 P2P 视频广播在带宽和延迟方面有更严格的实时性要求。文件下载程序 BitTorrent[3]的目标是下载完整的文件,对时限的要求并不高。BitTorrent 可以花费数天的时间进行文件下载,但对视频广播来讲这样的时延则不可能。VOIP 同样也有实时性要求,但视频广播的带宽需求更大,同时参与者数量非常多,而且参与者的动态性也更高。

近年来,人们对使用 P2P 技术进行视频广播提出大量的建议[4-8]。采用 P2P 技术进行视频广播的原因主要有两个,一是这种技术不需要路由器和网络基础设施的支持,因此成本非常低而且易于部署,二是在这种技术中,参与者不仅下载视频流,而且在观看节目的同时还上传视频流给其他参与者。因此,这种方式可以扩展到非常大的组规模,因为更大的需求同时也带来了更多的资源。而视频广播对资源的需求是非常巨大的。例如,采用 MPEG-4 编码的 1.5Mbps 标清电视,如果同时向 100 万用户播出节目,则总汇聚带宽超过了 1.5Tbps。

组播根据数据分发方式主要可分为两类[9]:“基于树的方式”(tree-based approach)和“基于数据驱动的随机方式”(data-driven randomized approach)。“基于树的方式”构造了一棵或多棵以节目源为根的组播分发树,按照转发规则采用“推”(push)的方式从父节点向每个子节点推送数据,典型例子如 ESM[5]、SplitStream[4]。但在波动环境下,“基于树的方式”的健壮性较差。另一种也被称为“基于网的随机方式”

① 973 计划(2007CB307102)资助项目。
② 男,1974 年生,博士,讲师;研究方向:覆盖网组播、网络管理系统;联系人,E-mail: hl.helei@gmail.com
(收稿日期:2008-04-09)

(mesh-Based randomized approach),它采用随机的方法对抗节点的随机失效,并采用“拉”(pull)的方法获取数据,如 CoolStreaming[7]、PPLive[8]。其主要问题是由于采用随机方式(Gossip-like)[7,12],不能保证视频广播的服务质量,导致频道切换和源端时延较长。例如,CoolStreaming 的切换时延长达 1min,PPLive 中的最短切换时延也超过 20s。

我们的设计称为 MTreeTV,设计思想来源于对一个实际 P2P 视频广播系统的流量模型的研究[3]。通过网(mesh)传送的数据传输路径大部分都符合一个或很少几个特定的树形结构,网中传送路径与树的相似性(定义为两者共同边的比例)能够高达 70%,覆盖网的性能因此和这些共同节点的构成和组织相关。如果预先构造多个高效的树形结构,并将临近父子节点组织成为一个网的方法,我们可以期望同时获得低时延、高效率和健壮性。针对上述问题,我们提出了一种新型的“数据驱动的多树方式”(data-driven multitree approach)。其关键思想是:使用自组织、支持邻近性和可扩展的基于多树的连接节点管理算法,构造支持临近性的网状节点连接关系,配合优化的调度算法,在保证健壮性的同时获得良好的用户体验。

2 MTreeTV 的设计

在 MTreeTV 中,视频流被分为大小相等的数据段,从节目源开始首先扩散到多个树的树根节点,然后到所有加入的节点。对该频道感兴趣的用户通过同时加入到多棵树的方法来加入一个频道,并通过“拉”(pull)的方法来获得满足实时性要求的数据段。除了数据源节点外,其他所有的节点既是数据段的接收者,又是数据段的转发者。而且,这些节点可以自由的加入和离开覆盖网。数据驱动的系统 MTreeTV 的研究主要有两方面的内容:(1)连接节点管理算法,能够支持临近性并提高自治节点情况下的可靠性;(2)优化的调度算法,避免数据冗余并最大化系统效率,并提供良好的用户体验。

2.1 连接节点管理算法

每个 MTreeTV 节点都拥有一个唯一的 128 位的节点标识符,并通过同时加入多棵树并将父子节点构成网的方法来管理连接节点关系。MTreeTV 构造于结构化 P2P 覆盖网 Pastry[11]之上,并利用了 Pastry 的健壮性、自组织性、临近性和可靠性。与基于树/多树的单向推送方式不同,MTreeTV 中的连接都是双向的。

MTreeTV 节点加入树的操作采用分布式方法,支持大规模及动态成员的管理。各树以自己汇聚点为根,采用反向路径转发(reverse path forwarding, RPF)机制构造。即:每个加入节点都要发出一个以汇聚点为目的地的加入消息,并由从该节点到树根节点的 Pastry 路由路径构成树。当每个节点都同时加入到多棵树时,就会拥有多个父亲及子女节点(图 1),这些父亲和子女就构成了候选连接节点集合。节点按照一定的标准从该集合中选择一定数量的节点并建立连接关系。图 1 显示了 MTreeTV 中由多棵树及由多棵树构成的支持临近性的网的过程。其中 A、B、C 分别是三棵树的树根(汇聚点),实线部分表示的是以 B 为树根的一棵树,虚线部分表示的是其它树。在多树的构造过程中,一个节点的连接节点数量可能会超过其出度限制。通过出度限制算法可以保证一个节点的连接节点数量不会超出其能力限制,从而减少开销并为数据段的扩散提供足够的资源。MTreeTV 节点采用与 Scribe[6]类似的出度限制算法,算法如下:

(1) 当一个节点探测到它超过了出度限制时,它就选择消耗资源最多的树,然后它根据邻近性参数选择该树中最远的子女节点。

(2) 父节点丢弃所选的子女节点,并发送消息包含其子女列表和各子女节点到自己的延迟。

(3) 当子女节点接收到消息,它就进行以下操作:

(a) 测量它自己与(从父节点接收的)孩子列表中各节点的延迟。

(b) 然后计算通过此节点到父节点的总延迟。

(c) 最后,发送加入消息给提供最小总延迟的节点。这样,它使得通过以前兄弟节点到父节点的总延迟最小。

临近性保证了 MTreeTV 能有效地传播数据:第一,由于 Pastry 短路由属性使得从汇聚点到每个成员转发数据的延迟变小了,降低了切换和源端时延。第二,Pastry 路由收敛属性保证了施加于物理网络的负载变小,这是因为大部分数据在接近叶的节点间交换,且数据经过的网络距离很短。MTreeTV 具有较短的源端时延和切换时延,量化的分析和仿真结果在第 3、4 节中给出。

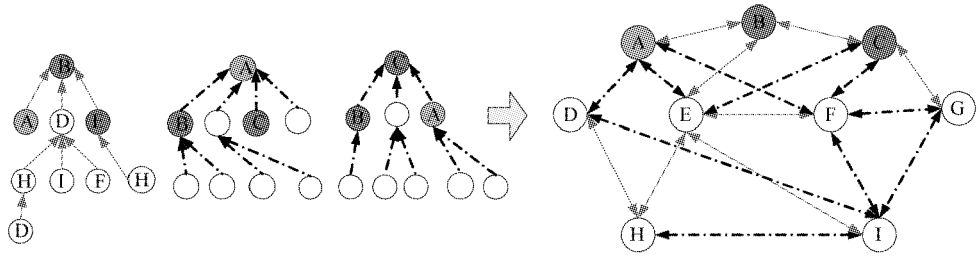

图 1　MTreeTV 基于多树的连接节点管理

2.2　节点退出

节点退出包括主动离开和突然失效(图 2):在前者情况时,退出节点会主动通知其连接节点;后者会因为失效节点在一定时间内没有消息交互而被其父子节点探测到。这两种情况都需要进行连接节点关系的修复。在覆盖网波动情况下,连接节点修复可以保证每个节点维持稳定的连接节点数量,算法具体如下:

(1) 当退出节点在某棵树中是父节点时,该退出节点的子节点可以调用 Pastry 路由一个消息到其树根节点,并发现一个新的父节点作为连接节点,从而修复了连接节点关系。

(2) 如果退出节点在其加入的所有树中都是叶子节点,该节点在各树中的父亲节点可以通过加入一棵新树的方法来发现一个新节点。

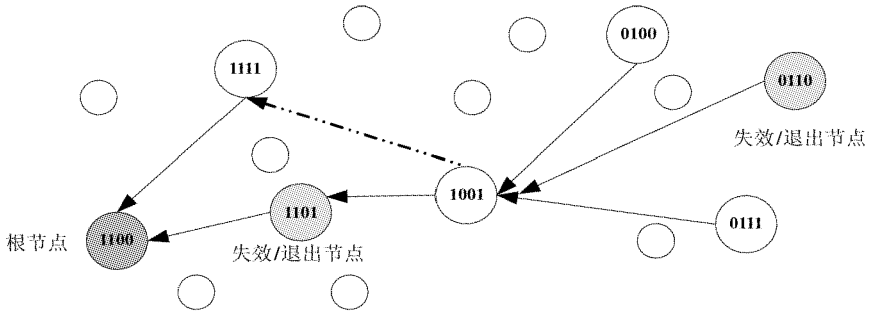

图 2　连接节点关系的修复过程

2.3　数据段调度算法

根据缓冲区位图和段需求情况,一个调度算法要决定如何从连接节点获取需要的段。调度算法必须满足两个限制:(1)每个数据段的播放时限;(2)不同连接节点的带宽限制。在保证一定播放连续性的前提下,缓冲区的填充时间应该尽可能的小。如果第一个限制条件不能被满足,则因为超出时限而被丢弃的段应该保持最少。这是并行机调度问题(parallel machine scheduling)的变形,为 NP 难问题[7]。考虑到此调度算法还必须快速适应高动态的网络环境时,则更难以找到最优的解决方法,因此我们采用了一种简单的、具有快速反应时间和较短时延的启发式调度算法。

为了满足上述的调度算法限制,MTreeTV 的调度策略如下:

(1) 由于媒体播放器即将播放“提交时限”前的数据段,所以先调度即将到达“提交时限”的段。

(2) 只有一个提供者的段比有多个提供者的段更难以获得,优先调度只有一个候选节点的段。

(3) 为了在同样时间内穿送更多的段,优先从“段平均传送时延”较小的候选节点调度其所拥有的段。

“段平均传送时延”为从某个连接节点一次请求多个可用段时,平均每段的传送时延。CoolStreaming 的调度算法只考虑候选节点数和节点带宽。而本调度算法综合考虑了时延、段数、带宽和候选节点数,调度效率更高;并且采用流水线的方法请求和传送数据段,大大降低了时延并提高了播放连续性。

3　性能分析

在本部分我们分析 MTreeTV 的覆盖网直径,覆盖网直径可以用段平均传送跳数或源端时延表示。我们的分析模型揭示了 MTreeTV 段传送跳数和加入节点规模之间的对数关系,这反映了 MTreeTV 具有良好的可扩展性。我们进一步量化分析了 MTreeTV

的每跳时延和源到端时延,结果表明在每跳时延和源端时延方面 MTreeTV 都优于基于网的随机方法。可用宽度优先搜索(breath-first search,BFS)树模型来模拟 MTreeTV 的段传播路径。在 BFS 树中,源节点是根节点位于第 0 级,在 k 跳内从源节点到达的节点位于第 k 级,考虑到网状结构中同时存在多条到达路径,所以每个节点可以在 BFS 树中出现多次。图 3 表示了一棵 BFS 树,其中灰色节点代表首次访问的节点,白色节点代表非首次访问的节点。覆盖网节点数量为 N,节点的平均出度为 2^b。

从源到每个目的节点的平均跳数为 d,可推导出[7]:

$$d < \log_{2^b-1} N + 3 \quad (1)$$

由公式(1)得知,MTreeTV 从源节点到所有目的节点的平均跳数为 $O(\log N)$,这表明 MTreeTV 可扩展性很好,能够扩展到非常大的节点规模。

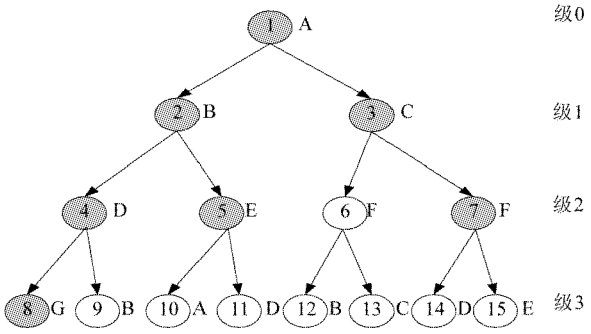

图 3 Breath-First Search(BFS)树

下面我们分析 MTreeTV 在邻近空间中每跳的期望时延(距离)。为了使得分析易于计算,我们假设所有节点都平均分布在一个圆圈(直径为 R)上,两节点间时延为它们在圆圈上的物理距离[10]。由于可从最近的连接节点中获取数据段,MTreeTV 第 h 跳时延在很大概率上小于 Pastry 第 h 跳的时延:

$$\text{popDelay}_1(h) \Leftarrow (R\cos^{-1}(1 - \frac{2^{b(h+1)+1}}{N})) \quad (2)$$

由于 CoolStreaming 的连接节点是随机选择的,则其每跳时延 popDelay_2 应为两点间的平均距离:

$$\text{popDelay}_2(h) = \frac{\pi R}{2} > \text{popDelay}_1(h) \quad (3)$$

CoolStreaming 源端时延为 $seDelay_2$, MTreeTV 源端时延为 $seDelay_1$:

$$seDelay_2 = \sum_{h=1}^{\log(N)} \text{popDelay}_2(h) > seDelay_1 \quad (4)$$

从公式(3)和(4)可知, MTreeTV 的每跳时延和源端时延都小于 CoolStreaming。由于源端时延小,MTreeTV 可使用较短缓冲时间就可以达到很高的播放连续性,这也降低了其频道切换时延。

4 仿 真

我们实现了一个数据段级、事件驱动的仿真程序,并进行了一系列的仿真试验。我们的仿真拓扑是使用 GT-ITM 拓扑生成器使用广泛采用的 transit-stub 拓扑模型生成,包括 100 路由器节点和 1000 个主机节点。主机到其直连路由器的时延为 1ms,根据统计路由器间平均时延为 216ms。我们仿真了 MTreeTV、Tree 和 CoolStreaming 的连接节点管理和调度算法,Tree 是 MTreeTV 生成的一棵单树。

我们分别在稳定和波动两种环境下运行仿真程序,节目流持续 200s,所有的主机节点在节目开始时全部加入到相应的覆盖网中,节目源节点从时刻 0 开始发送数据段。每次主机节点加入各系统的顺序都是随机生成的。为了去除偶然性的影响,所有统计都是 10 次运行结果的平均值。

下面介绍一下实验中使用到的评价指标,主要有切换时延、源端时延、播放连续性、控制开销等。

- 切换时延(setup delay):指各主机节点从收到本频道的第一个数据段开始,到开始播放节目时的平均时间间隔。
- 源端时延(source to end delay):从节目源发出数据段到各节点收到数据段的平均时间间隔。
- 播放连续性(playback continuity):各节点在提交时限内到达的数据段和应到达的数据段之间的平均比例。
- 控制开销(control overhead):平均各节点的控制消息总字节数/传送数据段的总字节数。
- 节点失效比例(failure percent):每时间间隔 T 内离开的主机节点占总节点数量的比例。
- 波动开销(churn overhead):平均各节点的总波动消息流量/总数据流量。

4.1 稳定环境仿真

在稳定环境中,所有参加节点在整个节目播出期间都保持加入状态。仿真程序缺省使用如下参数:节目编码速率 300kbps,主机节点带宽 1M bps,每节点平均有 7 个连接节点,段大小为 10kB,缓冲区缓冲 32s 的视频节目(即缓冲区共容纳 96 个数据段),各节点在完成 6s 节目缓冲后开始播放。

从图 4 和图 5 可以看出,MTreeTV 拥有最小的切换时延和源端时延。当节点数量为 1000 时,MTreeTV 覆盖网的平均每跳时延为 131ms,而 Cool-

Streaming 为 942ms,这表明 MTreeTV 具有良好的临近性。源端时延与段平均跳数及每跳时延相关。MTreeTV 由于每跳时延较小且调度算法效率较高,尽管其跳数略高于其它系统,其源端时延仍然最小。

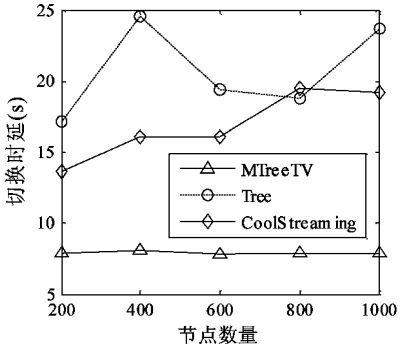

图 4　切换时延和节点数量的关系

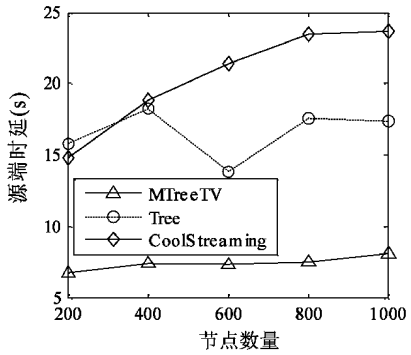

图 5　源端时延和节点数量的关系

从图 6 和图 7 可以看出,MTreeTV 连续性较高,开销较小。表 1 总结了三种系统在静态环境的性能,可以看出 MTreeTV 切换和源端时延最短,播放连续性最高,控制开销较小。

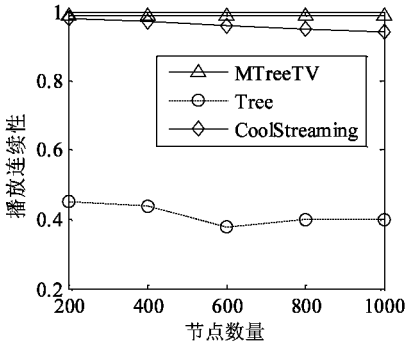

图 6　播放连续性和节点数量的关系

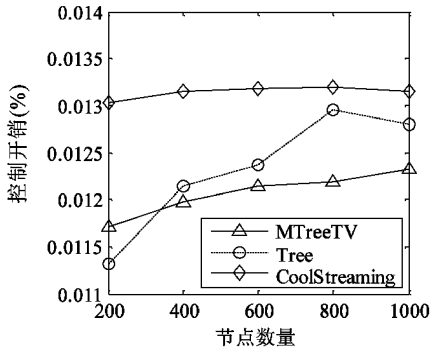

图 7　控制开销和节点数量的关系

表 1　三种系统在静态环境中的性能比较

	控制开销	切换时延	源端时延	播放连续性
MTreeTV	小(<1.3%)	短(约 8s)	短(约 8s)	高(>98%)
Tree	最小(1.3%)	长(>17s)	中(14~18s)	低(<50%)
CoolStreaming	小(<1.4%)	长(>14s)	长(14~24s)	高(>94%)

4.2　波动环境和不同参数仿真

在波动环境中,每时间间隔 T(4s)内都有一定比例的节点离开。节点波动对 MTreeTV 的切换时延影响不大(图 8),但对播放连续性影响较大(图 9),波动开销较小(图 10)。在三种系统中,MTreeTV 在波动环境下的性能最好。

连接节点数量、缓冲区缓冲时间、节目编码速率、节点带宽等参数对 MTreeTV 性能非常关键,其影响程度在表 2 中进行了总结。其中:●代表较小的影响程度,▲代表中等影响程度,★代表较大的影响程度。从表 2 中,我们可以得出如下结论:

• 由于充分利用了多树的邻近特性,主要从多个临近节点获取数据段,MTreeTV 在连接节点数量为 3 时就达到了很高的连续性(图 12)。

表 2　各种参数对性能的影响程度

	开销	切换时延	源端时延	播放连续性
覆盖网节点数量	●	●	▲	●
连接节点数量	●	★	▲	★
缓冲区缓冲时间	▲	★	★	★
节目编码速率	★	★	★	★
节点带宽	●	★	★	★
节点波动比例	▲	●	▲	★

• 适当地增加连接节点数量和节点带宽可以降低时延(图 12 和图 13)。因为缓冲时间对时延影响较大(图 14),在保证较高连续性的前提下,缓冲时间要尽可能小。

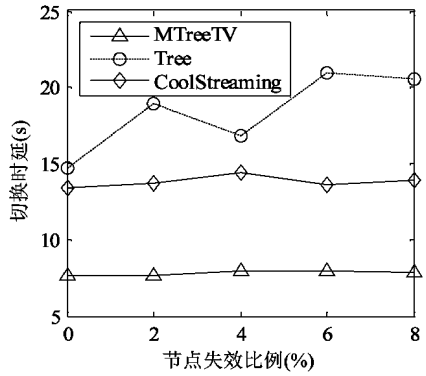

图 8 切换时延和失效比例的关系

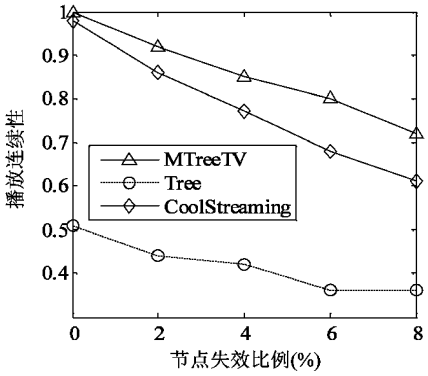

图 9 播放连续性和失效比例的关系

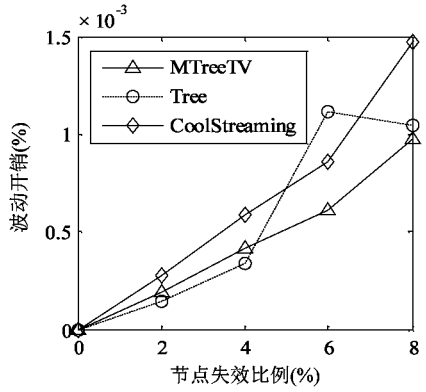

图 10 波动开销和失效比例的关系

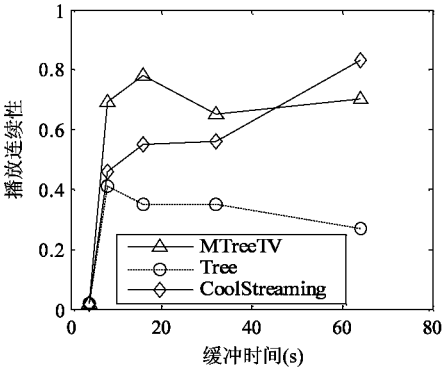

图 11 播放连续性和缓冲时间的关系

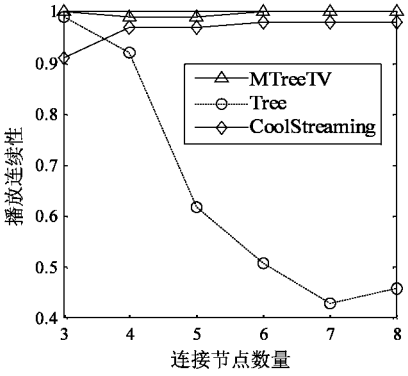

图 12 连接节点数量和连续性的关系

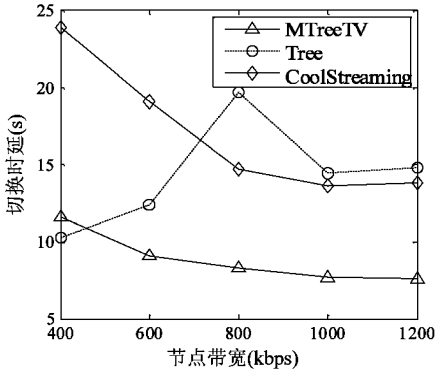

图 13 切换时延和节点带宽的关系
(10%失效比例情况下)

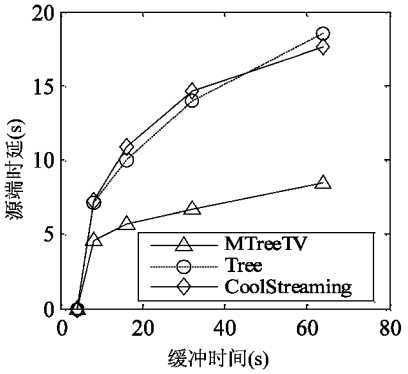

图 14 源端时延和缓冲时间的关系

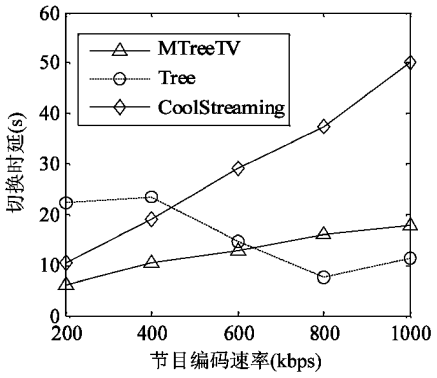

图 15 切换时延和编码速率的关系

• 高节目编码速率增加了时延(图 15),MTreeTV 因为效率较高,对高速率节目支持更好。

总之,无论是在静态或波动环境下,MTreeTV 性能都超过了 Tree 和 CoolStreaming。MTreeTV 能够以较小的控制开销(<1.3%)获得很高的播放连续性,其切换时延(约 8s)和源端时延(<9s)最小,用户体验较好。

5 结 论

经过多年的努力,因特网即将进入宽带多媒体时代,视频内容很快就会成为因特网的主要流量。在三种主要的视频分发模式(广播、视频流点播、文件下载)中,由于高扩展性、高带宽需求及实时性要求,视频广播的实现难度最大。基于 P2P 的方案由于其体系结构的自扩展性,是最有前途的因特网视频广播解决方案之一。本文针对当前流行 P2P 因特网视频广播系统频道切换慢、源到端时延长等问题,采用基于多树构造具有临近性网状节点连接关系的方法,提出了一种创新的具有较好用户体验的因特网视频广播系统 MTreeTV。理论分析和仿真表明,MTreeTV 在时延和播放连续性方面与 CoolStreaming 相比具有较大的优势。

未来可能的研究方向:改进 MTreeTV"基于网的多树"连接节点管理和数据段调度算法,进一步提高用户体验;研究各种 DSL 接入方法所带来的上下行带宽不均衡问题的影响;节点极端动态情况下或短时间内大量节点加入/退出时的 MTreeTV 系统性能;MTreeTV 系统中激励和公平问题的研究。

参考文献

[1] Deering S, Cheriton D R. Multicast routing in datagram internetworks and extended LANs. In: Proceedings of the ACM Special Interest Group on Data Communication, Stanford, CA, USA, 1988. 8. 85-110

[2] Danzig P B, Hall R S, Schwartz M F, et al. A case for caching file objects inside internetworks. In: Proceedings of the ACM Special Interest Group on Data Communication, San Francisco, CA, USA, 1993. 239-248

[3] Pouwelse J, Garbacki P, Epema D, et al. The bittorrent P2P file-sharing system: measurements and analysis. In: Proceedings of International Workshop on Peer-To-Peer Systems, Ithaca, NY, USA, 2005. 205-216

[4] Castro M, Druschel P, Kermarrec A M, et al. SplitStream: high-bandwidth multicast in cooperative environments. In: Proceedings of the ACM Symposium on Operating Systems Principles, New York, USA, 2003. 298-313

[5] Chu Y, Rao S G, Seshan S, et al. A case for end system multicast. In: Proceedings of ACM International Conference on Measurement and Modeling of Computer Systems, Santa Clara, California, USA, 2000. 1-12

[6] Rowstron A, Kermarrec A M, Castro M, et al. Scribe: the design of a large-scale event notification infrastructure. In: Proceedings of International Workshop Networked Group Communication, London, UK, 2001. 30-43

[7] Zhang X, Liu J, Li B, et al. CoolStreaming/DONet: a data-driven overlay network for peer-to-peer live Media Streaming. In: Proceedings of IEEE Conference on Computer Communications, Miami, FL, USA, 2005. 3. 2102-2111

[8] Hei X, Liang C, Liang J, et al. A measurement study of a large-scale P2P IPTV system. In: Proceedings of IPTV workshop, International World Wide Web Conference. Edinburgh, Scotland, UK, 2006

[9] Sentinelli A, Marfia G, Gerla M, et al. Will IPTV ride the peer-to-peer stream. *IEEE Communications Magazine*, 2007, 46(2): 86-92

[10] Ganesh A J, Kermarrec A M, Massoulie L. Peer-to-peer membership management for gossip-based protocols. *IEEE Transactions on Computers*, 2003, 52(2): 139-149

[11] Rowstron A, Druschel P. Pastry: scalable, distributed object location and routing for large-scale peer-to-peer systems. In: Proceedings of IFIP/ACM International Conference on Distributed Systems Platforms, Heidelberg, Germany, 2001. 329-350

[12] Castro M, Druschel P, Hu Y C, et al. Exploiting network proximity in peer-to-peer overlay networks: [technical report]. MSR-TR-2002-82 Microsoft Research. http://www.research.microsoft.om/~antr/pastry, 2002

A peer-to-peer internet video broadcast system with good user-experience

He Lei, Guo Yunfei, Zhang Weili, Liu Wenbo, Ma Hailong

(National Digital Switching System Engineering and Technological R&D Center, Zhengzhou 450002)

Abstract

The paper offers the MTreeTV, a new P2P Internet video broadcast system which uses a novel multitree /mesh design to reduce the long delay. The key idea is to construct a multitree-based mesh over a structured P2P overlay to manage the peership. The design utilizes the locality properties of Pastry to reduce the delay and accommodates node dynamics. The analysis and simulation show that the MTreeTV has less short setup delay and source to end delay, and can provide high playback continuity with little CONTROL overload(<2%). The paper also explores how the size of buffer, the number of partners, the bandwidth of node and the streaming rate can influence the MTreeTV's performance.

Key words: Internet video broadcast, structured P2P overlay, locality, multitree, user experience