ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.2, February 2007, pp.391−399
DOI: 10.1360/jos180391

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于 P2P 网络的大规模视频直播系统*

罗建光+, 张 萌, 赵 黎, 杨士强

(清华大学 计算机科学与技术系,北京 100084)

A Large-Scale Live Video Streaming System Based on P2P Networks

LUO Jian-Guang+, ZHANG Meng, ZHAO Li, YANG Shi-Qiang

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: Phn: +86-10-62772099, E-mail: luojg03@mails.tsinghua.edu.cn

Luo JG, Zhang M, Zhao L, Yang SQ. A large-scale live video streaming system based on P2P networks. *Journal of Software*, 2006,18(2):391−399. http://www.jos.org.cn/1000-9825/18/391.htm

Abstract: A P2P (peer-to-peer) network based large-scale live video streaming system called Gridmedia is presented in this paper. In this system, a gossip-based protocol is adopted to construct an unstructured application layer overlay. Each peer independently selects its own neighbors and uses a push-pull streaming method to fetch data from neighbors. Compared with the pure pull method of DONet, push-pull method greatly diminishes the accumulated latency observed at end users and efficiently reduces the control overhead of streaming system, both of which are evaluated by the experiments on PlanetLab. A prototype system of Gridmedia has been developed to broadcast the Spring Festival Gala Evening in 2005 over global Internet with 300Kbps video stream and attracted more than 500 000 users all over the world with the peak concurrent online users of 15 239 during the event.
Key words: peer-to-peer network; live video streaming; unstructured overlay network; push-pull method; low latency

摘 要: 介绍了一种基于 P2P(peer-to-peer)网络的大规模视频直播系统 Gridmedia.该系统采用 Gossip 协议构建无结构的应用层覆盖网络,每个节点可以独立地选择自己的伙伴节点.在覆盖网络上,每个节点通过一种推拉结合的流传输策略从邻居节点获取数据.与 DONet 中的纯拉策略相比,推拉结合策略大幅度减小了终端用户观看视频的延迟,并有效降低了直播系统的控制开销.PlanetLab 上的大量实验充分表明了该策略的有效性.Gridmedia 的原型系统通过 300Kbps 的视频码流对 2005 年春节联欢晚会进行了全球互联网直播.晚会期间,全球范围内有超过 500 000 人次通过系统观看了直播,最高在线人数达到了 15 239 人,充分验证了系统的性能.
关键词: P2P(peer-to-peer)网络;视频直播;无结构覆盖网络;推拉结合策略;低延迟
中图法分类号: TP393 文献标识码: A

近年来,基于 P2P(peer-to-peer)网络的应用层组播技术成为研究的热点.与基于代理服务器[1]或内容分发网络[2]的方案相比,P2P 组播不需要部署大量的专用服务器,可以节约大量的部署和管理成本.早期的 P2P 组播方

* Supported by the National Natural Science Foundation of China under Grant No.60432030 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.2006CB303103 (国家重点基础研究发展规划(973)); the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z321 (国家高技术研究发展计划(863))

Received 2005-10-13; Accepted 2006-02-23

案,如 ESM(end system multicast)[3],HMTP(host multicast tree protocol)[4],NICE[5],ZIGZAG[6]等,采用树形的应用层数据传输协议,不能很好地适应节点的动态性,接入网络的异构性和网络带宽的抖动.PRO(peer-to-peer receiver-driven overlay)[7],DONet(data-driven overlay network)[8]等通过 Gossip 协议来构建无结构的网状覆盖网络,并结合多发送者(multi-sender)的传输协议,有效改善了系统的稳定性,提高了系统的吞吐量.

Gridmedia 在无结构覆盖网络基础上采用了一种新颖的推拉结合的流传输策略,大幅度降低了终端用户观看视频的延迟,改善了直播系统的实时性,并有效降低了系统的控制开销.PlanetLab[9]上的大量实验充分验证了新策略的有效性.Gridmedia 的原型系统对 2005 年春节联欢晚会进行了全球互联网直播.晚会期间,有超过 500 000 人次通过系统观看了直播,峰值在线人数达到了 15 239 人,充分验证了系统的性能.

本文第 1 节概括叙述 Gridmedia 如何通过 Gossip 协议来构建无结构的网状覆盖网络.第 2 节首先分析 DONet 所采用的纯拉策略的不足,继而详细叙述本文提出的推拉结合策略.第 3 节和第 4 节分别给出 Gridmedia 在 PlanetLab 上的实验结果以及 2005 年春节联欢晚会期间系统运行的初步统计.第 5 节讨论研究背景和相关工作.第 6 节总结全文.

1 基于 Gossip 的覆盖网络构建方法

Gossip 是一种分布式协议,常被用来在 P2P 系统中分发消息.在 Gossip 协议中,节点首先将消息发送给周围的一组节点,周围节点在接收到消息后根据需要对消息进行转发.这样,消息就可以通过节点之间接力的方式进行传递.Gossip 中不存在集中控制,因而天生是一种可扩展性良好的协议.Gridmedia 采用 Gossip 协议在节点之间交换信息,以构建无结构的网状覆盖网络.

1.1 新节点的加入

在 Gridmedia 系统中,每个节点拥有一个全局唯一的标识符 *id*,维护一个部分在线节点的列表 *MemList* 以及一个系统同步时钟 t_{sys}.*MemList* 中的节点信息包括一个三元组:$\langle id,address,t_{live}\rangle$,其中,*id* 和 *address* 分别是节点的标识符和网络地址;t_{live} 是该条信息的生存期限,如果 $t_{sys}>t_{live}$,该节点信息将被从 *MemList* 中删除.

系统利用一个集中点服务器(rendezvous point,简称 RP)来启动新节点的加入过程.所有的节点都预先知道 RP 的地址信息,并且能够直接对它进行访问.新节点首先从 RP 获取 *MemList* 的初始列表,并校对时钟.之后,新节点开始与 *MemList* 中的节点通过 Gossip 协议交换信息,即完成加入过程.

1.2 节点*MemList*的维护

在 P2P 网络中,节点可以随时加入或退出系统,因此,各节点的 *MemList* 必须不断更新,以适应节点的动态性.在 Gridmedia 中,节点 *MemList* 的更新是通过 Gossip 协议来实现的.各个节点周期性地发送一个四元信息组$\langle id,address,t_{live},n_{hop}\rangle$,声明自己的存在.节点收到消息后的处理流程如图 1 所示.如果 $t_{live}<t_{sys}$,则说明该信息已经过期,直接丢弃;否则,节点更新其 *MemList*:(1) 如果节点 *id* 已经存在于 *MemList* 中,则更新 *MemList* 中对应项的 t_{live};(2) 否则,将收到的四元组中的前 3 项直接加入到其 *MemList* 中.如果 $n_{hop}>1$,则节点还将对消息进行转发.节点每隔一段时间就检查 *MemList* 中各节点的 t_{live},如果 $t_{live}<t_{sys}$,则将其从 *MemList* 中删除.

通过上述协议,Gridmedia 中每个节点可以不断更新其 *MemList*,避免节点因为其他节点的退出而出现被孤立的情况.这种方式完全依赖于节点之间的互相通信和协作,不需要中心节点的支持,具有很好的可扩展性.

1.3 节点的退出

在 P2P 网络中,节点的退出一般分为正常退出和非正常退出两种情况.正常退出是指程序按照正常步骤结束;非正常退出是指程序因出现异常而未按照正常步骤退出,包括程序崩溃、网络中断、强制结束进程等情况.

在 Gridmedia 中,正常退出的情况下,节点需要通过 Gossip 协议发送退出消息,其他节点在收到消息后将其从 *MemList* 中删除;在非正常退出的情况下,其他节点发现系统时钟 t_{sys} 大于失败节点的 t_{live} 时,将把其从 *MemList* 中删除.因此,非正常退出节点的信息不会长时间残留在系统中,从而保证了 *MemList* 中节点信息的有效性.

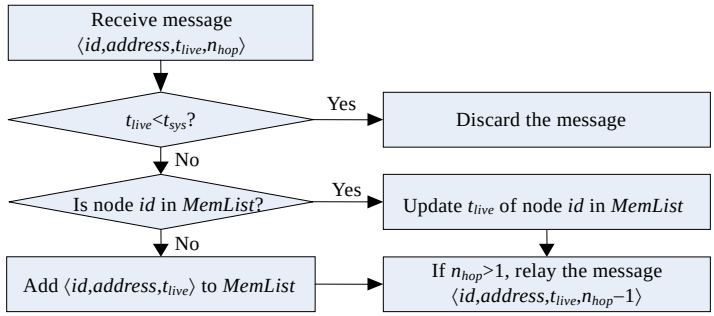


Fig.1 The process of handling message in Gridmedia

图 1 Gridmedia 中消息的处理过程

2 推拉结合的流传输策略

在基于 Gossip 的网状覆盖网络基础上,Gridmedia 采用了一种流传输策略.与 DONet 相比,Gridmedia 通过该策略大幅度减小了终端用户观看视频的延迟,改善了直播系统的实时性,并有效降低了控制开销.需要说明的是,Gridmedia 中的节点只与其 *MemList* 中的部分节点进行数据交换,并周期性地对这些节点进行更新,淘汰交换数据小于一定阈值的节点,补充新节点,以提高系统的吞吐量.

2.1 纯拉策略的不足

流传输策略是指如何在所构建的应用层覆盖网络上传输实时流媒体数据.DONet 采用了一种数据驱动的方式,节点之间首先互相交换各自缓存中存在的媒体数据的指示信息(buffer map,简称 BM),然后将自己的缓存中缺失的数据从其他声明拥有该数据的节点处请求过来,以补全自己缓存中的数据.在这种方式下,流媒体数据的传输是基于接收节点的主动请求,因此,本文中也称这种方法为纯拉策略.

在纯拉策略中,节点 *B* 从节点 *A* 处获取一个数据包 *p* 需要经过下面 3 个步骤:(1) *A* 发送 BM 给 *B*,声明数据包 *p* 存在自己的缓存中;(2) 如果 *B* 需要数据包 *p*,则请求 *A* 给自己发送数据包 *p*;(3) *A* 接收到请求后,将 *p* 发送给 *B*.可见,一个数据包的传递至少需要节点 *A* 和 *B* 进行 3 次通信.另外,考虑到效率问题,节点并不针对每个数据包发送 BM 和请求,而是将一组数据包的信息合并起来发送,这样,数据包在节点之间传递的平均时间延迟被进一步增大.数据包 *p* 从节点 *A* 传递到节点 *B* 的过程如图 2 所示.

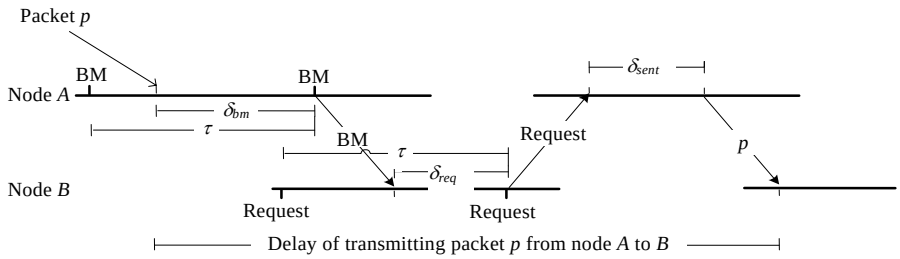


Fig.2 Delay of packet transmission using pure pull method

图 2 纯拉策略下数据包传递的延迟

假设节点 *A* 和节点 *B* 之间的平均传输延迟为 $\overline{\delta_{EED}}$,节点 BM 和请求的发送周期都为 τ,可知数据包交换 BM 和发送请求的平均等待时间分别为 $\overline{\delta_{BM}}=0.5\tau$, $\overline{\delta_{req}}=0.5\tau$.图 2 中 δ_{sent} 的存在是因为节点 *A* 在收到请求后,需要对请求的数据包按照一定的顺序进行发送,且发送过程在周期 τ 内完成,因此其平均值 $\overline{\delta_{sent}}=0.5\tau$.由图 2 可知,在纯拉策略下,数据包在两个节点之间传递的平均延迟 $\overline{T_{hop}}$ 可以表示为

$$\overline{T_{hop}}=\overline{\delta_{BM}}+\overline{\delta_{EED}}+\overline{\delta_{req}}+\overline{\delta_{EED}}+\overline{\delta_{sent}}+\overline{\delta_{EED}}=0.5\tau+\overline{\delta_{EED}}+0.5\tau+\overline{\delta_{EED}}+0.5\tau+\overline{\delta_{EED}}=1.5\tau+3\overline{\delta_{EED}}.$$

可以想象,如果节点 A 收到数据包 p 后能够立即将其转发至节点 B,那么数据包传递的平均时间仅为 $\overline{\delta_{EED}}$. 由此可见,纯拉策略会导致数据包传递的延迟大幅度增加.这在包括远程教学、视频直播等在内的很多对实时性要求较高的应用中是非常不利的.另外,在纯拉策略中,节点需要周期性地向邻居节点发送 BM 信息和请求,使得网络流量中控制信息的比重较高,导致系统的控制开销增大,这也是纯拉策略的不足之处.

2.2 推拉结合策略

为了解决纯拉策略在延迟上表现不佳和控制开销过大的问题,Gridmedia 采用了一种推拉结合的流传输策略.顾名思义,推拉结合策略中包含拉和推两种工作模式:拉模式与 DONet 中的纯拉策略相同;而推模式则截然不同.在推模式下,节点收到数据包后不需要向邻居节点发送 BM 或收到明确请求,而直接将数据包发送给需要该数据包的邻居节点,这样就使得数据包在节点之间的平均传输延迟降低到 $\overline{\delta_{EED}}$,并有效降低了系统的控制开销.这里需要解决的问题是:节点在接收到数据包 p 后,如何确定是否需要转发 p 以及向哪些邻居节点转发 p.

在 Gridmedia 中,节点将时间分成连续的时间片,在不同的时间片内分别采用拉模式或者推模式进行工作.通常在有新的邻居节点加入或者有邻居节点退出后的下一个时间片内,节点将工作在拉模式下,其他时间片内,节点工作在推模式下.在拉模式下,节点从邻居节点处把数据“拉”过来,并充分评价邻居节点的工作状态以及与自己的端到端链路状况;而在推模式下,节点根据上一时间片内对各邻居节点的评估,在时间片的一开始向各个邻居节点定制自己需要的数据.这样,邻居节点在收到新数据包后,就可以根据定制情况立即决定是否转发该数据包.节点向各个邻居节点定制的数据不应出现重复的情况,否则将收到重复数据,浪费网络带宽.在邻居节点间分配数据,可以通过对数据包序号作一个简单的哈希运算 mod(m)来实现(m 为一个较大的整数),对可能的结果 0,…,m−1,用轮盘赌的方法,按照比例向各个邻居节点进行定制.

在拉模式和推模式下都可能出现数据包没有按照预期到达的情况.在拉模式下,一次请求后没有收到的数据包可以在下一周期中继续请求,因此不需要对丢包情况进行特别处理;而推模式的情况相对复杂.直观上,在推模式的时间片内丢失的数据包也可以通过“拉”的方式再次向邻居节点发送请求来获得,但是,推模式下节点转发数据包的次序与它接收数据包的次序直接相关,不能确保按照数据包序号的大小顺序发送,因此难以根据接收到的数据包顺序来判断中间是否有数据包丢失的情况发生.比如按照定制情况,节点 A 负责向节点 B 转发数据包 1,3,5,7,9,…,但是,节点 A 收到数据包的顺序很可能为 1,3,9,7,5,….因为在推模式下,A 在收到数据包的同时立即向 B 转发数据,并不对数据进行重新排列,因此,B 收到的顺序也为 1,3,9,7,5,….这样,当 B 收到序号为 9 的数据包时,它就无法判断数据包 5 和数据包 7 是否真的丢失了.如果这时 B 采用拉模式对数据包 5 和数据包 7 向其他节点发送请求,则 B 很可能会收到双份的数据,从而浪费了带宽.为了避免这种情况的发生,节点 B 在向节点 A 定制数据的时候指定一个最大落后序号差 g.当 A 向 B 转发数据时,如果落后数据包的序号和已发送数据包的最大序号相差 g 以上,则不再对该包进行转发.同时,接收节点 B 也可以据此来判断是否需要将未到的数据包通过拉模式来获取.假如在上面例子中指定 g=3,则当节点 A 在发送了数据包 9 和数据包 7 以后,收到数据包 5 时,因为和已经发送数据包的最大序号 9 的差大于 g,因此不再转发数据包 5.而对于接收节点 B,当其接收到数据包 9 以后就可以确认 A 不会再向其转发数据包 5,因此,可以立即通过拉模式来获取数据包 5,这样就可以避免数据包的重复传递.需要注意的是,上述分析没有考虑 UDP(user datagram protocol)包在传输过程中因为路由振荡而引起的乱序问题,这种情况在互联网上发生的概率较小,对系统性能的影响可以忽略不计.

3 PlanetLab 实验及分析

为了验证推拉结合策略的有效性,我们在 PlanetLab[9]上进行了广泛的实验,结果证明了该策略在减小数据传输延迟和降低控制开销方面的有效性.

3.1 实验环境

因为 PlanetLab 的节点较多,在实验中我们采用分层的节点控制策略,将节点按照地域分布分成 12 组,每组

中选出一个头节点.当控制节点发送控制命令时,首先将控制命令发送给各个分组的头节点,然后头节点再将命令分发给组内的其他节点.

在实验中,为了模拟不同的节点动态性,我们采用两种节点行为方式:(1) 静态环境是指节点在进入系统以后一直停留在系统中,直到实验结束为止;(2) 动态环境是指节点不断地进行进入和退出系统的操作,其在线和离线时间分别符合平均值为 100s 和 10s 的指数分布.另外,为了验证系统在不同节点接入带宽下的性能,实验分两组进行,分别针对不限制节点上行带宽和限制上行带宽为 500Kbps 的情况.实验中采用的视频码流为 310Kbps,如果不特别说明,节点的邻居节点数量限制为 5 个.在拉模式中,节点之间 BM 和请求的发送周期 τ 为 1s.

3.2 延迟性能

为了比较策略的延迟性能,首先定义 3 个与传输延迟相关的性能指标:

(1) 绝对延迟(absolute-delay)——从源节点发出数据包到节点处理数据的延迟.

(2) 数据传递率(delivery-ratio)——截至某个延迟时间,节点收到的数据包的比例.

(3) α延迟(α-playback-delay)——当数据传递率达到一定值α ($0\le\alpha\le1$)时的绝对延迟.

图 3 给出了推拉结合策略在静态环境、不限制上行带宽的情况下,系统的 0.97-playback-delay 情况.从图中可以看到,在节点的加入过程中,在线节点的平均延迟基本呈减小的趋势,并且在所有节点加入以后很快稳定到 3s 左右.图 3 中的虚线给出的是最后一个加入的节点的 0.97-playback-delay,可以看到,它的延迟在很短时间(大约 20s)内就稳定到了整个系统的平均状态.节点在刚加入阶段的延迟快速减小的原因包括两个方面:(1) 邻居节点选择的逐步优化,即节点在加入覆盖网络后通过与其他节点之间的节点信息交换,逐步选择最优的邻居节点;(2) 找到稳定邻居节点后,节点之间传输策略从拉模式改变为推模式,有效减小了传输延迟.

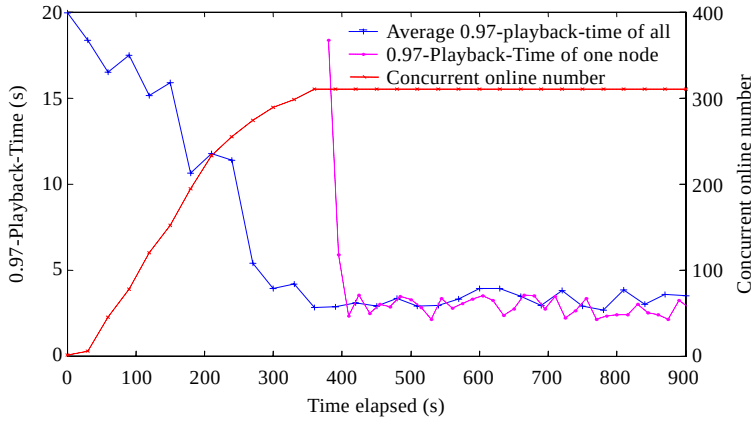


Fig.3 0.97-Playback-Delay of push-pull method in static environment

图 3 静态环境下推拉结合策略的 0.97-playback-delay

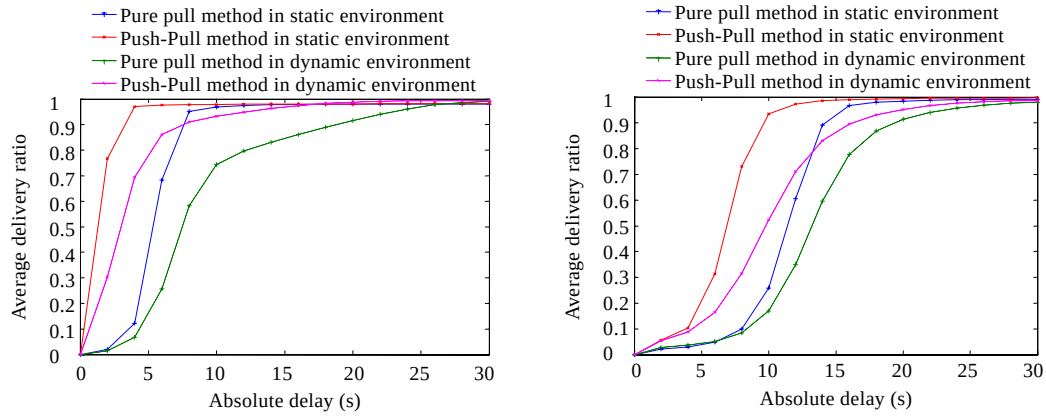
根据第 2.1 节中的分析,在纯拉策略中,节点之间每跳的延迟为 $\overline{T_{hop}}=1.5\tau+3\overline{\delta_{EED}}$.实验中 τ 为 1s,通过实验测得 PlanetLab 上节点之间的平均延迟 $\overline{\delta_{EED}}$ 约为 60ms,因此, $\overline{T_{hop}}$ 约为 1.68s.在节点的邻居节点数限制为 5 的情况下,相同延迟下数据传递率最大的应该是一棵最短树.树中只有根节点有 5 个子节点,其余节点最多有 4 个子节点.该最短树的数据传递率应为纯拉策略的严格上限,因此,当绝对延迟为 $i\times\overline{T_{hop}}$ 时,纯拉策略的数据传递率存在如下上限:

$$\overline{DR}(i\times\overline{T_{hop}})=\frac{Num(i)}{N}=\frac{\min\left(N,1+\sum_{j=1}^{i}[n\times(n-1)^{j-1}]\right)}{N},i\ge1,$$

其中:$Num(i)$为数据包经过 i 跳后可以到达的节点数;n=5,为节点的度;N=310,为覆盖网络的节点总数.因此,当

i=2,即绝对延迟为 3.36s 时,纯拉策略的数据传递率小于 10%.而根据图 3 可知,推拉结合策略在绝对延迟为 3.36s 时,系统的平均数据传递率可以达到 97%以上.

图 4(a)和图 4(b)分别给出了上行带宽无限制和限制为 500Kbps 时的节点平均数据传递率和绝对延迟之间的关系曲线.可以看到,无论是否对上行带宽进行了限制,推拉结合策略在静态和动态环境中的数据传递率都明显大于纯拉策略,证明了新策略的有效性.另外,通过图 4(a)和图 4(b)的对比我们可以发现,在对节点上行带宽进行限制以后,数据传输延迟明显增大,这是因为限制上行带宽造成了数据传输平均跳数的增加.



(a) Without uplink constraint (b) Uplink constrained to 500Kbps

(a) 上行带宽无限制 (b) 上行带宽限制为 500Kbps

Fig.4 Average delivery ratio vs. absolutely delay

图 4 不同绝对延迟下的平均数据传递率

3.3 协议通信开销

通信开销(overhead)是衡量协议效率的一个重要指标.本节给出我们对 Gridmedia 系统协议的额外开销的测量结果.我们定义额外开销为控制信息流量占节点全部流量的比例.图 5 给出了不同节点数下,纯拉策略和推拉结合策略在邻居节点数 n=3 和 n=5 时的通信开销.可以看到,两种策略的通信开销都不随着覆盖网络规模的增大而增大,这说明两种协议都具有很好的可扩展性.推拉结合策略因为减少了节点之间 BM 和请求的交换,其通信开销明显要小于纯拉策略,证明了推拉结合策略在降低系统控制开销上的有效性.

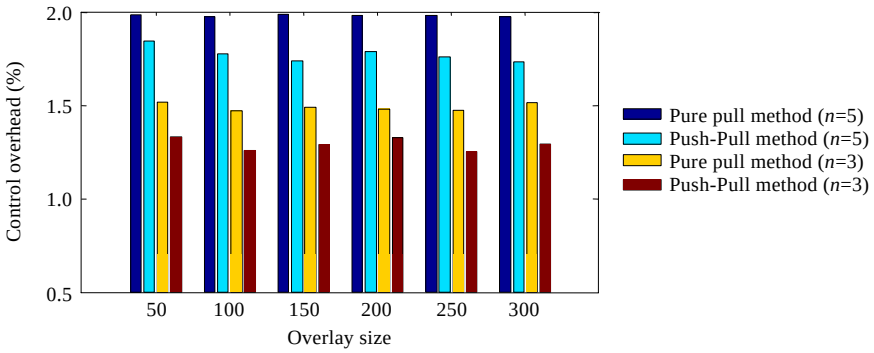


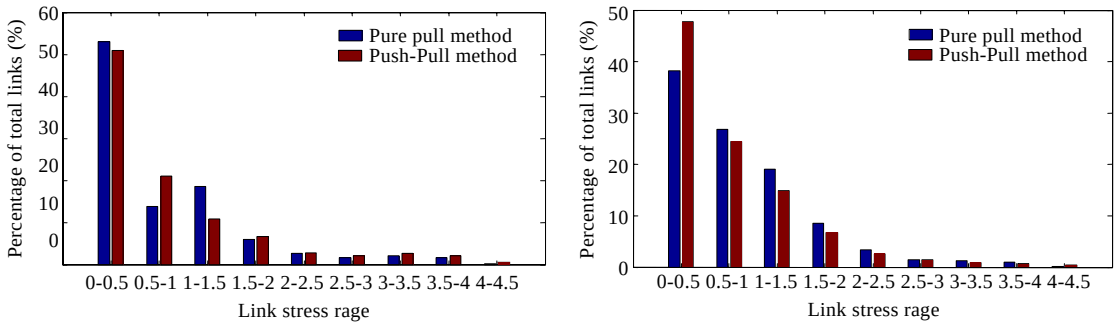
Fig.5 Control overhead of pure pull and push-pull method

图 5 纯拉策略和推拉结合策略的控制开销

3.4 链路负载

链路负载(link stress)是应用层组播的一个重要性能指标.ESM[3]中定义的链路负载是指在组播中,同一物

理链路上传递的同一数据包的次数,而 Gridmedia 是一种多发送者的结构,在一个物理链路上传递的数据包通常只是流媒体数据的一部分,该定义的指标不能直接用来反映链路负载情况.本文采用的链路负载指标是指物理链路上传递的数据流量和流媒体数据的码率之比.图 6(a)和图 6(b)分别给出了上行带宽无限制和限制为 500Kbps 时,Gridmedia 在静态环境下的链路负载情况.可以看到,纯拉策略和推拉结合策略的链路负载分布基本相同,这也说明了新策略并没有对链路负载状况产生明显影响.



(a) Without uplink constraint (a) 上行带宽无限制

(b) Uplink constrained to 500Kbps (b) 上行带宽限制为 500Kbps

Fig.6 Link stress in static environment

图 6 静态环境下的链路负载情况

4 实际系统运行情况

Gridmedia 的原型系统对 2005 年春节联欢晚会进行了全球网络直播.晚会期间,有来自全球 65 个国家和地区的超过 500 000 人次通过系统观看了直播.根据对用户 IP 的统计,当晚用户的地域分布见表 1.图 7 给出了晚会期间系统在线人数的统计,可以看到,在大多数时间内,系统的在线人数都超过了 10 000 人,最高在线人数达到了 15 329 人.需要说明的是,直播视频的平均码流为 300Kbps,这是考虑到中国有很多 ADSL 用户的接入带宽只有 512Kbps.通过我们的测试和用户的反馈来看,一般 ADSL 用户也可以正常收看直播视频.通过统计我们发现,有 60.8%的用户计算机在网络地址转换设备(network address translation,简称 NAT)后,其中至少有 16%的用户计算机是通过 ADSL 接入 Internet 的.这些统计信息为我们对系统的进一步改进提供了很好的参考.

Table 1 User location distribution during the Spring Festival Gala Evening in 2005

表 **1** 2005 年春节联欢晚会期间用户的地域分布

Country	China	UK	France	Canada	Japan	Singapore	USA	Germany	Others
Percentage (%)	77.7	3.3	1.2	4.5	2.9	1.0	4.1	1.4	3.9

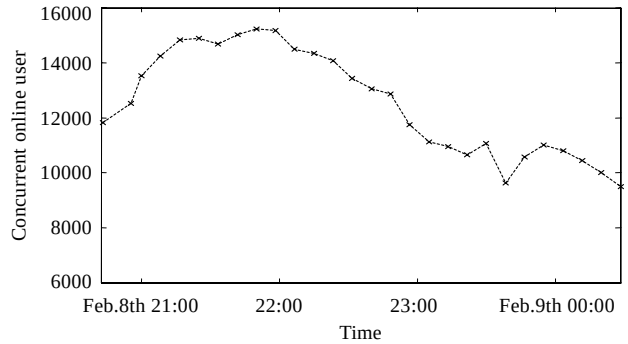


Fig.7 Number of concurrent online users during the Spring Festival Gala Evening in 2005

图 7 2005 年春节联欢晚会期间同时在线人数统计

5 研究背景和相关工作

目前,已经有很多工作对基于 P2P 网络的视频直播进行了分析研究,并进行了很多有益的尝试和实验,但是,能够在 Internet 上实际运行的系统却不多见.最早的实际系统是 ESM[3],但 ESM 的可扩展性不好,设计规模只有几百人,这主要是因为其覆盖网络拓扑维护和优化的代价过高.NICE[5]是一种可扩展的应用层组播方案,将端节点组织到不同的集群中,并在集群间形成层次结构的数据转发和拓扑管理体系,其贡献在于改善了应用层组播的可扩展性,并将协议负载控制在很低的程度.ZIGZAG[6]在 NICE 的基础上将集群的管理和数据转发功能分开,由不同的节点来完成,降低了高层节点的负载,进一步提高了系统的可靠性和扩展性.上述几种方案在传输流媒体数据时均采用树形的单发送者(single-sender)方式,对节点间端到端带宽的要求很高,并对节点动态性极其敏感,因而难以在 Internet 上广泛部署.CoopNet[10]和 SplitStream[11]采用的是一种基于多树的方案,在组成员之间同时维护多个组播树,利用 MDC(multiple description coding)[12]把视频编码成多个视频流,分别同时沿多个组播树进行传输,组成员只要收到这些视频流中的一部分,就可以独立地进行解码,因此可以较好地适应节点的动态性和节点之间带宽的抖动.多树方案与单树方案相比的另一个优点是,所有的节点都能够为系统贡献自己的上行带宽,有利于提高系统的吞吐量.但是,它们的不足之处在于,需要维护多个独立的组播树,协议的计算和通信开销较大.

DONet[8]采用了 Gossip 协议构建无结构的覆盖网络,并采用多发送者和数据驱动的方式进行流媒体数据传输,在很大程度上解决了上述方案中存在的不足,使其能够在 Internet 上开展大规模的服务.DONet 在一定程度上与流行的 P2P 文件共享系统 BitTorrent[13]具有相似之处,都是通过邻居节点之间的数据交换来实现内容分发.但是,BitTorrent 在下载数据分片时不需要按照顺序进行,而 DONet 因为传输的是流媒体数据,则必须按照一定的顺序来获取数据,以满足客户端实时播放的需求.虽然 DONet 可以说是第一个可以大规模部署的 P2P 视频直播系统,但是其采用的纯拉传输策略也导致了视频传输延迟和系统控制开销的增大.为此,本文提出了一种推拉结合的传输策略,有效减小了传输延迟和控制开销.

6 总 结

Gridmedia 使用 Gossip 协议构建无结构的网状应用层覆盖网络,并采用了一种推拉结合的流传输策略.与 DONet 中采用的纯拉策略相比,该策略大幅度减小了视频传输延迟,改善了直播系统的实时性,并有效降低了直播系统的控制开销.通过 PlanetLab 上的广泛实验,我们充分验证了该策略的有效性.Gridmedia 的原型系统对 2005 年春节联欢晚会进行了全球互联网直播,达到了预期的效果,这也进一步说明了 Gridmedia 系统在大规模视频直播中的性能和应用前景.

致谢 我们要特别感谢符文杰、张一飞、骆骥、张宇宙同志在 Gridmedia 的实现过程中做出的贡献,并感谢汤筠、张江、陈春晓同志在论文写作过程中给予的帮助.

References:

[1] Wang YW, Zhang ZL, Du DHC, Su DL. A network conscious approach to end-to-end video delivery over wide area networks using proxy servers. In: Guerin R, ed. Proc. of the IEEE INFOCOM. San Francisco: IEEE Press, 1998. 660−667.

[2] Vakali A, Pallis G. Content delivery networks: Status and trends. IEEE Internet Computing, 2003,7(6):68−74.

[3] Chu YH, Rao SG, Zhang H. A case for end system multicast. In: Brandwajn A, ed. Proc. of the ACM SIGMETRICS. Santa Clara: ACM Press, 2000. 1−12.

[4] Zhang B, Jamin S, Zhang L. Host multicast: A framework for delivering multicast to end users. In: Kermani P, ed. Proc. of the IEEE INFOCOM. New York: IEEE Press, 2002. 1366−1375.

[5] Banerjee S, Bhattacharjee B, Kommareddy C. Scalable application layer multicast. In: Steenkiste P, ed. Proc. of the SIGCOMM. Pittsburgh: ACM Press, 2002. 205−217.

[6] Tran DA, Hua KA, Do T. ZIGZAG: An efficient peer-to-peer scheme for media streaming. In: Bauer F, ed. Proc. of the IEEE

INFOCOM. San Francisco: IEEE Press, 2003. 1283−1292.
[7] Rejaie R, Stafford S. A framework for architecting peer-to-peer receiver-driven overlays. In: Padmanabhan V, ed. Proc. of the ACM NOSSDAV. Kinsale: ACM Press, 2004. 42−47.
[8] Zhang XY, Liu JC, Li B, Yum TSP. CoolStreaming/DONet: A data-driven overlay network for live media streaming. In: Znati T, ed. Proc. of the IEEE INFOCOM. Miami: IEEE Press, 2005. 2102−2111.
[9] PlanetLab. http://www.planet-lab.org/
[10] Padmanabhan VN, Sripanidkulchai K. The case for cooperative networking. In: Kaashoek F, ed. Proc. of the IPTPS. LNCS 2429, Heidelberg: Springer-Verlag, 2002. 178−190.
[11] Castro M, Druschel P, Kermarrec AM, Nandi A, Rowstron A, Singh A. SplitStream: High-Bandwidth content distribution in a cooperative environment. In: Castro M, ed. Proc. of the IPTPS. LNCS 2735, Heidelberg: Springer-Verlag, 2003. 292−303.
[12] Goyal VK. Multiple description coding: Compression meets the network. IEEE Signal Processing Magazine, 2001,18(5):74−93.
[13] BitTorrent. http://www.bittorrent.com/

罗建光(1981－),男,浙江嘉兴人,博士生,主要研究领域为多媒体网络,对等网络.

赵黎(1975－),男,博士,副教授,主要研究领域为多媒体信号处理,计算机网络传输,模式识别.

张萌(1982－),男,博士生,主要研究领域为多媒体网络,对等网络.

杨士强(1952－),男,教授,博士生导师,CCF 高级会员,主要研究领域为基于内容的多媒体检索,视频分析和流化技术,分布式多媒体系统,视频压缩,多媒体网络,嵌入式多媒体.