基于动态时间坐标系的搜索表拓扑组织方法

武广柱[1,2],王劲林[1]
WU Guang-zhu[1,2],WANG Jin-lin[1]

1.中国科学院 声学研究所,北京 100080
2.中国科学院 研究生院,北京 100080
1.Institute of Acoustics,Chinese Academy of Sciences,Beijing 100080,China
2.Graduate University of Chinese Academy of Sciences,Beijing 100080,China
E-mail:wugz@dsp.ac.cn

WU Guang-zhu,WANG Jin-lin.Search table topology structure based on dynamic time coordinate system.Computer Engineering and Applications,2008,44(15):82-84.

Abstract: Finding partners for a peer in P2P VoD systems is still a critical issue,especially when VCR functions are supported. Peers' cache are usually limited,and peers' play occasion may jump to any point of the stream at any time.So,it is very expensive to track buffer contents,which change constantly.This paper presents a search table topology structure based on dynamic time coordinate system for large-scale P2P VoD system.In this coordinate system,any peer's coordinate maintains constant unless the peer's play occasion jumps to another point of the stream.Thus a chord-like search table topology is designed.Simulations show that our design achieves good performance.

Key words: Peer-to-Peer;VoD;resource locating

摘 要: 在P2P点播系统中,如何快速发现合作节点这一资源定位问题是一个挑战。特别在用户进行VCR操作时,这一问题更显突出。播放点的随节点播放而连续前移和用户VCR操作造成的节点跳转使得系统追踪节点缓存信息代价很高。提出了一种基于动态时间坐标系的复杂度为O(log$M$)(M为节目的分段数)搜索表拓扑组织方法,以解决P2P VoD系统资源定位困难的问题。仿真结果标明,该方法具有很好地可扩展性和较高地查找效率。

关键词: Peer-to-Peer;VoD;资源定位
DOI: 10.3778/j.issn.1002-8331.2008.15.026 **文章编号:** 1002-8331(2008)15-0082-03 **文献标识码:** A **中图分类号:** TN915.03

1 引言

P2P流媒体系统[2-8]可以分为直播系统和点播系统。相对直播而言,P2P点播系统有以下特点:(1)异步性,同一节目不同用户会在任意时间从任意位置开始播放;(2)非连续性,用户在观看过程中会进行VCR操作,而不是像直播一样从头到尾的顺序观看;(3)冷热不均性,不同频道间存在冷播和热播现象。其中,异步性和非连续性的特点使得P2P点播系统的资源定位具有挑战性:结点的缓存状态难以追踪,搜索合作结点困难。

对于直播系统,处于同一频道的各个结点基本处于同一播放点,在结点缓存大小合适的条件下,这种同步性使得合作结点发现相对简单。索引服务器记录下频道内的所有结点,当新的结点加入时,索引服务器随机返回一些结点,这些结点都能够成为新加入结点的合作结点(本文不讨论选择合作结点的优化问题,而仅仅讨论结点间能否协作的问题)。因为在直播系统中结点不会进行快进、快退、拖动等VCR操作,结点无需经常性的报告自己的播放点也无需经常进行合作结点发现请求。故在直播系统中,索引服务器仅仅记录结点所属的频道即可,这种索引粒度为"频道"的索引方式以及无VCR操作使得服务器压力较小。

点播系统中的用户随时加入系统并从节目开始或者任意位置播放,此外用户还会在播放过程中进行快进、快退、拖动等VCR操作。在结点缓存受限的情况下,结点缓存一般仅保存几分钟的内容,两个结点要想协作,则其播放点必须相近。例如,一个频道中现有三个结点,P_a在节目的第1分钟播放,P_b在第2分钟位置播放,P_c在第30分钟位置播放;现在新加入的结点P_d要从第28分钟处播放,则只有结点P_c才可能成为P_d的合作结点。而P_d要发现P_c,索引粒度为"频道"是不行的,系统就必须提供一种较细粒度的索引机制。这种较细粒度的索引可能是分布式的也可能是集中式的。集中管理在结点数量不大时最为有效:结点定期向服务器报告其缓存位置,需要查找合作集的结点向索引服务器查询。但是,一旦大规模部署,服务器将难以承担,而且存在单点失效。一般的分布式解决可能采用Gen-

基金项目: 国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2005AA1032);中国下一代互联网示范项目(the China Next Generation Internet(CNGI) under Grant No.CNGI-04-15-2A)。
作者简介: 武广柱(1979-),男,博士,主要研究方向为宽带多媒体通信;王劲林(1964-),男,研究员,主要研究方向为宽带多媒体通信。
收稿日期: 2007-11-22 **修回日期:** 2008-02-25

eration、Session 等“关系内嵌”技术,将结点的播放点关系内嵌到结点的逻辑拓扑联系中。然而,通常的“关系内嵌”索引方式因需要逐级查询才能找到合作结点集,效率不高。

本文提出了动态时间坐标系的概念,系统中各频道都建立一个动态时间坐标系,在本坐标系下结点播放点坐标变成一个常数(不进行 VCR 操作时)。利用这一性质,并受 Chord[1]的启发,本文设计了基于动态时间坐标系的搜索表拓扑组织方法(DTCS-ST,Search Table based on Dynamic Time Coordinate System)。该方法属于 “关系内嵌”的拓扑组织方法,其查找复杂度 O(log$M$),M 为节目的分段数。

2 相关工作

P2P 直播系统已经走向成熟并投入商业运营。但 P2P 点播系统却因异步特性而仍然存在诸多挑战。资源定位问题便是其中的一个。下面对几个典型 P2P 点播系统的资源定位方法做一个回顾。

在 P2Cast[2]中,加入系统的时间相近的节点构成一个 session。对于每一个 session,媒体服务器和本 session 中的节点通过单播构成一棵应用层组播树,称为基础树。对于一个新加入的节点,如果其父节点没有缓存其所需要的内容片段,则节点直接从服务器下载, 也可以从本 seesion 内具有该片段的其它节点下载。这种打补丁的工作方式使得 P2Cast 比传统的 C/S 模式服务更多的用户。新节点 P 加入系统首先要联系服务器,以使得服务器能够追踪系统中的节点。如果 P 属于一个已经存在的 session,则 P 加入基础树并选择一个补丁服务节点,如果成功,这节点被接纳。否则,如果因节点带宽不足或者找不到合适的补丁服务节点进而需要服务器来提供补丁服务而服务器带宽又不足,则节点被拒绝。如果节点需要新开一个 session 而且服务器带宽允许,则允许节点加入,否则拒绝节点。P2Cast 的节点加入完全依赖于服务器进行索引, 是一种中心管理方式。P2Cast 并未对用户 VCR 操作时如何定位资源进行讨论。

在 P2VoD[3]中节点依据它们的加入时间组织成多等级群组,数据流沿重叠树进行转发。每个节点从其上级群组中的一个父亲节点接收数据并将数据转发到位于其低层群组的子节点。新节点加入系统可以尝试加入低的群组或形成一新的最低群组。如果它不能从组播树中找到一可用的父亲节点,且服务器有足够的带宽,则它直接连接到服务器。可见,P2VoD 采用的是一种关系内嵌式的索引方法。P2VoD 没有对用户 VCR 操作造成的节点跳跃索引做优化。如果一用户启动跳转请求,节点需要顺序地搜索其上群或下群组,由于群组数量很大,搜索开销是非常高的。

OBN[4]抛弃了 P2Cast 和 P2VoD 将资源定位内嵌到内容分发拓扑中的做法。OBN 构建了称作重叠缓存网络的拓扑结构,利用了节点流畅播放时各节点播放点的差相对固定这一性质进行资源定位。

Vmesh[5]将 DHT 索引引入了 P2P 点播系统。一个新加入节点可以使用 DHT 搜索它所关注的分段。然而,VMesh 应用 DHT 解决的是相对固定的硬盘数据索引。Vmesh 中的节点要存储一些频道媒体文件的数据块并保留较长时间,这些数据块并不随播放点的移动而动态变化。Vmesh 没有解决播放缓冲区的数据的索引。

3 动态时间坐标系

系统中每一个频道都有一个动态时间坐标系 DCS(Dynamic Coordinate System)。假设频道中有一个节点从某时间点开始循环播放本频道节目,这一假设节点当前正播放的位置称为本频道的虚拟播放点。动态时间坐标系的坐标原点和本频道的虚拟播放点相等。

设 C 为系统中所有频道的有限集。假设系统中的频道 $C_j \in C$ 的虚拟节点从绝对时间 TS_j 开始不断循环播放本频道的影片,本频道整个影片的播放时间是 T_j。对于 C_j,可以计算出当前时间的虚拟播放点播放循环播放的遍数 N_j 以及其播放位置 TP_{curr},TP_{curr} 即为 C_j 的动态时间坐标系原点 O_j:

$$N_j=floor((T_{curr}-TS_j)/T_j)$$

$$O_j=TP_{curr}=(T_{curr}-TS_j)\%T_j$$

每一个频道的影片都被按照一固定时间 TL 划分为多个片段,称为桶。设 B 为系统中所有桶的集合,$B_j \in B$ 是频道 $C_j \in C$ 的所有桶的集合。给定一 C_j 中任意播放点 P_j,设当前时间是 T_{curr},则 P_j 所属的桶为:

$$D_j=\begin{cases} TP_j-O_j & \text{if } N_i \text{ 为偶数} \\ TP_j-O_j+T_j & \text{else if } N_j \text{ 为奇数且 } TP_j<O_j \\ TP_j-O_j-T_j & \text{else} \end{cases}$$

$$B_j=floor(\frac{D_j}{TL})$$

其中,TP_j 是从影片开始到 P_j 的时长。

影片在通常方法可以划分为 $T_j/TL=M$ 个桶,则 B_j 的取值范围为[$-M$,$M-1$]中的整数,共 $2M$ 个。负桶表示节点播放点滞后于虚拟节点的播放点。

可以看出,在节点没有进行 VCR 操作的一段时间内,节点的播放点在这一个动态坐标系下的坐标是不变的。

4 基于动态时间坐标系的搜索表拓扑组织方法

基于动态时间坐标跳表的拓扑组织方法是一种 “关系内嵌”式的组织方法,它将不同结点的播放点的关系内嵌到了结点间的逻辑拓扑关系中。结点执行 VCR 操作时不需要向索引服务器请求资源列表,而是依据结点之间的逻辑关系进行合作结点查找并进行拓扑重构,从而减小了对索引服务器的依赖。

在动态坐标系下, 结点所在的桶在不进行 VCR 操作时是恒定的,并且,新加入的结点或者进行拖动操作的结点可以很容易的计算出其目标桶号;这样,结点的点播或者 VCR 操作就变为了加入某个目标桶。

为了依靠“关系内嵌”索引到目标桶,本文将桶按照桶号构成一个环, 并且按照 Chord 的组织方式将桶内的结点联系起来。为了编号方便,将 B_j 重新定义为 $B_j=floor(D_j/TL)+M$ 以不再出现负号的桶。每个结点除要维持其所在桶和相邻桶内一些结点的联系外,还要维持距离本桶 2^i($1 \leqslant i \leqslant n$,$2^n=M$,$M$ 为桶的个数)的桶内的若干结点的联系。本桶内或相邻桶内和本结点的播放点距离小于 TL 的结点构成的集合, 称为本结点的近邻结点集,其它结点构成的结合称为远邻结点集。图 1 示意了桶 0 内的结点按 Chord 组织时的路由表指针。

4.1 结点加入

当结点 P 点播某频道时,首先通过一索引服务器获知本频道内的某个结点 A 的信息,然后 P 计算其目标桶号,并向 A 发送加入请求。A 采用与 Chord 相似的搜索算法,将这一加入请

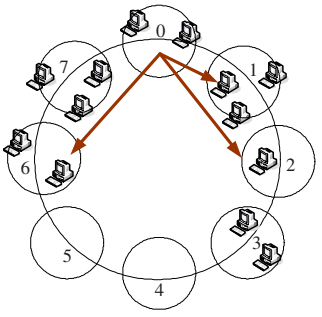

图 1 DTCS-ST 的路由表

求路由到目标桶内的某结点 B。B 将根据 P 的目标播放点,查抄其路由表内的属于 P 的邻居结点集的结点,并把这些结点通知给 P。至此,P 加入了本频道,并和其邻居结点相互协作。 如果 P 因目标桶内本没有结点并且其目标桶的相邻桶中也没有近邻结点;则 P 直接向媒体服务器请求片段。参考 Chord 的分析,可以得出本文方法的查询复杂度为 $O(\log M)$。如何从近邻结点集中选择最优的合作结点,片段如何分发,不是本文的讨论范围。

结点加入后,需要构建路由表,方法同 Chord 的结点加入后的路由表构建操作。

4.2 VCR 操作时的拓扑重构

当结点频道中的某结点 P 进行 VCR 操作时, 如果其目标桶和当前所在桶相同, 则 P 采用 Gossip 的方式向其近邻结点发送结点请求,目的是发现新坐标下的近邻结点。如果目标桶和当前桶不同,则结点 P 从其路由表中选择一个距离目标桶最近的桶内的结点 A,然后向 A 发送桶搜索请求,其操作同结点加入。当 P 加入目标桶后,P 需要重新构建其路由表,并不向其原路由表内的结点发送桶改变信息。当 P 收到其它结点 B 在拓扑维护中发来的心跳报文时,P 需要报告其桶号已经改变。B 需要根据需要将 P 放到新的路由表位置中或者将 P 从路由表中删除。

4.3 拓扑维护

系统中的结点 P 需要定期向其路由表中的结点发送心跳报文。当发现某结点异常离开时,P 需要将其从自己路由表中删除;当发现某结点改变了桶号时,P 需要更新路由表。如果路由表中某项缺失,则 P 需要运行桶发现操作,找到一个合适的结点并将其添加到路由表中。

4.4 结点离开

当结点退出某频道时,需要向其合作结点和路由表中的其它结点发送离开报告。系统依靠拓扑维护算法来维持路由表节点异常离开时逻辑拓扑的正确性。

5 仿真结果

为验证算法在大规模 P2P 点播系统中的性能,本文假设点播系统提供 1 000 套节目,每套节目的平均播放时间为 4 000 s,TL 为 200 s,系统中有 20 万用户,用户平均每 500 s 将进行一次拖动。网络中节点间的平均延迟为 100 ms,带宽为 10 M。索引服务器平均响应时间是 5 ms, 普通节点的平均响应时间是 20 ms。

本文对单纯使用索引服务器索引的 C/S 模式方法、P2VoD 采用的逐级查询的方法、基于动态时间坐标系的搜索表拓扑组织方法(DTCS-ST)进行了仿真。图 2 为节点系统对节点拖动和加入的响应时间(s)。可见,由于 C/S 模式下服务器不但要处理节点加入频道还要处理节点的拖动请求, 响应时间很长;P2VoD 采用的逐级搜索方式也需要较长的响应时间;DTCS-ST 因将索引负担从服务器分布到普通节点并且不需要逐级搜索从而使得响应速度较快。图 3 为索引服务器的负载情况(请求数/s)。

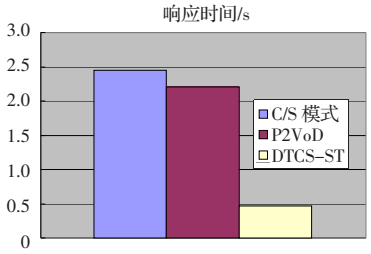

图 2 响应时间

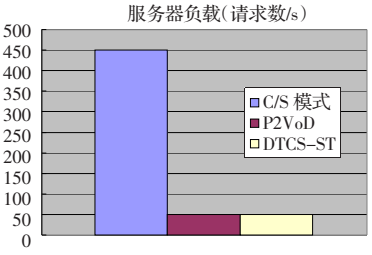

图 3 服务器负载

6 结束语

本文提出了一种用于大规模 P2P VoD 系统的组织方法,以解决 P2P VoD 系统索引困难的问题。仿真标明该方法在查找效率和降低服务器负载上都有较好的效果。

参考文献:

[1] Stoica I,Morris R,Liben-Nowell D,et al.Chord:a scalable Peer-to-Peer lookup protocol for Internet applications[J].IEEE/ACM Transactions on Networking,2003,11(1):11-32.

[2] Guo Y,Suh K,Kurose J,et al.P2Cast:Peer-to-Peer patching scheme for VoD service[C]//Proceedings of the 12th World Wide Web Conference(WWW'03),Budapest,Hungary,2003:301-309.

[3] Do T,Hua K A,Tantaoui M.P2VoD:providing fault tolerant video-on-demand streaming in Peer-to-Peer environment[C]//Proceedings of IEEE ICC'04,Paris,France,2004:1467-1472.

[4] Liao Chi-shiang,Sun Wen-hung,King Chung-ta,et al.OBN:peering for finding suppliers in P2P on-demand streaming systems[C]//Proc of 12th International Conference on Parallel and Distributed Systems,2006,1:235-242.

[5] Yiu K P K,Jin X,Chan S H G.Distributed storage to support user interactivity in Peer-to-Peer video streaming[C]//IEEE International Conference on Communications,2006,1:55-60.

[6] Zhang R M,Butt A R,Hu Y C.Topology-aware Peer-to-Peer on-demand Streaming[C]//Proceedings of 2005 IFIP Networking Conference,Waterloo,Ontario,Canada,2005:2-6.

[7] Cui Yi,Li Bao-chun,Nahrstedt K.oStream:asynchronous streaming multicast in application-layer overlay networks[J].IEEE Journal on Selected Areas in Communications,2004,22:91-106.

[8] Dana C,Li D,Harrison D,et al.BASS:BitTorrent assisted streaming system for video-on-demand[C]//2005 IEEE 7th Workshop on Multimedia Signal Processing,2005:1-4.