综述与评论 Review and Comment

P2P 流媒体系统模型及关键技术研究*

王 娟[1,2],黄鹏辉[1,2],朱艳琴[1,2]
(1.苏州大学 计算机科学与技术学院,江苏 苏州 215006;
2.江苏省计算机信息处理新技术重点实验室,江苏 苏州 215006)

摘 要:介绍了典型的 P2P 流媒体系统模型,并指出基于多播树协议的服务模型与基于 Gossip 协议的服务模型的区别。分析了对 P2P 流媒体系统的节点的调度算法、数据存储、资源发现、内容分发等关键技术,在此基础上指出了 P2P 流媒体系统进一步的研究方向。

关键词:对等网络;流媒体系统;数据存储;资源发现;内容分发
中图分类号:TP393 **文献标识码**:A

Research on P2P streaming media model and key techniques

WANG Juan[1,2], HUANG Peng Hui[1,2], ZHU Yan Qin[1,2]
(1.School of Computer Science and Technology, Soochow University, Suzhou 215006, China;
2.Provincial Key Laboratory for Computer Information Processing Technology, Suzhou 215006, China)

Abstract: This paper gives an introduction to typical P2P system models. Then it points out the differences between models based on Gossip protocol and multicast tree protocol. Then it shows some key technologies of typical P2P system models, such as the scheduling algorithm at one node, data storage, resource discovery, content distribution and so on. Finally, open issues in P2P media streaming systems for further study are discussed.

Key words: P2P; media streaming systems; data storage; resource discovery; content discovery

以 P2P(Peer-to-Peer)为代表的覆盖网络,以其独特的结构特点,可变集中处理和存储为分布处理和存储,充分挖掘 Internet 边缘的空闲资源,克服了传统的客户机/服务器(C/S)结构中服务器负载过高、网络带宽占用过大、服务器易形成单点失效等缺点,并改善了网络层组播(IP 组播)结构系统扩展性不好、内容分发网络 CDN(Content Delivery Network)部署成本高等缺点。基于这些优势,研究人员将 P2P 技术引入流媒体系统中,形成了 P2P 流媒体技术。

P2P 技术的快速发展为大规模流媒体应用提供了新的解决方案,许多实际运行的系统证明了将 P2P 技术应用于内容分发过程中的有效性,如提供文件共享服务功能的 BitTorrent 系统[1]、提供视频直播服务功能的 CoolStreaming[2]、PPLive[3]以及提供视频点播服务功能的 PPStream[4]等。本文对 P2P 流媒体系统模型及其关键技术进行了分析,并在此基础上指出了 P2P 流媒体技术的进一步的研究方向。

1 P2P 流媒体系统模型

P2P 流媒体系统按其工作方式大致可分为二类:基于树状多播系统和基于 Gossip 协议的网状多播系统。

基于树状多播的 P2P 流媒体系统将网络中所有节点组织成 1 棵多播树,如图 1 所示。树的根节点是媒体发布源,流媒体数据从多播树的父节点向其子节点传播直到叶节点。该方法可以最小化系统中多余的数据传播,并保证每个数据块能传播到系统中每个节点,其缺点是:(1)多播树极易分裂且维护的开销巨大;(2)多播树的父节点限制了其所在子树的最大输入带宽,因此多播树中带宽瓶颈节点到处存在;(3)各个节点的负载不

* 基金项目:国家自然科学基金资助项目(60673041);江苏省高校自然科学基础研究项目(08KJB520011);苏州市融合通信重点实验室资助(SZS0805)

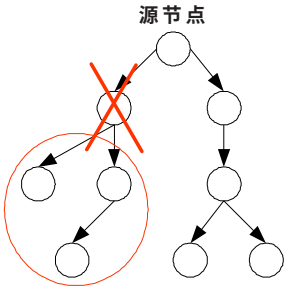

图 1 基于树状多播的工作方式

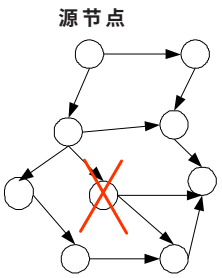

图 2 基于 Gossip 协议的工作方式

均衡，如叶节点只下载不上传，是纯粹的带宽消费者。典型代表有 ZIGZAG 系统[5]等。

近年来，基于 Gossip 协议的 P2P 流媒体系统已成为 P2P 流媒体系统的主流，如图 2 所示。基于 Gossip 协议的网状服务模型并没有依靠固定的拓扑结构把数据转发给接收节点，而是依靠数据有效信息来驱动数据在节点间流动，因此该结构又称为数据驱动化网络。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，消息就可以通过节点之间以接力的方式进行传递。邻居维护的灵活性与数据传播的随机性使得基于 Gossip 协议的 P2P 流媒体系统不会因节点失效而导致显著的性能下降，从而良好地适应了高动态性的互联网环境。因此，本文主要针对网状结构系统模型来分析 P2P 流媒体系统的主要组成部分及关键实现技术。

为了更好地理解 P2P 流媒体系统的框架结构，下面给出基于 Gossip 协议的 DONet 服务模型系统的主要功能模块。

(1)成员管理模块：实现成员节点的管理。在成员管理模块中维护一个 mCache (membership Cache)，该表包含当前系统中部分活动节点信息。当新节点加入系统时，首先连接节目源服务器，服务器从它的 mCache 中随机选择 1 个代理节点，并把新节点的加入请求重定向给该代理节点。新节点从代理节点获取 1 个成员列表作为备选节点集合，然后从该集合中选取部分节点作为自己的伙伴节点。

(2)伙伴关系管理模块：建立和维护节点间的伙伴关系，并通过交换缓存映射图 BM(Buffer Map)获取节点间的有效数据信息。整个视频流被分成长度相同的数据段 (segment)，节点缓存中的数据段有效性信息可通过 BM 来表示。节点定期地与其伙伴节点交换 BM，调度算法根据伙伴节点中的 BM 来确定获取哪个数据段。

(3)调度模块：负责把数据实时地传送到播放节点的缓存中，用以保证媒体播放的连续性。对于同构的静态网络，简单的轮询调度算法即可满足数据的调度；但对于异构的动态网络，则需要更加智能化的调度算法。调度应满足 2 个约束条件[6]：①每个数据段需在播放时限之前到达，错过时限的片段要尽可能少；②每个节点的带宽情况不同。该问题是并行机调度的一个变种，为 NP 难题。因此，想要寻找一个适合具体网络的调度算法，特别是适合动态网络环境的调度算法，将是非常困难的。下面给出一种可以实时连续观看流媒体内容的节点调度算法。

节点的调度算法

输入：deadline[i] 数据块 i 的截止期限
seg_size 数据块的大小
set_partners 伙伴节点集合
num_partners 节点的伙伴节点的数量
band[k] 伙伴节点 k 的带宽
bm[k] 伙伴节点 k 的缓存映射
fetch_set 需要取得的数据块集合
输出：supplier[i] 在 fetch_set 中不可用数据块 i 的提供者

(1)对于块 i∈fetch_set，执行
①n←0；
②对于 j 从 1 到 num_partners，执行
T[j，i]←deadline[i]−current_time； //传送到第 i 个//数据块可用的时间
n←n+bm[j，i]； //数据块 i 的潜在提供者数量
③若 n=1，则执行 //数据块只有 1 个潜在提供者
k←arg$_r${bm[r，i]=1}；
supplier[i]←k；
对于 fetch_set 中的每个块 j(j>k)，执行
t[k，j]←t[k，j]−seg_size/band[k]；
否则，执行
dup_set[n]←dup_set[n]Y{i}；
supplier[n]←null；
(2)对于 n 从 2 到 num_partners，执行
对于每个 i∈fetch_set[n]，执行 //数据块有 n 个//潜在提供者
k←arg$_r${band[r]>brand[r′]|t[r，i]>set_size/band[r]|，t[r′，i]>seg_size/band[r′]，r，r′∈set_partners}；
若 k 不为空，则
supplier[i]←k；
对于 fetch_set 中的每个块 j(j>k)，执行
t[k，j]←t[k，j]−seg_size/band[k]；
(3)返回 supplier[i]。

在节点调度算法中，首先计算每个数据块的潜在提供者的数量。因为当 1 个数据块有很少的潜在提供者时，要满足该数据块的截止期限的限制将会很困难。因为节点调度算法从开始只有单个提供者的块到具有 2 个提供者的数据块再到具有多个提供者的数据块的顺序来确定每个数据块的潜在提供者。在这些潜在提供者中，具有最高带宽、足够的可利用时间的提供者将会被

选择。算法被周期性地执行更新调度,调度完成后,同一个提供者的数据块被表示成 BM 的形式传送给相应的提供者,提供者通过一个实时的传输协议传输数据。

2 数据存储

媒体数据在系统中存储决定了系统中数据的可用性。这不仅对 P2P 直播系统中节点间播放的同步性有影响,而且对视频点播系统中交互性支持能力也有直接的影响。因此,好的数据存储策略对整个系统的性能的提高是至关重要的。

2.1 数据分块策略

单个节点的存储能力有限,这就要求对媒体数据进行分割,将其分散存储于系统中的多个节点中。CoolStreaming 首先把整个媒体文件分成大小相等的若干数据块,以连续编号进行标识,并且将整个视频流划分为一系列的子流,每个子流中存储一部分数据块。假设某媒体文件被分成 K 个子流,则第 K 个子流上存储的数据块为 $nK+i$,其中,n 是非负整数,i 是 $1\sim K$ 的正整数。

参考文献[3]指出,从资源调度和流媒体传输实时性角度考虑,媒体数据被划分的数据块数目越多越好,即数据块体积越小越好;而从网络开销角度考虑,媒体数据被划分的数据块数目越少越好,即数据块体积越大越好。因此,如何权衡这两方面的关系是一个很有价值的研究课题。

2.2 数据缓存及更新策略

缓存是指用户观看视频时把当前媒体数据暂时保存在系统内存或者外存中,它是一种被动的存储方式,存储内容由当前观看的视频内容决定。在 P2P 直播系统中,用户的观看过程基本同步,上游节点中的缓存内容可以很好地满足下游节点的要求。但在点播系统中,用户请求数据具有异步性,如何对分布于多个节点的媒体数据进行缓存和更新则需要更加复杂的策略。

通常的缓存策略是对正在下载播放的数据按时间顺序进行缓存,如果缓存空间已满,则采用 LRU(Least Recently Used Algorithms)或其他简单的缓存替换算法进行替换。该方法没有考虑缓存内容的流行度及其他节点的缓存情况,这样容易造成节点保存了较多流行度不高且在系统中已有足够副本的数据,而替换出了流行度高且缓存的副本数量不足的媒体数据。

参考文献[7]指出,一个缓存替换算法既要考虑到媒体数据的流行程度,也应关注到该流媒体块在其他节点中的缓存情况,其定义了流媒体块的使用价值 R=流媒体数据块的流行度 (F)/系统已缓存该流媒体块的副本数量(CN),并提出了相应的缓存替换算法。该替换算法的本质是替换出使用价值最小者,缓存使用价值最大者。

2.3 支持交互式的存储方法

为了支持视频点播系统中的 VCR (Video Cassette Recorder)操作,应采取相应的存储机制。数据预取机制可以为 VCR 操作备好所需内容,从而更加充分地利用节点的上行带宽,有效地减少交互操作时延。

在 VMesh 中,媒体数据被分成数块并以分布式的方式保存在多个节点中,这些节点通过结构化覆盖网络方式组织起来以支持 VCR 操作。该方法的性能取决于事先存储数据的受欢迎程度,因此 VMesh 是采用基于流行度的段存储方案。VMesh 假设数据块的流行度符合 Zipf 分布,把视频文件划分为 N 个数据块,每个数据块的播放时延为 Δt,第 i 个数据块的流行度 $p_i=\dfrac{1/i^{\alpha}}{\sum_{n=1}^{N}1/n^{\alpha}}$,其中 α 为一个常量。用户观看视频时从数据块 i 跳到数据块 j 的概率 $\sum_{j=1}^{N}q_{ij}=1-p_{seq}$,其中 p_{seq} 为用户观看完数据块紧接着播放数据块 $i+1$ 的概率。

3 资源定位机制

资源定位的结果是得到一个资源拥有者的列表,然后请求节点从该列表中选出期望能够提供良好服务的节点并与之直接连接。在 P2P 直播系统中,由于各节点的播放时间基本同步,节点间的数据传输关系相对稳定,因此,伙伴节点选择比较容易,通常采用基于 Gossip 协议的方式进行。基于 Gossip 协议的内容发现与定位方法不需要维护节点间的逻辑拓扑,但当节点内容更新较快时,通告消息发送频率低将导致内容定位准确性下降;而通告消息发送频率高时可能产生较大的控制流量。因此,找到一种较好的资源定位方案是重要的研究内容。类似地,结构化覆盖网络方法,如 DHT(Distributed Hash Table)机制,可以实现内容的快速查找,但系统动态性较强时结构难以维护。因此,尽管 GridMedia 系统[8]采用了无结构化网络结构,但并没有采用基于 Gossip 协议的节点发现策略,而是引入了 1 个集中点服务器 RP(Rendezvous Point)来维护覆盖网中所有节点的信息,它把合适的候选伙伴节点集合返回给需要资源定位的节点。

混合式 P2P 系统结构指的是在系统选择一些节点充当系统局部的中心,中心节点的邻居节点需要向中心节点报告其数据存储状态。中心节点相对稳定但并非一直不变,中心节点之间也需要进行一定的数据交换,从而使每个中心节点都可以获得全局的数据状态,尽管这个状态可能不完全准确。混合式系统结构在完全分布式的系统中引入了一定的结构化,有利于媒体内容的快速检索,同时又避免了维护固定网络拓扑的过重负担。现有的很多系统为了满足不同的信息检索需要采用混合式内容发现策略,其代表为 PPLive 系统[3]。

4 内容分发

P2P 网络中的绝大多数节点都是对等的,在某些网

络中会设置少量超级节点负责管理局部网络的事务。每个节点都可能对网络中的某些内容有兴趣，或者其所拥有的内容是其他节点感兴趣的。内容分发算法的目标是建立起从源到目标接收点满足播放质量的分发路径，下面分别从数据源的数量和数据交换技术二方面对内容分发进行介绍。

4.1 单源和多源分发策略

由于网络中资源的存放方式不同，分发策略可以分为单源的和多源的策略[9]。它们的数据传输方式分别如图 3、图 4 所示。

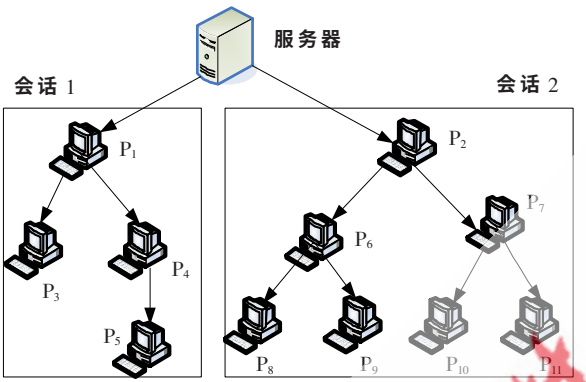

图 3 单源的 P2P 流媒体传输

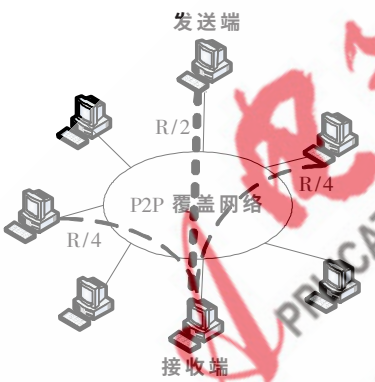

图 4 多源的 P2P 流媒体传输

单源分发策略常采用的网络拓扑结构为树型结构，基于应用层组播技术，由 1 个发送者向多个接收者发送数据，接收者有且仅有 1 个数据源。服务器和所有客户节点组织成组播树，组播树的中间节点接收来自父节点组播的媒体数据，同时将数据以组播的方式传送给其子节点。如图 3，P_2、P_6、P_7、P_8、P_9、P_{10}、P_{11} 请求同一媒体内容，服务器将其组织成 1 棵组播树，P_2 直接从服务器获取数据，同时又将数据传送给它的 2 个子节点 P_6 和 P_7。以此类推，P_6 又把数据传送给自己的子节点 P_8 和 P_9，同样地，P_7 又把数据传送给自己的子节点 P_{10} 和 P_{11}。以组播的方式传输流媒体，避免了单播 C/S 服务模式下为每个接收者单独发送信息的缺点，同时减轻了服务器的负载，节约了网络资源。但其缺点是，实际部署困难，并且节点的要求较高，如至少能发送 1 个完整的流媒体流，即上行带宽要足够大。此种分发方式的典型代表系统有 Promise、CoopNet、SplitStream 等。

在实际的网络环境中，各个节点之间在提供的带宽、存储空间以及 CPU 能力等方面存在着很大的差异。在当前的接入方式中，用户的上行带宽通常小于下行带宽，为了满足媒体数据播放的时间约束，通常采用如图 2 所示的多源发送的数据传输方式，以保证提供服务的所有节点出口带宽之和大于媒体流的编码速率 R。但该类型的数据传输方式所带来的问题在于怎样选择合适的发送节点、怎样协调多个发送节点之间的传输速率、如何分配各个发送节点的数据段等。

4.2 数据交换技术

根据媒体流传输的驱动端不同，内容分发方式又可以分为：接收端驱动，即“拉数据”；发送端驱动，即“推数据”。所谓“拉数据”，就是节点首先向另一个节点发出请求，另一节点再根据请求发送数据，这不需要节点之间任何层次性的关系，但是节点需要预先知道对方含有的数据；而所谓“推数据”，就是节点主动向另一个节点发送数据，这就需要节点之间有一种父与子的关系，父节点依据这种关系主动发送数据给子节点。

CoolStreaming 早期版本中流媒体数据的传输是基于接收节点的主动请求，即“拉数据”流传输策略。其缺点是将导致每个数据块传输都有一定的延迟，并且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，系统的控制开销增大。为了解决这些问题，Zhang 等人设计了一个“推拉结合”的 GridMedia[10] 系统，将 P2P 流媒体系统中的数据块分成二类：一类数据块只在被请求获取时才传播，称为“拉数据”；另一类数据块一旦节点收到就立即传播给邻居，称为“推数据”。GridMedia 系统主要的设计目标是减少数据传播时延，它间接地提高了播放连续度。但“推数据”的方法必然带来相当大的通信开销，而且也不能从本质上保证高播放连续度。

5 总结与展望

流媒体是今后互联网上主要应用之一，但传统的 C/S 服务模式存在可扩展性问题，使得流媒体技术无法实现大规模的应用。P2P 作为一种新型的网络模型，为流媒体的大规模应用提供了新的解决方案，基于 P2P 的流媒体服务系统已经引起了许多研究机构以及商业组织的重视。

本文主要对 P2P 流媒体系统中已有的资源存储、资源发现、内容分发等关键技术进行了详细的分析介绍，针对上述研究内容，目前 P2P 流媒体技术需要解决的主要问题有：

(1)如何将整个的视频进行更加合理地分块，怎样根据用户的行为特征进行数据的存储，如何进行数据的

更新以达到系统的负载均衡。

(2)在高度动态的网络环境中,如何设计出高效的资源发现算法对P2P流媒体系统来说仍是十分重要的内容;在VOD系统中,如何设计出支持用户频繁的VCR操作的资源定位算法,是评价系统优劣的一个重要指标。

(3)随着无线网络和各种各样手持设备的出现,无线流媒体的应用也变得越来越重要,尤其是3G解决了接入网的传输瓶颈。因此,在无线网络环境下进行P2P流媒体的研究是一个重要的研究方向。

(4)仅仅依靠上述研究内容还不足以在现实的网络环境中提供大规模的流媒体服务,还需要对目前的QoS保证机制、激励机制、容错机制、可靠性传输、安全机制和版权问题等作进一步深入研究。

参考文献

[1] BHARAMBE A R, HERLEY C, PADMANABHAN V N. Analyzing and improving a bitTorrent network's performance mMechanisms [C]. In: Proc. of IEEE INFOCOM'06, 2006.

[2] XIE Su Su, LI Bo, KEUNG G Y, et al. Coolstreaming: design, theory, and practice [C]. In: Proc. of IEEE Transactions on Multimedia, 2007(9):1661-1671.

[3] HUANG Yan, FU T Z J, CHIU D M, et al. Challenges, design and analysis of a large-scale P2P-VOD system[C]. In: Proc. of SIGCOMM'08, August, 2008.

[4] PPStream(PPS网络电视).http://www.ppstream.com, 2008-06.

[5] TRAN D A, HUA K A, Do T T. A peer-to-peer architecture for media streaming[J]. IEEE Journal on Selected Areas in Communications, 2004,22:1-14.

[6] ZHANG Xin Yan, LIU Jiang Chuan, LI Bo, et al. Coolstreaming/DONet: a data-driven overlay network for peer-to-peer live media streaming [C]. In: Proc. of IEEE Infocom,April, 2005.

[7] 杨传栋,余镇危,王行刚.混合P2P流媒体的缓存替换算法研究[J].计算机应用研究,2006(11):71-73.

[8] LI Zhao, LUO Jian Guang, ZHANG Meng, et al. Gridmedia: a practical peer-to-peer based live video streaming system [C]. In: Proc. of IEEE 7th Workshop on Multimedia Signal Processing, 2005:287-290.

[9] 郑常熠,王新,赵进,等.P2P视频点播内容分发策略[J].软件学报,2007,18(11):2942-2954.

[10] 罗建光,张萌,赵黎,等.基于P2P网络的大规模视频直播系统[J].软件学报,2007,18(2):391-399.

(收稿日期:2009-05-14)

APPLICATION OF ELECTRONIC TECHNIQUE www.chinaAET.com