应用层组播研究综述

章 淼[1],徐明伟[2],吴建平[2]
(1. 清华大学信息网络工程研究中心,北京 100084;2. 清华大学计算机科学与技术系,北京 100084)

摘 要: 组播是互联网研究的一个重要课题.最近的研究发现 IP 组播方案存在一些很难解决的问题.基于互联网的性质和应用的特点,在 IP 组播模型、Overlay Network 和 Peer-to-Peer 等技术的基础上,发展出了应用层组播技术.本文总结了目前应用层组播领域的主要算法,重点分析了其中的主要研究问题,概括了应用层组播算法研究中主要使用的评价方法,并对应用层组播的相关研究问题进行了讨论,并对未来的研究作了展望.

关键词: 组播; 应用层; Overlay; 互联网
中图分类号: TP393 **文献标识码**: A **文章编号**: 0372-2112 (2004) 12A-022-04

Survey on Application Layer Multicast

ZHANG Miao[1], XU Ming-wei[2], WU Jian-ping[2]
(1. *Network Engineering Research Center, Tsinghua University, Beijing* 100084, *China*;
2. *Dept. of Computer Science & Technology, Tsinghua University, Beijing* 100084, *China*)

Abstract: Multicast is an important research topic for the Internet. Recent research shows some intrinsic limitation in IP Multicast, and Application Layer Multicast (ALM) is proposed. In this article we discuss and classify some major algorithms proposed for ALM. The fundamental problems in ALM research are identified and the metrics for evaluation of ALM algorithms are summarized. Some topics (e.g., Overlay Network, media encoding method) are also discussed for their close relationship with ALM research. Finally, some possible directions for future research are discussed.

Key words: multicast; application layer; overlay; Internet

1 引言

组播是互联网研究的重要课题. IP 组播是对互联网的"单播、尽力转发"模型的重要扩充,组播的主要功能在路由器上实现,通过合并重复信息传输来减少带宽浪费和降低服务器的负担.由于 IP 组播在传输技术和管理上存在严重问题,目前没有在互联网中普遍采用.

最近出现了"应用层组播"(ALM: Application Layer Multicast)技术.它保持了互联网的"单播、尽力转发"模型,主要通过端系统来实现组播功能.应用层组播的系统框架和很多技术还在研究当中.媒体编码技术、Peer-to-Peer 和 Overlay Network 等技术的发展对应用层组播也有很大的促进.

2 IP 组播的回顾

IP 组播的主要思想是在互联网单播的框架上进行扩展,功能主要通过路由器来实现.组播适用于那些在时间上具有集中性、而在空间上具有分布性的应用. IP 组播适用于实时、不可靠的应用[1].

IP 组播存在以下问题[1~4]:(1)路由器必须为每个组播组保存状态,扩展性差;(2)要求所有路由器都支持,不利于推广使用;(3)用统一的模型来适应所有应用,算法设计困难;(4)组播组加入、退出和管理等开销大;(5)组播地址空间太小(针对 IPv4);(6)打破了传统的根据进入流量计费的机制. IP 组播在安全、拥塞控制等方面也存在问题.

3 应用层组播研究的概况

应用层组播的基本思想是保持互联网原有的简单、不可靠、单播的转发模型,由端系统实现组播转发功能.这也是"end-to-end argument"[5]所倡导的思想.它有一些假设:(1)网络的带宽和转发资源相对丰富,服务器能力是主要瓶颈;(2)组播组成员可贡献资源用于转发;(3)应用对性能要求并不苛刻,可容忍报文丢失和较大延迟.

应用层组播的优势有:只需改变端系统,便于实现和推广;便于针对特定应用优化.它的缺点为:一般比 IP 组播使用更多网络资源;端系统可能不稳定,导致组播的可靠性受影响;端系统性能无法保证,可能导致延迟、转发速率等性能的下降[4].

目前应用层组播研究集中于视频会议系统、媒体流的分发系统(如视频广播)和订阅/分发系统(Publish/Subscribe System)等.它主要用于实时的多媒体传输.这利用了多媒体信息的性质,即在传输链路质量下降时,用户仍可利用收到的低速率或者不完整的信息;也利用了组播"时间上集中、空间上分布"的特点.

4 应用层组播中的主要算法

用单一方法对应用层组播的算法分类很困难,这里从应用特点和算法设计方法这两个角度分类介绍.

收稿日期:2004-10-18;修回日期:2004-11-18
基金项目:国家自然科学基金(No. 90104002, 60373010, 60303006);973 项目(No. 2003CB324801)

4.1 小规模的多源组播方案

代表是 End System Multicast[4,6]和 ALMI[2],针对小规模、多数据源的情况,典型应用是视频会议系统. End System Multicast 的方案是:首先将组播组的成员组织成一个"网"(mesh),每个成员都维护所有组成员的列表,提高了组播组的可靠性;在 mesh 上以每个数据源为根各构造一个生成树(spanning tree),这样可针对每个数据源进行性能优化. 其缺点是系统开销比较大,降低了系统的可扩展性,适合小规模组播组的情况. ALMI 在组播成员之间维护一个"最小生成树"(MST: Minimum Spanning Tree),减小了维护开销,但从每个源出发传输开销无法单独优化. 生成树的维护开销限制了组播组的规模.

4.2 大规模的单源组播方案

代表是 NICE[7]和 Zigzag[8],它们解决在只有一个数据源时构造大规模组播树的问题,都使用了"分层"(Hierarchical)和"分群"(Cluster)的思路. 大部分组成员位于分层结构的底层,只和少量固定数目的节点存在联系,这样就大大降低了大部分组播成员的处理开销. 它们的不同是:NICE 把 Cluster 管理和数据分发这两个功能放在一个节点上;而在 ZigZag 中是由不同节点完成的,目的是提高系统的可靠性.

4.3 基于多树的方案

代表是 CoopNet[9]. 它假设网络带宽比较充足,而组播节点的稳定性不能保证. 主要思路是:在组成员之间同时维护多个组播树;利用 MDC 算法[10]把媒体编码成 N 个媒体流,沿多个组播树传播;组成员收到 N 个媒体流中的任何 M(M < N)个就可完成解码. 算法的问题是:维护多个组播树开销较大;在"多路同时传送"机制下,数据的同步是一个难点;MDC 编码和单路编码相比效率较低,对网络带宽的浪费较大.

4.4 应用层网关的方案

代表是 Scattercast[11]和 Overcast[12]. 主要思路是:在网络中部署一些专用服务器,在应用层构造一个实现组播转发功能的特殊网络. 这和完全基于用户主机的方案相比具有更高的稳定性;但使用专用的服务器降低了灵活性.

4.5 应用层组播和 IP 组播结合的方案

代表是 Yoid[1]和 Host Multicast[13]. 主要思想是在局部、小规模、支持 IP 组播的网络中使用 IP 组播,而在 IP 组播构成的"小岛"或没有 IP 组播支持的主机之间使用应用层方式连接. 它是一种混合方案,不受网络条件的限制,而且可以充分利用 IP 组播的优点.

4.6 基于特定逻辑结构的方案

代表是 Bayeux[14]和 CAN (Content-Addressable Network)[15]. 它们使用特殊的逻辑结构对组播节点映射或编址;组播转发可使用简单的规则实现,从而减少状态维护开销和转发开销,避免路由协议的使用.

Bayeux 基于 Tapestry[16]. 每个节点拥有全局唯一的 ID,并维护一个邻居表,这些邻居节点的 ID 和本节点的 ID 在一定数量的位上相同. 转发中第 n 跳节点 ID 和目的节点 ID 至少有 n 位相同. Bayeux 在 Tapestry 的基础上将组播树的状态信息保存在"中间节点"上. 其主要问题是这种维护状态的方式会限制算法的可扩展性.

CAN 组播是对 CAN[17]的扩展. CAN 将一个 d 维坐标空间划分成若干部分,每个节点拥有其中某部分. 两个直接相邻部分的坐标在 d-1 维上相同,而在另一维上不同. 转发报文时把报文发给邻居中和目标坐标最接近的节点. CAN 组播将组播组构造为 CAN,使用"洪泛"方法在 CAN 内转发报文. 这样可减少节点上维护的状态信息,提高数据传输的可靠性,但也会产生大量重复报文. 文献[15]试图用启发式方法减少重复报文,但无法完全避免.

这类方案还有 Delaunay Triangulation[18]等. 它们共同的问题是,逻辑空间中节点间的关系并不能对应实际网络中的关系,得到的报文转发路径很有可能在性能方面存在问题.

4.7 利用拓扑信息的方案

代表是 TAG[19]. 它把组播节点之间的拓扑信息用于组播树构造,减少报文的转发延迟和在同一链路上重复传递的报文数. 拓扑测量时它将延迟作为最重要的指标,也考虑带宽. 构造组播树时它使新加入的节点和父节点能够共用尽可能长的网络路径. TAG 使用拓扑信息获得性能提高,但它破坏了网络的分层结构;拓扑测量和网络性能测量还需要研究.

5 和应用层组播有关的其他技术

5.1 媒体编码传送技术

媒体编码传送技术对组播算法影响很大. 著名的 RLM[20]算法就是基于层次编码算法设计的. 目前媒体编码传送方案主要有[21]:

(1) 信息重复(Information Replication)的单速率方案:发送端对相同的媒体内容生成重复的、不同速率的数据流,根据接收端的情况决定使用某个数据流.

(2) 累积的分层(Cumulative Layering)传输方案:媒体被编码成一个必须的"基本层"(base layer)和若干个"增强层"(enhanced layer),"增强层"可根据接收端状况选择增加. 传统方案中"增强层"的粒度较粗,最近的 FGS 编码[22]可提供细粒度的调整能力.

(3) 非累积的分层(Non-cumulative Layering)传输方案:媒体被编码成若干层具有相同优先级的媒体流. MDC(Multiple Description Coding)编码[10]就属于这种方案.

表面上 FGS 和 MDC 比单速率方案具有优势,但文献[21]指出,FGS 和 MDC 的计算开销较大,且在相同质量下比单速率方案多消耗 20～40 %的带宽. 应用层组播算法和媒体编码方案如何配合很值得研究. 另外"累积的分层"方案只有完整收到某层信息才能获得相应的质量提高,这需要减少高优先级数据的丢失,如何在网络中满足这个需求也值得研究.

5.2 Peer-to-Peer, Overlay Network, Web Cache 和 CDN

Peer-to-Peer 是和 client-server 模型相对提出的概念. 其出发点是利用逐步提高的主机性能和网络带宽,通过增加主机之间的自组织,提高应用的灵活性,降低服务器的负担,解决 client-server 模型中的"单点故障"问题. 应用层组播算法的设计中引入了 Peer-to-Peer 的思想,但应用层组播的研究并不是 Peer-to-Peer 研究的一个子集.

Overlay Network 指出了网络研究的一个方向,它在互联网

单播路由转发的网络架构基础上,在服务器或主机之间构造一个应用层网络来实现某种应用.应用层组播的一些算法属于 Overlay Network 的范围.

Web Cache 和 CDN(content distribution network)是对 client-server 模型的扩展,目的是分散单个服务器的处理负担,减少网络中重复信息的传播.CDN 和应用层组播各有侧重.应用层组播偏重于实时的、多媒体信息的传输,而 CDN 偏重于非实时的、文件形式的传输.未来网络需要多种"上层"机制来支持不同的应用.

6 应用层组播中的关键技术

6.1 组播节点的组织方法

组播节点的组织方法决定了节点之间的关系,目前主要为"树"(Tree)、"网"(Mesh)和特定的逻辑结构.Tree 实现简单,维护开销小,扩展性好,但可靠性较差.Mesh 可靠性较高,但维护开销较大,扩展性较差.一般大组播组中使用 Tree,在中小组播组中使用 Mesh.如何结合 Tree 和 Mesh 的优点值得研究.特定的逻辑结构扩展性好,且不需要路由算法,但问题是建立的逻辑结构一般不能很好利用网络性能,这在目前的分析中很少涉及.

6.2 组播节点的维护方法

组播节点的维护包括节点的加入、退出和"失效"节点的检测.节点的加入指新的节点发现组播组的存在、加入到组播组中.目前大部分算法都假设存在"集中点"(RP:Rendezvous Point),通过 RP 完成加入,RP 很容易成为系统的瓶颈.节点退出时需要发出退出组播组的通知,有些算法要对节点的组织进行调整."失效"指节点没有发出退出组播组的通知但已无法正常工作.一般通过定期发送 keepalive 报文实现"失效"节点的检测.

6.3 安全

IP 组播中安全就十分重要,应用层组播中数据通过可信度不高的主机转发,安全更为重要.应用层组播中安全包括[1]:加入组播组控制;对读取组播组内传递的数据控制;避免转发数据被篡改.前两个问题可通过在 RP 增加机制解决.解决第三个问题比较困难,文献[1]建议在使用 Mesh 结构时,每个节点可把收到的数据做数字签名后传递给邻居节点,如果有节点恶意篡改,可以从其他节点发出的数字签名发现.

6.4 网络性能测量技术

网络性能测量技术对于应用层组播的实用十分重要,主要用于:(1)在 Mesh 中选择数据转发路径需考虑性能;(2)新节点加入时要选择性能较好的邻居节点(或父节点);(3)组播组结构调整时需测量性能.主要性能指标包括带宽、延迟和丢失率.大多数算法以带宽或延迟作为主要指标.网络性能测量目前仍是正在研究中的问题.

6.5 流量控制和拥塞控制

流量控制和拥塞控制是 IP 组播研究中的重要问题,难点是多个接收端速率不同,而且还要避免在单播中不存在的"反馈爆炸".IP 组播中的算法包括 RLM[20]、RLC[23]等,主要思想包括:使用接收端驱动方式;利用媒体分层编码传输的算法;根据报文丢失判断速率超过网络传输能力,控制分层媒体流的增加和减少.TCP Friendly 算法[24,25]出现后,出现一些相关的组播算法,如 MLDA[26]、PGMCC[27]等.

应用层在这方面还研究不多,可能思路包括:(1)借鉴 IP 组播的研究经验和思路;(2)针对和 IP 组播的不同提出新的思路,例如中间节点的功能可以更加复杂;(3)考虑媒体编码技术对于应用层组播算法的影响;(4)考虑和 TCP 流量的公平性问题.

7 应用层组播算法的评价方法

应用层组播算法的特点是它的设计针对某种应用进行优化,对其评价时没有统一的标准.常用的评价标准[3,4,6,13]包括:

(1)带宽的使用:应用层组播会比 IP 组播消耗更多的带宽.End System Multicast 消耗的网络带宽大约是 IP 组播的 2 倍[4].

(2)延迟:可以用数据传播的时间,大规模组播树时也可用树的深度来反映延迟.End System Multicast 的延迟大约是 IP 组播算法的 2.2～2.8 倍.

(3)可扩展性(Scalability):只在大规模组播算法时才需要考虑.

(4)鲁棒性(Robustness):应用层组播应提供一定的鲁棒性,并适应不同应用的要求.文献[3]提出建议:使用"洪泛"机制;在系统中增加冗余;提供快速错误发现和修复机制.

(5)易推广性(Ease of deployment):是应用层组播提出的一个推动力.不仅要不改变现有的网络体系结构,还要解决 NAT 和防火墙带来的问题.

(6)组播组维护的开销:包括保存状态信息的数量;组播节点之间为维护组播组的交互信息数量;组播节点加入、退出、失效等信息在组播组内同步的时间开销等.

(7)组播节点的"度"(degree):主机的资源有限,所以需要限制节点的"度".

8 总结和展望

本文对应用层组播研究的现状进行了总结,指出在目前研究中存在的很多问题.应用层组播有它的应用范围,它的"使用不同机制来支持不同应用"的思想为互联网下一步的发展指出了一个方向.

应用层组播的研究目前还相当不成熟.研究还集中在系统的框架设计和组播节点的组织上,对于一些深入的问题还涉及不多,如安全、拥塞控制、和媒体编码方式的结合等.在系统总体框架和组播节点组织上也还需进一步研究.应用层组播只是互联网应用模型研究的一部分,如何提供更多的应用模式为社会造福是下一阶段应重点研究的问题.

参考文献:

[1] P FRANCIS Yoid:extending the multicast internet architecture[EB/OL]. http://www.aciri.org/yoid,1999.

[2] PENDAKARIS D, SHI S. ALMI:an application level multicast infrastructure[A]. Anderson T. The 3rd USENIX Symposium on Internet Technologies and Systems[C]. San Francisco, CA, USA:USENIX As-

sociation ,2001. 49 - 60.

[3] EL-SAYED A ,ROCA V ,MATHY L. A survey of proposals for an alternative group communication service[J]. IEEE Network ,2003 ,17(1) : 46 - 51.

[4] CHU Y H ,RAO S G,SESHAN S ,ZHANG H. A case for end system multicast[J]. ACM SIGMETRICS Performance Evaluation Review , 2000 ,28(1) :1 - 12.

[5] SALTZER J ,REED D ,CLARK D. End-to-end arguments in system design[J]. ACM Transactions on Computer Systems ,1984 ,2(4) :195 - 206.

[6] CHU Y H ,RAO S G,SESHAN S ,ZHANG H. Enabling conferencing applications on the internet using an overlay multicast architecture[J]. ACM SIGCOMM Computer Communication Review ,2001 ,31(4) :55 - 67.

[7] BANERJEE S ,BHATTACHARJEE B ,KOMMAREDDY C. Scalable application layer multicast[J]. ACM SIGCOMM Computer Communication Review ,2002 ,32(4) :205 - 217.

[8] TRAN D A ,HUA K A ,DO T T. Zigzag: an efficient peer-to-peer scheme for media streaming[A]. BAUER F ,PUIGJANER R. IEEE INFOCOM 2003[C]. San Francisco ,CA ,USA: IEEE Press ,2003. 1283 - 1292.

[9] PADMANABHAN V N ,WANG H J ,CHOU P A ,SRIPANIDKULCHAI K. Distributing streaming media content using cooperative networking [A]. The 12th International Workshop on Network and Operating Systems Support For Digital Audio And Video[C]. Miami ,Florida ,USA: ACM Press ,2002. 177 - 186.

[10] GOYAL V K. Multiple description coding: compression meets the network[J]. IEEE Singal Processing Mag ,2001 ,18(5) :74 - 93.

[11] CHAWATHE Y. Scattercast: an architecture for internet broadcast distribution as an infrastructure service[D]. USA: University of California ,Berkeley ,2000.

[12] JANNOTTI J ,GIFFORD D K,JOHNSON K L. Overcast: reliable multicasting with an overlay network[A]. JONES M B ,KAASHOEK F. USENIX Symposium on Operating System Design and Implementation [C]. San Diego ,CA ,USA: USENIX Association ,2000. 197-212.

[13] ZHANG Bei-chuan ,JAMIN S ,ZHANG Li-xia. Host multicast: a framework for delivering multicast to end users[A]. KERMANI P. IEEE INFOCOM 2002[C]. New York ,NY ,USA: IEEE Press ,2002. 1366 - 1375.

[14] ZHUANG S Q ,ZHAO B Y,JOSEPH A D. Bayeux: an architecture for scalable and fault-tolerant wide-area data dissemination[A]. NIEH J , SCHULZRINNE H. The Eleventh International Workshop on Network and Operating System Support for Digital Audio and Video[C]. New York ,USA: ACM Press ,2001. 11 - 20.

[15] RATNASAMY S ,HANDLEY M ,KARP R ,SHENKER S. Application-level multicast using content-addressable networks[A]. CROWCROFT J ,HOFMANN M. Networked Group Communication ,Third International COST264 Workshop ,NGC 2001[C]. London ,UK: Springer ,2001. 14 - 29.

[16] ZHAO B Y, KUBIATOWICZ J D ,JOSEPH A D. Tapestry: an infrastructure for fault-tolerant wide-are location and routing[R]. Berkeley ,USA: University of California ,Computer Science Division ,2001.

[17] RATNASAMY S ,FRANCIS P ,HANDLEY M ,KARP R ,SHENKER S. A scalable content-addressable network[J]. ACM SIGCOMM Computer Communication Review ,2001 ,31(4) :161 - 172.

[18] LIEBEHERR J ,NAHAS M ,SI W. Application-layer multicasting with delaunaytriangulation overlays[J]. IEEE Journal on Selected Areas in Communications ,2002 ,20(8) :1472 - 1488.

[19] KWON M ,FAHMY S. Topology-aware overlay networks for group communication[A]. ALMEROTH K ,GRIFFIOEN J. The 12th International Workshop on Network and Operating Systems Support For Digital Audio And Video[C]. Miami ,Florida ,USA: ACM Press ,2002. 127 - 136.

[20] MCCANNE S ,JACOBSON V ,VETTERLI M. Receiver-driven Layered Multicast[J]. ACM SIGCOMM Computer Communication Review , 1996 ,26(4) :117 - 130.

[21] LI B ,LIU J C. Multirate video multicast over the internet: an overview [J]. IEEE Network ,2003 ,17(1) :24 - 29.

[22] LI Wei-ping. Overview of fine granularity scalability in mpeg-4 video standard[J]. IEEE Transactions on Circuits and Systems for Video Technology ,2001 ,11(3) :301 - 317.

[23] VICISANO L ,RIZZO L ,CROWCROFT J. TCP-like congestion control for layered multicast data transfer[A]. GUERIN R. IEEE INFOCOM 2003[C]. San Francisco ,CA ,USA: IEEE Press ,1998. 996 - 1003.

[24] FLOYD S ,HANDLEY M ,PADHYE J ,WIDMER J. Equation-based congestion control for unicast applications[J]. ACM Computer Communication Review ,2000 ,30(4) :43 - 56.

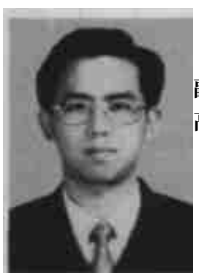
[25] WIDMER J ,DENDA R ,MAUVE M. A survey on tcp-friendly congestion control[J]. IEEE Network ,2001 ,15(3) :28 - 37.

[26] SISALEM D ,WOLISZ A MLDA. A tcp-friendly congestion control framework for heterogeneous multicast environments[A]. STEENKISTE P , ZHANG H. The 8th Intl. Workshop on Quality of Service (IWQoS 2000) [C]. Pittsburgh ,USA: IEEE Press ,2000. 65 - 74.

[27] Rizzo L pgmcc. A tcp-friendly single-rate multicast congestion control scheme[J]. ACM SIGCOMM Computer Communication Review ,2000 , 30(4) :17 - 28.

作者简介:

章　淼　男,1976 年出生于北京,博士,助理研究员,研究领域包括计算机网络体系结构、计算机网络性能测试、计算机网络管理等.



徐明伟　男,1971 年出生于辽宁朝阳,博士,副教授,主要研究领域为计算机网络体系结构、高速路由器体系结构、计算机网络协议测试等.