一种BT系统中节点通信HAVE协议的优化扩展

张 尧,李建春,黄道颖,李健勇
(郑州轻工业学院计算机与通信工程学院,河南郑州 450002)

摘 要: BT系统中节点之间通信共有11种消息,其中HAVE、REQUEST和PIECE消息是数据传输中流量的主要构成部分.为了提高网络传输效率,降低文件传输中的管理开销,对HAVE消息进行了研究,提出了对HAVE消息的改进方案——MultiHave消息.实验表明,MultiHave消息有效提高了系统的性能.

关键词: BT系统; 对等网; HAVE消息; MultiHave消息
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2009) 08-1816-04

An Optimization Extension of HAVE Message Among Peers on BT System Protocol

ZHANG Yao, LI Jian-chun, HUANG Dao-ying, LI Jian-yong
(*Institute of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, Henan* 450002, *China*)

Abstract: There are 11 kinds of communication messages among nodes in BT system. Flow sent by HAVE, REQUEST and PIECE messages is a main part of all. In order to improve the network transfer efficiency and reduce the management expenses in file's transferring, the research has focused on HAVE message, and proposed a improvement scheme to HAVE message—the Multi-Have message. It shows by the experiment that MultiHave message has improved systematic performance effectively.

Key words: BitTorrent; Peer-to-Peer; HAVE message; MultiHave message

1 引言

BitTorrent(简称BT)系统是一种P2P资源发布系统[1].BT系统中节点在利用下载的同时也为其他节点提供上传服务,每个节点既是客户端同时也是服务器,同时下载的节点越多,数据下载的速度就越快.BT技术和流媒体技术的结合在Internet上获得了广泛的应用.

研究表明,P2P网络的流量在Internet主干网上所占的比例已经从以前的10%上升到80%[2~4].根据Cachelogic公司提供的资料,2004年上半年,BT的流量在整个P2P流量中所占的比例由26%上升到52%,到2005年则达到了60%[2].一方面,大量的BT系统的使用造成局部网络经常被拥塞,致使许多ISP开始对BT系统的使用进行限制;另一方面,随着用户数的激增,以中心服务器的文件分发服务不堪重负,又急需BT系统的支持.有关BT系统的研究已成为P2P系统研究的一个热点.

BT系统产生的流量可分为两种:数据流量和管理流量.在一些应用中,管理流量可能达到系统流量的23%[5],严重地影响了系统的性能.BT系统中管理流量主要是由HAVE消息、REQUEST消息和BITFIELD消息产生的.通过对BT协议的深入分析,提出了一种HAVE消息改进方案——MultiHave消息.MultiHave消息既能完成了HAVE消息的功能,又能明显减少原来使用HAVE消息时的管理流量,提高了BT系统的性能.基于原BT系统协议,本文设计实现了MulitiHave消息;对分别采用HAVE消息和MultiHave消息的两种BT系统中的实验结果进行对比分析;最后进行总结并得出结论.

2 MultiHave消息

在BT系统中,节点之间通信共有11种消息,它们分别是:HANDSHAKE, KEEP ALIVE, CHOKE, UNCHOKE, INTERESTED, NOT INTERESTED, HAVE, BITFIELD, RE-RUEST, PIECE和CANCEL消息.其中,管理流量主要由HAVE、REQUET和BITFIELD消息产生.在HAVE、RE-QUST和BITFIELD消息中,REQUEST是节点向外请求数据的消息,它的数量与块和每块中子块的数量有关,若要减少REQUEST消息则需要调整块的大小和块中子块的大小.块的大小在制作torrent文件时已经确定,子块的大小是为了提高传输效率,使TCP的连接变得充分

收稿日期:2008-05-23;修回日期:2009-05-31
基金项目:国家科技支撑计划(No.2006BAK01A38);河南省杰出青年科学基金(No.0612000600)

饱和而设计的,因此,无法对 REQUEST 消息进行改进. BITFILED 消息的作用是节点间握手之后进行数据块信息的同步,在节点间连接建立后只需互发一次,也不能进行精简.HAVE 消息在节点建立连接后随着接收数据块的变化会不断产生并在节点间大量产生,减少 HAVE 消息的研究具有现实意义.

2.1 HAVE 消息的局限性

节点在下载并验证通过一个块后,会向所有与它建立连接的节点发送 HAVE 消息,以便其他节点同步更新连接建立以来有关本节点块位图信息的变化.在有许多个节点参与的 BT 系统中,随着追踪器(tracker)返回的节点列表数的增加,系统中 HAVE 消息的数量会呈迅速增加,从而在 BT 系统中形成"HAVE 消息风暴".网络带宽越大,下载数据块越快,HAVE 消息的交换越频繁,"HAVE 消息风暴"越强.实际上,频繁地告知其他节点本节点的某块数据已经下载完成,并不能加速其他节点下载的速度,反倒会因过于频繁的 HAVE 消息交换而影响系统的性能.因此,设法减少 HAVE 消息发出的频度既能降低节点发出 HAVE 消息的负担,又能减少 BT 系统的管理开销.

2.2 MultiHave 消息的结构

除 HANDSHAKE 和 KEEP ALIVE 消息外,其余 9 种消息都有 4 个字节的消息前缀和 1 个字节的消息标号.据此,给出 MultiHave 消息格式如表 1 所示.

表 1　MultiHave 消息格式

Message Name	Length prefix/4B	Message ID/1B	Payload
HAVE	0005	4	Integer/4B
MultiHave	Payload + 1	9	Variable length

表 1 中对比了 HAVE 消息与 MultiHave 消息的格式,其中 MultiHave 消息的由消息前缀、消息标志和负载 3 部分组成.

(1)消息前缀.长度为 4 字节,表示 MultiHave 消息中消息标号和负载的字节数,其取值应为 $n \times 4 + 1$,n 为负载中块的索引的个数.

(2)消息标号.长度为 1 字节,目前 BT 系统中节点通信消息标号最大为 8,这里将其设 9.

(3)负载.长度为 $n \times 4$ 字节,n 为块的个数,每 4 个字节为一个整数,代表某一个块的索引号.

3 MultiHave 消息的发送策略

HAVE 消息在每次得到一个新的数据块后都采用组播方式向其他节点进行通报.当节点连接数变大时,BT 系统中 HAVE 消息的数量会随节点连接数的增多而迅速增长.特别地,在高带宽的网络环境中,在一次阻塞转换周期(10s)或一次尝试性疏通的周期(30s)内,一个高速节点可能会接收到百兆级的数据.按照数据块大小为 256KB 来计算,接收数据的节点要在 10s 或 30s 内向其向所有连接的节点发送 400 或 1200 个 HAVE 消息,若其连接数为 50,则该节点发送的 HAVE 消息总数为 20000 或 60000 个,平均每秒 2000 个.显然,在 BT 系统的各节点上产生了严重的 HAVE 消息风暴.

假定每个节点都有相似的行为,它们之间的关系对称.在某一时段内,每个节点都均等的发送和接收数据,这样,上载和下载分享了整个带宽,每个节点接收到的数据也相应缩小一半,因此,其发送的 HAVE 消息频率也相应降低为每秒 1000 次.注意到,由于节点行为的对称性,在这个时段内,每节点要接收来自其他 50 个节点的各自的每秒 1000 个的 HAVE 消息,总计可达每秒 50000 个.显然,高密度的消息传递会严重影响系统的性能.

当多个高速节点和少量低速节点并存时,多个高速节点会在某一时段内发送大量的 HAVE 消息.对低速节点来说,HAVE 消息是必须接收并处理的消息,大量 HAVE 消息会占用其通信带宽,拥塞承载数据的 PIECE 消息,极端情况下会导致低速节点长时间内无法下载到任何数据.也就是说,在一个有大量高速节点不断加入的网络中,一个低速节点更易受到 HAVE 风暴的攻击.

在选择 MultiHave 消息的负载量时应慎重考虑 MultiHave 消息的发出频度,避免形成"MultiHave 消息风暴".MultiHave 消息的发出频度可以通过一个定时器来控制.当某节点发出 MultiHave 消息后,启动定时器,定时时间到后,对本节点定时期间接收到的所有新的数据块的生成一个 MultiHave 消息发出,同时开始下一次重新计时.

与 10s 的阻塞算法调度周期和 30s 的尝试性疏通周期相比,当选择一个长的 MultiHave 定时周期(如 30s)时,对于高速节点,可能由于在 10s 周期内未能及时得知节点间的新块的变化而产生某两个高速连接节点间互发送 NOT INTERESTED 消息而相互阻塞;而对于低速节点(如 56Kb 拨号用户),这个时间周期内可能还没有获得一个完整的数据块,这时 MultiHave 消息往往只会包含一个数据块的信息,即退化为 HAVE 消息.

因此,MultiHave 消息定时器间隔选定应遵循如下原则:

(1)定时间隔不能太短.避免产生新的"MultiHave 消息风暴".

(2)定时间隔不能太长.间隔超过阻塞算法的调度周期会产生高速节点间的非正常阻塞.

根据上述两条原则,可以选择一个小于 10s 的定时间隔,如 5s.

4 MultiHave 消息与 HAVE 消息的转换机制

BT 系统中,节点之间在建立连接后、发送第一个 HAVE 消息前会发送一个 BITFILED 消息,该消息会通知对方本节点当前拥有已下载完成的数据块的位视图,BT 节点依据其他节点的数据块位视图用 REQUEST 消息向对方请求某块的数据.在实际 BT 系统中,由于各节点加入的时机不同,新加入节点得到对方的数据块的位视图中往往已经有了多块数据,之后会不断再收到对方发来的 HAVE 消息或 MultiHave 消息.当 BT 系统采用 HAVE 消息时,各节点每下载并验证一个数据块,就向与其所有相连的节点组播 HAVE 消息;当 BT 系统采用 MultiHave 消息时,只能等到定时时间到才会向其他相连节点发送 MultiHave 消息.若从对方节点下载数据块的速度低于对方发出的 MultiHave 消息中的新增块速度,本节点中有关对方节点的待下载数据块队列就不会为空,MultiHave 消息与 HAVE 消息等同,不会带来延迟;反之,就会出现定时器间隔以内的延迟问题.

当与某节点连接的节点数目很多时,根据 BT 协议的节点调度算法,某节点的连接节点的集合不断动态更新,这时,新连入节点的 BITFILED 消息所携带数据块的位视图中往往含有大量可供下载的数据块信息,会形成很长的对方节点待下载数据块队列,从而减少 MultiHave 消息定时器间隔以内的延迟问题的出现.多个相连的节点也会降低各相连节点待下载数据块队列的消耗速度,从而减少 MultiHave 的延迟问题出现.

当与某节点连接的节点数目很少时,MultiHave 消息的延迟问题可能会影响系统下载数据的平稳性,这时,可以改用 HAVE 消息来通知相连节点.在使用 MultiHave 消息的 BT 系统中可以设置一个节点数目阈值,若节点数目低于该阈值时,不会出现 HAVE 消息风暴,BT 系统使用 HAVE 消息;否则使用 MultiHave 消息.这也正是设计 MultiHave 消息的初衷.

5 实验分析

实验采用了 BT 协议的发明者 Bram Cohen 开发的 BT 客户端.实验的主要参数如下:

(1)最大上载速率没有限制;

(2)向 Tracker 请求之前,集中节点≥20;

(3)节点初始最大连接数是 50;

(4)节点集中节点数目≤80;

(5)活动节点集包含的最优非阻塞节点数是 4;

(6)块的大小设为 2^{20}字节;MultiHave 定时间隔为 5s;

(7)下载的文件大小为 2.15GB,Torrent 文件大小 4.13KB,下载文件分成 2205 块.

严格测试 BT 系统的性能是非常复杂的,实验结果严重依赖于节点行为、种子的数目、BT 系统中的 leechers 数量和 tracker 随机返回的节点的数目等因素.通过选择一定量的不同节点,能够仿真 BT 协议的基本行为.在不考虑种子节点的情况下,具有不同数量的高速节点在某特定时段内采用 MulitHave 消息和 HAVE 消息的所发出消息数量和信息流量如表 2 所示.

表 2 最大传输速度下高速节点间 MultiHave 消息和 HAVE 消息的性能对比

节点数	MultiHave 数量/个·s^{-1}		HAVE 数量/个·s^{-1}		MultiHave 流量/B·s^{-1}		HAVE 流量/B·s^{-1}	
	发出	收到	发出	收到	发出	收到	发出	收到
2	0.2	0.2	20	20	125	125	980	980
3	0.2	0.4	40	80	205	365	1960	3920
4	0.2	0.6	60	180	285	765	2940	8820
5	0.2	0.8	80	320	365	1325	3920	15680
6	0.2	1.0	100	500	445	2045	4900	24500
7	0.2	1.2	120	720	525	2925	5880	35280
9	0.2	1.6	160	1280	685	5165	7840	62720
11	0.2	2.0	200	2000	845	8045	9800	98000
21	0.2	4.0	400	8000	1645	32045	19600	392000
31	0.2	6.0	600	18000	2445	72045	29400	882000
41	0.2	8.0	800	32000	3245	128045	39200	1568000
51	0.2	10.0	1000	50000	4045	200045	49000	2450000

表 2 所述的 BT 系统中 MultiHave 消息 HAVE 消息的发生频率如图 1 所示.

MultiHave 消息 HAVE 消息的传送流量如图 2 所示.

由图 1 和图 2 可以看出,相对于 HAVE 消息,BT 系统中 MultiHave 消息发生频率和数据流量均大幅降低.MultiHave 减少了 BT 系统中主要的管理消息 Have 数量,

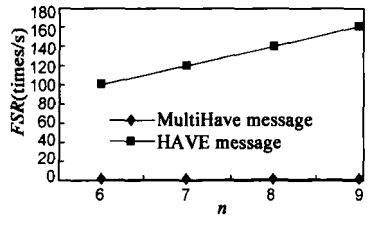


图1　MultiHave消息和HAVE消息的频率对比

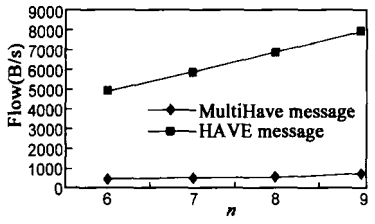


图2　MultiHave和HAVE 消息的流量对比

降低了系统的管理代价.

上述实验结果是在高带宽,节点行为一致的情况下得到的.在实际应用中,多个节点的不同的带宽、高带宽和低带宽并存、各节点随机加入等因素都会使多节点失去行为一致性.对于高带宽节点,由于加入的时机不同,彼此之间传送的速率会严重下降,造成 HAVE 消息的减少,使得 HAVE 消息和 MultiHave 消息比率下降,当这个比率下降到 1 的时候,MultiHave 消息退化为 HAVE 消息.当低带宽节点下载一个块的时间超过计时器时间时,MultiHave 消息也会退化为 HAVE 消息.

6　结论

MultiHave 消息由多个 HAVE 消息组成,通过调整计时器间隙,可以调整 MultiHave 消息的大小,在高速网络环境下,MultiHave 消息可以显著降低 HAVE 消息的流量,有效的抑制"HAVE 消息风暴"的产生,降低 BT 系统通信中的管理流量.在低速环境下,MultiHave 消息退化成 HAVE 消息,不会降低 BT 系统的性能.另外,通过设定 BT 节点连接数阀值,可实现 BT 系统中 HAVE 消息和 MultiHave 消息的转换,从而解决 MultiHave 消息可能带来的定时间隔内的延迟效应.

参考文献:

[1] Cohen B. BitTorrent protocol specification [DB/OL]. http://www.bittorrent.org/protocol.html,2006-02-15.

[2] Parker A. The true picture of peer-to-peer filesharing [EB/OL]. http://www.cachelogic.com/ research/slide9.php, 2005-12-05.

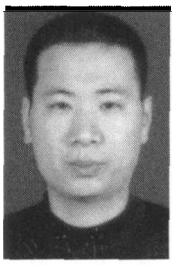
[3] Karagiannis T, Broido A, Faloutsos M, et al. Transport layer identification of P2P traffic[A]. In Proceedings of ACM IMC [C]. Taormina, Sicily, Italy, 2004. 121-134.

[4] Karagiannis T, Broido A, Brownlee N, et al. Is P2P dying or *just hiding*? [A]. Proceeding of IEEE Globecom[C]. Dalla, Texas, USA, 2004. 1532-1538.

[5] Legout A, Urvoy-Keller G, Michiardi P. Understanding BitTorrent: An Experimental Perspective[EB/OL]. http://hal.inria.fr/inria-0000 0156/en, 2005-12-05.

作者简介:

张　尧　男,1951 生于河南洛阳,郑州轻工业学院副教授,主要从事计算机控制和通信工程研究.
E-mail: zhangyao@zzuli.edu.cn



李建春　男,1976 年生于河南新乡,郑州轻工业学院讲师,主要从事对等网体系结构研究.
E-mail: lijianchun@zzuli.edu.cn

黄道颖(通信作者)　男,1967 年生于河南信阳,郑州轻工业学院教授,主要从事计算机网络体系结构研究.
E-mail: dyhuang@zzuli.edu.cn