最小化出口流量花费的接入级P2P缓存容量设计方法

翟海滨,张 鸿,刘欣然,王 勇,沈时军,李正民
(国家计算机网络应急技术处理协调中心,北京 100029)

摘 要: 接入级P2P(Peer-to-Peer)缓存容量设计回答在接入级ISP(Internet Service Provider)出口部署多大容量缓存的问题,目前还没有最优P2P缓存容量设计方法被提出.本文提出一种权衡存储成本和带宽成本的P2P缓存容量设计方法,以最小化ISP出口流量总花费为目标,将最优缓存容量设计问题描述为整数规划问题,其目标函数形式为单调阶梯函数,通过理论推导得出最优缓存容量计算公式指导接入级ISP进行缓存容量设计.将本文所提方法与Median和"20-80 Rule"等几种ISP常用的容量设计方法进行性能比较,结果表明,本文所提方法明显优于已有方法,与目前ISP最认可的"20-80 Rule"相比,应用本文所提方法的ISP出口流量总花费最多可降低7.5%.

关键词: P2P缓存技术;ISP网络;缓存部署;容量设计
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2015)05-0879-09
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.05.007

A P2P Cache Capacity Design Method to Minimize the Total Traffic Cost of Access ISPs

ZHAI Hai-bin, ZHANG Hong, LIU Xin-ran, WANG Yong, SHEN Shi-jun, LI Zheng-min
(*National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing* 100029, *China*)

Abstract: By deploying Peer-to-Peer (P2P) caches at the edge of access ISPs, cross-ISP P2P traffic can be reduced. The problem is how to design an optimal cache capacity? Up to now, no optimal P2P cache capacity design method has been proposed yet. In this paper, an optimal P2P cache capacity design method is proposed based on the storage versus bandwidth cost tradeoff. In order to minimize the total traffic cost of access ISPs, the cache capacity design problem is formulated into an integer programming problem whose objective function is piecewise continuous. The formula of optimal cache capacity can be derived through theoretical derivation. Experimental results suggest that compared with several simple design rules including No Cache, Median and "20-80 Rule", ISPs can achieve significant cost saving using our method. For example, compared with "20-80 Rule" that is most commonly used by ISPs nowadays, the cost saving can be as much as 7.5%.

Key words: P2P(Peer-to-Peer) caching technology; ISP (Internet service provider) network; cache deployment; capacity design

1 引言

目前,P2P流量已成为互联网流量的最主要组成部分,给网络服务提供商(ISP,Internet Service Provider)带来前所未有的运营压力.在国际上,P2P流量占据互联网流量的60%~75%[1],甚至高达90%[2];在国内,以中国电信为例,2010年电信骨干网确知的P2P流量已超过55%[3].P2P应用的广泛流行使ISP骨干网络流量负载成倍增加,运营压力显著提高.P2P缓存[4]是解决上述P2P流量问题最有效的方法之一.在P2P缓存中,通过在接入网出口部署缓存设备,将P2P流量进行缓存并服务于后续请求,实现了P2P流量的本地化,从而达到减少骨干网络P2P流量及降低ISP运营压力的目标.

接入级ISP指需要从上层ISP购买接入服务的ISP[5],其部署P2P缓存的主要动机是最小化出口P2P流量花费.另一类ISP是骨干级ISP,指无需购买接入服

收稿日期:2013-10-23;修回日期:2015-01-29;责任编辑:马兰英
基金项目:国家973重点基础研究发展规划(No.2011CB302605)

务的 ISP,主要通过对等互联等方式实现互联网接入.P2P 应用的广泛流行,使得接入级 ISP 的大量 P2P 用户需要向其他 ISP 用户节点请求资源,导致本 ISP 出口流量花费大增.通过在接入级 ISP 出口部署缓存设备为本地 P2P 用户服务,可以减少 P2P 出口流量,同时节省出口流量花费.接入级 P2P 缓存容量设计用于决策"在接入级 ISP 出口部署多大容量的缓存".如图 1 所示,缓存容量越大,可节省带宽开销,但导致更多存储开销;缓存容量越小,可节省存储开销,但导致更多的带宽开销.

目前关于 P2P 缓存的研究已经较为广泛,包括分析 P2P 流量特征[6,7],提出新的 P2P 缓存置换算法[4, 7],以及探讨 P2P 缓存协作机制[6, 8]等方面.相比之下,同样作为 P2P 缓存技术关键内容的 P2P 缓存部署问题研究尚处于起步阶段.文献[9]提出一种在骨干级 ISP 链路上部署 P2P 缓存以最小化骨干网络负载的最优部署策略.文献[10]设计了基于中心系数的缓存部署算法,实验结果表明该算法可以明显降低骨干网络流量负载.文献[11]提出一种在接入级 ISP 出口部署 P2P 缓存的最优策略,达到最小化接入级 ISP 流入骨干级 ISP 的 P2P 流量的目的.文献[12]提出一种协作式 P2P 缓存容量设计理论模型,重点关注跨 ISP 流量花费的最小化问题.然而,上述研究都是以缓存容量已知为前提的.

与本文工作最相似的是文献[13],首次从经济开销角度对 Web 缓存服务进行研究,该文在已知流量负载、存储和带宽开销的条件下,提出一种权衡内存和带宽成本的 Web 缓存最优容量设计方法.该文所提方法需要预知每个文件的大小和热度,以及请求队列信息.然而,请求队列信息在实际环境下是不断变化的,因此该文所提方法对参数输入的要求过高,影响其实用性.此外,该方法并不能直接应用于 P2P 缓存场景:传统 Web 应用请求与 P2P 内容请求的访问模式具有明显差异,比如传统应用中内容请求目的地一般为位置确知的源服务器,而 P2P 应用中的内容请求目的地为位置分散且随机的 peer 节点,在缓存部署方案设计过程中需要对 P2P 内容请求模式进行考虑.

截至目前还没有文章回答最小化出口流量花费的最优 P2P 缓存容量是否存在这一问题.因此,本文的研究问题是:

(1)能否通过存储成本与带宽成本的权衡实现 ISP 出口流量总花费最小化?

(2)如果问题(1)的答案是肯定的,那么如何设计最小化出口流量总花费的最优 P2P 缓存容量?

本文提出一种权衡存储成本和带宽成本的 P2P 缓存容量设计方法,以最小化 ISP 出口流量总花费为目标,通过理论推导得出最优缓存容量计算公式指导接入级 ISP 进行缓存容量设计.将本文所提方法与 Median、"20-80 Rule"等几种 ISP 常用的容量设计方法进行性能比较,结果表明,本文所提方法明显优于已有方法.

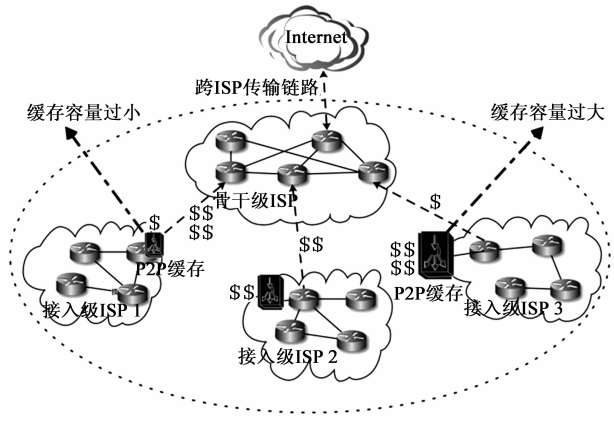

图1 接入级P2P缓存部署示意图

2 最小化出口流量花费的 P2P 缓存容量设计

最小化出口流量花费的 P2P 缓存容量设计方法包括问题形式化和问题求解两个步骤.首先,对 P2P 缓存容量设计问题进行形式化:将 ISP 的出口流量总花费成本函数定义为带宽成本与存储成本之和,其形式为单调阶梯函数,以最小化 ISP 出口流量总花费为目标,将接入级 P2P 缓存容量设计问题形式化为整数规划问题;其次,通过理论推导对该问题进行求解:利用近似变换和函数求导得出原问题可能最优解,并将其与最优阶梯区间边界值进行比较,最终得出原问题最优解.

2.1 参数定义

接入级 P2P 缓存容量设计方法使用的符号定义如表 1 所示.设 ISP 所部署缓存总容量为 S,容量为 S 时的缓存命中率为 $h(S)$,S 是整数.设网络中内容平均大小为 B 字节,求得的最优缓存容量为 S^*,则最终 ISP 需要投入的缓存容量为 BS^* 字节.ISP 出口流量总花费为 $C(S)$,包含带宽成本 $C_{bw}(S)$和存储成本 $C_{storage}(S)$两部分,均为缓存总容量 S 的函数,并且单位均为 \$.而单位带宽成本为 P_{bw}(\$/字节),单位存储成本为 P_{s_driver}(\$/字节/月).在实际环境中,缓存的部署需要配套硬件的支持,设单个缓存配套硬件的存储容量为 $S_{\text{enclosure}}$,$S_{\text{enclosure}}$与 S 一样也是整数,部署容量为 S 的缓存可能需要多台配套硬件.单个缓存配套硬件的成本定义为 $P_{\text{s\_enclosure}}$,单位为 \$.存储硬件的使用寿命定义为 T_c,单位为月.设稳定状态下整个 P2P 网络中内容总数为 M,用户总数为 N,其中热度排名为 i 的内容热度为 $p(i)$,跨 ISP 流量占总流量比例为 η,每月产生的请求数为 R,每月每个用户平均产生的内容数为 R_o,而新产生内容生存时间为 T_o 月.

表 1　P2P 缓存最优容量设计方法符号定义说明表

符号	定义
M	稳定状态下的 P2P 内容总数
N	系统中用户总数
$p(i)$	热度排名为 i 的内容热度
S	ISP 所部属缓存总容量
B	内容平均大小(字节)
$C(S)$	ISP 出口流量总花费($)
$C_{bw}(S)$	带宽成本($)
$C_{\text{storage}}(S)$	存储成本($)
P_{bw}	单位带宽成本($/字节)
P_{s_driver}	单位存储成本($/字节/月)
$S_{\text{enclosure}}$	单个缓存配套硬件的存储容量
$P_{s\_enclosure}$	单个缓存配套硬件成本($)
T_c	存储硬件的使用寿命(月)
η	跨 ISP 流量占总流量比例
$h(S)$	容量为 S 时的缓存命中率
R	用户请求率(请求数/月)
R_o	内容请求生成速率(内容数/用户/月)
T_o	新产生内容的生存时间(月)
d_p	用户节点之间的平均距离(跳数)

2.2　问题形式化

ISP 出口流量总花费 $C(S)$包含两部分:带宽成本 $C_{bw}(S)$和存储成本 $C_{\text{storage}}(S)$,均为缓存总容量 S 的函数:

$$C(S) = C_{bw}(S) + C_{\text{storage}}(S) \tag{1}$$

接入级 P2P 缓存容量设计问题可形式化为:

$$\min\ C(S) = C_{bw}(S) + C_{\text{storage}}(S) \tag{2}$$
$$\text{s.t.}\quad 0 \leqslant S \leqslant M, S:\text{integer}$$

带宽成本 $C_{bw}(S)$又可形式化为:

$$C_{bw}(S) = P_{bw}BS + T_cP_{bw}\eta BR(1 - h(S)) \tag{3}$$

公式(3)中带宽成本 $C_{bw}(S)$实际为内容未命中产生的带宽成本.式(3)的第一部分表示内容首次获取产生的带宽成本:当缓存存储空间未占满时,如果请求到来而缓存并未存储该请求所需内容,则需要首先进行内容获取并存储至缓存,称该部分操作带来的成本为内容首次获取产生的带宽成本.该成本是单位带宽成本 P_{bw}、内容平均大小 B 以及缓存存储容量 S 的乘积.

式(3)的第二部分表示内容非首次获取产生的带宽成本:当缓存存储空间已经占满后,需要根据指定的缓存替换算法如 LRU(Least Recently Used)、LFU(Least Frequently Used)等对已有内容进行替换,才能得到足够存储空间缓存新内容.请求命中时不会带来额外的带宽开销,这也是缓存部署效果的体现;请求未命中时,缓存需要首先获取该内容,进而利用该内容替换已有内容才能为后续请求服务,称该部分成本为内容非首次获取产生的带宽成本.内容非首次获取产生的带宽成本为单个内容请求产生的带宽成本与未命中内容请求总数的乘积:

(1)单个内容请求产生的带宽成本为 $P_{bw}\eta B$,是单位带宽成本、内容平均大小和 η 的乘积.η 是跨 ISP 流量占总流量比例,该参数是 P2P 缓存设计与 Web 缓存设计的最主要区别之一.对与 Web 缓存而言,内容会被整体缓存,因此 $\eta = 1$;而对于 P2P 缓存而言,通常只有内容的一部分被缓存,因而 $\eta \leqslant 1$.

(2)未命中内容请求总数为每月未命中请求数与存储使用寿命的乘积.已知每月 ISP 产生的用户请求数目为 R,则发生未命中的请求数为 $R(1 - h(S))$.其中 $h(S)$是容量为 S 的缓存的命中率,由缓存所使用的替换算法决定.比如根据 LFU 或者 LSB(Least Sent Byte)算法[4],$h(S) = \sum_{j=1}^{S} p(j)$,其中 $p(j)$为排名为 j 的内容热度,已有文献发现 P2P 内容热度服从 Mandelbrot-Zipf 分布[6,8],而 $p(j) = a/(j+q)^b$,a 是归一化常系数,比如 $a = 1/\sum_{j=1}^{|M|} p(j)$,$b$ 为偏度系数,q 为高原因子.因此,对于使用寿命为 T_c 的存储而言,未命中的内容请求总数为 $T_cR(1 - h(S))$.

从而,内容非首次获取产生的带宽成本为单个内容请求产生的带宽成本乘以未命中内容请求总数,即 $T_cP_{bw}\eta BR(1 - h(S))$.

存储成本 $C_{\text{storage}}(S)$可以形式化为:

$$C_{\text{storage}}(S) = T_cP_{s\_\text{driver}}BS + \left\lceil \frac{S}{S_{\text{enclosure}}} \right\rceil P_{s\_\text{enclosure}} \tag{4}$$

上式中第一项 $T_cP_{s\_\text{driver}}BS$ 为存储硬件成本,是存储使用寿命 T_c、单位存储成本 P_{s_driver}、内容平均大小 B 与存储总容量 S 的乘积.第二项为存储配套硬件成本,其中 $\left\lceil \frac{S}{S_{\text{enclosure}}} \right\rceil$ 是所需存储配套硬件数目,$P_{s_\text{enclosure}}$则是单个存储配套硬件的成本.

2.3　问题求解

如式(2)所示的接入级 P2P 缓存容量设计问题的完整形式可表示如下:

$$\min\ C(S) = (P_{bw} + T_cP_{\text{storage}})BS + T_cP_{bw}\eta BR(1 - \sum_{i=1}^{S} p(i)) + \left\lceil \frac{S}{S_{\text{enclosure}}} \right\rceil P_{s\_\text{enclosure}}$$
$$\text{s.t.}\quad 0 \leqslant S \leqslant M, S:\text{integer} \tag{5}$$

该问题为整数规划问题,其目标函数为单调阶梯函数,可进一步转换为如下形式:

$$\min\ C(S)=\begin{cases} g(0), & \text{if } S=0 \\ g(S)+P_{s\_\text{enclosure}}, & \text{if } 0<S\leqslant S_{\text{enclosure}} \\ \cdots & \\ g(S)+kP_{s\_\text{enclosure}}, & \text{if } (k-1)S_{\text{enclosure}}<S\leqslant kS_{\text{enclosure}} \\ \cdots & \\ g(S)+k_{\max}P_{s\_\text{enclosure}}, & \text{if } (k_{\max}-1)S_{\text{enclosure}}<S\leqslant k_{\max}S_{\text{enclosure}} \end{cases}$$

$$\text{s.t.}\quad S:\text{integer} \tag{6}$$

其中 $k_{\max}=\left\lceil \frac{M}{S_{\text{enclosure}}} \right\rceil$;

$g(S)=(P_{bw}+T_cP_{\text{storage}})BS+T_cP_{bw}\eta BR(1-\sum_{i=1}^{S}p(i))$.

针对问题(6),求解步骤如下:

步骤 1 为方便求解,对目标函数进行近似变换:利用 $\int_1^S p(S)$ 代替 $\sum_{i=1}^{S}p(i)$.

步骤 2 对目标函数求导,得出原问题的可能最优解 S^*,详细过程如下:

经过变换后,目标函数 $C(S)$ 的一二阶导数为:

$$\frac{\mathrm{d}C(S)}{\mathrm{d}S}=(P_{bw}+T_cP_{\text{storage}})B-T_cP_{bw}\eta BRp(S) \tag{7}$$

$$\frac{\mathrm{d}^2C(S)}{\mathrm{d}S^2}=-T_cP_{bw}\eta BR\frac{dp(S)}{dS}=abT_cP_{bw}\eta BR\frac{1}{(S+q)^{b+1}} \tag{8}$$

上述二阶导数始终大于零,说明目标函数为凸函数,近似变换后的问题(6)最优解为使得 $C(S)$ 的一阶导数为零的值:

$$\frac{\mathrm{d}C(S)}{\mathrm{d}S}=(P_{bw}+T_cP_{storage})B-T_cP_{bw}\eta BRp(S)=0$$

$$\Rightarrow\quad p(S)=\frac{a}{(S+q)^b\sum_{i=1}^{M}p(i)}=\frac{P_{bw}+T_cP_{\text{storage}}}{T_cP_{bw}\eta R} \tag{9}$$

令 S^* 为最接近上述最优解的整数,则 S^* 有可能就是原问题的最优解,不妨假设 S^* 所在的区间为 $[k-1)S_{\text{enclosure}},kS_{\text{enclosure}}]$. 之所以说为可能最优解,是因为原问题为单调阶梯函数,最优解还可能是区间边界值,下面需要找出最优区间边界值.

步骤 3 对目标函数进行近似变换,得出原问题最优区间边界值 $S^{\perp}$,详细过程如下:

首先,将目标函数 $C(S)$ 变换为 $\tilde{C}(S)$,问题(6)就变换为如下形式:

$$\min\ \tilde{C}(S)=(P_{bw}+T_cP_{\text{storage}})BS+T_cP_{bw}\eta BR(1-\int_1^S p(S))+\frac{S}{S_{\text{enclosure}}}P_{s\_\text{enclosure}} \tag{10}$$

上述形式去掉了问题(6)中 S 为整数的约束条件,并且将 $\left\lceil \frac{S}{S_{\text{enclosure}}} \right\rceil$ 变为 $\frac{S}{S_{\text{enclosure}}}$,降低了求解难度. $\tilde{C}(S)$ 的一二阶导数为:

$$\frac{\mathrm{d}\tilde{C}(S)}{\mathrm{d}S}=(P_{bw}+T_cP_{\text{storage}})B-T_cP_{bw}\eta BRp(S)+\frac{P_{s\_\text{enclosure}}}{S_{\text{enclosure}}} \tag{11}$$

$$\frac{\mathrm{d}^2\tilde{C}(S)}{\mathrm{d}S^2}=-T_cP_{bw}\eta BR\frac{dp(S)}{dS}=abT_cP_{bw}\eta BR\frac{1}{(S+q)^{b+1}} \tag{12}$$

上述二阶导数始终大于零,说明目标函数为凸函数,问题(10)的最优解为使得 $\tilde{C}(S)$ 一阶导数为零的值:

$$\frac{\mathrm{d}\tilde{C}(S)}{\mathrm{d}S}=(P_{bw}+T_cP_{\text{storage}})B-T_cP_{bw}\eta BRp(S)+\frac{P_{s\_\text{enclosure}}}{S_{\text{enclosure}}}=0$$

$$\Rightarrow p(S)=\frac{a}{(S+q)^b\sum_{i=1}^{M}p(i)}=\frac{P_{bw}+T_cP_{\text{storage}}+\frac{P_{s\_\text{enclosure}}}{BS_{\text{enclosure}}}}{T_cP_{bw}\eta R} \tag{13}$$

令 $S^{\perp}$ 为问题(10)的最优解,则最接近 $S^{\perp}$ 的区间边界就是最优区间边界值,表示如下:

$$S^{\perp}=\begin{cases} S_{\text{lower}}{}^{\perp}, & \text{if } C(S_{\text{lower}}{}^{\perp})\leqslant C(S_{\text{upper}}{}^{\perp}) \\ S_{\text{upper}}{}^{\perp}, & \text{if } C(S_{\text{lower}}{}^{\perp})>C(S_{\text{upper}}{}^{\perp}) \end{cases} \tag{14}$$

其中 $S_{\text{lower}}{}^{\perp}=\left\lfloor \frac{S^{\perp}}{S_{\text{enclosure}}} \right\rfloor S_{\text{enclosure}}$; $S_{\text{upper}}{}^{\perp}=\left\lceil \frac{S^{\perp}}{S_{\text{enclosure}}} \right\rceil S_{\text{enclosure}}$.

步骤 4 通过比较 S^* 和 $S^{\perp}$,得出原问题的最优解 S^{OPT}:

$$S^{OPT}=\begin{cases} S^*, & \text{if } C(S^*)\leqslant C(S^{\perp}) \\ S^{\perp}, & \text{if } C(S^*)>C(S^{\perp}) \end{cases} \tag{15}$$

通过上述理论推导过程,即可得出 ISP 的最优缓存容量计算公式,ISP 只需输入上式所需参数即可快速求解出所需部署的最优缓存容量.

3 实验结果与评价

首先,对本文所提方法与已有容量设计方法 No

Cache、Median、"20-80 Rule"的性能进行比较;其次,分析存储成本、带宽成本等关键参数对最优缓存容量的影响.本文所提方法适用于接入级 ISP,该类 ISP 部署 P2P 缓存的拓扑架构如图 1 所示.接入级 ISP 通过在骨干出口部署缓存设备为本地 P2P 用户服务,可以减少 P2P 出口流量,同时节省出口流量花费.接入级 ISP 1、接入级 ISP 2 或接入级 ISP 3 均可使用本文所提方法进行最优缓存容量计算.由于本文所提为一种容量设计方法,因此下述实验是模拟环境下的离线实验,但带宽、存储等实验参数全部来自真实的数据.

3.1 参数设置

(1)P2P 系统参数设置

本实验所采用的 P2P 系统相关参数全部为来自 Gnutella[7]和 eDonkey、Bittorrent[14]的真实数据,Hefeeda 等人认为不同 P2P 系统中的用户和内容特征基本类似[7,8],因此本文中将混合使用来自上述三个 P2P 系统的参数,详细参数列表如表 2 所示.

表 2　P2P 系统参数设置说明表

用户数目 N	稳定状态下内容数目 M	用户请求速率 R	内容平均大小	内容热度分布
7012	456481	35060	24M	M-Zipf
40000	2604000	200000	24M	M-Zipf
100000	6510000	500000	24M	M-Zipf

已有工作[7]和[14]并没有给出稳定状态下 P2P 系统中的内容总数,但是给出了新产生内容的生存周期 $T_o = 3.5$ 月和新内容生成速率 $R_o = 18.6$ 个/月,据此我们可以计算出稳态下网络中的内容数 M 为:

$$M = R_o N T_o \tag{16}$$

其中 N 为网络用户总数,与文献[14]类似,本文中我们将展示用户数分别为 7012,40000 和 100000 时的实验结果.文献[7]对 Gnutella 系统八个月的日志进行分析发现,内容热度服从 Mandelbrot-Zipf 分布,并且偏度系数因子 $b = 0.75$、高原因子 $q = 3$,本实验中内容热度分布相关参数参照上述设置.在 eDonkey 系统中,内容平均大小为 24 MB,而 BitTorrent 系统中,内容平均大小为 400MB.在类 Bittorrent 系统中,用户从同一 ISP 内 peer 节点下载的块数约占总块数的 10% ~ 30%[15],因此我们设置 η 的值为 0.8.文献[7]还指出,每个用户平均每月下载文件数为 5 个,因此一个月请求总数 R 为 $5N$,即可得出如表 2 所示的参数值.

(2)单位带宽成本设置

本实验中,我们根据美国多个不同种类的 ISP 宽带服务定价数据[16]进行带宽成本参数设置,详细设置如表 3 所示.

表 3　单位带宽成本设置

带宽类型	每月带宽成本($)	单位带宽成本 P_{bw}($/字节)
GigE	2e + 5	6.15e - 10
STM4-OC12	4e + 5	1.23e - 9
FastE	5e + 5	1.5375e - 9

假设 ISP 平均所需带宽为 10Gbps,则每月可传输字节总数约为 3.25e + 14.单位带宽成本 P_{bw} = 每月带宽成本/每月可传输字节总数.比如,GigE 的每月带宽成本为 $2e + 5,STM4-OC12 的每月带宽成本为 $4e + 5,FastE 的每月带宽成本为 $5e + 5.因此,GigE 单位带宽成本 P_{bw} = $2e + 5/3.25e + 14 = 6.15e - 10 $,STM4-OC12 的单位带宽成本为 1.23e - 9 $/字节,FastE 的单位带宽成本为 1.5375e - 9 $/字节.

(3)单位存储成本

本实验中,我们根据 PeerApp 公司的 UltraBand 缓存产品数据[17]进行存储成本参数设置,详细设置如表 4 所示.

目前,主流 P2P 缓存产品均使用 SAN(Storage Area Network)存储架构,因为 SAN 存储具有管理简便、灵活性高和故障恢复速度快速等特点,例如 PeerApp 公司的缓存产品 UltraBand 就是采用 Dell/EMC CX300[18]作为其 SAN 存储.本文以 Dell/EMC CX300 为例进行存储成本参数设置.存储成本包含两部分存储配套硬件成本和存储成本两部分.根据文献[18]可知,Dell/EMC CX300 的成本为 $18000,这也就是 $P_{s_\text{enclosure}}$的值.此外,单个存储配套硬件的容量是 38 TB,因此 $S_{\text{enclosure}}$的值为 $\frac{38e+6}{B}$.

Dell/EMC CX300 可以使用 FC (Fibre Channel)存储或 SATA(Serial Advanced Technology Attachment)存储,前者具有较高的性能,但是价格较高,而后者价格较低,但是性能低于前者.在 SAN 中,容量为 750 GB 和 500 GB 的存储最为常用.以 Seagate 为例[19],上述两种容量规格的 SATA 存储价格分别为 $62 和 $49 每块,因此每 GB SATA 存储的价格约为 0.085 $.Seagate 没有容量为 750 GB 和 500 GB 的 FC 存储,但是有容量分别为 600 GB、450 GB 和 300 GB 的存储,价格分别为 $595、$495 和 $304,因此每 GB FC 存储的价格约为 1 $/GB.

表 4　单位存储成本设置

存储类型	每 GB 带宽成本($)	单位存储成本 $P_{s\_\text{enclosure}}$($/字节/月)
Fibre Channel	1	2.8e - 11
SATA	0.085	2.36e - 12

根据文献[20],存储使用寿命 T_c 设置为 3 年或者

说 *36* 个月.单位存储成本 P_{s_driver} = 每 GB 带宽成本/存储使用寿命,正如表 4 所示,SATA 的单位存储成本为 0.085 $/GB/$T_c$ = 2.36e - 12 $/字节/月;FC 的单位存储成本为 1 $/GB/$T_c$ = 2.8e - 11 $/字节/月.

3.2 与已有容量设计方法的对比分析

我们用 OPT 表示利用本文方法得出的 ISP 最优缓存容量值.在实际应用中,许多 ISP 根据经验规律来设定 P2P 缓存容量值,目前没有最优缓存容量计算方法提出.因此,本实验中我们仅与一些比较简单的容量设计方法进行比较,目的不只在于反映本方法的优势,同时展示应用这些简单方法时会给 ISP 带来较多无谓花费的增加.

第一个方法是不部署任何缓存(No Cache).

第二个方法是简单地将缓存容量设置为所有内容总容量的一半(Median)[13,21].

首先,将带宽类型固定为 GigE,存储类型分别为 SATA 和 FC 时根据 OPT、No Cache 和 Median 得出的 ISP 出口流量总花费对比情况如图 2 和图 3 所示.当使用其他带宽类型时,得到的结论相同.可以发现,无论采用何种带宽类型与存储类型,OPT 的出口流量总花费均明显低于其他两种方法.

其次,我们将存储类型固定为 FC,使用 SATA 时得到的结论相同.类型分别为 GigE、STM-OC12 和 FastE 时根据 OPT、No Cache 和 Median 得出的 ISP 出口流量总花费对比情况分别如图 3、图 4 和图 5 展示.使用 OPT 的总花费均明显低于其他两种方法.可见,如果仅适用一些简单的容量设计方法,会造成 ISP 出口流量总花费的较大浪费.

最后,我们将 OPT 与"20-80 Rule"进行对比分析."20-80 Rule"是目前 ISP 最认可的缓存容量设计方法,认为用 20% 的存储即可服务 80% 的流量,因而通常会将缓存容量设计为所有内容总大小的 20%.在本文实验环境下,表 5 和图 6 所示为带宽类型设定为 GigE 时最优缓存容量的值.

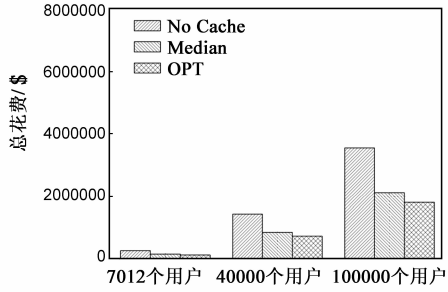

图2 使用GigE和SATA时不同方法的总花费对比

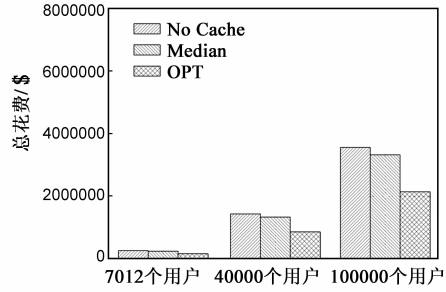

图3 使用GigE和FC时不同方法的总花费对比

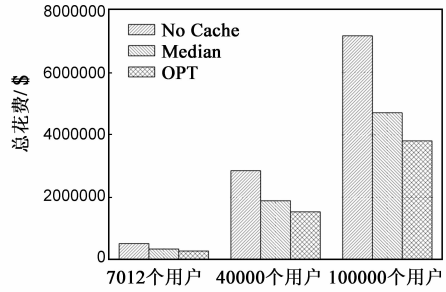

图4 使用STM-OC12和FC时不同方法的总花费对比

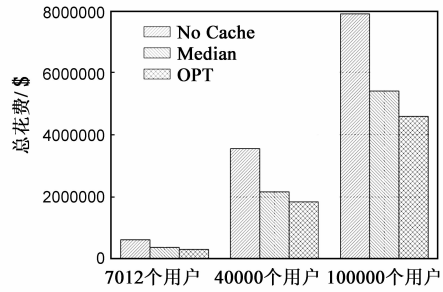

图5 使用FastE和FC时不同方法的总花费对比

表 5 带宽类型为 GigE 时最优缓存容量值

(用户数目,存储类型)	最小花费($)	存储容量占内容总数的比例
(7012,FC)	161270	13.4%
(40000,FC)	862210	10.9%
(100000,FC)	2138480	8.8%
(7012,SATA)	129701	20.8%
(40000,SATA)	731137	21.9%
(100000,SATA)	1817530	20.4%

我们发现"20-80 Rule"具有一定的可行性,但仅适用于 SATA 存储,不适用于 FC 存储:对于 SATA 存储而言,最优缓存容量占内容总数的比例为 20% 左右,此时应用"20-80 Rule"与应用本文所提方法的 ISP 出口流量总花费基本相同;对于 FC 存储而言,最优缓存容量占内容总数的比例为 10% 左右,与"20-80 Rule"相比,应用本文所提方法的 ISP 出口流量总花费可降低约 7.5%.可见,本文所提方法的性能优于"20-80 Rule",详细比较结果如图 6 所示.同时,我们发现应用 SATA 存储的最优缓存容量始终大于应用 FC 存储的最优缓存

容量,因为 SATA 价格较低,ISP 多部署 SATA 存储可以抵消更多带宽开销,从而降低出口流量总花费.

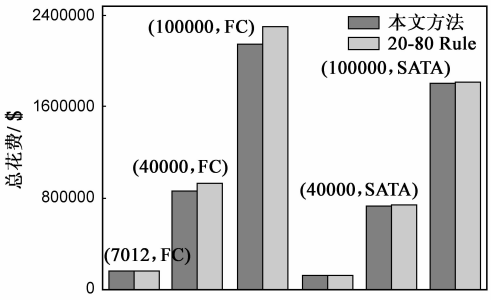

图6　本文所提方法与"20-80 Rule"对比

3.3　带宽固定时最优缓存容量的变化情况

首先,不妨将带宽类型固定为 GigE,进而可以分析最优缓存容量随用户数目和存储成本变化情况.我们将本文所提方法求得结果与穷举法所求解进行对比,结果发现所求结果与最优解一致,说明利用本方法所求得的解就是最优 P2P 缓存容量.同时我们还发现,最优缓存容量的变化情况与内容平均大小无关,即不同内容平均大小下最优缓存容量变化曲线相同,因此,本文将内容平均大小固定为 400 MB,只展示该内容大小下的实验结果.

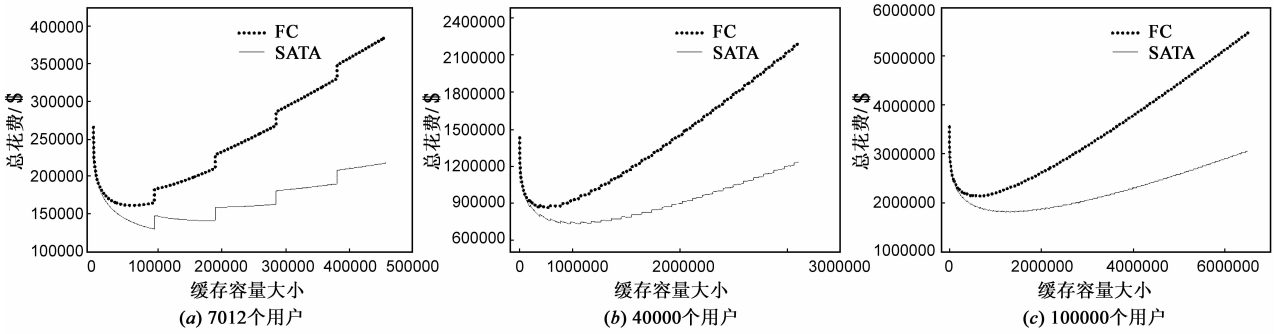

图7　不同用户数目下ISP总花费的变化情况

带宽固定为 GigE 时最优缓存容量的变化情况如图 7 所示,可以得出如下结论:

①随着用户数目的增多,最优缓存容量也在增加.用户数的增多意味着 P2P 流量的增长,为应对流量增长,缓存容量必然增加.

②应用 SATA 存储的最优缓存容量大于应用 FC 存储的最优缓存容量.当 ISP 使用价格较低的存储如 SATA 时,应多部署存储以抵消更多带宽开销.

③存在"总花费突增"现象.由于存储配套硬件容量有限,当一台配套硬件容量用完时,需要新部署一台,自然带来成本的剧增,同时在新添加硬件容量用完前,可能会出现总花费随容量增加而短暂降低的现象.

3.4　存储固定时最优缓存容量的变化情况

当用户数目为 4000、存储类型分别固定为 FC 和 SATA 时,最优缓存容量随带宽成本的变化情况分别如图 8 和图 9 所示,可以得出以下结论:

①最优缓存容量确实存在.回答了"能否通过缓存存储成本与带宽成本的权衡实现 ISP 出口流量总花费的最小化"的问题,利用本文所提方法可求解出最优缓存容量.

②最优缓存容量随带宽成本的增加而增加.说明当带宽成本增加时,ISP 应部署更多缓存以减小对出口带宽的占用,这样就可以降低 ISP 出口流量总花费.

图 7 到图 9 中,随着缓存容量的不断增加,ISP 出口流量总花费首先出现急剧下降,直至到达最低点,进而总花费出现缓慢的上升.这种现象的产生是可以用目标函数 $C(S)$的一阶导数值来解释,而由于目标函数为单调阶梯函数,无法直接求导,因此我们用其近似函数 $\tilde{C}(S)$的一阶导数来寻找规律.$\tilde{C}(S)$的一阶导数如式(11)所示,为关于容量 S 的函数.当用户数目为 40000,带宽类型为 GigE,存储类型为 FC 时,$\tilde{C}(S)$的一阶导数变化情况如图 10 所示.该图中,横坐标表示缓存容量大小,起始值为 30000,当 $S<30000$ 时 $\tilde{C}(S)$的一阶导数为更小的负数.

图 10 中,目标函数在最优值处的一阶导数为零,意味着在容量为该值时 ISP 出口流量总花费最低.由于一阶导数在值为零的附近变化缓慢,说明当用户数目为 40000、带宽类型为 GigE、存储类型为 FC 时,如果缓存容量与最优容量差别不大,则该容量产生的 ISP 出口流量总花费与 ISP 最小花费差距较小.因此,如果所部署缓存容量与最优容量存在较小幅度的误差,ISP 是可以容忍的.

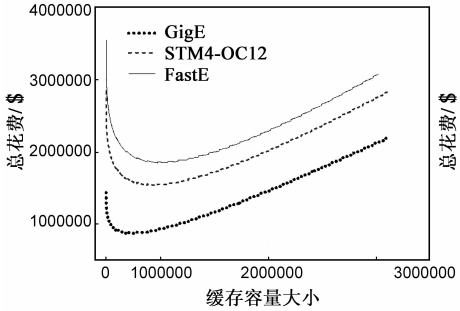

图8 存储为FC时不同类型带宽成本对比

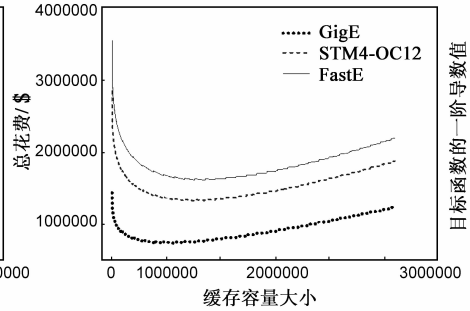

图9 存储为SATA时不同类型带宽成本对比

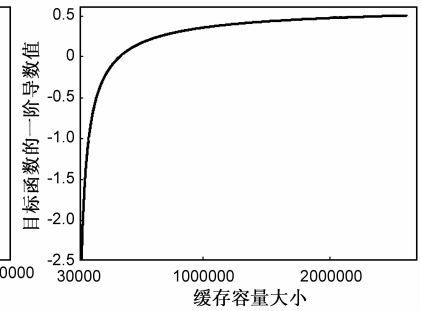

图10 目标函数的近似一阶导数变化情况

4 结论

针对接入级 ISP,本文给出一种 P2P 缓存容量设计方法,在容量设计过程中权衡缓存存储成本与带宽成本,达到最小化接入级 ISP 出口流量花费的目标.

对 P2P 缓存容量设计问题进行形式化:将 ISP 的出口流量总花费成本函数定义为带宽成本与存储成本之和,进而以最小化 ISP 出口流量总花费为目标,将接入级 P2P 缓存容量设计问题形式化为整数规划问题.通过理论推导对该问题进行求解,得出最优缓存容量计算公式指导接入级 ISP 进行 P2P 缓存容量设计.

通过实验将本文所提方法与已有容量设计方法 No Cache、Median、"20-80 Rule"进行性能比较.结果表明,所提方法明显优于已有方法,可使接入级 ISP 的 P2P 出口流量总花费最小.本文实验环境下,与目前 ISP 最认可的"20-80 Rule"相比,应用本文所提方法的 ISP 出口流量总花费最多可降低 7.5%.

近年来,诸如视频流媒体(Http Streaming)、P2P 流媒体(P2P Streaming)等应用日益广泛,本文所提的容量设计方法也适用于这些应用.但是在带宽成本的计算过程中,需要综合考虑这些应用的特点进行缓存命中率的模型更新和参数选择.

参考文献

[1] Slyck News. CacheLogic study-P2P is changing[EB/OL]. http://www.slyck.com/story914_CacheLogic_Study_P2P_is_Changing,2005-09-16.

[2] Ipoque. Ipoque Company Internet study 2008/2009[EB/OL]. http://www.ipoque.com/sites/default/files/mediafiles/documents/internet-study-2008-2009.pdf,2009-04-29.

[3] 韦乐平.电信业和电信技术发展的趋势和挑战[EB/OL]. http://wenku.baidu.com/view/be139cc78bd63186bcebbcdc.html,2010-10-15.

[4] Wierzbicki A, Leibowitz N, Ripeanu M, et al. Cache replacement policies revisited: the case of P2P traffic[A]. Proceedings of the 2004 IEEE International Symposium on Cluster Computing and the Grid[C]. Chicago, Illinois, USA: IEEE Press, 2004.182-189.

[5] 丁珂.中国互联网骨干网市场问题分析与政策建议[J].广东通信技术,2007,27(6):10-14.

[6] Gummadi K, Dunn R, et al. Measurement, modeling, and analysis of a Peer-to-Peer file-sharing workload[A]. Proceedings of the 19th ACM Symposium on Operating Systems Principles[C]. New York, USA: ACM Press, 2003.314-329.

[7] Hefeeda M, Saleh O. Traffic modeling and proportional partial caching for Peer-to-Peer systems[J]. IEEE Transactions on Networking, 2008, 16(6): 1447-1460.

[8] Hefeeda M, Noorizadeh B. On the benefits of cooperative proxy caching for Peer-to-Peer traffic[J]. IEEE Transactions on Parallel and Distributed Systems, 2010, 21(7): 998-1010.

[9] Ye Mingjiang, Wu Jianping, Xu Ke. Caching the P2P traffic in ISP network[A]. Proceedings of the 2008 IEEE International Conference on Communications[C]. Beijing, China: IEEE Press, 2008.5876-5880.

[10] Kamiyama N, Mori T, Kawahara R, et al. ISP-Operated CDN[A]. Proceedings of the 28th Conference on Computer Communications[C]. Rio de Janeiro, Brazil: IEEE Press, 2009.1-6.

[11] Kamiyama N, Mori T, Kawahara R, et al. Analyzing influence of network topology on designing ISP-operated CDN[J]. Telecommunication Systems, 2013, 52(2): 969-977.

[12] Dai J, Li B, Liu F, Li B, Jin H. On the efficiency of collaborative caching in ISP-aware P2P networks[A]. Proceedings of the 30th IEEE International Conference on Computer Communications[C]. Shanghai, China: IEEE Press, 2011. 1224-1232.

[13] Kelly T, Reeves D. Optimal Web cache sizing: Scalable methods for exact solutions[J]. Computer Communications, 2001, 24(2): 163-173.

[14] Carlinet Y, Debar H, et al. Caching P2P Traffic: What are the Benefits for an ISP[A]. Proceedings of Ninth International Conference on Networks[C]. Menuires, France: IEEE Press, 2010.5876-5880.

[15] Karagiannis T, Rodriguez P, Papagiannaki K. Should internet

service provid-ers fear peer-assisted content distribution[A]. Proceedings of the ACM 2005 Conference on Internet Measurement[C]. New Orleans, USA: ACM Press, 2005. 63 – 76.

[16] Easyt1. Easyt1 Products[EB/OL]. http://www.easyt1.net/, 2013-08-10.

[17] PeerApp. PeerApp UltraBand Products[EB/OL]. http://www.gzhowe.com/product_detail.asp? name = P2P Cache, 2013-01-08.

[18] Dell. Dell Products[EB/OL]. http://www1.la.dell.com/content/products/productdetails.aspx/pvaul_cx300? c = la&l = en&s = corp&~ck = p, 2013-08-09.

[19] Seagate. Seagate Products[EB/OL]. http://www.seagate.com/, 2013-07-05.

[20] Pinheiro E, Weber W, et al. Failure trends in a large disk drive population[A]. Proceedings of USENIX Conference on File and Storage Technologies[C]. San Jose, USA: ACM Press, 2007. 17 – 29.

[21] Rasmussen A, Kiciman E, et al. Improving the responsiveness of internet services with automatic cache placement[A]. Proceedings of the 4th ACM European conference on Computer systems[C]. Nuremberg, Germany: ACM Press, 2009. 27 – 32.

作者简介

翟海滨(通信作者)　男, 1983 年 10 月出生,山东淄博人. 2013 年 7 月毕业于中国科学院计算技术研究所,工学博士,主要研究方向为 P2P 网络、缓存技术、分布式计算等.

E-mail: zhb@cert.org.cn

张　鸿　男, 1976 年出生,陕西西安人,工学博士,高级工程师,主要研究方向为云计算技术、计算机网络、信息安全.

E-mail: zhangh@isc.org.cn