一种实现高效副本发布与查询的DHT覆盖网

李文翔[1,2] 熊庆国[1] 杨林涛[2]

(武汉科技大学信息科学与工程学院 武汉 430081)[1] (武汉大学电子信息学院 武汉 430079)[2]

摘 要 在对等计算应用中,副本复制技术是提升查询命中率、提升查询速度、维护负载均衡的一种有效方法,然而它也提升了在存储空间和流量上的代价。研究如何在结构化的P2P覆盖网中,通过拓扑优化手段减少复制中的冗余流量和冗余副本。首先在网络中选择支配集节点作为超级节点,设计一个层次化的、体现节点邻近度的P2P覆盖网,然后基于多hash函数,开发对应的复制技术以实现低代价的副本查询。该方法能够有效地在网络中分散副本,提升查询命中率,减少冗余消息和所需存储空间。给出了性能指标的理论分析,并通过仿真验证了该方法的优越性。

关键词 P2P,副本,覆盖网,冗余,Chord,支配集

中图法分类号 TP393 **文献标识码** A

DHT Overlay for Efficient Replicas Distribution and Query

LI Wen-xiang[1,2] XIONG Qing-guo[1] YANG Lin-tao[2]

(School of Information Science & Engineering, Wuhan University of Science & Technology, Wuhan 430081, China)[1]

(School of Electronics Information, Wuhan University, Wuhan 430079, China)[2]

Abstract Replication technique is an effective method for improving the availability of data, enhancing performance in terms of query latency and load balance, while replication also brings significant costs in storage space and traffic. We studied how to decrease the redundant traffic cost for replication in structured P2P overlay by topology optimization. We developed a new Hierarchical Proximity-Aware P2P overlay with dominating set nodes acting as super peers, and designed a corresponding replication technique with multiple hash functions for costless query. Our method can efficiently disseminate replicas across the network, increase query hit ratio, and decrease redundant query messages and storage spaces required. We gave theoretical analysis for performance metrics, and by simulation, we verified the superiority of our method.

Keywords P2P, Replica, Overlay, Redundancy, Chord, Dominating set

在对等通信技术的应用中,通过将资源的副本发布到网络各处,并将查询消息发送到最近的副本所在节点,获得副本的几率将大大提升。复制技术已成为提升数据可用率、减少平均查询时延的有效方法,然而复制也带来很大的运行开销,覆盖网和物理网的拓扑结构不一致会产生冗余的物理网查询消息,副本物理分布不合理(过于集中)会降低网络运行效率[1],以及额外的副本维护开销。需要开发合适的机制,在满足可用性和性能指标的前提下,减少存储空间和带宽的占用,减少冗余消息流量,优化副本的分布。

本文主要贡献为:1) 设计一种层次化的(Hierarchical)、体现邻近度的(Proximity-Aware)P2P覆盖网,记为HPA;2) 提出在一簇物理距离接近的节点中,而非单个节点上,放置特定资源的至多一个副本的思想,以避免覆盖网和物理网中的副本冗余;3) 在HPA覆盖网上,提出基于多Hash函数的副本发布与查询技术,以及新的ID分配方案,以实现高效的查询。

1 问题描述

离散hash表提供了高效的定位位置未知的数据的方法,但是基于Hash ID的路由未考虑实际的物理网路径是否也是最短的,一条查询消息可能多次穿过一条链路,产生不必要的流量。如图1所示,查询消息从节点N1发向N3,经过N2的转发,而实际的物理路径为b→d→e→a→e→d→c,并通过域间链路L多次,子网之间的网关节点可能承担较大的负载。

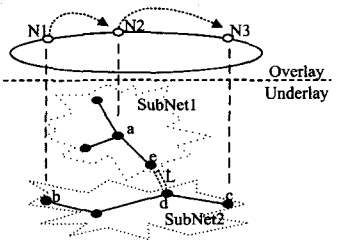

图1 冗余消息示意

另一方面,用基本的一致hash方案,节点和资源都被映射到1维的环空间上,资源的副本被分配到顺时针方向上的最近一个节点,这样就有可能将一个资源的两个副本分配到物理邻近的两个节点上,从而在局部区域中产生冗余副本。

到稿日期:2009-08-10 返修日期:2009-10-05 本文受武汉市重点科技攻关项目(200750499172)资助。

李文翔(1979—),男,讲师,主要研究方向为对等通信,E-mail:liwx2001@yahoo.com.cn;熊庆国(1956—),男,教授,主要研究方向为信号处理通信;杨林涛(1983—),男,博士生,主要研究方向为对等通信。

然而理想的情况应是副本在物理网中也是分散的。

2 相关工作

目前关于副本发布的方法有如下几类：1)将副本分散到邻近节点，在无结构[2,3]和结构化[4]覆盖网中都有实现；2)将副本放置在该资源的查询路径沿途节点上，这种方法既可用在无结构覆盖网中[5]，又可用在结构化覆盖网中；3)依系统性能要求(如最小时延、最大容量)，基于运筹优化理论，在无结构[6]和结构化[7]覆盖网中选择满足要求的节点放置副本；4)在结构化覆盖网中，指定多个 hash 函数，借助它们对一个资源得到相应数目的副本地址[8,9]。前两种方法按需创建副本，副本不能很快在网络中分布开来，且原资源拥有者需要专门记录这些副本的位置以进行副本更新。第三种方法一般需要全局的网络信息，以及各个节点的特征信息，这种方法难以适应网络的快速变化。第四种方法能主动地将副本分散在覆盖网的不同地方，且借助多个 hash 函数，资源所有者能够很方便地定位各个副本以进行更新。

本文的工作针对拓扑不匹配和副本分布不均匀这两个问题，从覆盖网构建、副本分布策略、副本查询策略这几个方面研究解决方法。特别之处在于提出在一簇物理位置接近的节点中至多放置一个副本的策略，实现物理网中副本分布的分散化；基于多 hash 函数开发副本查询技术，每次仅用一条消息寻找离当前节点最近的副本，实现了最小的查询代价。在传统的两层覆盖网的基础之上，用稳定的支配集节点构建超级节点层，以进行拓扑优化，提升覆盖网的路由效率。

3 系统设计

3.1 覆盖网体系

HPA 覆盖网包括两层，一层是由所有超级节点(SP)构成的一个 Chord 覆盖网，被称为 DSLayer，另一层包含多个被称为 PhyCluster 的 Chord 覆盖网，每一个 PhyCluster 包含一个 SP 和其邻近的普通节点(CP)。

图 2 所示为从传统的平面式 Chord 到 HPA 覆盖网的转换，图中标记为黑色的节点为 SP，标记为白色的节点为 CP，节点记为 N_{xy}，x 代表其所在的 PhyCluster，y 代表该节点在自己的 PhyCluster 内的索引。首先选择稳定的 d-hop 支配节点作为 SP，然后，每个 CP 选择最近的一个 SP，物理上邻近的节点组簇，最后，各个 PhyCluster 及 DSLayer 中形成 Chord 覆盖网。一个 PhyCluster 内的所有节点在物理网中的位置都相近，每个 SP 能够直接和该 PhyCluster 中的其它 CP 以较短的物理跳数通信。而在平面式 Chord 中，这些物理位置相近的节点可能依 Hash 值分散到覆盖网 ID 空间中的各个位置。

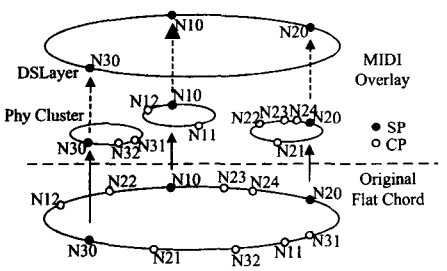

图 2 从平面式 Chord 到 HPA 覆盖网的转换

3.2 超级节点选择

定义 1 图 G 中顶点集 S 被称为 d-hop 支配集，如果图 G 中的每个顶点要么在 S 中，要么离 S 中的某个节点距离最多为 d 跳。

通过对节点的布尔运算操作[10]，可以得到不同 d 取值下的极小支配集。将 G 中的每个顶点 v_i 当作一个布尔变量，其 d-hop 邻节点集合表示为 $Adj(v_i)$，定义两个布尔表达式：

$$\varphi_i = v_i + \sum_{v_j \in Adj(v_i)} v_j \tag{1}$$

$$\psi = \varphi_1 \varphi_2 \cdots \varphi_n = \prod_{v_i \in V} \varphi_i \tag{2}$$

定义逻辑运算规则为：$v_i + v_i = v_i$，$v_i v_i = v_i$，$v_i(v_i + v_j) = v_i$，$v_i + v_i v_j = v_i$，式(2)的最终结果中每个乘积项即可作为一组 d-hop 支配集节点。例如，基于图 3 设 $d=1$，得到 $\varphi_1 = v_1 + v_2 + v_3 + v_4$，$\varphi_2 = v_1 + v_2 + v_4$，$\varphi_3 = v_1 + v_3 + v_4$，$\varphi_4 = v_1 + v_2 + v_3 + v_4 + v_5 + v_6$，$\varphi_5 = v_4 + v_5 + v_6$，$\varphi_6 = v_4 + v_5 + v_6$，最终结果为 $\psi = \varphi_1\varphi_2\varphi_3\varphi_4\varphi_5\varphi_6 = v_1v_5 + v_1v_6 + v_4 + v_2v_3v_5 + v_2v_3v_6$，所得极小 d-hop 支配集为 $\{v_1, v_5\}$，$\{v_1, v_6\}$，$\{v_4\}$，$\{v_2, v_3, v_5\}$，$\{v_2, v_3, v_6\}$。

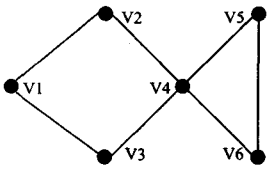

图 3 SP 选择示例图

将节点 v_i 的稳定度(用节点能力、在线时间等来衡量)表示为 $St(v_i)$，定义每个候选的极小支配集 $DS(j)$ 的稳定度为：

$$St(DS(j)) = \frac{1}{K}\sum_{i=1}^{K} St(v_i), K = Sizeof(DS(j)) \tag{3}$$

所得的最稳定支配集即作为 SP 集合，其数目记为 N_{SP}。

3.3 副本发布方法

令 h_0 为用于节点定位的 Hash 函数，它将节点映射到 DSLayer 和 PhyCluster 上，其中 DSLayer 上的路由表规模将远小于平面式 Chord 上的，每个 SP 维护 DSLayer 上的路由表，并将其所在 PhyCluster 上的成员信息记录到 CPList 中，每个 CP 记录其 PhyCluster 上的 SP 信息。

因为节点依据位置聚类为 PhyCluster，所以不同于以单个节点为单元，以 SP 代表的 PhyCluster 成为放置副本的基本单元，在一簇位置相近的节点中，没有 2 个以上的副本存在，这样就消除了一个区域中的多余副本，使各个副本在物理网络中分散。每个副本有两个 ID 地址，一个是其在 DSLayer 上的映射地址，n 个一致认可的 Hash 函数($h_1, h_2, h_3, \cdots$)被用来创建一个资源的 n 个副本的地址，而另一个地址是其在 PhyCluster 上的映射地址，这里只需要一个 Hash 函数(h_0)来计算地址，因为在 PhyCluster 中最多只有特定资源的一个副本。

假设每个资源具有相同的最大副本数目，记为 R_m。对资源 O_i 的第 j 个副本，其在 DSLayer 和 PhyCluster 上的地址分别表示为 $ID(h_j(O_i))$和 $ID(h_0(O_i))$，那么在 $DSLayer$ 中顺时针方向离 $ID(h_j(O_i))$最近的 SP 就代表该副本所在的 PhyCluster，而在 PhyCluster 中顺时针方向离 $ID(h_0(O_i))$最近的节点则是该副本的存放节点。

3.4 副本查询方法

因为可以借助 Hash 函数定位覆盖网中的最近副本，所

以本方法采用的查询策略是一次仅用一条消息查找一个最近的副本，若找不到则从当前查询所在节点出发，通过计算其它副本到当前节点的距离，选择下一个最近的副本再次查找，而不必同时发出多个查询查找多个副本，且不需和查询的最初发起者进行沟通，这样冗余消息的数目最少。当 SP 收到发向其所在 PhyCluster 的查询，它直接把查询发向对应的 CP，而不再在 PhyCluster 内多跳转发，因此 PhyCluster 内的冗余消息减少到最小。

具体的查询过程为：首先查询被送到 DSLayer，之后查询所在节点(*curNode*)从当前可用的 Hash 函数之中选择一个对应最近副本的函数(*curHash*)，可用的 Hash 函数信息封装在查询的数据包中，之后查询逐跳在 DSLayer 上传递。当查询发送到一个 SP 上时，对此 SP 进行判断，如果发现 curHash 对应的副本还在前方，需确定下一跳 SP 继续传递查询；如果查询消息已经跳过目标副本所在 PhyCluster 对应的 SP，其原因可能是那个 SP 不在线，此时 *curNode* 应移除消息中无用的 *curHash*，并选择离 *curNode* 最近的副本所对应的 Hash 函数；如果查询消息已经送到目标 SP，则该目标 SP 将把查询直接送到目标 CP，如果目标 CP 不在线，目标 SP 将移除消息中无用的 *curHash*，并选择离 *curNode* 最近的副本所对应的 Hash 函数展开下一轮查询。

4 性能分析

设网络中节点总数为 N，分析本文提出的方法在动态网络中的性能(包括查询成功率、覆盖网与物理网各自的查询跳数、所需存储空间)，并将其与文献[8,9]中提出的平面式 Chord 中的多 Hash 复制技术进行比较。

4.1 存储空间

平面式 Chord 的路由表项数为 $\log_2 N$，整网用于定位的信息存储空间表示为 $S_{flat}=N\log_2 N$，而在 HPA 覆盖网中，每个 SP 维护路由表、*CPList*，每个 CP 维护其 SP 的信息，令第 i 个 PhyCluster 所含的节点数为 $Size_i$，所需的对应存储空间为：

$$S_{HPA}=\sum_{i=1}^{N_{SP}}(\log_2 N_{SP}+Size_i)+\sum_{i=1}^{N-N_{SP}}(1)$$
$$=N_{SP}\log_2 N_{SP}+2N-N_{SP} \quad (4)$$

除非 $N_{SP}\to N$，S_{HPA} 将远小于 S_{flat}，而随着 N_{SP} 或 N 增长，S_{HPA} 单调递增。

4.2 查询跳数

假定查询路径经过的节点为$\{N_1,N_2,N_3,\cdots,N_n\}$，则覆盖网中的查询跳数为 $n-1$，物理网中的实际跳数为$\sum_{i=1}^{n-1}d_{i,i+1}$。

当 $IR_{flat}=0.0$ 时，在平面式 Chord 中的最大查询跳数为 $Hop_{flat}=\log_2 N$。而在 HPA 覆盖网中，查询跳数为 CP 到 SP 的一跳(如果资源请求者为 CP，对应的概率为 $1-N_{SP}/N$)，加 DSLayer 中的跳数，加目标 SP 到目标 CP 的一跳(如果副本在 CP 上)，故最大跳数为：

$$Hop_{HPA}=\frac{N_{SP}}{N}\cdot\frac{N_{SP}}{N}\cdot(0+\log_2 N_{SP}+0)+\frac{N_{SP}}{N}\cdot(1-\frac{N_{SP}}{N})\cdot(0+\log_2 N_{SP}+1)+(1-\frac{N_{SP}}{N})\cdot\frac{N_{SP}}{N}\cdot(1+\log_2 N_{SP}+0)+(1-\frac{N_{SP}}{N})\cdot(1-\frac{N_{SP}}{N})\cdot(1+\log_2 N_{SP}+1)=\log_2 N_{SP}+2-2N_{SP}/N \quad (5)$$

显然 Hop_{HPA} 比 Hop_{flat} 小很多，当 $N_{SP}<N/2$ 时，Hop_{HPA} 随 N_{SP} 或 N 增加而单调递增。

4.3 查询成功率

因为节点频繁地加入退出网络，一些查询可能无法送至目标节点，所以在不考虑对网络中关键节点故意攻击的情况下，采用节点随机失效模型。假设 CP 失效的概率为 IR，SP 失效的概率为 IR_{SP}，因为 SP 层由最稳定的极小支配集节点构成，所以 $IR_{SP}<IR$。在 R_m 个副本中，其中有 i 个被放到 SP 上的概率为 $C_{R_m}^{i}\prod_{m=0}^{i-1}\frac{N_{SP}-m}{N-m}\cdot\prod_{n=0}^{R_m-i-1}\frac{N-N_{SP}-n}{N-n}$，所有放置这些副本的节点(包括 SP 和 CP)都失效的概率为 $IR_{SP}^{i}IR^{R_m-i}$，对一个查询，基于 HPA 覆盖网和本文的查询算法的查询成功率为：

$$P_{HPA}=1-\sum_{i=0}^{R_m}C_{R_m}^{i}\left(\prod_{m=0}^{i-1}\frac{N_{SP}-m}{N-m}\right)IR_{SP}^{i}\cdot\left(\prod_{n=0}^{R_m-i-1}\frac{N-N_{SP}-n}{N-n}\right)IR^{R_m-i} \quad (6)$$

假设平面式 Chord 中失效节点总数与 HPA 覆盖网中的相同，有 $IR_{SP}\cdot N_{SP}+IR\cdot(N-N_{SP})=IR_{flat}\cdot N$，于是平面式 Chord 中的节点失效概率为 $IR_{flat}=(IR_{SP}\cdot N_{SP}+IR\cdot(N-N_{SP}))/N$，对应在平面式 Chord 中的查询成功率为：

$$P_{flat}=1-\left(\frac{N_{SP}}{N}IR_{SP}+\frac{N-N_{SP}}{N}IR\right)^{R_m}$$
$$=1-\sum_{i=0}^{R_m}C_{R_m}^{i}\left(\frac{N_{SP}}{N}IR_{SP}\right)^{i}\cdot\left(\frac{N-N_{SP}}{N}IR\right)^{R_m-i} \quad (7)$$

比较式(6)和式(7)，因为 $\frac{N-N_{SP}-n}{N-n}\leqslant\frac{N-N_{SP}}{N}$，$\frac{N_{SP}-m}{N-m}\leqslant\frac{N_{SP}}{N}$，显然 $P_{HPA}>P_{flat}$。

5 仿真

5.1 仿真设置

用 Inet 3.0 生成一个含 6000 节点的网络拓扑，基于 C++ 开发仿真程序，其中包括：HPA 覆盖网构建、副本发布、失效节点随机选择、8000 个随机查询事件生成和查询执行。以 R_m，d 和 IR 作为变量，计算上述的性能指标，并进行平面式 Chord 和 HPA 覆盖网之间的性能比较，研究指标随参数变化的规律。

5.2 仿真结果与分析

首先，基于拓扑得到 d 分别为 1,2,3 时的稳定支配集节点集合，构建 HPA 覆盖网以及平面式 Chord，*SP* 集合大小和用于定位所存储的节点信息条目如表 1 所列。可以看出平面式 Chord 所需存储的节点信息大约是 HPA 覆盖网的 6 倍，在 HPA 覆盖网中，随着 N_{SP} 增大，所需节点条目略微增加。

表 1 用于路由的节点信息条目

Overlay	Nsp	Route Table	CPList	SP for each CP	S
HPA,d=1	35	189	6000	6000-35	12154
HPA,d=2	47	280	6000	6000-47	12233
HPA,d=3	71	457	6000	6000-71	12386
Chord	6000	77072	0	0	77072

接着，获得 IR 分别为 0.1 和 0.2 时的查询命中率，这两种情况下的结果基本一致，从图 4 可以看出，HPA 覆盖网的查询命中率总是高于平面式 Chord 中的，这归因于稳定的 DSLayer，随着 R_m 增大($R_m\geqslant 4$)，查询命中率很快接近 100%，且 d 越大，查询命中率相对越高。

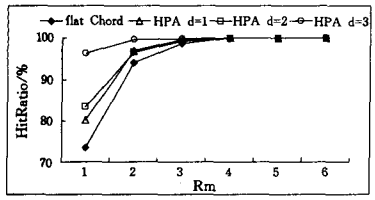

图4 不同 IR 和 d 下的查询命中率（IR=0.2）

然后，从单个查询的视角分析，得到8000个查询事件在 d=1时覆盖网和物理网的查询跳数，如图5和图6所示。d=2和3时的结果和结论类似，可以看出：1）物理网和覆盖网的查询跳数曲线的分布和相对位置是基本一致的；2）给定 IR，无论是物理查询跳数还是覆盖网中的查询跳数，在平面式 Chord 中的要比 HPA 覆盖网中的高很多；3）除去 R_m=1 的情况，对每根曲线而言，当 R_m=6 时可以获得最小的查询跳数；4）当 IR=0.0 时，查询跳数随 R_m 增大而单调减小；而当 IR=0.1 或 0.2 时，在 R_m=2 或 3 时获得查询跳数的最大值，这意味着当某个副本未被找到时，查询继续进行，查询消息将被发送到下一个离当前节点最近的副本提供者，因此查询跳数增加；当 $R_m \geqslant 4$ 时，所需平均查询跳数渐渐减小，这是因为可用副本增多，副本查找失败的几率相对下降。与 R_m=1 的情况相比，基于多 hash 函数的复制技术（R_m>1）能获得更高的查询命中率，不过会带来更长的查询路径。

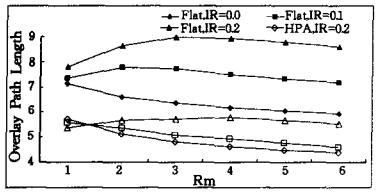

图5 覆盖网中的查询跳数（d=1）

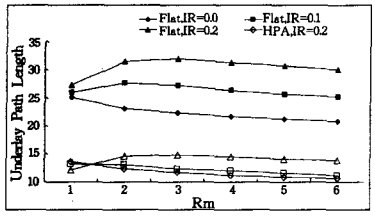

图6 物理网中的物理查询跳数（d=1）

结束语 本文讨论了一个结构化的、实现了拓扑优化的覆盖网，以及在此覆盖网中基于多 Hash 函数的复制发布与查询策略。本文提出的方法能有效地减少冗余流量，合理地发布副本，通过理论分析和仿真，证实了这些方法的有效性。深入的工作包括考虑节点的动态行为、负载均衡，对路由方法展开优化，在保证现有性能的基础上加快查询速度，以及设计分布式的 SP 选择算法等。

参 考 文 献

[1] Chen G H, Li Z H. Peer-to-Peer network structure, application and design[M]. Beijing: Tsinghua University Publishing House, 2007

[2] Wan K H, Chris L. An overlay network for replica placement within a P2P VoD network[J]. International journal of high performance computing and networking, 2005, 3(5/6): 320-335

[3] Mondal A, Madria S K, Kitsuregawa M. EcoRep: an economic model for efficient dynamic replication in mobile-P2P networks [C]// Proceedings of 13th international conference on management of data. Dec. 2006: 185-196

[4] Min C, Ann C. A peer-to-peer replica location service based on a distributed hash table[C]// Proceedings of ACM/IEEE conference on supercomputing. Nov. 2004: 56-67

[5] Lv Q, Li K, Cao P, et al. Search and replication in unstructured peer-to-peer networks[C]// Proceedings of the 16th international conference on supercomputing. New York, USA, June 2002: 84-95

[6] Zhou J, Zhang X. Clustered K-Center effective replica placement in peer-to-peer systems[C]// Proceedings of Global telecommunications conference 2007. Nov. 2007: 2008-2013

[7] Bassam A A, Chen W, Zhou B B, et al. Effects of Replica Placement Algorithms on Performance of structured Overlay Networks[C]// Proceedings of IEEE International Parallel and Distributed Processing Symposium. 2007: 460-467

[8] Marcel W, Paul H, Daniel B. Dynamic replica management in distributed hash tables[R]. RZ-3502. IBM, July 2003

[9] Xia Y, Chen S G, Cho C. Algorithms and performance of load-balancing with multiple hash functions in massive content distribution[J]. Computer Networks, 2009, 53(1): 110-125

[10] Lu K C, Lu H M. Graphics Theory and its Applications[M]. Beijing: Tsinghua University Press, 2005

（上接第61页）

参 考 文 献

[1] Larzon L-A, Degermark M, Pink S, et al. The Lightweight User Datagram Protocol (UDP-Lite)[S]. RFC 3828, Internet Engineering Task Force, July 2004

[2] Zhang X, Li Ch, Li X, et al. Multicast congestion control on many-to-many videoconferencing[C]// The Second International Conference on Future Generation Communication and Networking. 2008: 260-263

[3] Zhang Zaichen, Li O K. Network-Supported Layered Multicast Transport Control for Streaming Media[J]. IEEE Trransactions on PAarallel and Distributed Systems, 2007, 18(9)

[4] Vanit-Anunchai S, Billington J, Kongprakaiwoot T. Discovering chatter and incompleteness in the Datagram Congestion Control Protocol[C]// Proc. 25th IFIP International Conference on Formal Techniques for Networked and Distributed Systems. Oct. 2005

[5] Hong Yang, Yang W W. Design of Adaptive PI Rate Controller for Best-Effort Traffic in the Internet Based on Phase Margin [J]. IEEE Transactions on Parallel and Distributed Systems, 2007, 18(4)

[6] Voice T. Stability of Multi-Path Dual Congestion Control Algorithms[J]. IEEE/ACM Transactions on Networking, 2007, 15 (6)

[7] Kamal A E, Sankaran S. A Combined Delay and Throughput Proportional Scheduling Scheme for Differentiated Services[J]. Journal Computer Communications, 2006, 29: 1754-1771

[8] IEEE Std 802. 1ad™-2005[S]. IEEE Standard for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks-Amendment 4: Provider Bridges. 2005

[9] 李士勇. 模糊控制. 神经控制和智能控制论[M]. 哈尔滨：哈尔滨工业大学出版社，1996：42-43