ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.7, July 2009, pp.1943−1952
doi: 10.3724/SP.J.1001.2009.03307

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于重复链路检测的 P2P 网络拓扑一致性方案*

于 婧+, 汪斌强

(国家数字交换系统工程技术研究中心,河南 郑州 450002)

Links Overlapped Detecting Based Scheme to Make P2P Network Topology-Aware

YU Jing+, WANG Bin-Qiang

(National Digital Switching System Research Center, Zhengzhou 450002, China)

+ Corresponding author: E-mail: yujing@mail.ndsc.com.cn

Yu J, Wang BQ. Links overlapped detecting based scheme to make p2p network topology-aware. *Journal of Software*, 2009,20(7):1943−1952. http://www.jos.org.cn/1000-9825/3307.htm

Abstract: Topology mismatching between the overlay network and physical network is a main factor which affects the routing performance of structured P2P network. A detecting and decreasing links overlapped scheme (DDL) is proposed. It examines the overlapped physical links caused by overlay routing, and on the appropriate condition, sends the redirect messages to decrease the physical links crossed. DDL solves the topology mismatching problem on the physical network level, and it can be used in any structured P2P network without the limitation of overlay structures. According to different definitions of links overlapped, backward and forward DDL schemes are described in detail. Through performance analysis and simulation, DDL scheme can dramatically improve the topology consistency.

Key words: peer-to-peer network; overlay network; topology aware; redirect mechanism

摘 要: 结构化 P2P 覆盖网络与实际物理网络的拓扑不匹配问题是影响结构化 P2P 网络路由性能的重要因素.提出了检测并降低重复链路使用的拓扑一致性解决方案 DDL(detecting and decreasing links overlapped scheme).DDL 从实际物理网络路由出发,通过检测覆盖网络三点路由经历的实际物理链路重复利用的情况,在适当的条件下,通过重定向报文的发送,降低路由经历的物理链路数.根据不同的重复链路的定义,给出了后向和前向两种检测重复链路的方法.DDL 是一种从物理链路层面解决 P2P 网络拓扑一致性问题的方法,可以使用在任何结构化 P2P 网络中,不受限于覆盖网络层的组织方式.性能分析及仿真实验结果表明,使用 DDL 方案能够显著提高覆盖网络与物理网络的拓扑一致性.

关键词: 对等网络;覆盖网络;拓扑一致性;重定向机制

中图法分类号: TP393 **文献标识码**: A

结构化 P2P 网络[1−4]的路由是在覆盖网络层面进行的,而覆盖网络是建立在实际物理网络之上的虚拟网络,拓扑的构建没有考虑底层物理网络的实际情况,节点之间的连接不等同于实际的物理连接,很有可能在覆盖网络上的两个相邻节点在实际物理网络上相距很远.路由转发的过程最终要落到实际物理网络层进行,由此导致

* Supported by the National Basic Research Program of China under Grant No.2007CB307102 (国家重点基础研究发展计划(973))
Received 2008-01-14; Accepted 2008-03-06

P2P 节点资源定位开销的增加,这个问题称为拓扑不匹配[5](topology mismatching).对拓扑不匹配问题进行研究和改进将会降低网络流量,提高路由效率.

本文基于拓扑一致性问题研究,第 1 节分析当前拓扑一致性问题的解决方案及其优缺点.第 2 节根据覆盖网络层路由引发实际物理网络路由产生物理链路重复使用的问题,以实际网络路由为基本出发点,提出检测并降低重复链路使用的拓扑一致性解决方案(detecting and decreasing links overlapped,简称 DDL)的基本思想.根据重复链路定义的不同,分别于第 3 节、第 4 节给出了后向和前向两种重复链路检测机制的具体步骤,并对之进行性能分析及比较.最后是全文总结.

1 相关工作

解决拓扑不匹配问题一般分为两个步骤[6,7]:首先要获取节点的临近信息;然后根据节点的临近信息调整覆盖网络构造,达到覆盖网络与实际物理网络拓扑一致.

1.1 获取节点的临近信息

当前,针对拓扑不匹配问题,出现了很多获取节点临近信息的解决方案,总的来说可以归结为以下两种.

1.1.1 基于 IP 地址的解决方案

基于 IP 地址的解决方案是基于当前 Internet 网络 IP 地址分配原则,认为 IP 地址处于同一个网段的节点在物理位置上是相近的.Krishnamurthy 提出的 Network-Ware Clustering[8]技术通过获取路由器 BGP 更新报文,从中获取处于同一网段或同一自治域下的节点信息,并将节点连接成簇.然而,BGP 路由表信息不易获取,并且基于自治域或 IP 网段的簇都是粗略估计.另外,随着网络中防火墙及 VLAN 的广泛使用, 限制了基于 IP 地址解决方案的应用.

1.1.2 基于节点间时延的解决方案

基于节点间时延的解决方案的基本原则认为,节点间传输时延越短意味着节点在物理位置上越近.目前,解决拓扑不匹配问题的多数方案都是基于测量节点间时延的.主要的有以下几种:

Ratnasamy 提出的 Binning[9]方案将一些知名服务器作为路标(landmarking),每个 P2P 节点测量自己与所有路标的往返时间(round trip time,简称 RTT),形成一个 RTT 向量.依据向量相近的节点物理位置也相近的思想,将向量相近的节点组织到一个 bin 中.该方案虽然实现简单,但对路标的位置有一定的要求,而且大量的时延测量会使路标成为网络的瓶颈.

由此,Winter 提出了 Random landmarking[10]方案.在该方案中,landmarking 不再是固定的服务器,而是一些指定的节点标识符对应的节点.由于 P2P 网络节点的动态性,在任意两个时刻,节点标识符相同的节点并不一定是同一个节点,从而实现了 random landmarking.而也正是由于 landmarking 的动态性,在任意两个时刻同一个节点计算的 RTT 向量也可能是不同的,因此该方案需要较准确的时间同步.

Liu 提出位置相关拓扑匹配(location-aware topology matching,简称 LTM)[11]算法.LTM 算法基于 *TTL* 洪泛,获取 *TTL* 范围内邻居 IP 地址及延迟信息,根据延迟信息动态删减无效和冗余的连接,将离源节点延迟最小的节点作为源节点的直接邻居.LTM 技术能够分布式地构建与底层物理网络匹配的覆盖网络拓扑,但也引入了一定的开销,并需要 P2P 系统中的每个节点始终保持同步.

基于时延的解决方案最大的缺陷在于节点间时延并不正比于节点间物理距离.如节点间通过拨号连接或卫星连接,虽然时延较大,但并不代表两点间的距离一定较远.这也是限制基于时延的解决方案的一个重要问题.

1.2 根据节点的临近信息调整覆盖网络构造

获取到节点的临近信息之后,根据节点的临近信息动态地调整覆盖网络的构造存在两种解决方案.

1.2.1 节点标识符生成或调整

节点标识符生成方案[12]是指节点在进入系统前获取临近信息,并据此生成节点标识符进入系统.节点标识

符调整是指节点已处于系统内,在获取到临近信息后,根据临近信息动态调整节点标识符,使得物理临近的节点在覆盖网络层面上也是临近的.节点标识符调整的一种具体应用见文献[6].节点间通过节点标识符交换(swap)重建覆盖网络.然而,在对节点标识符进行调整的同时,需要对节点路由表、邻居表等存储内容进行交换及重新整合,在交换频繁时,无疑,极大地增加了网络负担.

1.2.2 修改路由表或邻居表信息

在获取临近信息之后,节点可以把与自己物理相近的节点加入到路由表或邻居表中,也就是相当于添加了新的基于物理拓扑的连接信息.在路由的过程中,首先考虑邻居节点,从而以另外一种方式达到了重构覆盖网络的目的.修改路由表或邻居表信息的方案比调整节点标识符更简单、高效,避免了节点标识符调整或互换引发的一系列节点信息的传输,节省了网络带宽.

2 重复链路检测工作原理

2.1 设计思路

基于节点间时延的解决方案以降低实际路由的时延为目标,而没有根据实际 Internet 网络拓扑构造考虑实际物理网络路由的具体情况.在实际物理网络中,节点间通过物理链路连接,物理链路的使用情况直接影响网络的性能.DDL 正是基于实际 Internet 网络,通过降低覆盖网络路由导致的实际物理链路的重复使用,达到解决拓扑不匹配问题的目的.

如图 1 所示,上层是覆盖网络层,下层是实际物理网络层.在覆盖网络层,节点 $N1$ 发起对 $N4$ 的路由,路径是 $N1\rightarrow N2\rightarrow N3\rightarrow N4$,对应的实际物理网络层的路径则是 $N1\rightarrow R1\rightarrow R2\rightarrow N2\rightarrow R2\rightarrow R1\rightarrow N3\rightarrow R1\rightarrow R3\rightarrow N4$.显而易见,链路 $R2\leftrightarrow R1$ 在此次路由中被途经了两次,浪费了网络的带宽.特别是在节点间是跨骨干网络连接的情况下,不匹配问题将会导致大量的骨干网链路因 P2P 路由而重复使用,极大地浪费了网络资源.因此,若能降低路由中物理链路的重复使用,就可以节省网络带宽,提高路由效率,同时解决覆盖网络与物理网络的拓扑不匹配问题.

图 2 是对图 1 进行重复物理链路检测后的路由示意图.可以看出,优化后的路由 $N1\rightarrow R1\rightarrow N3\rightarrow R1\rightarrow R3\rightarrow N4$.经历的物理链路数显著降低,降低路由经历的物理链路数也就缩短了查询时延.

实现检测并降低重复链路使用的拓扑一致性方案 DDL 的首要问题就是要进行重复链路检测,只有检测到路由过程中存在重复链路的问题,才能采取相应的措施降低重复链路的使用.

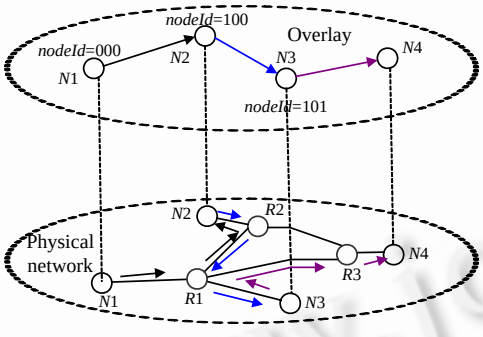

Fig.1 Physical links overlapped illustration

图 1 物理链路重复使用示意图

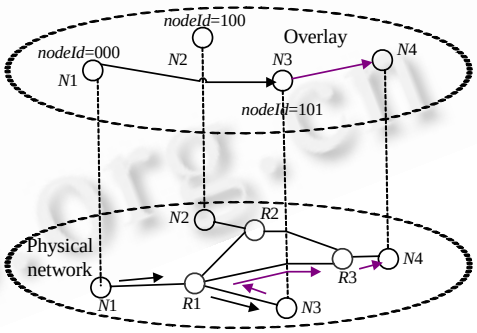

Fig.2 Optimized routing after DDL

图 2 链路重复检测后的优化路由示意图

2.2 工作原理

重复链路检测机制是基于 ICMP(Internet control message protocol)重定向机制[13]原理得到的.ICMP 重定向机制的工作原理是当路由器进行报文转发时,根据报文目的地址查找下一跳的地址,若发现下一跳与报文的源处于同一个子网,则向源发送重定向(redirect)消息,同时正常转发该报文到下一跳.重定向消息的目的是告诉主

机到达目的地的最短路径,下次发送到此目的地的报文直接走这条路径.ICMP 重定向机制保证主机去往目的地的路径是最优的.

由此得到重复链路检测的基本工作原理:当节点收到查询请求时,首先根据覆盖网络层路由表查找下一跳节点,并将查询请求向下一跳转发;与此同时,比较查询请求到来的网络层路径与发往下一跳的网络层路径,根据这两条路径途经的相同链路的比例来决定是否发送重定向消息到发来查询请求的上一跳节点.根据重复链路定义的不同,重复链路检测分为前向和后向两种,下面将详细介绍两种重复链路检测的具体步骤及性能.

3 后向重复链路检测

后向重复链路检测是指当前节点对比查询请求到来的网络层路径的后端与它发往下一跳的网络层路径前端路径,也就是临近当前节点部分的路径,计算这两条路径存在相同路径的比例,据此决定是否发送重定向消息.如图 3 所示,路径 $S\rightarrow N$ 与 $N\rightarrow D$ 经历相同的路径 $M\leftrightarrow N$.

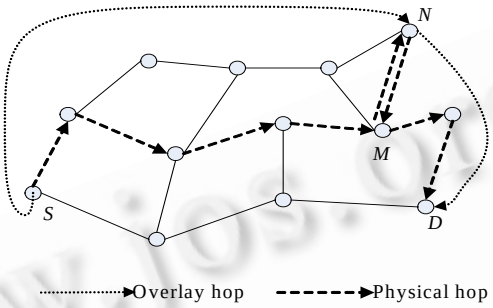

Fig.3 Backward DDL illustration

图 3 后向重复链路检测示意图

3.1 符号描述

在具体描述之前,先给出一些相关定义:

- 源节点S,当前节点N,目的节点D;

这里定义的源节点和目的节点并不是查询请求实际的发起节点和最终目的节点,而是针对当前节点而言查询请求到来的上一跳节点和转发到的下一跳节点.当前节点是指当前处理查询请求的节点.

- 源节点到当前节点经历的网络层路径($P_{S\rightarrow N}$);
- 当前节点到目的节点的网络层路径($P_{N\rightarrow D}$);
- $P_{S\rightarrow N}$和$P_{N\rightarrow D}$的重复路径P_{shared};

以TTL作为衡量网络层路径长度的参数,则有

- $P_{S\rightarrow N}$,$P_{N\rightarrow D}$,P_{shared}的长度:$TTL_{S\rightarrow N}$,$TTL_{N\rightarrow D}$,TTL_{shared};
- 重复路径系数$l_{overlap}$:

$$l_{overlap}=\frac{TTL_{shared}}{TTL_{S\rightarrow N}} \tag{1}$$

3.2 执行步骤

后向重复链路检测的实现需要完成以下 5 个方面:

(1) 当前节点获取从源节点 S 到当前节点 N 的网络层路径

从源节点到当前节点路由的过程中要记录途经节点的地址信息.在查询请求到达当前节点时,当前节点能够从查询请求报文中获取从源节点 S 到当前节点 N 的网络层路径 $P_{S\rightarrow N}$ 以及对应的 $TTL_{S\rightarrow N}$.

(2) 获知从本节点 N 到下一跳节点的网络层路径

当前节点 N 收到查询请求后,根据查询请求中的目标键值 K 进行覆盖网络层面的路由,找到下一跳的地址,

并封装成新的查询请求报文发送至下一跳节点.与此同时,当前节点 N 以目的节点 D 为目标,以 $TTL_{S\to N}$ 为生存期探测当前节点 N 到目的节点 D 的网络层路径信息,记作 $P'_{N-\to D}$.$P'_{N\to D}$ 的路径长度等于 $TTL_{S\to N}$.

(3) 对这两条路径进行比较,决定是否发送重定向消息

此时,当前节点拥有 $P_{S\to N}$ 和 $P'_{N\to D}$ 两条路径的信息.节点 N 反转 $P_{S\to N}$ 得到 $P_{S\to N}$ 的反转路径 $P_{S\to N}^{reverse}$.比较 $P_{S\to N}^{reverse}$ 与 $P'_{N\to D}$,得到两条路径的最长匹配路径 P_{shared} 和对应的路径长度 TTL_{shared},根据式(1)计算重复路径系数 $l_{overlap}$.

定义 $l_{overlop}^{thresh}$ 为重复路径系数 $l_{overlap}$ 的门限值.若 $l_{overlop}\geq l_{overlop}^{thresh}$,则说明两条路径重复率较高,当前节点 N 向源节点 S 发送重定向消息,否则不予发送.

(4) 重定向消息的内容

当前节点向原节点发送重定向消息,用于指示源节点对于该目标键值的查询不需要再经过当前节点,而直接发送到目的节点 D 即可.因此,重定向消息中要包含目的节点 D 的地址信息.重定向消息的基本格式见表 1.

Table 1 Redirect message format

表 1 重定向消息报文格式

Redirect message ID	Dst nodeId	Dst IP address
Original searching request message		

(5) 上一跳节点收到重定向消息后的操作

DDL 方案中采用修改路由表或邻居表信息的方法重构覆盖网络.源节点 S 在收到当前节点 N 发来的重定向消息时,更新目的键值 K 对应的路由表项.

DDL 实施中记录沿途节点的地址信息以及处理路径探测消息可能会对节点造成压力,这里考虑只在主要节点,如边界路由器上进行操作.记录地址信息与处理探测消息路由器的分布及要达到的拓扑一致性的程度相对应.分布越密集,得到的拓扑一致性程度越高.当前很多路由器,如思科,都具有 DPI(deep packet inspection,深度包检测)功能,可以对 P2P 报文进行有效区分,从而也可以进行简单的处理.

3.3 性能分析

3.3.1 链路使用

重复链路检测的主要目的是降低路由过程中链路的重复利用,因此,该方案性能的评价标准就是方案实施前后链路使用的对比.定义

- 原来经历的链路数为TTL_{old},则

$$TTL_{old}=TTL_{S\to N}+TTL_{N\to D} \tag{2}$$

- 经过重复链路检测后经历的链路数为TTL_{new}:

$$TTL_{new}=TTL_{S\to D}\leq TTL_{S\to N}-TTL_{shared}+TTL_{N\to D}-TTL_{shared}=TTL_{old}-2\times TTL_{shared} \tag{3}$$

- 两者的比值称为链路使用优化率,记作R,则

$$R=\frac{TTL_{old}}{TTL_{new}} \tag{4}$$

由上述 3 式,可以得到 R 的最终表达式:

$$R=\frac{TTL_{old}}{TTL_{new}}\geq\frac{TTL_{S\to N}+TTL_{N\to D}}{TTL_{S\to N}+TTL_{N\to D}-2\times TTL_{shared}}=\frac{1+\gamma}{1+\gamma-2\times l_{overlap}} \tag{5}$$

其中,

$$\gamma=\frac{TTL_{N\to D}}{TTL_{S\to N}}.$$

由于 $TTL_{N\to D}\geq TTL_{shared}$,故 $\gamma\geq l_{overlap}$.

由式(5)可以看出,链路使用优化率 R 的大小取决于重复路径系数 $l_{overlap}$ 及两段路径长度的比值γ的大小.由式(5)可以看出,随着 $l_{overlap}$ 的增大,R 是不断增大的;而当 $l_{overlap}$ 固定时,γ的增大则会导致 R 的降低;当 $l_{overlap}=\gamma$

时,R 达到对应的最大值,当 $l_{overlap}=\gamma=1$ 时,R 极值趋向无限大.因此,要想链路使用优化率有所提高,除了 $l_{overlap}$ 值要大,也就是重复链路比例增大以外,还要求两段路径长度的比值接近 $l_{overlap}$.根据上述分析,提出了如下修正的重定向报文发送条件:

(1) $l_{overlap} \ge l_{overlap}^{thresh}$;

(2) $\gamma - l_{overlap} \le \varepsilon$,

则重复链路检测实施中步骤(2),(3)重新描述为:

(1) 获取 $P'_{N\to D}$ 时,最大跳数设置为 $(1+\varepsilon)\times TTL_{S\to N}$;

(2) 当 $P'_{N\to D}$ 第 1 次出现与 $P_{S\to N}^{reverse}$ 不相同的路径时,计算 $l_{overlap}$;

(3) 若 $l_{overlap} \ge l_{overlap}^{thresh}$,转到步骤(4);否则,停止探测;

(4) 继续增大 TTL,若跳数小于或等于最大跳数 $(1+\varepsilon)\times TTL_{S\to N}$,返回目的节点信息,说明满足发送重定向报文发送条件(2),发送重定向报文;否则,认为不满足,这时即使满足条件(1)也不发送重定向报文.

引入上述发送重定向报文的条件,由式(5)推导得到:

$$R \ge \frac{1+\gamma}{1+\gamma-2\times l_{overlap}} \ge \frac{1+l_{overlap}+\varepsilon}{1-l_{overlap}+\varepsilon} \tag{6}$$

由式(6)R 与 $l_{overlap}$ 及 ε 的关系可以看出,在 $l_{overlap}$ 相同的情况下,随着 ε 的增大,R 不断降低.在方案实施的过程中,$l_{overlap}$ 和 ε 的取值对系统性能起着非常重要的作用,需要在实际运行过程中积累经验,确定参量的取值.

3.3.2 消息数量

系统中产生的消息分为两种:一种是探测 $P'_{N\to D}$ 引起的,另外则是产生的重定向报文.重定向报文的数量也就是查询时满足发送重定向报文条件的查询数量,它取决于实际网络路由情况以及 $l_{overlap}$ 和 ε 的取值.

根据修正的重定向报文发送条件,若 $l_{overlap} < l_{overlap}^{thresh}$,对 $P'_{N\to D}$ 的探测产生的消息个数为 TTL_{shared};否则,消息个数为 $\min(TTL_{N\to D},(1+\varepsilon)\times TTL_{S\to N})$,因此,探测 $P'_{N\to D}$ 的报文数与 $l_{overlap}$,ε 以及两条路径的长度有关.

3.4 实验仿真

仿真中使用的网络是采用 GT-ITM[14,15]生成的较能代表当前 Internet 结构的穿通-末端(transit-stub)TS 模型的随机拓扑图[16,17],网络中有 1 024 个节点,节点间路径通过最短路径算法计算得到.

3.4.1 链路使用优化率

通过对 1 024 个节点构成的网络进行 10 000 次查询的结果进行统计,得到链路使用优化率 R 与 $l_{overlap}$ 和 ε 的关系曲线如图 4 所示.可以看出,实验值走向基本符合 R 与 $l_{overlap}$ 及 ε 的理论值关系曲线.实验结果显示了重复链路检测对链路使用率有显著的提高作用,比如在 $l_{overlap}=0.8$,$\varepsilon=0.2$ 时,R 达到了 9 以上.

3.4.2 消息数量

$l_{overlap}$ 和 ε 的取值情况直接影响到引发产生重定向报文的条件,从而影响系统中发送重定向报文的数量.发送重定向报文比例与 $l_{overlap}$ 及 ε 的实际测试关系曲线如图 5 所示.可以看出,在 ε 不变的情况下,随着 $l_{overlap}$ 的增大,重定向报文的发送数量有所降低;而在 $l_{overlap}$ 相同时,ε 的增大则会使得发送重定向报文的数量有所增加.

重复链路检测产生的平均消息数与 $l_{overlap}$ 及 ε 的实际测试关系曲线如图 6 所示.可以看出,在 ε 不变的情况下,随着 $l_{overlap}$ 的增大,消息数量有所降低;而在 $l_{overlap}$ 相同时,ε 的增大则会使得消息的数量有所增加.

综上所述,从仿真实验得到的结果可以看出,随着 $l_{overlap}$ 的增大,ε 的减小,链路使用优化率 R 增大,而由于重复链路检测产生的消息数量降低,这与第 3.3 节中的理论分析相吻合.因此,在进行重复链路检测方案时,要采用较大的 $l_{overlap}$,较小的 ε 来达到较高的性能优化.在上面的实验中,当 $l_{overlap}=0.8$,$\varepsilon=0.2$ 时,$R=9.142857$,对应的平均消息数则为 1.544 6,系统的整体性能较高.

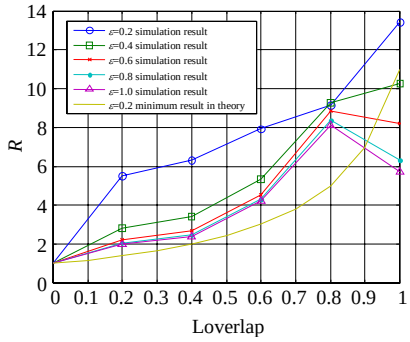

Fig.4 Simulation result of relation of R, $l_{overlap}$ and ε

图 4 R 与 $l_{overlap}$,ε关系的仿真结果

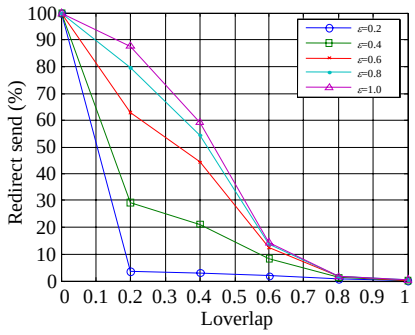

Fig.5 Simulation result of relation of redirect messages number, $l_{overlap}$ and ε

图 5 重定向报文数与 $l_{overlap}$,ε关系的仿真结果

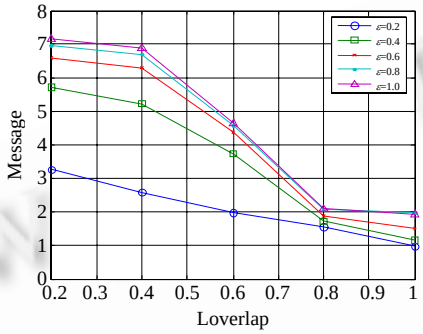

Fig.6 Simulation result of relation of average DDL messages number, $l_{overlap}$ and ε

图 6 DDL 平均消息数与 $l_{overlap}$,ε关系的仿真结果

4 前向重复链路检测

尽管后向 DDL 机制能够提高覆盖网络与物理网络的拓扑一致性,但其作用范围仍存在一定的局限性.如图 7 所示,后向 DDL 在检测到 $P_{S\to N}$ 与 $P_{N\to D}$ 在节点 N 就不具有共同路径,故认为该 3 点路由不存在重复链路.而从图 7 可以看出,在节点 M,$P_{S\to N}$ 与 $P_{N\to D}$ 重新进行了汇合.因此,为了提高重复链路检测性能,将重复链路的定义由后向 DDL 中往返于节点间完全相同的链路,如图 3 中所示的 $M\to N$ 和 $N\to M$,扩展为实际路由过程中两节点间的往返链路的集合,如图 7 中 $M\to R\to N$ 与 $N\to M$.

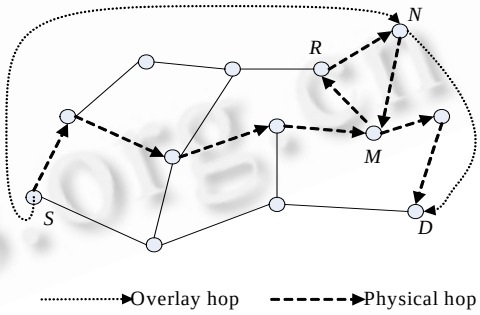

Fig.7 Forward DDL illustration

图 7 前向重复链路检测示意图

前向重复链路检测以扩展重复链路定义为基础.当前节点对比查询请求到来的网络层路径的前端与其发往下一跳的网络层路径后端路径,也就是临近源节点和目的节点部分的路径,计算这两条路径存在相同路径的比例,据此决定是否发送重定向消息.

4.1 符号描述

下面给出前向重复链路检测的符号描述:

- $P_{S\to N}$与$P_{N\to D}$离节点S和D最近的相同节点定义为M;

数学形式描述为:当 $P_{S\to N}=p_0\to p_1\to\ldots\to p_m\to M\to\ldots\to N, P_{N\to D}=N\to\ldots\to M\to q_0\to q_1\to\ldots\to q_n\to D$ 时,$\forall i\in[0,m]$, $j\in[0,n], p_i\neq q_j$;

- $P_{S\to N}$和$P_{N\to D}$的重复路径P_{shared};
- 网络层路径$P_{S\to M}, P_{M\to N}, P_{N\to M}, P_{M\to D}$,对应的路径长度为$TTL_{S\to M}, TTL_{M\to N}, TTL_{N\to M}, TTL_{M\to D}$.

4.2 执行步骤

- 当前节点N获取从源节点S到当前节点N的网络层路径$P_{S\to N}$以及获取当前节点N到目的节点D的网络层路径$P_{N\to D}$的方法与后向重复链路检测相同.
- 进行路径比较,决定是否发送重定向消息.步骤如下:

(1) 当前节点反转$P_{N\to D}$得到$P_{N\to D}$的反转路径 $P_{N\to D}^{reverse}$.

(2) 比较 $P_{N\to D}^{reverse}$ 与$P_{S\to N}$,得到离节点S和D最近的相同节点,计算对应的网络层路径长度.

(3) 定义 $\rho_1=\frac{TTL_{M\to N}}{TTL_{S\to N}}$, $\rho_2=\frac{TTL_{N\to M}}{TTL_{N\to D}}$,则$0\le\rho_1,\rho_2\le 1$.定义$\rho_{thresh}$为$\rho_1,\rho_2$的门限值.当$\rho_1\ge\rho_{thresh}$且$\rho_2\ge\rho_{thresh}$时,当前节点$N$向源节点$S$发送重定向消息,否则不予发送.

4.3 性能分析

根据式(4)得到链路使用优化率 R 的计算式如下:

$$R=\frac{TTL_{old}}{TTL_{new}}\ge\frac{TTL_{S\to N}+TTL_{N\to D}}{TTL_{S\to N}+TTL_{N\to D}-TTL_{M\to N}-TTL_{N\to M}}=\frac{1}{1-\frac{\rho_1+\rho_2\gamma}{1+\gamma}} \tag{7}$$

其中,$\gamma=\frac{TTL_{N\to D}}{TTL_{S\to N}}$.

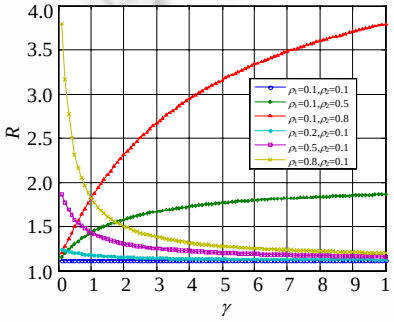

Fig.8 Relation of R, ρ_1, ρ_2 and γ

图 8 R 与ρ_1,ρ_2 以及γ的关系曲线

由式(7)可以看出,随着ρ_1,ρ_2 的增大,链路使用优化率 R 是不断增大的,当$\rho_1\ge\rho_{thresh}$ 且$\rho_2\ge\rho_{thresh}$ 时, $R_{\min}=\frac{1}{1-\rho_{thresh}}$.

图 8 给出了 R 与ρ_1,ρ_2 以及γ的关系图.当$\rho_1=\rho_2$ 时,R 的大小和γ无关;而当$\rho_1>\rho_2$ 时,R 随着γ的增大而递减;当$\rho_1<\rho_2$ 时,R 随着γ的增大而增大.另外,当$\rho_1>\rho_2, 0<\gamma\le 1$ 时 R 的值较大,当$\rho_1<\rho_2, \gamma\ge 1$ 后 R 的值较大;而在$\gamma>3$ 后,无论ρ_1,ρ_2 大小如何,随着γ的增加,R 的变化趋于平缓.

因此,要想获得较大的链路优化率,必须考虑γ对 R 的影响.故重新修正前向重复链路检测重定向报文发送条件:

1) $\rho_1\ge\rho_{thresh}$ 且$\rho_2\ge\rho_{thresh}$;

2) $\lambda_1\times\frac{\rho_2}{\rho_1}\le\gamma\le\lambda_2\times\frac{\rho_2}{\rho_1}$, $\lambda_1\le\lambda_2$取值大于零实数.

4.4 实验仿真

通过对 1 024 个节点构成的网络进行的 10 000 次查询的结果进行统计,得到链路使用优化率 R 以及发送重定向报文个数与ρ_{thresh} 和λ_1,λ_2 的关系曲线如图 9、图 10 所示.可以看出,随着ρ_{thresh} 的增大,R 不断增大,而发送重定向报文的数量有所降低;γ取值范围的扩大使得满足发送重定向报文的条件放宽,也使得重定向报文的个数增加,但由于较大的γ值对应的 R 值比小的γ对应的 R 值要低,从而使得 R 的平均值有所降低.

由于 $l_{overlap}$ 与ρ_{thresh} 表示的都是检测到的重复链路与原路径比值的门限,将前向重复链路检测与后向重复链路检测在$\rho_{thresh}=l_{overlap}, \varepsilon=0.2, \lambda_1=1, \lambda_2=2$ 时的仿真结果进行比较.可以看出,前向重复链路检测到重复链路并发送重定向消息的数量是对应相同 $l_{overlap}$ 的后向重复链路检测的 4 倍左右,而在链路使用优化率 R 上,平均高出

43%.因此,前向重复链路检测比后向重复链路检测更能提高系统的路由性能.

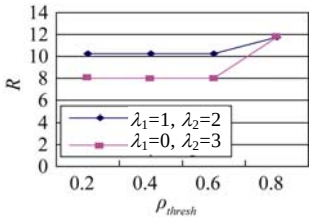

Fig.9 Simulation result of relation of R,ρ_{thresh} and λ_1,λ_2

图 9 R 与ρ_{thresh} 和λ_1,λ_2 关系的仿真结果

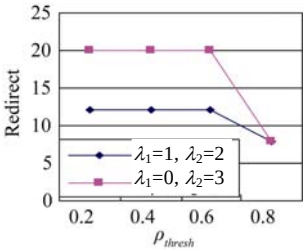

Fig.10 Simulation result of relation of redirect messages number, ρ_{thresh} and λ_1,λ_2

图 10 重定向报文数与ρ_{thresh},λ_1,λ_2 关系的仿真结果

5 总 结

检测并降低重复链路使用的拓扑一致性解决方案 DDL 是从实际物理网络路由情况出发,检测覆盖网络路由导致的物理链路重复使用的情况,在适当的参数下,通过重定向报文的发送,降低路由经历的链路数.根据重复链路定义的不同,给出了后向和前向两种重复链路检测机制,仿真实验结果表明,重复链路检测能够以较小的消息数量为代价,显著地提高系统链路使用优化率.而在相同的条件下,前向重复链路检测比后向重复链路检测更能提高系统的路由性能.

DDL 机制是一种全新的从物理链路层面解决 P2P 网络拓扑一致性问题的方法,与以前基于 IP 地址和时延的方法不同,它更注重于对实际网络路由情况下的拓扑一致性问题的解决,它采用了基于途经链路数来衡量路由性能的方式,解决重复链路带来的资源消耗问题.并且,DDL 是一种可以应用于任何结构化 P2P 网络中的拓扑一致性解决方案,不受限于覆盖网络层的组织方式.

References:

[1] Zhao BY, Kubiatowicz JD, Joseph AD. Tapestry: An infrastructure for fault-resilient wide-area location and routing. Technical Report, UCB//CSD-011141, Berkeley: University of California, 2001.

[2] Rowstron A, Druschel P. Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems. In: Guerraoui R, ed. Proc. of the 18th IFIP/ACM Int'l Conf. on Distributed Systems Platforms (Middleware 2001). Heidelberg: Springer-Verlag, 2001. 329−350.

[3] Stoica I, Morris R, Karger D, Kaashoek MF, Balakrishnan H. Chord: A scalable peer-to-peer lookup service for Internet applications. Technical Report, TR-819, New York: MIT, 2001.

[4] Ratnasamy S, Francis P, Handley M, Karp R, Schenker S. A scalable content-addressable network. In: Proc. of the ACM SIGCOMM. New York: ACM Press, 2001. 161−172. http://berkeley.intel-research.net/sylvia/cans.pdf

[5] Ren SS, Guo L, Jiang S, Zhang XD. SAT-Match: A self-adaptive topology matching method to achieve low lookup latency in structured P2P overlay networks. In: Proc. of the 18th Int'l Parallel and Distributed Processing Symp. (IPDPS 2004). Santa Fe, New Mexico, New York: IEEE Press, 2004. 83−91. http://www.cse.ohio-state.edu/hpcs/WWW/HTML/publications/papers/TR-04-8.pdf

[6] Qiu T, Wu F, Chen G. A generic approach to make structured P2P systems topology-aware. In: Proc. of the 3rd Int'l Symp. on Parallel and Distributed Processing and Applications (ISPA 2005). Berlin, Heidelberg: Springer-Verlag, 2005. 816−826. http://www.cse.buffalo.edu/~fwu2/res/ISPA2005-qiu.pdf

[7] Xu Z, Tang C, Zhang Z. Building topology-aware overlays using global soft-state. In: Proc. of the 23rd Int'l Conf. on Distributed Computing Systems (ICDCS 2003). New York: IEEE Press, 2003. 500−508. http://pages.cs.wisc.edu/~zhichen/290xu.pdf

[8] Krishnamurthy B, Wang J. On network aware clustering of Web clients. In: Proc. of the SIGCOMM 2000. Stockholm: IEEE Press, 2000. 97−110. http://www.research.att.com/~bala/papers/sigcomm2k.ps

[9] Ratnasamy S, Handley M, Karp R, Shenker, S. Topologically-Aware overlay construction and server selection. In: Proc. of the INFOCOM 2002. New York: IEEE Press, 2002. 1190−1199. http://berkeley.intel-research.net/sylvia/infocom02.pdf

[10] WInter R, Zahn T, Schiller J. Random landmarking in mobile, topology-aware peer-to-peer networks. In: Proc. of the 10th IEEE Int'l Workshop on Future Trends of Distributed Computing Systems (FTDCS 2004). New York: IEEE Press, 2004. 319−324. http://www2.computer.org/portal/web/csdl/doi?doc=abs/proceedings/ftdcs/2004/2118/00/21180319abs.htm

[11] Liu Y, Liu X, Xiao L, Ni LM, Zhang X. Location-Aware topology matching in P2P systems. IEEE Trans. on Parallel and Distributed Systems, 2005,16(2):163−174.

[12] Waldvogel M, Rinaldi R. Efficient topology-aware overlay network. ACM SIGCOMM Computer Communication Review, 2003,33(1):101−106.

[13] Braden R. Requirements for Internet Hosts—Communication Layers. STD 3, RFC 1122, USC/Information Sciences Institute, 1989.

[14] GT-ITM Homepage. http://www.cc.gatech.edu/projects/gtitm/

[15] Medina A, Lakhina A, Matta I, Byers J. BRITE: An approach to universal topology generation. In: Proc. of the Int'l Workshop on Modeling, Analysis and Simulation of Computer and Telecommunications Systems (MASCOTS). Washington: IEEE Computer Society, 2001. 346−353. http://www.cs.bu.edu/brite/publications/BriteMascots.pdf

[16] Palmer CR, Steffan JG. Generating network topologies that obey power laws. In: Kero TEF, ed. Proc. of the IEEE Global Telecommunications Conf. San Francisco: IEEE Computer Society Press, 2000. 434−438.

[17] Zegura E, Calvert KL, Donahoo M. A quantitative comparison of graph-based models for Internet topology. IEEE/ACM Trans. on Networking, 1997,5(6):770−783.

于婧(1979－),女,山东威海人,博士,讲师,主要研究领域为对等网络体系结构及路由.

汪斌强(1963－),男,博士,教授,博士生导师,主要研究领域为宽带 IP 网络体系结构.