ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.12, December 2007, pp.3080−3091
DOI: 10.1360/jos183080

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

一种基于故障时间的可调域间出口选择算法*

刘亚萍+, 龚正虎

(国防科学技术大学 计算机学院,湖南 长沙 410073)

A Tunable Interdomain Egress Selection Algorithm Based on the Failure Duration

LIU Ya-Ping+, GONG Zheng-Hu

(School of Computer, National University of Defense Technology, Changsha 410073, China)

+ Corresponding author: Phn: +86-731-4575818, E-mail: ypliu73@yahoo.com.cn, http://www.nudt.edu.cn

Liu YP, Gong ZH. A tunable interdomain egress selection algorithm based on the failure duration. *Journal of Software*, 2007,18(12):3080−3091. http://www.jos.org.cn/1000-9825/18/3080.htm

Abstract: Hot-Potato routing is a mechanism widely employed in the border gateway protocol (BGP) interdomain egress selection in large internet service provider (ISP). Recent work has shown that hot-potato routing is convoluted, restrictive so that it can impact the robustness of interdomain routing. Though a lot of research have been done to replace it with new mechanisms, these methods often ignore the issue of link failures or the failure duration, which arise as part of everyday network operations. In this paper, a tunable interdomain egress selection algorithm based on the IP link failure duration is proposed. The algorithm is tunable with the change of traffic engineering goals and routing stability in routers. It can also satisfy the purpose of real time in routers. Simulation results show that the algorithm can reach good balance among multiple goals.

Key words: BGP (border gateway protocol); traffic engineering; route prefix; forwarding path; routing stability

摘 要: 在大型 Internet 服务提供商中,BGP (border gateway protocol)出口路径选择常常采用“热土豆”机制.然而研究表明,该机制具有相当大的局限性以及出口调节的间接性,它容易影响域间路由的健壮性.针对“热土豆”机制的缺点,出现了一些新的 BGP 出口路径选择机制和算法.然而,这些方法在解决问题时往往忽略网络运行过程中经常出现的 IP 链路故障或故障持续时间的影响.提出了一种基于故障时间的可调域间出口路径选择算法,该算法能够根据流量工程的目标、路由稳定性等要求进行动态的调整,同时满足路由变化的实时性.模拟实验结果表明,该算法能够有效地在多个目标之间达到平衡.

关键词: BGP(border gateway protocol);流量工程;路由项;转发路径;路由稳定性

中图法分类号: TP393 **文献标识码**: A

BGP(border gateway protocol)路由协议是当前 Internet 域间路由协议的事实标准[1].当到某一目的地存在多条路由时,BGP 选择过程通常根据规则的优先级进行冲突消解来选择最优路由,例如,根据 local preference,as path length 等进行出口选择.图 1 是 CISCO 采用的规则及其优先级定义[2].当存在多条路由且前 7 条规则对应属

* Supported by the National Basic Research Program of China under Grant No.2003CB3148020 (国家重点基础研究发展计划(973)); the National Natural Science Foundation of China under Grant No.90204005 (国家自然科学基金)

Received 2006-03-07; Accepted 2006-10-31

性相同时,根据规则 8,选择距离最近的出口,这就是热土豆算法.在大型的传输 ISP(Internet service provider)中,60%的目标地址对应的 BGP 路径选择需要使用热土豆算法[3].然而,热土豆算法主要存在下述问题:域内链路的微小变化容易导致 BGP 路径选择结果的改变,从而导致 BGP 路由的大量变化以及本域和邻居域的流量模式的动荡[4].测量发现,在邻居域的流量模式发生变化中,25%的变化流量是由于域内 IGP(interior gateway protocol)的变化引起的.但是,当前 BGP 出口选择往往重点研究如何使得出口的选择满足流量的平衡[5],这些算法往往将 BGP 出口选择归结为一种静态的分派问题,而忽略域内拓扑变化以及拓扑变化持续时间的影响.

1 Largest weight
2 Largest local preference
3 Local paths over remote paths
4 Shortest AS path
5 Lowest origin type
6 Lowest MED
7 eBGP over iBGP paths
8 Lowest IGP metric
9 Oldest path
10 Lowest router ID
11 Minimum cluster ID length
12 Lowest neighbor IP address

Fig.1 Cisco BGP decision process

图 1 Cisco BGP 路径选择规则

若将一个域的拓扑结构用一个无向图 $G(N,L)$表示,节点表示互连的路由器,N 是节点的集合;边表示互连的链路,L 是边的集合.域内拓扑变化主要包括增加节点、删除节点、增加边、删除边、改变边的度量等.边度量的变化是网络管理员预先可知的,可以通过某些策略的设置消除路由不稳定性的影响.边的增、删在网络运行中是经常出现的现象,它对应 IP 链路或接口的 up/down.节点增、删对应路由器 up/down,可以通过增、删与该节点相连的所有边表示.本文研究的故障主要集中于边的删除.测量发现,大多数的 IP 链路故障是短暂故障[6,7](故障恢复时间小于 10 分钟的概率大于 81%),所以,研究 BGP 出口选择需要考虑域内故障及其持续时间.

本文通过引入两个新的度量——控制稳定性(表示域内拓扑变化下 BGP 路由的变化率)和流量累积效应(表示链路故障下流量平衡度量对故障持续时间的累积效应),提出了一种基于故障时间的可调域间出口选择算法 TIE_TF.本文第 1 节是相关研究的介绍.第 2 节对所研究的问题进行分析与描述,阐述了引入控制稳定性和流量累积效应两个新度量的原因.第 3 节详细描述了算法 TIE_TF,并对算法进行了详细的分析.第 4 节根据该算法,采用 Internet 中实际网络的拓扑结构、路由数据和流量数据进行了模拟实验与性能比较.第 5 节是本文的总结和下一步的工作.

1 相关工作

与本文相关的主要研究工作有热土豆算法弊病的研究、BGP 最优出口选择的研究以及流量工程的相关研究.热土豆算法弊病的研究主要通过测量手段研究热土豆算法引起的路由长时间收敛、流量的大幅度振荡以及 BGP 路由的不稳定性等问题.Teixeira 指出,域内链路的微小变化易导致 BGP 路径选择结果的改变[3],从而导致 BGP 路由的大量变化以及邻居域流量模式的动荡.在邻居域的流量模式发生变化的过程中,25%的变化流量是由于域内 IGP 的变化引起的.这种流量矩阵的变化使得域内流量工程的实施不易达到预定的目标.Agarwal 的研究表明,BGP 路径选择结果改变后,重新进行域内流量工程的结果与不考虑热土豆算法的影响情况下进行域内流量工程,其最大链路利用率可以减少 20%[8],但是前者将会大幅度增加域内流量工程的复杂度.

由于热土豆算法存在诸多缺点,并且 BGP 最优出口选择问题是 NP 难问题[5],研究人员提出了各类算法:基于流量分布最优均衡性的启发式算法[9]、基于多目标的遗传算法[10]等.但是,上述算法并没有考虑域内路由变化会导致算法的结果偏离预期值.Teixeira 研究了域内路由变化的影响,提出了一种可调出口选择算法 TIE[11].该算法通过一个线性表达式对 IGP 度量的控制来适应域内路由的变化,但是没有考虑故障持续时间的影响,即出口的选择随故障持续时间的变化来调节.

流量工程的相关研究包括通过 IGP 链路度量的调节来调整域内流量的平衡[12]、通过 BGP 策略的实施来调整 BGP 出口选择以及入口控制实现对某个域的流量调整[13]、MPLS 流量工程[14]、负载敏感的路由协议[15]等.但是这些方法不能很好地解决 BGP 出口选择的稳定性和流量工程的折衷问题.与本文最接近的是 TIE 的研究, TIE_TF 与 TIE 的不同之处是,TIE_TF 能够针对故障所呈现的不同持续时间特性,使得 BGP 出口选择能够保持在短暂故障下的稳定性.

2 问题描述

为了对 BGP 出口选择问题建模,需要对一些细节进行简化和假设:

1) 所有的 BGP speakers 组织成全互联的 iBGP 结构,所有的 BGP 路由器都知道到达某一目标网络的所有域间路由,本文暂且不考虑 iBGP 组织为层次结构的情况;

2) BGP 路由器的配置满足避免环和冲突的条件[16];

3) BGP 策略和 eBGP 路由是稳定的;

4) 域间流量需求是稳定的;

5) 网络中的拓扑变化只考虑 IP 链路的 up/down.上述假设条件使得 BGP 出口选择问题的建模集中考虑域内拓扑变化的影响.

在 BGP 出口选择问题中,有 3 类比较典型的机制:热土豆算法、固定出口算法[11]和 TIE 算法[11].例如在图 2 中,*AS* 1 有 5 个路由器(*A*,*B*,*C*,*D*,*E*),路由器 *A*,*B* 均是可达目标网络 *p* 的出口路由器且具有相同的路径属性.路由器 *A*,*B* 将到达 *p* 的最好路径通告给路由器 *C*.

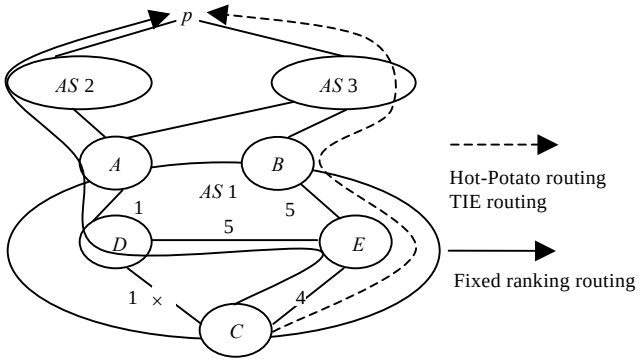

Fig.2 Different routings under link failure

图 2 链路发生故障下的不同路由选择算法

如果出口选择采用热土豆算法,对于目标网络 *p*,路由器 *C* 将选择 *A* 作为出口,如果链路 *C*-*D* 出现故障,那么路由器 *C* 将选择 *B* 作为出口.如果采用固定出口选择算法,那么路由器 *C* 在链路 *C*-*D* 出现故障时,仍选择 *A* 作为出口.如果采用 TIE 算法,不妨取参数 *t*=2(表示当拓扑发生变化时,到原有出口 *e* 的 IGP 距离如果小于在初始拓扑下到出口 *e* 距离的 2 倍,则仍选择 *e* 为出口),那么在初始情况下,路由器 *C* 将选择 *A* 作为出口,当链路 *C*-*D* 出现故障时,由于当前路由器 *C* 到 *A* 的距离为 10,大于在初始拓扑结构下 *C* 到 *A* 距离的 2 倍,所以路由器 *C* 将选择 *B* 作为出口.

然而,上述算法均存在问题.如果链路 *C*-*D* 的故障是短暂故障,则热土豆算法和 TIE 算法的选择易导致 BGP 路由变化的速度太快,引起路由的不稳定.如果链路 *C*-*D* 的故障是长时间故障,则固定出口选择算法将会长时间增加报文传输的延迟或者引起流量分布的不均衡.所以需要引入新的度量,衡量链路故障和故障时间对路由稳定性及流量分布的影响.为方便讨论,引入表 1 所示的基本符号定义.

Table 1 Summary of notation

表 1 符号列表

Notations	Description
G	Undirected graph
ΔG	Change of the graph
V	Set of nodes
P	Set of network prefixes for transit routing
$P(\delta)$	Probability of the graph change δ
$T(\delta)$	Duration time of the graph change δ
E	The mapping of prefixes to egress sets

定义 1(控制稳定性). 表示当出现域内链路故障时,网络中路由器的路由平均变化与故障持续时间之比.具体见式(1)中 s^{RM} 的定义.其中,$R_i(G,N)$表示将图 G 划分为 N 个区域,每个区域 $R_i(G,N)$表示以点 i 为根节点的最短路径树.如果 N 表示目标网络 p 的可达出口集合,那么,$RI(G,N,v)$表示节点 v 到目标网络 p 的出口选择.$H(G,N,v,\delta)$表示当网络拓扑发生变化后,节点 v 到目标网络 p 的出口选择是否发生变化,若发生变化,$H(G,N,v,\delta)$的值为 1,否则为 0.s^{RM} 与控制剖面敏感性 σ^{RM}[17](见式(2))的区别在于 σ^{RM} 未考虑故障时间的影响.

$$R_i(G,N)=\{v|\forall v\in V,m(v,i)\le m(v,i'),\forall i'\in N,i\ne i'\},$$

$$RI(G,N,v)=\{i|\forall i\in N,v\in R_i(G,N)\},$$

$$H(G,N,v,\delta)=\begin{cases}1, & RI(G,N,v)\ne RI(\delta(G),N,v)\\ 0, & \text{Otherwise}\end{cases},$$

$$s^{RM}(\delta)=\frac{1}{|P||V|}\sum_{p\in P}\sum_{v\in V}\frac{H(G,E(p),v,\delta)}{T(\delta)}P(\delta).$$

$$s^{RM}=\sum_{\delta\in\Delta G}s^{RM}(\delta)=\frac{1}{|P||V|}\sum_{\delta\in\Delta G}\sum_{p\in P}\sum_{v\in V}\frac{H(G,E(p),v,\delta)}{T(\delta)}P(\delta) \tag{1}$$

$$\sigma^{RM}=\frac{1}{|P||V|}\sum_{\delta\in\Delta G}\sum_{p\in P}\sum_{v\in V}H(G,E(p),v,\delta)P(\delta) \tag{2}$$

定义 2(流量累积效应). 流量累积效应表示当发生故障 δ 时,传统的流量度量与故障 δ 的持续时间和故障 δ 的发生概率的乘积,用 $ste(\delta)$表示(见式(3)).其中,$u(l)$表示链路 l 的利用率,$\Phi(u(l))$表示链路利用率的惩罚函数[12](见式(4)),te 是传统的流量平衡度量,$te(\delta)$表示故障 δ 下的流量平衡度量.用 ate 表示平均流量累积效应,其中,$p(G)$和 $T(G)$是网络处于正常拓扑结构下的概率和持续时间,$te(G)$表示在正常初始拓扑结构下的流量平衡度量.

$$te(\delta)=\sum_{l\in L}\Phi(u(l))(\delta),\delta\in\Delta G.$$

$$ste(\delta)=te(\delta)\times p(\delta)\times T(\delta) \tag{3}$$

$$ate=p(G)\times T(G)\times te(G)+\sum_{\delta\in\Delta G}ste(\delta).$$

根据定义 1 和定义 2,BGP 出口选择问题可以描述为如何寻找出口选择算法 β,使得在满足约束条件(5)和约束条件(6)时 s^{RM} 最小.约束条件(5)表示 β 的平均流量累积效应与热土豆算法的平均流量累积效应小于一定的阈值 γ.约束条件(6)表示在任意故障 δ 下,在 β 的流量平衡度量与热土豆算法的流量平衡度量大于阈值 γ 条件下所允许的最大累积效应,其中,ω 表示 β 能够容忍的流量平衡最大偏差阈值,TT 表示在流量平衡最大偏差下的最长持续时间.假定在初始拓扑结构下采用热土豆算法.如果将 $P(\delta)$,$T(\delta)$看成常数,则该问题是整数规划问题,是一种 NP 难问题[18].

$$\phi(u(l))=\begin{cases}u(l) & u(l)\in[0,1/3)\\ 3\cdot u(l)-2/3 & u(l)\in[1/3,2/3)\\ 10\cdot u(l)-16/3 & u(l)\in[2/3,9/10)\\ 70\cdot u(l)-178/3 & u(l)\in[9/10,1)\\ 500\cdot u(l)-1468/3 & u(l)\in[1,11/10)\\ 5000\cdot u(l)-16318/3 & u(l)\in[11/10,\infty)\end{cases} \tag{4}$$

$$\min\ s^{RM}$$

s.t.

$$\frac{p(G)\times T(G)\times te(G)+\sum_{\delta}ste_{\beta}(\delta)}{p(G)\times T(G)\times te(G)+\sum_{\delta}ste_{hot-potato}(\delta)}<\gamma,\gamma>1 \tag{5}$$

$$\text{if } \frac{te_{\beta}(\delta)}{te_{hot-potato}(\delta)} > \gamma,$$
$$\frac{ste_{\beta}(\delta)}{te_{hot-potato}(\delta)} \le \omega \times TT, \omega > \gamma, TT > 0 \tag{6}$$

3 TIE_TF 算法

采用固定出口选择算法,s^{RM} 值达到最小,但是可能在某些δ下,不满足不等式(5)和不等式(6).采用热土豆算法,不等式(5)和不等式(6)一定满足,但是 IGP 变化会直接影响出口选择,s^{RM} 值将比固定出口选择算法下的值大得多.TIE_TF 算法的目的是选择一种合适的方法,使得满足不等式(5)和不等式(6)的同时,s^{RM} 值尽量接近固定出口选择算法的值.为避免控制粒度太小而导致巨大的计算与存储开销,TIE_TF 算法约定在 IP 链路故障发生时,目标地址的出口选择采用相同的策略,TIE_TF 算法采用贪心算法思想,尽量在短暂故障下采用固定出口选择.

3.1 算法描述

TIE_TF 算法包括两部分:在线部分和离线计算部分,如图 3 所示.离线计算部分预先根据拓扑结构,可能的拓扑变化计算出可能发生故障对应的参数 $f_{\delta}(t)$.在线部分根据参数 $f_{\delta}(t)$求出当前到可选出口的距离度量 m,BGP 路径选择过程按照略加改进的决策过程进行选择,即图 1 中规则 8 的 IGP 距离改为新的距离度量 m,其他规则和顺序不变.任意两点的距离度量 $m(i,e)$的计算见式(7).其中,$d_G(i,e)$表示初始拓扑结构下节点 i 和 e 间的 IGP 距离,$f_{\delta}(i,e)$表示图发生变化δ时节点 i 和 e 之间的 IGP 距离,参数 $f_{\delta}(t)$取值为 0 或 1.离线计算部分需要使用故障统计历史信息与分析的结果.故障统计与分析定时地收集 IGP 协议中关于链路变化的事件,并对各种类型的故障进行统计分析,当分析结果与上次结果有明显不同时,将通告参数计算模块,重新进行参数的计算.TIE_TF 算法的描述包括两部分:参数的计算和改进的 BGP 路径选择过程,分别记为 TIE_TF off-line 和 TIE_TF on-line.

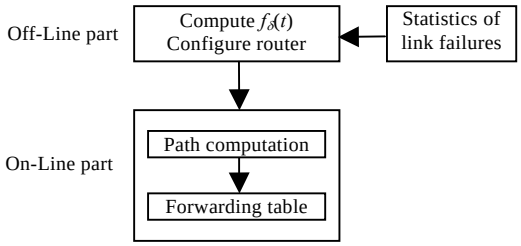

Fig.3 Architecture of TIE_TF

图 3 TIE_TF 的结构

$$\begin{cases} m_{\delta}(i,e) = f_{\delta}(t)d_{\delta}(i,e) + (1 - f_{\delta}(t))d_G(i,e) \\ m_G(i,e) = d_G(i,e) \end{cases}, i \in N,\ e \in N,\ \delta \in \Delta G \tag{7}$$

由于大多数故障是单链路故障和路由器相关故障,且路由器相关故障中以两条链路同时故障为最多[6,7],所以,预先计算参数部分包括所有单链路故障和两条链路故障的路由器相关故障以及历史统计信息中出现的其他故障.设单链路故障用δ_1 表示,两条链路故障的路由器相关故障用δ_2 表示,历史统计信息中出现的其他故障用δ_3表示,其他故障用δ_4表示.令$\Delta G=\{\delta_1,\delta_2,\delta_3\}$,固定出口选择算法用 Fix 表示,热土豆算法用 Hot 表示,~$T(\delta)$表示故障修复估计时间.图 4 是 TIE_TF off-line 算法的伪码描述.

TIE_TF off-line 算法中的步骤 1)表示初始化参数.步骤 2)表示通过计算ΔG 中每种δ下的固定出口选择算法与热土豆算法的流量平衡度量之比,确定某些参数的取值,以保证算法一定满足约束条件(6).其中,步骤 2.1)表示,若固定出口选择算法的流量平衡度量不大于热土豆算法下的对应值,则对应 $f_{\delta}(t)$一定取为 0;步骤 2.2)表示,如果固定出口选择算法不能满足约束条件(6),则对应 $f_{\delta}(t)$一定取为 1.然后,将其他δ按照 $(s^{RM}_{Hot}(\delta) - s^{RM}_{Fix}(\delta))/(ste_{Fix}(\delta) - ste_{Hot}(\delta))$比值按由大到小的顺序放入队列 q,等待进一步处理.步骤 3)表示采用贪心策略保证算法在满

足约束条件(5)时,s^{RM}值尽量接近最小.

//input params: $p(\delta)$,~$T(\delta)$,γ, ω, TT, traffic demand;
output: $f_\delta(t)$, $type(\delta)$.
1) init all $f_\delta(t)\leftarrow 1$; $type(\delta)\leftarrow 2$;
2) **while** (ΔG is not empty) **do**
if ($\delta\in\Delta G$) **then**
2.1) **if** ($te_{Fix}(\delta)\le te_{Hot}(\delta)$) **then** //satisfy (6)
$f_\delta(t)\leftarrow 0$; $type(\delta)\leftarrow 1$;
2.2) **else**
$$T\leftarrow\frac{\omega\cdot TT\cdot te_{Hot}(\delta)}{p(\delta)\cdot te_{Fix}(\delta)};$$
if $\left(\frac{te_{Fix}(\delta)}{te_{Hot}(\delta)}>\gamma \,\&\&\sim T(\delta)>T\right)$ **then**
$f_\delta(t)\leftarrow 1$; $type(\delta)\leftarrow 4$;
else put δ to q and sort with decrement according to $\frac{s^{RM}_{Hot}(\delta)-s^{RM}_{Fix}(\delta)}{ste_{Fix}(\delta)-ste_{Hot}(\delta)}$;
Delete δ from ΔG;
3) Compute r, which is the left side of inequation (5); //satisfy (5)
while (q is not empty) **do**
get the top element δ from q and delete it from q;
recompute r using $te_{Fix}(\delta)$ to replace $te_{Hot}(\delta)$;
if ($r<\gamma$) **then**
$f_\delta(t)\leftarrow 0$;
if $\left(\frac{te_{Fix}(\delta)}{te_{Hot}(\delta)}>\gamma\right)$ **then** $type(\delta)\leftarrow 3$;
else $type(\delta)\leftarrow 1$;

Fig.4 Off-Line part of TIE_TF

图 4 TIE_TF off-line 部分

假设用链表 fq 记录当前正在发生的故障δ,故障δ发生时间和对应 $f_\delta(t)$的实际取值.$I2Bq$ 表示 IGP 协议和 BGP 协议间的消息队列,$\Lambda\delta$表示故障恢复消息.如果发生δ或$\Lambda\delta$,IGP 将通过 $I2Bq$ 通告给 BGP 协议.图 5 是 TIE_TF on-line 算法的伪码描述,该算法描述如何计算两点之间新的距离度量.其基本思想是,当发生故障δ时,根据式(7)和当前$f_\delta(t)$的值计算两点间新的距离度量.如果故障δ不属于ΔG,则令$f_\delta(t)$为 1,即若遇到在预计算部分未考虑的故障,则缺省采用热土豆算法.由于ΔG 内都是发生概率较高的故障,该情况发生的概率较小.如果 BGP 收到$\Lambda\delta$消息且该故障恢复只是部分故障的恢复,网络拓扑结构还未恢复到初始拓扑状态,则令当前的 $f_\delta(t)$为 1,即按热土豆算法处理.此后,若再收到δ或$\Lambda\delta$消息,只要网络拓扑结构未恢复到初始拓扑,均按 $f_\delta(t)$为 1 的热土豆算法处理.如果故障恢复是完全恢复,则当前的 $f_\delta(t)$恢复为 0,BGP 出口选择恢复到与初始拓扑状态下的出口选择一致.

TIE_TF on-line 算法中的步骤 2.1)表示收到一个故障消息的处理.步骤 2.1.1)表示,当仍然存在未恢复的故障时,则将故障进行合并为一种故障处理.如果合并后的故障不属于预先考虑的故障集合ΔG,则取其对应的 $f_\delta(t)$为 1;如果合并后的故障属于ΔG,但是未恢复的故障对应$f_\delta(t)$的取值是 1,则当前$f_\delta(t)$取为 1,否则$f_\delta(t)$取对应预计算的值.步骤 2.1.2)表示,当前只有故障δ发生,如果该故障属于ΔG,则$f_\delta(t)$的取值按照 TIE_TF off-line 计算结果取值,否则其对应的 $f_\delta(t)$为 1.步骤 2.2)表示,当收到一个故障恢复消息的处理过程.其中,步骤 2.2.1)表示,该故障恢复消息使得网络拓扑恢复到初始拓扑结构.步骤 2.2.2)表示该故障恢复消息使得网络得到部分恢复,此时,根据原先 $f_\delta(t)$的取值进行处理,如果为 0,则 BGP 出口选择不变;如果为 1,则需要重新计算任意两点间的距离来进行 BGP 出口选择.步骤 2.2)保证了在故障完全恢复或部分恢复的情况下,BGP 出口选择的结果与 $f_\delta(t)$取值的一致性.步骤2.3)表示,当δ的类型是3时,如果发生超时则表示对应故障仍未恢复.这说明当前发生故障的修复时间与预先估计不符合,是一个长时间故障,因此,将其对应$f_\delta(t)$改为 1,重新计算两点间的距离来进行 BGP 出口选择.

1) In the original graph, $m_G(i,e)=d_G(i,e)$; For every prefix p, select the egress point;
2) **while** (1) **do**
  if ($I2Bq$ is not empty) **then**
2.1) **if** (δ is true) **then**
  put δ to fq;
2.1.1) **if** (fq is not empty) **then**
  $\delta \leftarrow \{\delta, fq\}$;
  if ($\delta \notin \Delta G$) **then** $f_\delta(t) \leftarrow 1$;
  else if ($f_\delta(t)$==1) **then** $f_\delta(t) \leftarrow 1$;
  else goto empty;
2.1.2) **else**
empty: record the begin time of δ;
  if ($\delta \in \Delta G$) **then**
  get $f_\delta(t)$;
  if ($type(\delta)$==3) **then** put $T(\delta)$ to timer list;
  else $f_\delta(t) \leftarrow 1$;
reselect: **if** ($f_\delta(t)$==1) **then** recompute $m_\delta(i,e)$ and reselect the egress point;
2.2) **else** //receive recover message
  reset timer $T(\delta)$ according to $\Lambda\delta$ and delete δ in the fq;
2.2.1) **if** (fq is empty) **then**
  if (the original $f_\delta(t)$==1) **then** select as in the original graph;
2.2.2) **else** //part recover
  if (the original $f_\delta(t)$==1) **then** $\delta' \leftarrow \{fq\}$; $f_\delta(t) \leftarrow 1$; goto reselect;
  else $\delta' \leftarrow \{fq\}$; goto empty;
2.3) **if** (timer is timeout) **then**
  if (Find the corresponding δ in fq) **then** $f_\delta(t) \leftarrow 1$; goto reselect;

Fig.5 On-Line part of TIE_TF

图 5 TIE_TF 的 on-line 部分

3.2 算法分析

本节分别对 TIE_TF 算法的正确性.计算复杂性以及性能进行分析.

3.2.1 TIE_TF 算法的正确性分析

TIE_TF 算法一定满足约束条件(5)和约束条件(6),并且在 IP 链路发生故障时,目标地址的出口选择采用相同的策略下,TIE_TF 算法的近似度不大于 2.这可以通过定理 1~定理 3 的证明来加以保证.

定理 1. TIE_TF 算法一定满足约束条件(5).

证明:TIE_TF off-line 算法执行完步骤 2)后,显然有下面不等式(8)成立.如果此时队列 q 为空或队列 q 中每个δ对应的$f_\delta(t)$取为 1,则不等式(9)成立.算法执行完步骤 3)后,由于不等式(8)和不等式(9)成立,因此根据ΔG中每个 $f_\delta(t)$的取值,不等式(10)成立.在执行 TIE_TF 算法的 on-line 过程中,如果链路故障是ΔG 中的故障且当 $type(\delta)$=3 时故障估计时间不小于实际故障时间,则发生故障时 $f_\delta(t)$值不变,显然不等式(10)仍然成立.如果 $type(\delta)$=3 且故障估计时间小于实际故障时间,则对应$f_\delta(t)$取值由 0 变为 1,不等式(10)一定仍然成立.如果链路故障不属于ΔG,则根据图 5 描述的步骤 2.1.1)和步骤 2.1.2),对应的$f_\delta(t)$取值为 1,不等式(11)成立.根据不等式(10)、不等式(11)以及不等式的性质,不等式(5)一定成立. □

$$\frac{\sum_{Type(\delta)=1,Type(\delta)=4} ste_{TIE\_TF}(\delta)}{\sum_{Type(\delta)=1,Type(\delta)=4} ste_{hot-potato}(\delta)} \le 1 < \gamma \tag{8}$$

$$\frac{\sum_{Type(\delta)=1,Type(\delta)=2,Type(\delta)=3} ste_{TIE\_TF}(\delta)}{\sum_{Type(\delta)=1,Type(\delta)=2,Type(\delta)=3} ste_{hot-potato}(\delta)} < \gamma \tag{9}$$

$$\frac{\sum_{\delta\in\Delta G} ste_{TIE\_TF}(\delta)}{\sum_{\delta\in\Delta G} ste_{hot-potato}(\delta)}<\gamma \tag{10}$$

$$\frac{\sum_{\delta\notin\Delta G} ste_{TIE\_TF}(\delta)}{\sum_{\delta\notin\Delta G} ste_{hot-potato}(\delta)}=1<\gamma \tag{11}$$

定理 2. TIE_TF 算法一定满足约束条件(6).

证明:如果链路故障不属于ΔG,则由于对应 $f_\delta(t)$取值为 1,显然满足约束条件(6).如果链路故障属于ΔG,则根据 TIE_TF off-line 算法描述,算法执行完后,ΔG 中的故障($type(\delta)$等于 1,2,4 和 3 且故障估计时间不小于实际故障时间)一定满足约束条件(6).如果 $type(\delta)$=3 且故障估计时间小于实际故障时间,则对应 $f_\delta(t)$取值由 0 变为 1,仍然满足约束条件(6).综上所述,TIE_TF 算法一定满足约束条件(6). □

定理 3. 已知 IP 链路故障集合ΔG,若 IP 链路故障发生时,目标地址的出口选择采用相同的策略,若用热土豆算法的 s^{RM} 值与待研究算法的 s^{RM} 之差表示优化目标值,则 TIE_TF 算法的近似度是 2.

证明:设用 OPT 表示在 IP 链路故障发生时,目标地址的出口选择采用相同策略的前提下,满足约束条件(5)和约束条件(6)的最优选择算法.显然在 OPT 算法下,当δ符合 $te_{Fix}(\delta)/te_{Hot}(\delta)>\gamma$且 $ste_{Fix}(\delta)/te_{Hot\text{-}potato}(\delta)>\omega\times TT$ 时,为了满足不等式(6),$f_\delta(t)$取值为 1.当δ符合 $te_{Fix}(\delta)\le te_{Hot}(\delta)$,$f_\delta(t)$取值为 0.这与 TIE_TF 算法的选择结果一致.其他属于ΔG 的δ构成的集合就是 TIE_TF 算法描述中队列 q 中元素构成的集合.显然,队列 q 中的所有元素都满足不等式(6).由于 $te_{Hot}(\delta)$和 $te_{Fix}(\delta)$可以预先计算出来,那么,约束条件(5)可以转换为下列不等式(12),其中,L 是根据不等式(5)计算出的值.

$$\sum_{\delta\in q}(ste_\beta(\delta)-ste_{Hot}(\delta))<L \tag{12}$$

若用热土豆算法的 s^{RM} 值与待研究算法的 s^{RM} 之差表示优化目标,则优化目标可以表示为:

$$\begin{cases} C=\sum_{\delta\in q} s^{RM}_{Hot}(\delta) \\ \max C-\left(C-\sum_{\delta\in q}(s^{RM}_{Hot}(\delta)-s^{RM}_{\beta}(\delta))\right) \end{cases} \tag{13}$$

即优化目标是

$$\max\sum_{\delta\in q}(s^{RM}_{Hot}(\delta)-s^{RM}_{\beta}(\delta)) \tag{14}$$

如果将 q 中的 n 个δ对应 n 个物品 $s_1,s_2,\ldots,s_n$, $s^{RM}_{Hot}(\delta)-s^{RM}_{Fix}(\delta)$ 分别对应每个物品的价值,记为 $p_1,p_2,\ldots,p_n$,$ste_{Fix}(\delta)-ste_{Hot}(\delta)$对应每个物品的总量,记为 $w_1,w_2,\ldots,w_n$,则该问题实际上是一个背包问题.令$\rho_j=p_j/w_j$ 表示 s_j 的价值密度,TIE_TF 算法是一种按价值密度由大到小的顺序依次将物品装入背包的贪心算法.

设 TIE_TF 算法的解为 $\tilde{S}=(\tilde{s}_1,\tilde{s}_2,...,\tilde{s}_k)$,最优解为 $S^*=(s_1^*,s_2^*,...,s_l^*)$,不妨设它们的元素同样是按照$\rho$的值由大到小排列的.设 s_m^* 是 S^*中第 1 个不属于 $\tilde{S}$ 的物品(s_m^* 必存在,否则,近似解即为最优解,结论自然成立),从而有 $s_1^*,s_2^*,...,s_{m-1}^*$ 属于 $\tilde{S}$.设 $\tilde{W}$ 是 $\tilde{S}$ 中价值密度大于 $\tilde{\rho}_m$ 的所有物品的总量之和,即 $\tilde{W}=\sum_{j=1}^{r}\tilde{w}_j$,则 $w_m^*>L-\tilde{W}$,否则,s_m^* 要被选入近似解.记 $W^*=\sum_{j=1}^{m-1}w_j^*$,设 I 是任一实例,$TIE_TF(I)$和 $OPT(I)$分别表示对应算法下按照式(14)求得的优化目标值,则有下式成立:

$$TF\_TIE(I)=\sum_{j=1}^{k}\tilde{p}_j\ge\sum_{j=1}^{m-1}p_j^*+(\tilde{W}-W^*)\frac{P_m^*}{w_m^*},$$

$$OPT(I)=\sum_{j=1}^{l}p_j^*=\sum_{j=1}^{m-1}p_j^*+\sum_{j=m}^{l}p_j^*\le\sum_{j=1}^{m-1}p_j^*+(L-W^*)\frac{P_m^*}{w_m^*}$$
$$=\sum_{j=1}^{m-1}p_j^*+(\tilde{W}-W^*)\frac{p_m^*}{w_m^*}+(L-\tilde{W})\frac{P_m^*}{w_m^*}\le TF\_TIE(I)+p_m^*\le 2TF\_TIE(I).$$

所以,TIE_TF 算法的近似度是 2,结论成立. □

3.2.2 TIE_TF 算法复杂性分析

TIE_TF off-line 部分的复杂性取决于图 4 所描述的步骤 2).如果用$|\Delta G|$表示ΔG 所包含的故障数目,$|V|$表示节点的个数,$|L|$表示边的个数,最坏情况下,TIE_TF off-line 的复杂性是 $O(|V|^3|\Delta G|)$.根据当前算法中ΔG 的定义可知,TIE_TF off-line 的复杂性是 $O(|V|^3|L|^2)$.如果ΔG 包含所有可能的故障情况,则算法的复杂性是 $O(|V|^3 2^{|L|})$.如果不考虑域内路由的计算,则 TIE_TF on-line 部分的复杂性是 $O(1)$.因此,该算法能够满足实时性的要求.

3.2.3 TIE_TF 算法性能分析

假设存在$\delta(f_\delta(t)=0)$,满足下式:

$$\frac{te_{fixed}(\delta)}{te_{hot-potato}(\delta)}=k\gamma,k>1.$$

根据约束条件(6),不等式(15)成立.

$$T(\delta)\le\frac{\omega\times TT}{p(\delta)\times k\gamma},\omega>\gamma,TT>0 \tag{15}$$

根据不等式(15)以及等式(1),可得不等式(16)成立.

$$\frac{\Delta s^{RM}}{\Delta TT}<0 \tag{16}$$

所以,TT 的设置决定了 s^{RM} 可以达到的最小值,TT 越大,s^{RM} 越小.说明在很多δ下,即使采用固定出口选择算法,也能满足约束条件(6).参数 TT 体现了一个网络能够容忍的流量最大不平衡性的最长持续时间,它反映了网络中节点(路由器)对转发报文的存储能力.

令 $c=\omega/\gamma$,同理,根据不等式(15)以及等式(1),可得不等式(17)成立.

$$\frac{\Delta s^{RM}}{\Delta c}<0 \tag{17}$$

所以,c 的设置决定了 s^{RM} 可以达到的最小值,c 越大,s^{RM} 越小.说明在很多δ下,采用固定出口选择算法,能够满足约束条件(5)和约束条件(6).参数 c 反映了一个网络能够容忍的流量平衡性的最大偏差,它体现了网络带宽对所承载流量的过度供给比例因子.

尽管σ^{RM}的定义与故障时间无关,然而由于参数 TT 越大,c 越大,TIE_TF 算法在很多δ下能够采用固定出口选择,所以相应σ^{RM}的值也越小.

4 实 验

为了验证 TIE_TF 算法的有效性,实验采用 Abilence[19]的拓扑结构及其 2005 年 1 月 1 日的 BGP 路由信息和流量信息作为实验数据.链路故障按概率产生[7].首先采用 CBGP[20]模拟 BGP 行为,在 CBGP 中扩展 TIE_TF 算法的 on-line 部分以及固定出口选择算法对应的策略.然后在 totem tool[21]平台上扩展 TIE_TF 算法的 off-line 部分,并增加 TIE_TF 算法与固定出口选择算法访问方式.同时,增加控制稳定性 s^{RM}、控制剖面敏感性σ^{RM}、平均流量累积效应 ate 等算法性能指标的计算,然后分别计算 TIE_TF 算法,固定出口选择算法以及热土豆算法在给定路由信息和流量信息下的 s^{RM},σ^{RM} 和 ate 的值.实验中,T_1 表示故障的持续时间满足 $P(T(\delta)<T_1)=0.7$,T_2 表示故障的持续时间满足 $P(T(\delta)<T_2)=0.8$,T_3 表示故障的持续时间满足 $P(T(\delta)<T_3)=0.9$,T_4 表示故障的平均持续时间.实验中分别取 T_1,T_2,T_3,T_4 作为~$T(\delta)$的值,γ的取值分别为 1.000 001~1.01,ω的取值满足$\omega-1/\gamma-1=10$.

图 6 是在 4 种路由算法(热土豆算法、固定出口选择算法、TIE 算法(t=2)、TIE_TF 算法)下的σ^{RM}值,其中,X 轴表示γ的取值,Y 轴表示σ^{RM}值.图 7 是在 4 种算法下的 s^{RM} 值.图 8 是在 4 种算法下的 ate 值.

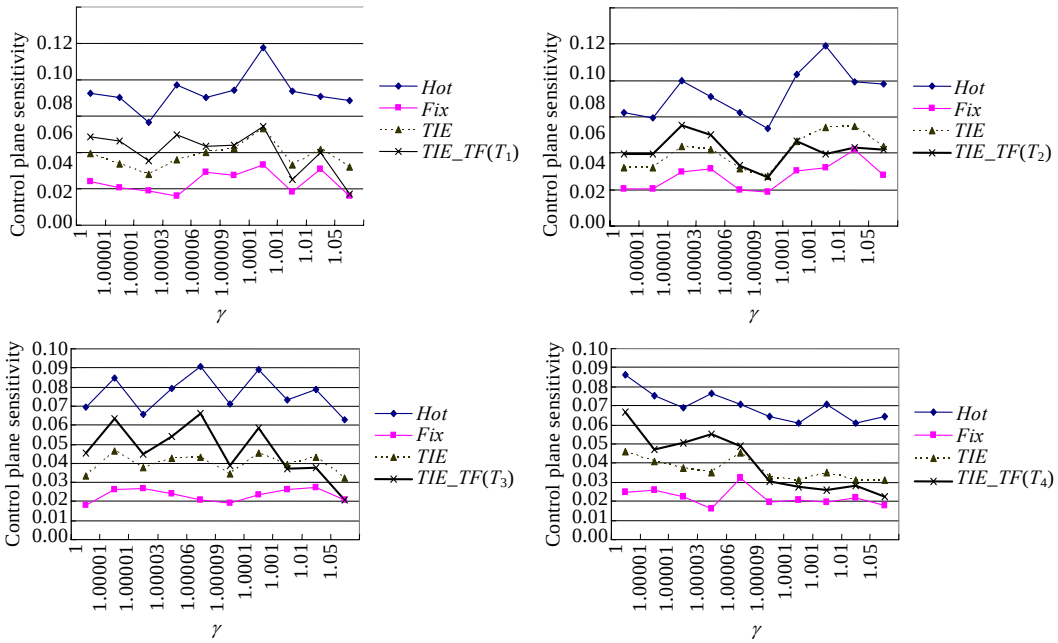

Fig.6 σ^{RM} under different algorithms

图 6 不同算法下的σ^{RM}

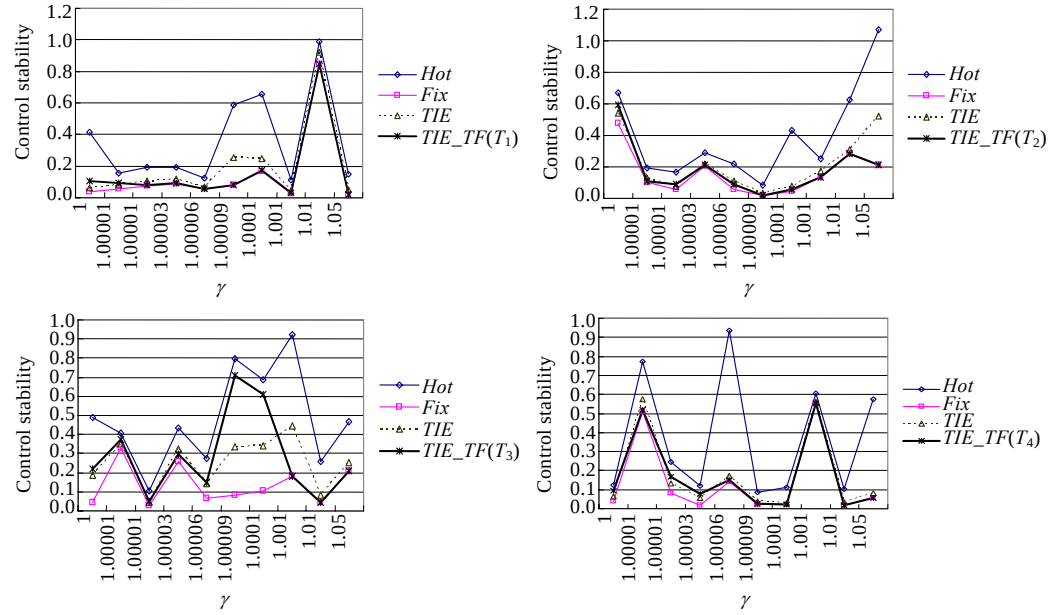

Fig.7 s^{RM} under different algorithms

图 7 不同算法下的 s^{RM}

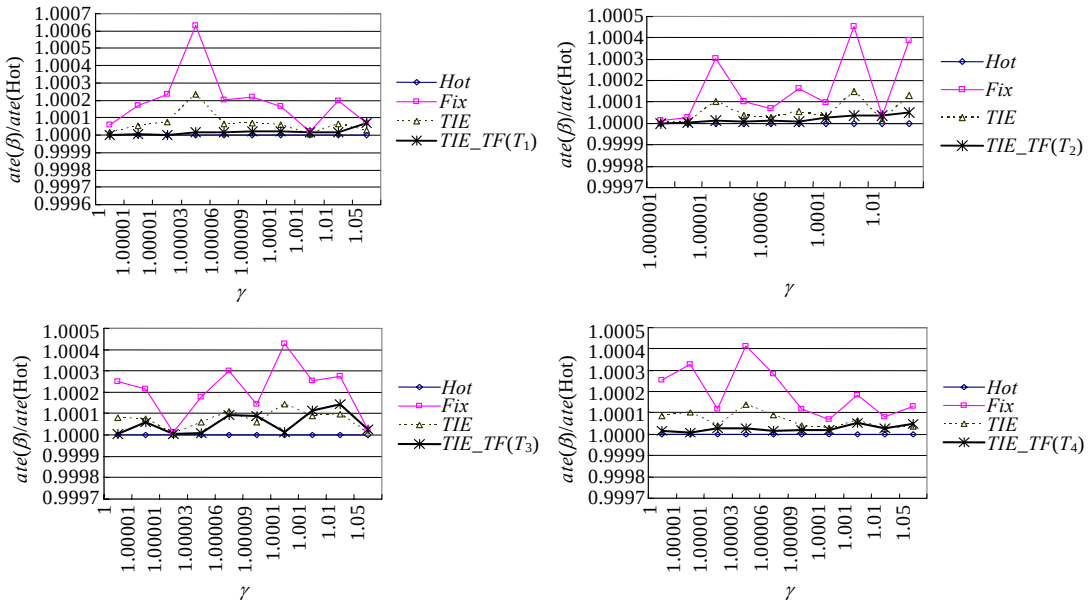

Fig.8 *ate* under different algorithms

图 8 不同算法下的 *ate*

模拟结果表明:

(1) TIE 算法、TIE_TF 算法的σ^{RM},s^{RM},*ate* 的值介于热土豆算法与固定出口选择算法之间.TIE_TF 算法的 s^{RM} 的值小于 TIE 算法的 s^{RM} 值的次数,大于 TIE_TF 算法的σ^{RM} 值小于 TIE 算法的σ^{RM} 值的次数,这是因为 TIE_TF 算法考虑了故障持续时间的影响.

(2) TIE_TF 算法的σ^{RM} 和 s^{RM} 值随γ的增大呈下降趋势,但是,TIE 算法的σ^{RM},s^{RM} 与γ间变化趋势不明显,这种差别是由 TIE_TF 算法的固有性质决定的.

(3) 在大多数情况下,TIE_TF 算法的 *ate* 值小于 TIE 算法的 *ate* 值,这是因为 TIE_TF 算法考虑了故障持续时间的影响.

(4) 故障的估计时间越接近实际的故障时间,TIE_TF 算法的性能越好,因此,取 T_4 时 TIE_TF 算法性能要好.

上述实验表明,TIE_TF 算法在控制稳定性和流量累积效应的平衡上优于热土豆算法、固定出口选择算法以及 TIE 算法.

5 结论及下一步工作

本文提出了 TIE_TF 算法用于解决 BGP 出口选择在发生链路故障时的路由稳定性和流量平衡的折衷,为此,提出了两个新的度量——控制稳定性和流量累积效应.TIE_TF 算法与 TIE 算法相比,在控制稳定性和流量累积效应等方面优于该算法.然而,本文的工作只考虑了域内拓扑变化对 BGP 出口选择的影响,事实上,还有很多因素可以影响 BGP 出口选择,如流量的变化、域间路由的变化等.所以,本文下一步的工作是要继续深入研究在考虑多种因素下的 BGP 出口选择问题.

References:

[1] ReKhter Y, Li T. A Border gateway protocol. RFC1771, 1995.

[2] Cisco Networks. BGP case studies. 2004. http://www.cisco.com/warp/public/459/bgp-toc.pdf

[3] Teixeira R, Shaikh A, Griffin T, Rexford J. Dynamics of hot-potato routing in IP networks. In: Liu Z, Merchant A, eds. Proc. of the ACM SIGMETRICS. New York: ACM Press, 2004. 307−319.

[4] Teixeira R, Duffield N, Rexford J, Roughan M. Traffic matrix reloaded: Impact of routing changes. In: Barakat C, ed. Proc. of the Passive and Active Measurement Workshop. Boston: Springer-Verlag, 2005. 251−264.

[5] Uhlig S. Implications of characteristics on interdomain traffic engineering [Ph.D. Thesis]. Belgium: University Catholique de Louvain, 2004.

[6] Iannaccone G, Chuah CN, Bhattacharyya S, Diot C. Feasibility of IP restoration in a Tier-1 backbone. IEEE Network Magazine, Special Issue on Protection, Restoration and Disaster Recovery, 2004,18(2):13−19.

[7] Markopoulou A, Iannaccone G, Bhattacharyya S, Chuah CN, Diot C. Characterization of failures in an IP backbone. In: Zhang ZS, Low S, eds. Proc. of the IEEE INFOCOM. Hong Kong: IEEE, 2004. 2307−2317.

[8] Agarwal S, Nucci A, Bhattacharyya S. Measuring the shared fate of IGP engineering and interdomain traffic. In: Gouda M, Matta I, eds. Proc. of the 13th IEEE Int'l Conf. on Network Protocols. Bosto: IEEE, 2005. 236−245.

[9] Bressoud T, Rastogi R, Smith M. Optimal configuration of BGP route selection. In: Matta I, ed. Proc. of the IEEE INFOCOM. San Francisco: IEEE, 2003.

[10] Uhlig S. A multiple-objectives evolutionary perspective to interdomain traffic engineering in the Internet. In: Lozano JA, Burke E, Smith J, eds. Proc. of the Workshop on Nature Inspired Approaches to Networks and Telecommunications. Birmingham, 2004. http://totem.info.ucl.ac.be/publications/papers-elec-versions/niant-121-uhlig.pdf

[11] Teixeira R, Griffn T, Resende M, Rexford J. TIE breaking: Tunable interdomain egress selection. In: Owezarski P, ed. Proc. of the CoNEXT 2005. ACM, 2005. http://www-cse.ucsd.edu/users/teixeira/teixeira-cv.pdf

[12] Fortz B, Thorup M. Internet traffic engineering by optimizing OSPF weights. In: Cohen R, Pitt D, eds. Proc. of the IEEE INFOCOM. Tel-Aviv: IEEE, 2000. 519−528.

[13] Bonaventure O, Cnodder SD, Haas J, Quoitin B, White R. Controlling the redistribution of BGP routes. Internet Draft, draft-ietf-grow-bgp-redistribution-00, Work in Progress, 2003.

[14] Liu H, Bai D, Ding W. A heuristic adaptive genetic algorithm for load balancing in MPLS networks. Journal of China Institute of Communications, 2003,24(10):39−45 (in Chinese with English abstract).

[15] Elwalid A, Jin C, Low SH, Widjaja I. MATE: MPLS adaptive traffic engineering. In: Bauer F, Cavendish D, eds. Proc. of the IEEE INFOCOM. Anchorage: IEEE, 2001. 1300−1309.

[16] Griffn TG, Wilfong G. On the correctness of IBGP configuration. In: Paxson V, Balakrishnan H, eds. Proc. of the ACM SIGCOMM. Pittsburgh: ACM, 2002. 17−29.

[17] Teixeira R, Griffn T, Shaikh A, Voelker G. Network sensitivity to hot-potato disruptions. In: Zegura E, Rexford J, eds. Proc. of the ACM SIGCOMM. Potland: ACM, 2004. 231−244.

[18] Xie Z. Algorithms for Networks and the Theory of Complexity. 2nd ed., Changsha: Press of National University of Technology, 2003. 242−245 (in Chinese).

[19] Abilene backbone network. 2005. http://abilene.internet2.edu/

[20] C-BGP—An efficient BGP simulator. 2005. http://cbgp.info.ucl.ac.be/#section_description

[21] TOTEM project toolbox for traffic engineering methods. 2005. http://totem.run.montefiore.ulg.ac.be/download.html

附中文参考文献:

[14] 刘红,白栋,丁炜.应用于 MPLS 网络负载均衡的启发式自适应遗传算法研究.通信学报,2003,24(10):39−45.

[18] 谢政.网络算法与复杂性理论.第 2 版,长沙:国防科学技术大学出版社,2003.242−245.

刘亚萍(1973－),女,广西桂林人,博士,副教授,主要研究领域为计算机网络技术.

龚正虎(1945－),男,教授,博士生导师,主要研究领域为计算机网络技术.