

# 一种基于链路状态的域间出口优化选择框架及关键算法

刘亚萍 龚正虎 何俊峰

(国防科技大学计算机学院 长沙 410073)

**摘 要** 随着 Internet 的快速发展,域间路由变得越来越重要,域间出口选择优化问题成为域间路由协议研究的一个重要问题.当前的域间出口选择机制通常缺乏灵活性和有效性,例如,这些机制往往忽略路由的稳定性、网络的动态性、选择的实时性、流量工程等诸多因素.基于以上因素,作者提出了一种基于链路状态变化的高效的 BGP 出口选择框架.该框架能够根据 AS 的多目标提供一种灵活的路由优化方法.基于控制规则和当前的链路状态,每个 BGP 路由器能够在线选择合适的出口.该框架具有灵活性、可扩展性以及健壮性.在此基础上,讨论了其中的一个基于链路故障的关键算法.模拟实验表明,该文所提出的机制对网络管理者是灵活而有效的.

**关键词** BGP;流量工程;路由稳定性;路由优化

**中图法分类号** TP393

## A Framework and Critical Algorithm of Interdomain Egress Selection Optimization Based on Link States

LIU Ya-Ping GONG Zheng-Hu HE Jun-Feng

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

**Abstract** With the rapid development of Internet, interdomain routing becomes more important. Interdomain egress selection optimization is one of the important problems in the research of interdomain routing protocol. Current mechanisms of interdomain egress selection are often inflexible or ineffective with ignoring many factors such as routing stability, network dynamics, the demand of real time, traffic engineering and so on. This paper proposes and evaluates a framework to facilitate efficient selection of Border Gateway Protocol (BGP) egress for Autonomous System (AS) when Interior Gateway Protocol (IGP) link state changes. It can provide a flexible means for AS to optimize routing according to their multiple goals. Based on control rules and current link state, every BGP router can select appreciate egress points online. The framework is extensible, flexible, and robust. A critical algorithm based on link failures is applied to illustrate the main idea of the framework. Simulation results demonstrate that this solution is feasible and expressive for the network administrators.

**Keywords** BGP; traffic engineering; routing stability; routing optimization

收稿日期:2007-03-29;最终修改稿收到日期:2007-06-28. 本课题得到国家“九七三”重点基础研究发展规划项目基金(2003CB314802)、国家自然科学基金(90204005)资助. 刘亚萍,女,1973年生,博士,副教授,研究方向为网络体系结构、网络优化等. E-mail: ypliu73@yahoo.com.cn. 龚正虎,男,1945年生,教授,博士生导师,研究领域为并行与分布处理、计算机网络技术等. 何俊峰,男,1984年生,博士研究生,研究方向为计算机网络体系结构技术.

# 1 引言

在 Internet 中,边界路由器通过 BGP (Border Gateway Protocol) 协议学习域间路由<sup>[1]</sup>. BGP 出口路由的选择主要有两种方法:(1)通过路由的 LOCAL\_PREF 属性选择;(2)通过比较路由的 IGP (Interior Gateway Protocol)度量选择.后者主要针对来自 IBGP 的路由.本文的工作针对的是后一种.

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,当存在多条属性相同的路由时,将选择距离最近的作为出口,这就是“Hot-potato”算法.在大型的传输 AS (Autonomous System) 中,60% 的目标地址对应的 BGP 路径选择需要使用“Hot-potato”算法<sup>[2]</sup>.然而,“Hot-potato”算法存在若干问题,例如,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流量模式的动荡<sup>[3]</sup>,”Hot-potato”算法没有提供对流量平衡的支持等.

针对“Hot-potato”算法的若干缺点,出现了一些改进的机制.例如,研究 BGP 出口选择能否满足流量的平衡<sup>[4]</sup>.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对路由稳定性的影响,忽略机制的实现代价以及这些算法是否具有随网络变化而调整的实时性要求.有些研究考虑了域内路由变化的影响,提出了一种可调出口选择机制 TIE<sup>[5]</sup>,然而该机制也同样具有考虑问题上的片面性和缺乏灵活性等缺点.例如,该机制未考虑链路故障持续时间的影响,未考虑其实现的可行性等.

针对当前 BGP 出口选择机制的不足之处,本文提出了一种灵活的出口路径选择优化框架 BGP-ROS (BGP Route Optimization Service). 该机制一方面具有随网络变化而调整的实时性的特点,另一方面能根据 AS 的要求的变化、当前网络变化的特征对出口路径进行调整. BGP-ROS 不改变两个通信的 BGP 路由协议实体间的行为,但是增强了 BGP 出口选择的智能性和灵活性.

通常入口边界路由器的路由发生改变后,会把新路由传递给邻居自治系统的路由器,这样将给互联网的路由稳定性带来负面影响,而本文提出的方法考虑了这方面的问题.在本文的方法中,当入口边界路由器的路由发生改变后,并不会马上把新路由

传递给邻居自治系统的路由器,而是由 ROS 预先计算出的控制规则来决定是否把新路由传递给邻居自治系统的路由器,从而增强了路由的稳定性.

本文的主要贡献是:(1)提出了一种新的域间出口路径选择优化框架 BGP-ROS;(2)全面描述了 BGP 出口选择优化问题,并以 AS 的不同要求为例,讨论了其中的核心算法;(3)采用 Abilence 的拓扑结构和实际网络中的路由数据、流量数据进行了模拟实验与性能比较,实验表明 BGP-ROS 比传统机制在灵活性、多目标优化间的平衡性、实用性上具有优势.

## 2 问题求解模型

BGP 出口选择问题是指在 BGP 协议的决策过程中,当存在多出口路径时,如何根据路径信息、网络的拓扑结构、流量信息等选择出最优路径,并满足 AS 的要求.由于“Hot-potato”算法在域内拓扑发生变化时呈现的问题较多,因此本文对问题的描述主要集中考虑域内拓扑变化下,如何寻找合适的选择算法.为简化问题所涉及的某些细节,我们假设:(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的域间路由,本文暂且不考虑 iBGP 组织为层次结构的情况;(2)BGP 路由器的配置正确,能避免环和冲突;(3)BGP 策略和 eBGP 路由是稳定的;(4)域间的流量需求是稳定的;(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链路故障变化的影响.

设 AS 希望达到的优化目标用集合  $GL$  表示,  $GL = \{gl_1, gl_2, gl_3, \dots, gl_n\}$ ,  $gl_i (i \in \{1 \dots n\})$  是具体的优化目标.例如,下面是一个 AS 的  $GL$ .

$$GL = \{gl_1, gl_2, gl_3, gl_4\},$$

其中,

$gl_1 = \{\text{转发的流总体占用本网络资源越小越好}\};$

$gl_2 = \{\text{采用技术所带来的代价越小越好}\};$

$gl_3 = \{\text{流量分布越接近平衡越好}\};$

$gl_4 = \{\text{BGP 路由越稳定越好}\}.$

显然,根据该  $GL$  的定义,该问题是一个多目标规划问题,并且多个目标之间往往是矛盾的.例如  $gl_1, gl_4$  是矛盾的,当拓扑发生变化时,如果优化  $gl_4$ , BGP 出口最好不变,但是这就导致  $gl_1$  变坏;同理,  $gl_2$  与  $gl_3$ 、 $gl_2$  与  $gl_4$  之间均存在矛盾.当前广泛

采用的“Hot-potato”算法,它的  $gl_1$  最优,由于在正确配置下能保证与转发的一致性,  $gl_2$  最优;但是  $gl_3, gl_4$  可能不一定最优. 因此,有部分的 BGP 出口选择问题的研究是基于如何优化  $gl_3$  的,而这也是一种静态优化问题. 但是如前所述,“Hot-potato”算法的主要缺点是在拓扑发生变化时,易导致 BGP 路由的不稳定,因而,本文主要从路由稳定性的角度研究出口的选择. 为便于建立该问题的数学模型,需要引入符号(如表 1 所示)和测度.

表 1 符号定义

符号	描述
$G$	无向图
$\Delta G$	拓扑变化的集合
$P$	目标地址集合
$V$	路由器集合
$E(p)$	$p$ 的可行出口链路的集合
$\delta$	某类 IP 链路故障
$p(\delta)$	发生 IP 链路故障 $\delta$ 的概率
$T(\delta)$	发生 IP 链路故障 $\delta$ 的持续时间
$\Phi(u(l))(\delta)$	在故障 $\delta$ 下的链路 $l$ 的利用率的惩罚函数
$d(i, j)$	节点 $i$ 和 $j$ 间的 IGP 距离

**定义 1.** 网络资源消耗代价<sup>[6]</sup>. 它表示在某种 BGP 出口选择情况下,在给定流量需求下,AS 传输所有流的总体距离. 设用  $d(i, j)$  表示路由器  $i, j$  之间 IGP 距离,  $t(i, p)$  表示从路由器  $i$  进入,目标地址为  $p$  的流量,  $\delta$  表示网络拓扑发生的某种变化,  $\Delta G$  表示网络拓扑变化集,  $x_{ip}^j$  是一个 0-1 变量,它表示如果出口选择为  $j$  则该变量为 1,否则为 0.  $tcost_\beta(\delta)$  表示了,在  $\delta$  下,当 BGP 出口选择采用某种机制(假设为  $\beta$  机制)下的网络资源消耗代价:

$$tcost_\beta(\delta) = \sum_{p \in P} \sum_{i \in N} \sum_{j \in E(p)} x_{ip}^j \cdot d(i, j) \cdot t(i, p) \quad (1)$$

**定义 2.** 转发一致性代价. 它表示在某种 BGP 出口选择情况下,维护 BGP 出口选择与实际转发路径一致性的代价. 由于这种维护机制通常需要建立虚通道或类似的绑定机制,所以转发一致性代价定义为需要建立虚通道的个数或需要建立虚通道的对应  $\delta$  的个数,可以用变量  $icost$  表示.

**定义 3.** 流量累积效应<sup>[11]</sup>. 它表示当发生故障  $\delta$  下,传统的流量度量与故障  $\delta$  的持续时间、故障  $\delta$  的发生概率的乘积,用  $ste(\delta)$  表示. 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<sup>[12]</sup>,  $te$  表示传统的流量平衡度量<sup>[11]</sup>.

$$te(\delta) = \sum_{l \in L} \Phi(u(l))(\delta), \delta \in \Delta G, \\ ste(\delta) = te(\delta) \times p(\delta) \times T(\delta) \quad (2)$$

**定义 4.** 控制稳定性<sup>[11]</sup>. 它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,用  $s^{RM}$  来表示. 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  $\sigma^{RM}$ <sup>[13]</sup> (式(4)所示)的区别是  $\sigma^{RM}$  未考虑故障时间的影响.

$$R_i(G, N) = \{v | \forall v \in V, d(v, i) \leq d(v, i'), \\ \forall i' \in N, i \neq i'\}, \\ RI(G, N, v) = \{i | \forall i \in N, v \in R_i(G, N)\}, \\ H(G, N, v, \delta) = \begin{cases} 1, RI(G, N, v) \neq RI(\delta(G), N, v) \\ 0, \text{其它} \end{cases}, \\ s^{RM} = \frac{1}{|P|} \frac{1}{|V|} \sum_{\delta \in \Delta G} \sum_{p \in P} \sum_{v \in V} \frac{H(G, E(p), v, \delta)}{T(\delta)} P(\delta) \quad (3)$$

$$\sigma^{RM} = \frac{1}{|P|} \frac{1}{|V|} \sum_{\delta \in \Delta G} \sum_{p \in P} \sum_{v \in V} H(G, E(p), v, \delta) P(\delta) \quad (4)$$

如果将  $GL$  表示为目标规划问题,即使可以求出非劣解集,仍然需要在非劣解集中,根据目标之间的优先权重,选出满意解. 所以将该问题转换为多约束条件下的单目标规划问题可以加速搜索的速度. 假设在初始拓扑结构下采用“Hot-potato”算法,那么,该问题可以表述为式(5)~(8)所示的形式. 其中,优化目标是控制稳定性最小,不等式(5)表示在  $\delta$  下所选算法的网络资源消耗代价不能超过“Hot-potato”算法的  $\gamma_1$ ,不等式(6)和(7)表示算法所需满足的流量平衡的约束,不等式(7)表示可允许的最大转发一致性代价. TIE<sup>[5]</sup>、RTF\_TIE<sup>[11]</sup>、固定出口选择算法<sup>[5]</sup>、Hot-potato 算法均可归结为该问题在特定条件下的求解结果. 例如,对于 TIE,  $t$  取 2 的算法对应问题的数学模型是  $\gamma_1$  等于 2,  $\gamma_2$  和  $\gamma_3$  取无穷大;对于 RTF\_TIE,对应问题的数学模型是  $\gamma_1$  和  $\gamma_3$  取无穷大;对于固定出口选择算法,对应问题的数学模型是  $\gamma_1, \gamma_2$  和  $\gamma_3$  取无穷大;Hot-potato 算法对应问题的数学模型是  $\gamma_1$  等于 1,  $\gamma_3$  等于 0.

$$\min s^{RM} \quad \text{s. t.} \\ \frac{tcost(\delta)}{tcost_{\text{hot-potato}}(\delta)} \leq \gamma_1 \quad (5)$$

$$\frac{te(\delta)}{te_{\text{opt}}(\delta)} \leq \gamma_2 \quad (6)$$

$$\text{If } \frac{te(\delta)}{te_{opt}(\delta)} > \gamma_2, \frac{te(\delta) \cdot P(\delta) \cdot T(\delta)}{te_{opt}(\delta)} \leq \omega \cdot T \quad (7)$$
$$icost \leq \gamma_3 \quad (8)$$
$$\gamma_1, \gamma_2, \omega > 1, \omega > \gamma_2, \gamma_3 \geq 0.$$

如果将  $P(\delta), T(\delta)$  看成常量, 上述问题是一个 0-1 整数规划问题, 是 NP 难问题<sup>[13]</sup>.

3 BGP-ROS 框架

图 1 是 BGP-ROS(BGP Routing Optimization Service)框架的逻辑结构. 它可以分成两部分: 离线部分 ROS(Routing Optimization Service)和在线部分. BGP-ROS 的离线部分需要收集有关链路状态、流量需求、网络拓扑等的统计信息以及网络管理员定义的优化目标、经验知识等. 同时它通过与所有的 BGP 路由器建立连接, 获取所有的 BGP 可达信息. 根据上述信息预先计算出控制规则, 并将控制规则通告给每一个 BGP 路由器. BGP-ROS 的离线部分同时负责事件的检测, 并将发生的事件通告给每个 BGP 路由器, 这样可以避免 BGP 优化选择的实现依赖于 IGP(如 OSPF、IS-IS)的实现. 每个 BGP 路由器运行 BGP-ROS 的在线部分. 它根据设定的控制规则以及收到的网络事件进行处理. 尽管 BGP-ROS 的离线部分是一种集中的机制, 但是如果 BGP-ROS 的离线部分发生故障, 也不会影响每个 BGP 路由器的正常工作.

学习机制为整个系统提供了一种反馈控制, 使得 BGP 出口选择不断向最优化逼近. 例如, 通过不断地收集链路故障信息, 基于链路故障的控制规则的计算将越精确.

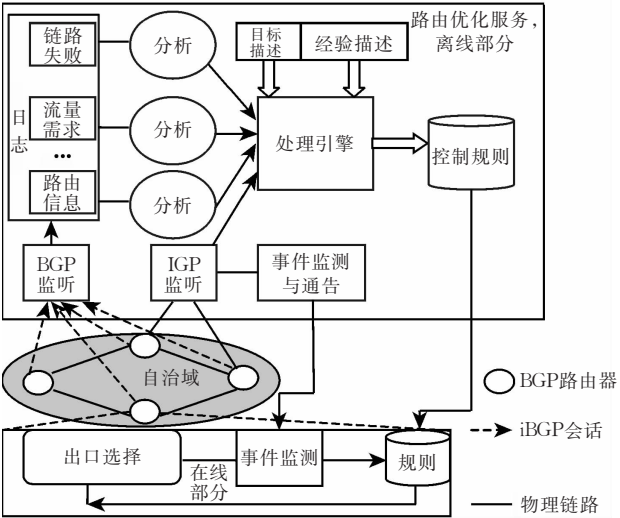


图 2 BGP-ROS 的详细结构

IGP 监听模块与 IGP 路由器建立邻接关系, 并配置为被动接收状态. 这样, IGP 监听模块可以随时监听到整个 AS 网络的拓扑信息与变化, 并将该信息提供给处理引擎使用. 为了保证 IGP 监听模块的健壮性, 我们可以与多个不同的 IGP 路由器建立邻接关系.

BGP 监听模块与所有的 BGP 路由器建立 iBGP 会话关系, 并配置为被动接收状态. 这样, BGP 监听模块能够获取所有的 BGP 信息, 同时它不需要向其它的 BGP 路由器播报路由信息, 并且不需要进行数据报文的转发. BGP 监听模块根据 BGP 选择过程常用的规则优先级(使用最短 IGP 距离规则之前的规则集)进行冲突消解来选择最优路由, 并将所选的路由作为候选路由集交给信息统计与分析模块处理, 该模块将统计的信息提供给处理引擎使用.

目标与约束描述模块描述网络管理员定义的优化目标或者相应的约束条件, 根据这些目标和约束条件的定义, 建立相应的数学模型, 如上述的式(5)~(8)所示. 目标与约束描述模块的信息将提供给处理引擎使用. 经验知识描述模块用来描述网络管理员在网络运营过程中总结出的一些经验知识, 这些知识有助于确定具体的参数取值. 例如, 可定义如果当前的网络负载是重载(例如链路利用率大于 2/3), 路由的稳定性将优先于负载平衡考虑; 如果当前的网络负载是轻载(例如链路利用率大于 1/10), 网络剩余的带宽足以吸收流量的变化, 则主

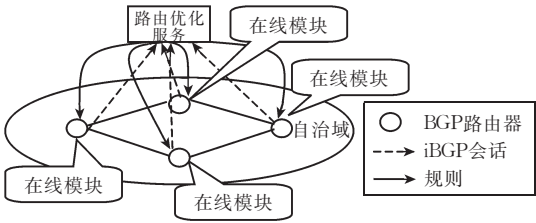


图 1 BGP-ROS 离线部分逻辑框图

图 2 是 BGP-ROS 的详细结构图. 其中离线部分 ROS 包括 8 个模块: 信息统计与分析模块、IGP 监听模块、BGP 监听模块、目标与约束描述模块、经验知识描述模块、处理引擎、控制规则集以及事件监测与通告模块.

信息统计与分析模块记录与跟踪所有的流量需求特性、链路故障特性、网络负载特性、BGP 路由特性, 并对其进行分析. BGP-ROS 框架能够灵活支持新的信息记录与分析模块的加入. 例如, 可以根据问题的需要加入域间关系分析模块. 信息统计与分析模块向处理引擎提供了必要的信息, 并通过这种自

要考虑路由的稳定性, 上述可以表示为式(9). 其中,  $te(G)$  表示在初始拓扑结构下流量平衡度量,  $|L|$  表示链路的个数.

$$\text{if } te(G) \geq \frac{4 \cdot |L|}{3} \text{ or } te(G) \leq \frac{|L|}{10}, \gamma_2 = \infty \quad (9)$$

处理引擎是 ROS 中最核心的部分, 它根据信息统计与分析模块、IGP 监听模块、目标与约束描述模块、经验知识描述模块提供的信息计算出一组控制规则, 并将这组控制规则分发给每个 BGP 路由器. 控制规则是处理引擎的输出, 可以将控制规则以路由器配置命令的形式实现, 例如可以采用类似 CISCO 路由器的 route-map 命令的形式. 这样, 就可以直接用 TELNET 命令将这些控制规则分发到每个路由器中.

事件监测与通告模块用于监测网络的变化, 并将这些变化以事件的形式通告给每个 BGP 路由器. 在 ROS 通告模块与每个 BGP 路由器之间可通过 SNMP(Simple Network Management Protocol) 协议进行交互. 当 ROS 监测到网络状态的变化后, 它将产生一个特定的 trap 消息. BGP 路由器收到 trap 消息后, 将根据消息的类型, 将相应的控制规则设定为活跃或非活跃状态. BGP 路由器上的 BGP 出口选择过程将根据活跃的控制规则进行出口的优化选择.

BGP-ROS 的在线部分由事件监测、控制规则以及出口选择三个模块组成. 其中事件监测用于接收 ROS 事件通告模块发出的事件, 并根据事件的类型对控制规则的状态进行设置; 控制规则用于保存 ROS 发布的控制规则; 出口选择模块只是对当前标准的 BGP 选择过程进行略微修改.

图 3 表示 BGP-ROS 进行路由出口优化选择的一个过程. 其中核心的部分就是处理引擎如何计算

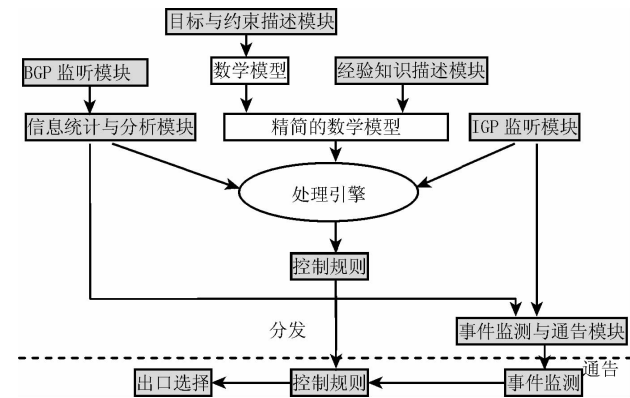


图 3 BGP-ROS 优化过程

控制规则以及在线部分如何利用控制规则进行出口选择的过程. 该算法是 BGP-ROS 的关键算法.

## 4 BGP-ROS 关键算法

“Hot-potato”算法最主要的缺点是路由的稳定性在域内拓扑发生变化时难以保证. 如果当前的出口选择与上次拓扑结构下的选择保持一致, 则路由的稳定性可以保证. 这说明上次的选择结果对当前的选择是有影响的. 在对 Sprint IP 骨干网的研究中发现, 80% 的链路故障不超过 10min, 50% 的链路故障在 1min 内恢复<sup>[15-16]</sup>. 这说明初始拓扑结构下的出口选择对当前的选择有影响. 同时, 如果需要在某种拓扑结构、流量需求下达到流量分布的平衡, BGP 出口的选择可以通过重新进行基于流量工程的优化计算来决定出口的分派. 所以, 可以用新的距离度量  $m$  来替代现有的 IGP 距离, 新的距离度量  $m$  的计算与上次的选择、初始拓扑结构下的选择、基于流量工程的优化选择相关.  $m$  的计算可以用式(10)表示. 其中,  $d_G(i, e)$  表示在初始拓扑结构下点  $i$  与  $e$  间的 IGP 距离;  $d_{\text{last}}(i, e)$  表示上次拓扑变化(非初始拓扑结构)下  $i$  与  $e$  间的 IGP 距离;  $d_{\text{current}}(i, e)$  表示当前拓扑结构下  $i$  与  $e$  间的 IGP 距离. 参数  $f_1(i, p, e, \delta, t)$  表示初始拓扑结构的影响; 参数  $f_2(i, p, e, \delta, t)$  表示上次拓扑结构的影响; 参数  $f_3(i, p, e, \delta, t)$  表示当前拓扑结构的影响; 参数  $x(i, p, e, \delta, t)$  表示流量平衡措施的影响. 为方便问题的求解, 不妨约定对于某个路由器  $i$ , 某个目标网络  $p$ , 这 4 个参数中最多有 1 个参数取值非 0.

$$m(i, p, e) =$$

$$f_1(i, p, e, \delta, t) \cdot d_G(i, e) + f_2(i, p, e, \delta, t) \cdot d_{\text{last}}(i, e) + f_3(i, p, e, \delta, t) \cdot d_{\text{current}}(i, e) + x(i, p, e, \delta, t) \quad (10)$$

BGP-ROS 关键算法包括两部分: (1) 处理引擎完成的离线部分; (2) 每个路由器的在线选择过程. 其中离线部分完成  $f_1(i, p, e, \delta, t)$ ,  $f_2(i, p, e, \delta, t)$ ,  $f_3(i, p, e, \delta, t)$  和  $x(i, p, e, \delta, t)$  的计算; 每个路由器的在线选择过程将根据  $f_1(i, p, e, \delta, t)$ ,  $f_2(i, p, e, \delta, t)$ ,  $f_3(i, p, e, \delta, t)$  和  $x(i, p, e, \delta, t)$  的取值, 计算距离度量  $m$  并替代现有的 IGP 距离进行 BGP 出口选择的过程.

图 4 给出了处理引擎完成的离线部分通用算法的伪码描述. 其中,  $\Delta G$  表示根据信息统计与分析模块提供的出现概率较高的链路故障集合. 图 4 的步 1 和步 2 表示根据目标与约束描述模块、经验知

识描述模块,初始化所定义的数学模型参数的取值以及输出的各类参数值;步 3 表示  $\Delta G$  中的  $\delta$  按照  $s_{\text{hot}}^{\text{RM}}(\delta) - s_{\text{fix}}^{\text{RM}}(\delta)$  由大到小的顺序排序,采用贪心算法的思想;步 4 表示当“Hot-potato”算法满足式(6)、式(7)时的算法描述;步 5 和步 6 表示在通用条

件下的算法描述. 处理引擎的输出是控制规则. 控制规则采用以 routemap 命令的形式实现. 其中, scenario 表示当前的负载情况,处理引擎可根据不同的负载计算不同的控制规则集.

```

//input:  $p(\delta), T(\delta)$ , traffic demand,  $\gamma_1, \gamma_2, \gamma_3, \omega, T$ 
//output:  $f_i(i, p, e, \delta, t) (i=1, 2, 3), x(i, p, e, \delta, t)$ 
1. init every  $f_1(i, p, e, \delta, t), f_2(i, p, e, \delta, t), x(i, p, e, \delta, t)$  to 0 and  $f_3(i, p, e, \delta, t)$  to 1;
2. set  $\gamma_2, f_1(i, p, e, \delta, t), f_2(i, p, e, \delta, t), x(i, p, e, \delta, t)$  if we use formula (9);
3. compute all  $s_{\text{Hot}}^{\text{RM}}(\delta) - s_{\text{Fix}}^{\text{RM}}(\delta)$  in  $\Delta G$  and sort them with decrement, put  $\delta$  whose  $te_{\text{Hot-potato}}(\delta)$  can satisfy condition (6) or (7) to  $\Delta G1$ , else to  $\Delta G2$ ;
4. while ( $\delta \leftarrow \text{DEQUEUE}(\Delta G1) \neq \text{NIL}$ ) do {
    if ( $\delta$ 's last state has little effect) then
        {  $f_2(i, p, e, \delta, t) \leftarrow 0$ ;  $workp(i, p, e, \delta, t) \leftarrow \&f_1(i, p, e, \delta, t)$ ;  $type \leftarrow 0$  }
    else
        {  $f_1(i, p, e, \delta, t) \leftarrow 0$ ;  $workp(i, p, e, \delta, t) \leftarrow \&f_2(i, p, e, \delta, t)$ ;  $type \leftarrow 1$  };
    for ( $i \in V, p \in L, E(p) > 1$ ) {
        for ( $e \in E(p)$ ) {  $f_3(i, p, e, \delta, t) \leftarrow 0$ ;  $*workp(i, p, e, \delta, t) \leftarrow 1$ ; }
        if (condition (5) is false or (condition (6) and (7) are false)) then { //check condition (5), (6), (7)
            reset  $f_3(i, p, e, \delta, t)$  and  $*workp(i, p, e, \delta, t)$ ;
            continue;
        }
        if (condition (8) is false) then { //check condition (8)
            reset  $f_3(i, p, e, \delta, t)$  and  $*workp(i, p, e, \delta, t)$ ;
            goto  $G2\_process$ ;
        }
    } //end for
} //end while
 $G2\_process$ :  $\Delta G2' \leftarrow \Delta G2$ ;
5. while ( $\delta \leftarrow \text{DEQUEUE}(\Delta G2) \neq \text{NIL}$ ) do {
    {select an algorithm who can satisfy condition (6) or (7),  $f_i(i, p, e, \delta, t) \leftarrow 0$ ;
    if ( $e$  is the egress) then  $x(i, p, e, \delta, t) \leftarrow 0$ ; else  $x(i, p, e, \delta, t) \leftarrow 1$ ; }
    if (condition (5) is false) then { //check condition (5)
        for ( $i \in V, p \in P, E(p) > 1$ ) {
            for ( $e \in E(p)$ ) {  $f_3(i, p, e, \delta, t) \leftarrow 1$ ;  $x(i, p, e, \delta, t) \leftarrow 0$ ; }
            if (condition (6) and (7) is false) then {
                reset  $f_3(i, p, e, \delta, t)$  and  $x(i, p, e, \delta, t)$ ;
                continue;
            }
            if (condition (5) is true) then goto Next;
        } //end for
    } //end if
    Next: if (condition (5) is false) then goto Error;
} //end while
if (condition (8) is false) then goto Error;
6. while ( $\delta \leftarrow \text{DEQUEUE}(\Delta G2') \neq \text{NIL}$ ) do { //check if we can reduce  $s^{\text{RM}}$ 
    for ( $i \in V, p \in P, E(p) > 1$ ) {
        if (for every  $e, f_3(i, p, e, \delta, t) \leftarrow 1$ ) then {deal with it as in step 4;}
    } //end for
} //end while
return the results;
7. Error: return ("can not find a solution");

```

图 4 处理引擎算法

图 4 表示的是通用的算法处理过程. 实际上, 信息统计与分析模块提供的信息往往可以进一步简化参数的计算. 例如, Sprint IP 骨干网的研究表明大多数的故障是短暂的, 经过几分钟, 拓扑结构就恢复为初始状态. 如果链路故障满足这种特性, 就可以忽略上次拓扑变化的影响, 令所有的  $f_2(i, p, e, \delta, t)$

为 0. 如果缺省的“Hot-potato”算法能够在链路故障时满足流量平衡的要求(即满足式(6), (7)), 则忽略流量平衡措施的影响, 令所有的  $x(i, p, e, \delta, t)$  为 0.

BGP-ROS 关键算法的复杂性主要取决于离线部分. 如果不考虑域内路由的计算, 在线部分的复杂性取决于规则的匹配过程. 如果规则采用数组方

式的存储,在线部分的复杂性是  $O(1)$ ,因此该算法能满足实时性的要求.如果用  $|\Delta G|$  表示  $\Delta G$  所包含的故障数目,  $|P|$  表示目标地址的个数,用  $|V|$  表示节点的个数,需要比传统的 Hot-potato 算法最多增加  $|\Delta G| |P| + 1/2 |V| (|V| - 1)$  的存储空间.

如果  $|L|$  表示边的个数,最坏情况下, BGP-RO-ES 的复杂性是  $O(|V|^3 |\Delta G| |P|)$ . 如果  $\Delta G$  包含所有可能的故障情况,则算法的复杂性是  $O(|V|^3 2^{|L|} |P|)$ . 所以,  $\Delta G$  的取值与算法的复杂度密切相关.在本文所取的  $\Delta G$  中,主要考虑发生概率较高的单故障以及 2 条链路同时发生故障的路由器相关故障,所以算法的复杂性是  $O(|V|^3 |L|^2 |P|)$ .

## 5 模拟实验

为验证 BGP-ROS 的有效性,实验采用 Abilence<sup>①</sup> 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.链路故障根据 Markopoulou<sup>[16]</sup> 的结果按概率产生.  $T(\delta)$  的取值采用平均故障的持续时间.我们采用 CBGP<sup>②</sup> 模拟 BGP 行为,通过扩展 totem tool<sup>③</sup> 的功能来计算新定义的测度. BGP-ROS 的关键算法采用 BGP-RO-ES 算法.同时,将该算法与 TIE 算法、Hot-

patoato 算法、固定出口选择算法、RTF\_TIE 算法进行比较.在如下的实验中,坐标  $X$  表示  $\gamma_3$  的取值,坐标  $Y$  表示控制稳定性 (control stability)  $s^{RM}$  的取值,并且图(a)表示  $icost$  为需要建立虚通道的个数 ( $icostType$  为 0),图(b)表示  $icost$  为需要建立虚通道的对应  $\delta$  的个数 ( $icostType$  为 1).

在实验 1 中,  $\gamma_1$  设置为无穷大,  $\gamma_2$  设置为 1.001,  $\omega$  设置为 1.01. 图 5(a) 和图 5(b) 表示实验结果.从图中,可以看出 Hot-potato 算法的  $s^{RM}$  的值是 RTF\_TIE 算法的 1.23 倍.然而,如果取  $icostType$  为 0, BGP-RO-ES 算法只需要花费 50% 的  $icost$ ,其  $s^{RM}$  的值仅比 RTF\_TIE 算法下该值大 7.6%;如果取  $icostType$  为 1, BGP-RO-ES 算法只需要花费 50% 的  $icost$ ,其  $s^{RM}$  的值仅比 RTF\_TIE 算法下该值大 2.3%.在当前参数设置下, Hot-potato 算法、RTF\_TIE 算法、BGP-RO-ES 算法可以满足条件式(6)和(7),但是固定出口选择算法有约 13% 的链路故障不满足条件式(6)和(7).

在实验 2 中,根据式(9)来设置参数  $\gamma_2$ , 其它的参数设置同实验 1. 图 6(a) 和图 6(b) 表示实验结果.在该条件下  $\gamma_2$  实际为无穷大.条件式(6)和(7)不起作用. RTF\_TIE 算法退化为固定出口选择算法.

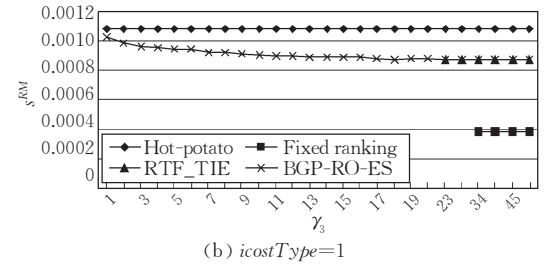
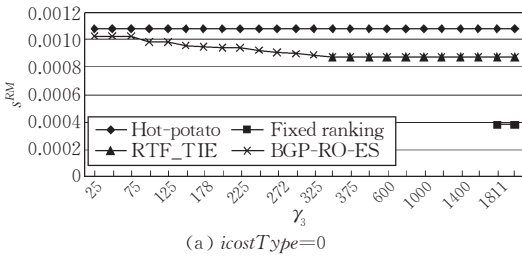


图 5 实验 1 不同算法下的  $s^{RM}$  比较

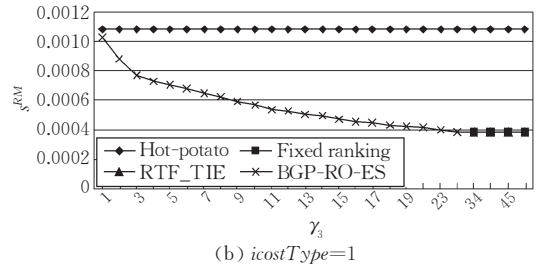
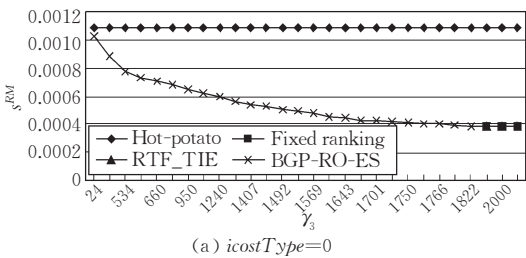


图 6 实验 2 不同算法下的  $s^{RM}$  比较

在实验 3 中,  $\gamma_1$  设置为 2,  $\gamma_2$  设置为 1.001,  $\omega$  设置为 1.01, 比较 BGP-RO-ES 算法、固定出口选择算法、Hot-potato 算法以及 TIE( $t=2$ ) 算法下  $s^{RM}$  的值. 图 7(a) 和图 7(b) 表示实验结果. 图中, 可以看出 Hot-potato 算法的  $s^{RM}$  的值是 TIE 算法的 1.64 倍.

然而, 如果取  $icostType$  为 0, BGP-RO-ES 算法只

- ① Abilene Backbone Network. <http://abilene.internet2.edu/>
- ② C-BGP- An efficient BGP simulator. [http://cbgp.info.ucl.ac.be/#section\\_description](http://cbgp.info.ucl.ac.be/#section_description).
- ③ TOTEM Project Toolbox for Traffic Engineering Methods. <http://totem.run.montefiore.ulg.ac.be/download.html>, 2005.



需要花费 50% 的  $icost$ , 其  $s^{RM}$  的值仅比 TIE 算法下该值大 20.8%; 如果取  $icostType$  为 1, BGP-RO-ES 算法只需要花费 50% 的  $icost$ , 其  $s^{RM}$  的值仅比 TIE 算法下该值大 16.8%. 实验 3 的结果不如实

验 1 和实验 2 的结果好, 其原因是 BGP-RO-ES 算法需要保证条件式(5)是否满足. 在上述设置下, 固定出口选择算法是不能满足条件式(5)的.

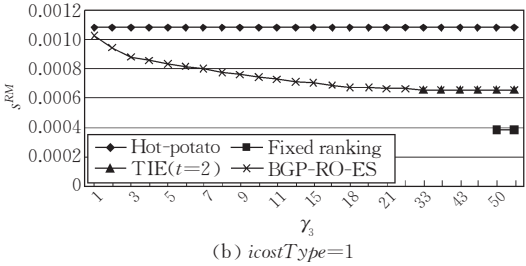
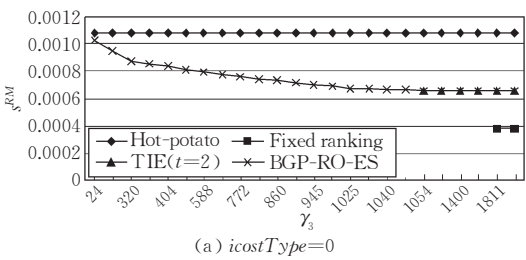


图 7 实验 3 不同算法下的  $s^{RM}$  比较

模拟实验表明, 当  $\gamma_3$  较大时, 固定出口选择算法、TIE 算法、RTF\_TIE 算法才能满足条件式(7). 然而, BGP-RO-ES 算法在能满足问题所需的所有条件时, 它只需花费约 50% 的  $icost$  代价, 就能获得与 TIE 算法或者 RTF\_TIE 算法较接近的  $s^{RM}$  的值.

6 结论及下一步的工作

本文提出了一种灵活的出口路径选择优化框架 BGP-ROS. 该机制一方面具有随网络变化而调整的实时性的特点, 另一方面能根据 AS 的要求的变化、当前网络变化的特征对出口路径进行调整. 由于 BGP-ROS 不改变 BGP 路由协议实体间的行为, 因此该机制具有可行性、灵活性、高效性、健壮性等特点. 本文详细描述了其中的关键算法, 模拟实验表明, 该算法能够获得与 TIE 算法或者 RTF\_TIE 算法较接近的  $s^{RM}$  的值, 而只需花费约 50% 的  $icost$  代价, 能够满足 AS 所提出的多种约束条件. 然而, 本文的工作实际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, 如域间路由的变化等. 所以, 下一步的工作, 将继续深入研究在多种因素下 BGP 出口的选择问题.

参 考 文 献

[1] ReKhter Y, Li T. A Border Gateway Protocol 4 (BGP-4). RFC1771, March 1995

[2] Teixeira R, Shaikh A, Griffin T, Rexford J. Dynamics of hot-potato routing in IP networks//Proceedings of the ACM SIGMETRICS, New York, NY, USA, 2004: 307-319

[3] Teixeira R, Duffield N, Rexford J, Roughan M. Traffic matrix reloaded: Impact of routing changes//Proceedings of the Passive and Active Measurement Workshop. Boston, 2005: 251-264

[4] Uhlig Steve. Implications of characteristics on interdomain traffic engineering[Ph. D. dissertation]. University Catholique de Louvain, 2004

[5] Teixeira R, Griffin T, Resende M, Rexford J. TIE Breaking: Tunable interdomain egress selection. AT&T Labs Research; Technical Report TD-69EJBE, 2005

[6] Bressoud T, Rastogi R, Smith M. Optimal configuration for BGP route selection//Proceedings of the INFOCOM'03. San Francisco, 2003

[7] Uhlig Steve. A multiple-objectives evolutionary perspective to interdomain traffic engineering in the Internet. International Journal of Computational Intelligence and Applications, 2005, 5(2): 215-230

[8] Bonaventure O, Cnodder S D, Haas J, Quoitin B, White R. Controlling the redistribution of BGP routes. IETF draft, work in progress, draft-ietf-grow-bgp-redistribution-00. txt, April 2003

[9] Mahajan R, Wetherall D, Anderson T. Towards coordinated interdomain traffic engineering//Proceedings of the 3rd Workshop on Hot Topics in Networks (HotNets-III). San Diego, CA, 2004

[10] Caesar M, Caldwell D, Feamster N, Rexford J, Shaikh A, van der Merwe Jacobus. Design and implementation of a routing control platform//Proceedings of the 2nd Conference on Symposium on Networked Systems Design & Implementation. Boston, Massachusetts, USA, 2005

[11] Liu Ya-Ping, Gong Zheng-Hu, Zhao Feng. RTF\_TIE: A tunable interdomain egress selection algorithm robust to transient link failures//Proceedings of the IEEE HPSR. Poznan, Poland, 2006: 279-286

[12] Fortz B, Thorup M. Internet traffic engineering by optimizing OSPF weights//Reuven Cohen, Dan Pitt eds. Proceedings of the IEEE INFOCOM. Tel-Aviv, Israel, 2000



[13] Teixeira R, Griffn T, Shaikh A, Voelker G. Network sensitivity to hot-potato disruptions//Proceedings of the ACM SIGCOMM. Portland, DR, USA, 2004: 231-244

[14] Xie Zhen. Algorithms for Networks and the Theory of Complexity. 2nd Edition. Changsha: Press National University of Technology, 2003(in Chinese)  
(谢政. 网络算法与复杂性理论. 第 2 版. 长沙:国防科技大学出版社, 2003)

[15] Iannaccone Gianluca, Chuah Chen-Nee, Supratik Bhattacharyya, Christophe Diot. Feasibility of IP restoration in a Tier-1 backbone. Network, IEEE, 2004, 18(2): 13-19

[16] Markopoulou Athina, Iannaccone Gianluca, Bhattacharyya Supratik, Chuah Chen-Nee, Diot Christophe. Characterization of failures in an IP backbone//Proceedings of the Infocom'04. Hong Kong, China, 2004



**LIU Ya-Ping**, born in 1973, Ph. D. , associate professor. Her research interests include network architectures, optimization of network systems and so on.

**GONG Zheng-Hu**, born in 1945, professor and Ph. D. supervisor. His current research interests include parallel and distributed computing, computer network and so on.

**HE Jun-Feng**, born in 1984, Ph. D. candidate. His research interests focus on network architectures.

Background

The research problem of this paper is the egress selection optimization problem which is one of the important problems in the research of interdomain routing protocol. Presently, the most common egress selection algorithm that we used in transit AS(Autonomous System) is hot-potato algorithm. However, the hot-potato algorithm has two main problems; First, a small intradomain link change may be easy to trigger big problem of BGP(Border Gateway Protocol) routing stability. Second, it is not easy to support traffic engineering.

Prior research has considered a range of possible techniques. Teixeira proposed a TIE ( Tunable Interdomain Egress Selection) mechanism in 2005. Bressoud developed heuristic solutions to determine an optimal selection of outgoing links and associated border routers in 2003. Uhlig proposed a multiple objectives evolutionary algorithm to solve the potential conflicting nature of the traffic engineering objectives in 2004. There are many other techniques what are not listed here.

However, current mechanisms of interdomain egress selection are often inflexible or ineffective with ignoring many factors such as routing stability, network dynamics, the demand of real time, traffic engineering and so on. In this paper, the authors propose and evaluate a framework to facili-

tate efficient selection of BGP (Border Gateway Protocol) egress for AS (Autonomous System) when IGP (Interior Gateway Protocol) link state changes.

The research belongs to the project of "Theory of Switching and Routing for the Next Generation Internet", supported by the National Basic Research Program (973 Program) of China under grant No. 2003CB314802. The project aims at creating new switching and routing mechanism for the next generation Internet to achieve scalability, security, controllability of the network. The work of this paper is an important part of the interdomain routing optimization, which researches how to select interdomain egress point optimally.

The authors are working towards research of the fuzzy switching and interdomain routing optimization. More than 20 papers are published in the magazines and important international conferences. Three of them have been included in SCI (Science Citation Index), such as the paper titled "A Routing Optimization Algorithm for BGP Egress Selection". And eight of them have been included in EI (The Engineering Index), such as the paper titled "A Tunable Interdomain Egress Selection Algorithm Robust to Transient Link Failures". The others have been included in core magazines of china. Moreover, they are applying for a patent.