

一种带有路由反射的 BGP 路由预测算法

徐鑫¹ 吴静² 高远¹

(东北大学信息科学与工程学院 沈阳 110004)¹ (辽宁医学院网络中心 锦州 121000)²

摘要 由于路由反射器的影响,路由器可能只知道非常有限的到达外部目的地的路由的子集,BGP 路径选择过程并不会形成一个确定的路径等级顺序,使得路径预测变得非常困难。给出了一种算法,该算法不需要对路由协议动态进行复杂的模拟,只需要静态的路由数据就可以计算出带有路由反射的路由选择结果,并在一个模拟的大型 ISP 上验证了预测算法的正确性。结果显示配置的改变可以引起路由表的变化。

关键词 BGP,路由反射器,路径预测

中图分类号 TP393 **文献标识码** A

Algorithm for Predicting the BGP Routes in iBGP Routing of Route Reflection Graph

XU Xin¹ WU Jing² GAO Yuan¹

(School of Information Science and Engineering, Northeastern University, Shenyang 110004, China)¹

(China Center. of Network, Liaoning Medical University, Jinzhou 121000, China)²

Abstract Predicting the BGP routes is particularly difficult when the BGP selection process does not form a deterministic ranking of the routes, due to the route reflection. This paper presented provably correct algorithms for computing the outcome of the BGP route-selection process for each router in a network, without simulating the complex details of BGP message passing. The Algorithms require only static inputs. The prediction algorithm was verified in a simulated large-scale ISP. The results show that configuration adjustment cause can the routing table changes.

Keywords BGP, Route reflection, Predicting the BGP routes

1 引言

在一个 AS 内部实现 iBGP 全网,需要在 AS 内部建立 $n(n-1)/2$ 个 iBGP 会话,因此要引进路由反射来限制大型网络的 iBGP 会话数量。AS 内部的路由反射定义了 BGP 路由器之间的两种关系:客户和非客户。BGP 对等体之间的这些关系在路由器之间定义了一个松散的等级关系,从没有客户的底层路由器到不是其他任何路由器的客户的大型路由反射器。iBGP 中路由的通告依据一些明确界定的规则。如果路由器不充当路由反射器的角色,即它没有“客户”对等体,那么如果是从 eBGP 会话中学到这条路由的,路由器将会把这条路由通告给其所有 iBGP 对等体,如果是经过 iBGP 会话学到的将不会通告给任何 iBGP 对等体。另一方面,对于路由反射器^[1],如果路由是从客户对等体(或 eBGP 对等体)处学到的,路由反射器重新通告这条路由到其所有的客户和非客户对等体(通告这条路由的对等体除外);如果路由是从非客户对等体处学到的,路由反射器只将这条路由重新通告给其客户对等体。

为了控制通信在网络中的流动,网络经营者需要经常改变网络配置。在现实网络中修改配置之前,网络经营者必须

预测路由协议配置的改变会带来的影响。一个路由器选择的路径常常要依赖于其他路由器所选的路径^[2],因此预测每个路由器最终选择的路径是很困难的。本文给出的路由预测算法不需要对路由协议动态进行复杂的模拟,只需要静态的路由数据就可以计算出使用了路由反射器的路由选择结果,并在一个模拟的大型 ISP 上验证了算法的正确性。

本文第 2 节介绍了网络模型及约束条件;第 3 节给出了预测算法,并证明了算法的正确性,利用 C-BGP 模拟了算法,结果显示算法是正确的;最后得出结论并讨论了今后的工作。

2 网络模型和约束条件

定义 1 Internet 可以抽象地表示为一个无向图 $G=(V, E)$,其中 V 是所有 AS 的边界路由器的集合; E 是所有 EBGP 会话的集合。用 N 代表 BGP 下一跳的集合, R 代表 AS 内部运行 BGP 的路由器的集合^[3-6]。

假设 E 中包含的是 AS 中每个路由器实施完本地入口策略(可能会过滤掉它接收到的一些路径集,及设置或修改其他路径的属性)之后的 eBGP-learned 路径。

定义 2 当使用路由反射器时,AS 中的路由器形成一个 iBGP 会话的有向图 $G_{iBGP}=(R, T)$,每条边 $a=(u, v) \in T$,其

到稿日期:2009-05-09 返修日期:2009-07-01 本文受国家自然科学基金资助项目(60673159),东北大学“985 工程”——信息化基础结构关键技术科技创新平台资金(985-2-DC-C09),辽宁省教育厅基金资助项目(2008005)资助。

徐鑫(1966—),男,博士生,教授,主要研究方向为网络路由管理, E-mail: xuxin@bhu.edu.cn; 吴静,女,副教授; 高远,教授,博士生导师。

中 $u, v \in R$ 对应于一对路由器之间的 iBGP 会话。如果 v 是 u 的一个路由反射器客户, 则 $a \in down$; 如果 u 是 v 的一个路由反射器客户, 则 $a \in up$; 如果 u 和 v 之间有正规的 iBGP 会话, 则 $a \in over$ 。图 1 给出了一个信令图实例。在全网配置中, 每个 eBGP-speaking 路由器与 AS 中所有其他路由器都有一条边在 $over$ 中, 且集合 up 和 $down$ 都是空的。

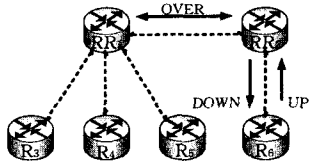


图 1 iBGP 信令图实例

定义 3 当 $r \in R, E_r \subseteq E$ 表示路由器 r 当前通过 eBGP-learned 所拥有的路由, 在任意时刻, 路由器 r 有零个或多个 iBGP-learned 路径 $I_r \subseteq E$ 。

定义 4 令 $V = \{r | r \in R\}$, 一个 BGP 系统的状态定义为 $S = \langle E_r \rangle$, 其中 E_r 表示路由器 r 当前时刻所拥有的路由。如果 $E_r = \Phi$, 则表明路由器 r 尚未从其邻居那里学到任何路由信息。状态 $\langle E_r^0 \rangle$ 称为初始状态, 并记为 S_0 。

定义 5 如果在某一时刻, 路由器 r 收到其对等体发送的更新报文, 并进行路由选择, 使得路由器 r 的状态由 E_r 变为 E'_r , 则称路由器 r 为活动的, 并称 BGP 系统经历了一次活动。

定义 6 如果 $S' = S_0$, 或存在一个可达状态 $S = \langle E_r \rangle$ 以及一个非空的活动点集 $A \subseteq V$, 使得 $S \xrightarrow{A} S'$, 则称状态 $S' = \langle E'_r \rangle$ 是可达的。

定义 7 一个活动序列定义为一个 BGP 系统所经历的一系列的活动 $\sigma = \{A_1, A_2, A_3, \dots\}$, $\sigma(i) = A_i$ 表示活动序列 σ 中的第 i 次活动。活动序列 σ 可以是有限的也可以是无限的。

一个 BGP 系统的动态行为可以通过活动序列 σ 来描述, 一个 BGP 系统将按照该活动序列从其初始状态 S_0 开始进行动态变化, 即: $S_0 \xrightarrow{\sigma(1)} S_1 \xrightarrow{\sigma(2)} S_2 \xrightarrow{\sigma(3)} \dots$

定义 8 如果存在状态 S , 使得对任意 $\sigma \in V$ 有 $S \xrightarrow{\sigma} S$, 则称 S 为系统的最终状态。如果一个 BGP 系统存在最终状态, 则称这个 BGP 系统是可解的。

定义 9 函数 $best_r$ 按照 BGP 路径选择过程在路由器 r 接收到的路径集中找出路由器 r 上的最佳路径。

定义 10 函数 $select$ 按照 BGP 路径选择过程的前四步在 BGP 路径集中选出同样好的路径集。

假定系统满足如下 4 个约束:

约束 1 (BGP 系统的活动是缓慢的) eBGP-learned 路径遵照网络工程决定的时间量程而缓慢变化。任何 eBGP 路由改变都可以被看作为一个单独的问题实例。

约束 2 (BGP 系统是可解的) 已知稳定的 eBGP 学习路径和确定的 iBGP 和 IGP 拓扑, AS 内的每个路由器都收敛于一个唯一的 routing 选择。

约束 3 (活动信息顺序的独立性) 每个路由器上的路由决定只依赖于从邻居那里接收到的路径, 而不是这些路由信息的顺序或时间。

约束 4 (1) $\forall u, v, w \in R, (u, v) \in down$ and $(u, w) \notin$

$down \Rightarrow best_u(\{p_v, p_w\}) = p_v$, 其中 p_v 代表从 v 中学到的任意路径, p_w 代表从 w 中学到的任意路径; (2) up 中的边是非循环式的。

这时我们才有可能有效地计算由于配置或拓扑改变带来的路由选择的结果。带有约束的预测算法可以不用考虑路由信息的不同顺序是否会产生不同的结果, 可以用仿真某一特定的信息顺序来防止算法重新访问那些已经做出预测的路由器。

3 路径预测算法

3.1 问题描述

路由反射使 AS 中的路由器不再需要形成一个全网 iBGP 拓扑图, 这意味着一些路由器可能不会学到最优路径中的所有路径。第一, 算法不能简单地指派给 non-eBGP-speaking 路由器一条来自“最近的” eBGP-speaking router 中的路径, 因为前者路由器可能从来没学到过那条路径, 如图 2 中的网络。R₁, R₂ 和 R₃ 是路由反射器 RR 的客户, R₄ 是 R₃ 的常规 iBGP 对等体。R₂ 和 R₃ 之间有一条短的 IGP 路径, 但它们并没有通过 iBGP 会话直接相连。通过路径选择过程的前四步, 路由器 R₁, R₂ 和 R₄ 有相同的 eBGP 路径, 并已选择了它们自己的 eBGP-learned 路径。网络中, R₃ 最近的出口点是 R₂, 但是 R₃ 选择了 R₁, 因为 RR 的最近出口路由器是 R₁。第二, 一些 non-eBGP-speaking 路由器上常常没有可能的出口路由器的一致排列。例如, 图 2 中, RR 喜欢出口路由器 R₁, 因为它到 R₁ 的路径花费最短。路由器 R₃ 对可能的出口路径的喜爱依赖于其他路径的存在与缺少。如果 AS 经由路由器 R₂ 和 R₄ 上的 eBGP 会话学习到某些目的地的路径, 那么路由器 R₃ 喜欢用 R₂ 作为其出口路由器。另一方面, 如果 AS 在路由器 R₁, R₂ 和 R₄ 上学习路径, 那么 R₃ 喜欢使用 R₄, 这意味着 R₃ 喜欢 R₄ 优于 R₂, 与当只有 R₂ 和 R₄ 是可达的路由器时 R₃ 的选择不一致。

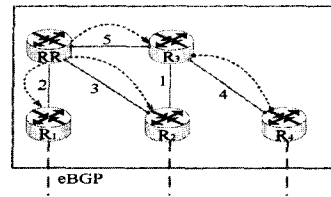


图 2 当使用路由反射器时, 一些路由器可能不会学到最优路径中的所有路径

3.2 算法描述

为了克服一些路径在一些路由器上不可见的影响, 设计了一个算法来仿真某一特定活动序列, 使路径在每个路由器可能的地方分配, 并将这些决定的作用传播给其他路由器, 不用重新访问任何分配。算法首先从路由反射器的底部分层次地向上激活路由器, 这保证了每个路由器选择一条 $down$ 路径, 如约束 4(1)中所要求的。由于算法从等级的底部向上移动, 当知道了来自客户的所有路径之后, 算法为每个路由反射器进行计算; 这些路由器的计算从不需要重新访问, 凭借约束 4, 路由器总是喜欢来自“孩子”(即客户)的路径优于来自对等体或父母的。在 $down$ 方向上访问路由器保证了算法使用来自 up 和 $over$ 集合中的所有可达路径对剩余路由器实行计算。算法基于 up 和 $over$ 集合中的元素定义了两种路由器的

部分顺序。我们能够定义这两种部分顺序是因为约束 4(2) 要求信令图不能有这些边的任何循环,所以每个部分顺序必须有一个顶部和低端元素。下面是带有路径反射但不考虑 MED 属性的计算网络中每个路由器最佳路径的算法的伪码描述:

```
// Proceed up the hierarchy, assigning best routes.
// Find a router for which all children are activated.
SelectBest_eBGP_RR(E,R)
A ← ∅
While ∃ r ∈ R s. t. r ∉ A and c ∈ A ∀ c ∈ DOWN(r)
    I_r ← ∪_{c ∈ DOWN(r)} b_c
    b_r ← best_r(I_r ∪ E_r)
    A ← A ∪ r
// Proceed down the hierarchy.
// Find a router for which all parents are activated.
A ← ∅
While ∃ r ∈ R s. t. r ∉ A and c ∈ A ∀ c ∈ UP(r)
    I_r ← ∪_{c ∈ UP(r) ∪ OVER(r)} b_c
    b_r ← best_r(I_r ∪ b_r)
    A ← A ∪ r
```

在图 2 的例子中应用这个算法,路由器 R_1, R_2 和 R_4 在第 1 步中选择了最佳路径,因为这些路由器都在等级的低端,所以它们在 *down* 上的所有邻居都已经被激活(因为它们什么都没有)。 R_3 被激活了,但它没有在这点选择路径,因为它在 *down* 上没有邻居。这 4 个路由器处于等级中的同一层,因此它们可以以任何顺序被激活。之后 RR 被激活;它应用 $best_{RR}(\{r_{R1}, r_{R2}\})$ 并选择 r_{R1} ,因为它的 IGP 路径花费最小。在向下的方向上,所有的路由器又一次被激活。 R_3 从 RR 处接收 r_{R1} ,并将其与到达目的地的最佳路径 r_{R4} 做比较。 R_2 和 R_4 也接收到了 r_{R1} ,但是还继续选择它们自己的路径,因为 $best_r$ 喜欢 eBGP 路径优于 iBGP 路径。由此证明算法是正确的。

定理 1 如果每个路由器都可以对候选路径集合范围内的所有路径形成一个总的顺序,那么算法为所有的路由器 $r \in R$ 正确计算了 BGP 路径选择过程的结果 b_r 。

证明:假设有路由器 r 选择了路径 b_r ,与算法分配的路径 b_r' 不同。这种不匹配可以发生于以下两种情况:(1)当 b_r 是由 *down* 中的会话学到的;(2)当 b_r 不是由 *down* 中的会话学到的(即从 *up* 或 *over*)。

考虑情形(1),其中 b_r 是由 *down* 中的会话学到的。 b_r' 错误计算的第一个例子(即算法已经正确计算出等级中 r 以下的的所有路由器的最佳路径);因为我们检查第一个这种不匹配 I_r 是正确的。如果 b_r' 也在 *down* 上,当向上运行算法时 $b_r' = best_r(I_r \cup E_r)$,这意味着按照 BGP 路径选择过程 b_r' 比 b_r 好,而且 r 将实际选择 b_r' 。如果 b_r' 在 *up* 或 *over* 上,按照 BGP 路径选择过程,它比 *down* 上的 b_r 好。但是之后,通过定义 $best_r$,路由器 r 也将选择 b_r' 。因此算法为所有的从 *down* 中选择最佳路径的路由器 r 正确计算 b_r 。

考虑情形(2), b_r 是由 *up* 或 *over* 会话学到的。从证明的前半部分可知算法从 *down* 中选择最佳路径的所有的路由器 r 正确计算 b_r ,所以叫作 b_r' 错误匹配的第二个例子,即一些从 *up* 或 *over* (即算法为等级中比 r 高的所有路由器正确分配 b_r) 中选择最佳路径的路由器。因为考虑第一个这种不匹配,

所以我们知道 I_r 是正确的。如果算法选择的路径 b_r' 在 *down* 上,那么由约束 4(1),BGP 不会选择 b_r ,因此产生矛盾。如果 b_r 和 b_r' 都是从 *up* 和 *over* 会话中学到的,那么两个都在 I_r 中,按照算法的 $best_r(I_r \cup b_r)$ 这一步及 $best_r$ 的定义,算法和 BGP 路径选择过程都会选择相同的路径。

3.3 算法的实现

仿真实验工具采用开放源代码的 C-BGP,它可以很方便地模拟大型 ISP 的域内及域间的 BGP 路由协议,它的最大特点是支持路由策略,配置命令与 CISCO 路由器的命令非常相似。我们构造了一个如图 3 所示的拓扑,IGP 拓扑是算法的固定输入。为了模拟一些内部网络故障(当 IGP 链路失效时)和维修操作(当路由器失效时),在运行算法之前可以对 IGP 图进行修改。其中 AS1 中配置了 3 个路由反射器 RR1, RR2, RR3。AS1 与 AS2 和 AS3 之间配置了路由策略来保证系统的收敛。仿真过程是:在系统收敛的情况下,保持 AS1 与 AS2 和 AS3 之间的路由策略不变,改变 AS1 中的拓扑。首先仿真了路由反射器的故障,3 个路由反射器有一个失效的情况下,对系统路由选择结果的影响,重点研究了如何配置路由反射器的冗余备份问题;其次仿真了路由反射器之间的链路故障,对系统路由选择结果的影响;最后仿真了路由反射器与其它路由器之间的链路故障对系统路由选择结果的影响。

在 DELL650 的工作站上运行了模拟程序,模拟路由反射器故障时发现图 3 所示的拓扑是有缺陷的,3 个路由反射器中的任何一个出现故障均会导致致命的问题。因此,把图 3 的拓扑改成图 4 所示的拓扑,对路由反射器做 iBGP 全网,为了避免路由循环与路由震荡,需要在路由器上配置路由策略,对入、出路由进行过滤。实验数据显示,对模拟的 ISP 内部路由器的配置调整都会引起 BGP 路由表的变化,如表 1 所列。

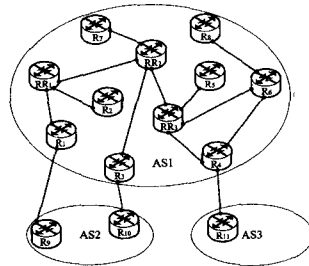


图 3 模拟的 ISP 拓扑

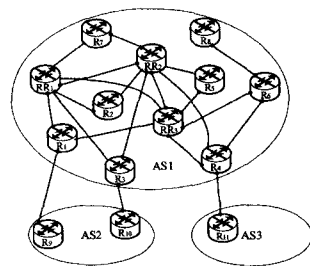


图 4 配置了路由反射器冗余备份的模拟拓扑

表 1 稳定 BGP 拓扑的路由表变化

AS1 内部调整的项目	AS1 内部路由器路由表的变化(%)	AS1 边界路由器的 BGP 表的变化(%)
IGP 链路	24.6	35.8
内部路由器故障	23.4	76.5
边界路由器故障	63.7	64.3
路由反射器故障	68.3	51.9
入口过滤策略	43.5	43.6
出口过滤策略	0	17.3

计算复杂度:算法两次遍历路由反射器的层次。该算法的运行时间是 $O(N + |S|)$,其中 N 是 eBGP-learned 路径的数量, $|S|$ 是 iBGP 会话的数量。从路由反射器等级的低端开始,算法必须在 N 条路径之间实行比较,以确定层级低端的路由器选择的路径(层级低端路由器的数量无关紧要;这种比

(下转第 128 页)

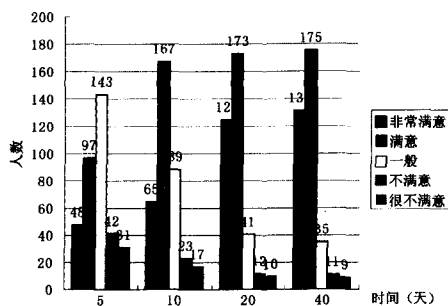


图6 学习者满意程度调查结果

10天后,64.27%的学习者对该系统满意,说明该系统已经得到大部分学习者的认可和喜爱;24.65%的学习者感觉一般,这部分学习者有的是原来就感觉一般的,有的是本来感觉不满意,经过更长时间的试用,感觉变好了的。

20天后,满意该系统的学习者达到82.55%,感觉不满意和很不满意的仅占8.86%。这说明经过20天的使用,绝大部分学习者都已经感受到系统的优势,已经接受并喜欢该系统。

40天后,84.76%的学习者对该系统满意,感觉一般和不满意的分别占9.70%和5.54%。与前一次调查相比,满意的人数有增,不满意的人数有减,但是变化不大。这说明该系统的用户群已经确定。之所以还有5.54%的学习者对该系统不满意,我们分析原因有3点:①系统中的知识点还不是十分完善,系统所提供的知识点并不是他们所需的,而他们想学习的知识点系统中还没有建立起来;②他们不适应网络学习模式;③系统界面不够友好,或者他们真的不喜欢该系统。对于第①种情况,可通过加强知识点维护,或者允许学习者自己有针对性地维护知识点来解决;对于第②种情况,可以通过对他们进行定向培训的方式解决;对于第③种情况,需要我们完善系统的界面,并有针对性地满足他们的需求。

结束语 本文针对在当今的现代远程教育领域,教育资源重用难,学习内容选择难现状,提出了以知识点本体来组织教学内容,生成个性化课程。论文构建了计算机科学与技

术学科的知识点本体,以知识点本体为核心的知识管理不但利于教育工作者组织教学内容,更适合于学习者的自主学习,还有利于知识的组织和管理以及教育资源的发布。

参考文献

- [1] 陈智,隋光远,皮秀云.论知识点是人的认知单位[J].心理科学,2002,25(3):369-370
- [2] 谢深泉.知识点及其网络的特性分析[J].软件学报,1998,9(10):785-789
- [3] 施岳定,张树有,项春.网络课程中知识点的表示与关联技术研究[J].浙江大学学报:工学版,2003,37(5):508-511
- [4] Engel G, Roberts E. Computing Curricula 2001[R]. IEEE Computer Society, Association for Computing Machinery, 2001
- [5] Jean - Marc R, Gracia G. Experimental study on the reuse of learning objects and teaching practices[C]//Proc. Int'l Conf. on Education and Technology. Calgary, Canada, 2005:107-112
- [6] 郭东,胡亮,郭冰心,等.基于应用偏好模糊聚类的网格资源选择[J].仪器仪表学报,2008,29(7):1403-1407
- [7] Sanchez - Alonso S, Garcia - Barriocanal E. Making use of upper ontologies to foster interoperability between SKOS concept schemes[J]. Online Information Review, 2006, 30(3): 263-277
- [8] 向昉,周旭.基于知识点模型的通用ICAI系统[J].湘潭大学自然科学学报,2006,22(2):17-21
- [9] 丁雅娴.学科分类研究与应用[M].北京:中国标准出版社,1994:54
- [10] Jovanovic J, Gašević D, Devedžić V. Dynamic assembly of personalized learning content on the semantic Web[C]// Proc. of the 3rd European Semantic Web Conference 2006 (ESWC2006). Budva, Montenegro: Springer Verlag, 2006:544-558
- [11] Felder R M, Silverman L K. Learning and teaching styles in engineering education[J]. Journal of Engineering Education, 1988, 78(7):674-681
- [12] 朱郑州,吴中福,吴开贵.基于用户满意度的学习服务发现算法[J].计算机研究与发展,2008,45(7):1161-1168

(上接第99页)

较可以通过从源 N 路径中构建一个 M 路径的子集来执行,可以对 N 路径进行一次扫描。之后算法可以将 M 路径的选择传播给等级中的下一层,那里 S_l 比较必须在下一个最高层的路由器中进行,其中 S_l 是 l 层中iBGP会话的数量。在等级中向上重复这一过程总共的运行时间是 $O(N+|S|)$ 。

带路由反射的算法计算BGP路径选择过程结果的运行时间不再复杂,尽管计算结果的过程涉及的更多。在带有路由反射的iBGP拓扑中,会话数目 $|S|$ 实际将比iBGP全网时的 $|R|^2$ 少;因此,算法的运行时间从路由反射器减少了iBGP拓扑中会话数量这一事实中受益。

结束语 本文给出了一种不需模拟BGP协议的复杂操作,预测了带有路由反射器的BGP路由的方法。利用C-BGP模拟了一个大型AS,并运行了我们的预测算法。结果显示,在现实的网络中如果进行配置的变动有可能引起系统的不收敛。这个算法可以应用于运营商在修改网络配置之前进行风险评估,可以减少由于管理员的误操作而引起的网络故障。

参考文献

- [1] Oran D. OSI IS-IS Intra-domain Routing Protocol[S]. Internet Engineering Task Force, RFC 1142, Feb. 1990
- [2] Feamster N, Rexford J. Network - Wide Prediction of BGP Routes[R]. georgia Tech TR GT-CSS-2006-003
- [3] Griffin T, Shepherd F B, Wilfong G. The Stable Paths Problem and Interdomain Routing[J]. IEEE/ACM Transactions on Networking, 2002, 10(1): 232-243
- [4] Gao L, Rexford J. Stable Internet routing without global coordination[J]. IEEE/ACM Transactions on Networking, Dec. 2001: 681-692
- [5] Feldmann A, Greenberg A, Lund C, et al. Deriving Traffic Demands for Operational IP Networks: Methodology and Experience[J]. IEEE/ACM Transactions on Networking, 2001, 9(3): 257-270
- [6] Griffin T, Wilfong G. On the Correctness of IBGP Configuration [C]//Proc. ACM SIGCOMM. Pittsburgh, PA, Aug. 2002:17-29