DOI:10.13295/j.cnki.jlut.2009.05.036
第35卷 第5期 兰州理工大学学报 Vol.35 No.5
2009年10月 Journal of Lanzhou University of Technology Oct.2009

文章编号:1673-5196(2009)05-0099-04

BGP网络故障模型与检测算法

包广斌,袁占亭,张秋余,邱 剑

(兰州理工大学 计算机与通信学院,甘肃 兰州 730050)

摘要:BGP网络故障严重影响自治系统间的网络性能,根据Internet中观测点获得的BGP路由信息,描述域间路由系统的BGP网络模型,建立BGP故障事件模型,为BGP网络的拓扑变化提供一种简明的描述方式.根据高度时间相关性的BGP路由事件触发的网络拓扑变化,提出查找故障链路的近似算法.提出的BGP网络故障查找模型和算法能够比较准确地检测Internet域间路由系统中的路由异常事件.

关键词:域间路由;网络故障;网络拓扑;检测算法

中图分类号:TP393 **文献标识码**:A

Model of BGP network fault and its detection algorithm

BAO Guang-bin, YUAN Zhan-ting, ZHANG Qiu-yu, QIU Jian

(College of Computer and Communication, Lanzhou Univ. of Tech., Lanzhou 730050, China)

Abstract: BGP network fault seriously affected internet performance among the autonomous systems. According to the BGP routing information acquired from the internet access to observation points, the BGP network model of inter-domain routing system was described and the BGP fault event model was established, which provided a concise description of the BGP network topology changes. On the basis of the topology change triggered by the routing events with high time correlation, an approximate algorithm was proposed in order to find the fault links. The BGP network fault model and the algorithm could detect the abnormal events in the inter-domain routing system.

Key words: inter-domain routing; network fault; network topology; detection algorithm

Internet路由系统是互联网的核心基础设施,它分为域内路由和域间路由.域间路由协议是自治系统之间联系的纽带,提供其他自治系统网络的可达信息,对端到端的服务质量有非常重要的作用.BGP(border gateway protocol)目前已成为事实上的Internet域间路由协议,把全球25 000多个自治系统连为一体,组成了全球最大的BGP路由网络.故障管理对于一个网络的可用性和鲁棒性具有十分重要的意义,找到一种有效的故障检测技术已成为网络管理的关键.

要进行Internet域间路由系统的故障检测最好是依靠BGP协议本身.BGP协议采用增量更新机制进行路由信息的更新和传播,当且仅当一个BGP路由器发现了网络的路由变动后,它才会向其他BGP路由器发布路由更新消息来通告路由信息的变化.因此BGP路由器本身就是一个故障检测器.在本地网络中发生的路由变化可能会经由BGP系统传播到整个Internet.而在Internet骨干网中的BGP路由器会接收到来自整个Internet的路由更新消息.这些路由更新消息是由网络中的路由事件触发的.本文通过分析BGP路由更新消息的特征来反向推测触发这些消息的路由事件的可能行为和位置,从而实现Internet域间路由系统的故障检测.

1 BGP路由运行机制和通告原则

1.1 BGP协议的运行机制

BGP协议[1]有4种报文:OPEN,UPDATE,KEEPALIVE,NOTIFICATION.当2个BGP路由器配置成对等体后,其中一个BGP路由器首先发送OPEN报文请求建立BGP连接,然后双方协商

收稿日期:2009-05-05

基金项目:国家自然科学基金(50877034),甘肃省自然科学基金(2007GS04107)

作者简介:包广斌(1975-),男,甘肃兰州人,博士生,副教授.

BGP 会话参数. 若协商失败,将发送 NOTIFICATION 报文通告错误信息,同时关闭连接. 若协商成功,BGP 会话正确建立,BGP 对等体间交换全部已有的路由信息. 只有在网络路由发生变化的情况下,才通过 UPDATE 报文发送路由更新消息,或者添加一条新的网络前缀,或者撤销一条无效的路由前缀. 当没有可交换的路由信息时,对等路由器之间交换 KEEPALIVE 报文来维持 BGP 会话.

1.2 BGP 路由通告原则

自治系统根据商业关系制定路由策略,一个 AS(autonomous system)只有愿意为某个 AS 承载到目的网络的流量时,才向该 AS 通告路由. Huston 等人[2]给出了配置路由输出策略时需要遵守的原则:

1) 输出给一个提供者:当一个客户向提供者通告路由信息时,作为客户的自治系统可以输出自己的路由和它客户的路由,但不能输出从其他提供者或对等者获得的路由.

2) 输出给一个客户:当提供者向客户通告路由信息时,作为提供者的自治系统可以输出自己的路由和它客户的路由,也可以输出从其他提供者或对等者获得的路由.

3) 输出给一个对等者:当与对等者交换路由信息时,可以输出自己的路由和客户的路由,但不能输出从其他提供者或对等者获得的路由.

2 网络模型

通过 Internet 骨干网路由器上获得的 BGP 路由信息,可以建立 BGP 网络模型,如图 1 所示,Internet 网络中 BGP 数据收集器(简称 SJQ)和自治系统 A 的 Internet 骨干网路由器建立 E-BGP 连接. 自治系统 A 向 SJQ 输出所有的最佳路由. 本文的故障检测方法对 SJQ 从自治系统 A 接收的 BGP 路由更新消息进行分析,自治系统 A 称为被观测自治系统(简称 GCD).

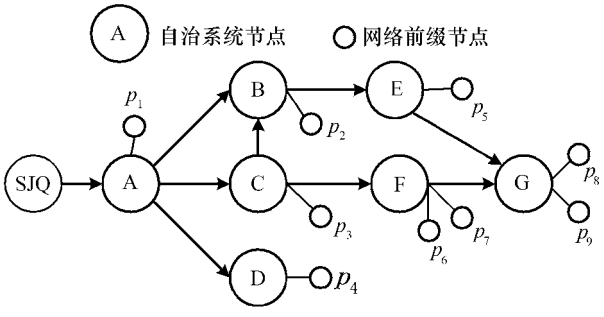

图 1 BGP 网络拓扑

Fig. 1 BGP network topology

BGP 是一种基于策略的路径向量协议,每一条 BGP 路由或路由更新消息一般都包含了地址前缀 p 和相关的自治系统路径 P,该路径指明了从本地自治系统到达地址前缀 p 需要顺序经过的自治系统序列. 一条路径开始于 GCD,结束于目的网络前缀 p. 有 2 类 BGP 路由更新消息,一种为路由宣告消息,用于携带到达相应地址前缀的有效路由信息;另一种为路由撤销消息,用于通知相应地址前缀不可达. 如果地址前缀 p 不存在于路由表中,或者关于 p 的路由更新消息是一个路由撤销消息,p 的路径 P 定义为空路径. 一个自治系统节点代表一个自治系统,一个前缀节点代表相应自治系统拥有的网络前缀. 自治系统链路和前缀链路都是有向边,其方向由导出这些链路的路径的方向决定. 给定了若干 GCD 的 BGP 路由,其导出的节点和链路可生成一个有向图[3] $G=(V, E)$,相应的路由表如表 1 所示.

表 1 BGP 网络路由表

Tab. 1 BGP network routing table

网络前缀节点	网络路径
p^1	A-i
p^2	A-B-i
p^3	A-C-i
p^4	A-D-i
p^5	A-C-B-E-i
p^6, p^7	A-C-F-i
p^8	A-C-F-G-i
p^9	A-C-B-E-G-i

有向图 G 构建在 BGP 的路径信息基础之上,该网络拓扑只是一个建立在下层路由器层次网络之上的"虚拟"自治系统层次的网络. 为了和下层的路由器组成的物理网络区分,本文称该网络为 BGP 网络. 但 BGP 网络并不代表完整的 Internet 自治系统拓扑,该网络代表了整个 Internet 拓扑中能被 GCD 观察到的部分.

3 网络故障模型

所谓网络故障检测就是发现路由器层次网络中的路由事件. 由于 BGP 路由更新消息中通常只包含某网络前缀的可达信息和相应的自治系统路径信息,能够从中获得的仅仅只是 BGP 网络的拓扑变化. 因此不可能从 BGP 路由更新消息中直接推断出路由事件[4]. 本文计划建立一个 BGP 网络故障事件模型来描述 BGP 网络的拓扑变化,并用 BGP 网络故障事件来进一步推断相应的路由事件.

3.1 路径问题

路由事件是发生在 Internet 中的一个路由拓扑的变化,该变化被相应的 BGP 路由器检测到,并触

发了路由更新消息.被触发的路由更新消息揭示了路由事件的存在.一个路由事件可以由软件引起,如:路由软件故障,路由策略的错误配置;也可以由硬件引起,如:路由器停机,通信线缆断裂等.

文献[5]称在 Internet 的一个自治系统中,同时发生多个路由事件的概率极低.因此,本文认为如下假设在绝大多数情况下都成立:在一个足够短的时间内,BGP 网络中的一条链路或节点上不可能同时发生多个独立的路由事件.

当 BGP 网络通向某个前缀的路径上发生了路由事件,相关的路由器将会宣告路由更新消息来调整到达该前缀的路由,这个过程称为路由收敛.在收敛过程中,由于缺乏全局的路由拓扑信息,单个 BGP 路由器往往会依次尝试所有其已知的到达该前缀的路径,从而会向外宣告多个瞬态路由更新消息,直至最终找到最佳的路由[6].这个过程导致了 BGP 的慢收敛.这些瞬态更新消息是 BGP 路由器在路由动态调整的过程中生成的,并不一定包含有效的路径信息,因此不能用来推断 BGP 网络的拓扑变化.

假设在一个 BGP 网络中,有若干个路由事件在时刻 T 几乎同时发生.在路由收敛过程中,被观测点 A 向数据收集器 SJQ 依次宣告关于前缀 p 的一系列路径 $P_1, P_2, \cdots, P_n$.这里称前缀 p 受到这些路由事件的影响,即受影响前缀在时刻 T 前,前缀 p 的路径为 P_0.当收敛过程结束后,p 的路径为 P_n.前缀 p 及其稳态路径 P_0 和 P_n 描述了一个该前缀的路径变化,P_0 和 P_n 可以是空路径也可以是有效路径.

本文研究的路径变化集合中包含的路径变化都是由若干具有高度时间相关性的路由事件触发的.在这样的情况下,一个受影响前缀只可能有一个路径变化.受影响前缀和路径变化之间可以建立一一对应的关系.因此,一个路由事件影响的前缀可以唯一代表在路由事件中该前缀的路径变化,反之亦然;一个路由事件影响的前缀集合可以代表这些前缀的路径变化组成的集合,反之亦然.

3.2　链路状态

假设若干路由事件在时刻 T 发生,前缀 p 是一个受影响的前缀,其路径从 P_0 变为 P_n.对于路径 P_0 的所有导出链路,前缀 p 称为它们的 P_0 前缀,表示这些链路中的任意一条失效都会导致前缀 p 放弃路径 P_0.对于路径 P_n 的所有导出链路,前缀 p 称为它们的 P_n 前缀,这些链路中的任意一条可用都会促使前缀 p 采用 P_n.因此,一个链路本身与 P_0 集合、P_n 集合、稳定集合构成了一个链路变化.所有这些链路在路由事件中的链路变化构成了 BGP 网络的拓扑变化.路由事件是在下层路由器层次网络发生的变化,在上层的 BGP 网络中相应的是和该路由时间起源相关联的链路,这样的链路称为故障链路.显然,一条故障链路必定是变化链路,但是一条变化链路却不一定是故障链路.

一条前缀链路只包含了一个前缀.因此,在路由事件中前缀链路只能按 3 种方式变化:消失、重现和波动.对于一条自治系统链路,它可能包含了若干条前缀,变化方式会比较复杂[7],如表 2 所示.

表 2　域间路由的链路状态

Tab.2　Link state of inter-domain routing

类型	定义
未变化	$A^e=\Phi$ 且 $A^s\neq\Phi$
完全变化	$A^e\neq\Phi$ 且 $A^s=\Phi$
部分变化	其他
撤销	$A^D\neq\Phi$ 且 $A^U=\Phi$
恢复	$A^D=\Phi$ 且 $A^U\neq\Phi$
波动	$A^D\neq A^U$
撤销-波动	$A^U\subset A^D$ 且 $A^U\neq\Phi$
恢复-波动	$A^D\subset A^U$ 且 $A^D\neq\Phi$

3.3　BGP 故障事件

文献[7]研究了 Sprint Link IP 骨干网上链路中断故障的特征,发现大约 20%的链路中断故障是由于定期的网络维护造成的;剩余的 80%都是突发的意外故障,其中大约 30%是并发故障,是由于这些链路共享的路由器或光缆出现故障造成的,另外 70%的故障才是单独链路的故障.基于稀疏性假设,如果若干条共享了起始或终止节点的故障链路具有相似的变化类型,这些故障链路很可能都是由于同一路由事件造成的,因此这些链路需要聚集起来以对该路由事件进行刻画.

BGP 网络中单个链路变化可以组成一个 BGP 故障事件,BGP 故障事件是 BGP 网络中故障点的抽象,它指出对应的路由事件的可能位置.因此可以在 BGP 网络中的 BGP 故障事件和实际路由器层次网络中发生的路由事件之间建立一一对应的关系.

一个 BGP 故障事件是若干个路径变化的组合,是拓扑变化的另外一种描述方式.但 BGP 故障事件把多个路径变化聚集在一起,以一种更加简洁的方式描述了 BGP 网络的拓扑变化.本文通过建立 BGP 网络的描述,更好地刻画了 BGP 系统中的动态变化.

BGP 故障事件只描述了 BGP 网络的拓扑变化,网络故障检测实质上感兴趣的是下层路由器层次网络的路由事件.路由事件和 BGP 故障事件之间

存在因果关系. 给定一个 BGP 故障事件 f, 其描述的拓扑变化是一个或若干个路由事件单独作用的结果. 这些路由事件组成一个等价类 E_f, 使得对于 $\forall e \in E_f$, 其在域间路由系统中发生后引发的拓扑变化都是 f. 给定一个 BGP 故障事件 f, 在已知下层网络的拓扑和相关信息的前提下, 可以推测出对应的等价类 E_f, 可以缩小待求的路由事件的范围, 进而确定相应的路由事件. 因此本模型所能检测的故障是 BGP 网络层次的 BGP 故障事件, 可根据 BGP 故障事件以及其他先验网络知识推断相应的路由事件.

4 近似算法

假设 BGP 系统中若干个路由事件在时刻 T 几乎同时发生, 从接收到的具有很强时间相关性的路由更新消息中可以获得所有这些路由事件引发的路径变化, 进而得到相应的链路变化和 BGP 网络的拓扑变化. 通过路径变化和拓扑结构的相关性, 发现隐含的 BGP 故障事件. 其形式化描述如下: 假设若干个路由事件于时刻 T 几乎同时独立地发生, 已知所有可疑链路的路径变化, 找出 k 个 BGP 故障事件 $S=\{f_1, f_2, \cdots, f_k\}$ 满足 $\bigcup_{i=1}^{k} R_{f_i} = \Delta C$ (其中 R_f 是 BGP 故障事件的变化前缀集合, ΔC 表示路径变化集合), 并且使得 $R_{f_i} \cap R_{f_j}$ 最小. 这是一个 NP 完全问题[8], 无法找到多项式时间的算法, 本文采用近似算法得到优化解.

第一步, 把可疑链路按最大化原则构造候选 BGP 故障事件. 具体如下: 设 $|V_a|$ 代表自治系统节点集合, 对每个和可疑链路相连的自治系统节点, 根据相连的可疑链路的变化类型划分成 3 类, 每一类的链路有相似的链路变化类型, 组成一个候选 BGP 故障事件, 最多产生 $3|V_a|$ 个候选 BGP 故障事件.

第二步, 按照贪婪策略寻找可行解. 首先按照候选 BGP 故障事件的主前缀集合的大小, 把候选 BGP 故障事件进行排序, 从最小的 BGP 故障事件开始, 逐一对 BGP 故障事件 f_i 进行筛选. 若 f_i 和另一 BGP 故障事件 f_j 等价, 将 f_i 放入 f_j 的等价类, 并删除 f_i; 若 f_i 被 BGP 故障事件 f_j 包含, 则删除 f_i; 若 f_i 和 f_j 有重叠前缀, 则删除 f_i 中的链路, 使重叠前缀最小. 算法如下所示:

```
{
for i=n to 1   //n是候选故障事件个数
for j =1 to i-1
if R_{f_i}=R_{f_j}
equivalent_class(f_j)=equivalent_class(f_j)∪{f_i},
remove f_i
else if R_{f_i}⊂R_{f_j}
remove f_i
else if R_{f_i}∩R_{f_j}≠∅
foreach uv∈f_i
if A_u⊂R_{f_j}
f_i=f_i-{uv}
}
```

算法主体部分由 2 个循环构成, 候选故障事件个数 $n \leqslant 3|V_a|$, 所以本算法的复杂度为 $O(|V_a|^2)$. 这是一个近似算法, 输出结果包括等价解, 其结果有进一步优化的空间.

5 结论

BGP 故障事件是 Internet 中路由事件的抽象, BGP 故障事件能够较好地反映 Internet 中的路由事件. 本文根据 Internet 中观察点获得的 BGP 路由信息, 建立了 BGP 故障事件模型, 用 Java 实现了整个 BGP 故障检测近似算法, 对 CERNET (AS4538) 的 BGP 试验数据进行分析, 发现能够比较准确地检测域间路由故障事件, 为下一步 BGP 网络故障检测算法优化奠定了基础.

参考文献:

[1] REKHTER Y, LI T. A border gateway protocol 4 (BGP-4), RFC 4271 [EB/OL]. (2006-11) [2009-04]. http://www.cnpaf.net/Class/Rfcen/200611/18136_10.html.

[2] HUSTON G. Analyzing the Internet's BGP pouting table [J]. The Internet Protocol Journal, 2001, 4(1): 2-16.

[3] 兰家隆, 刘 军. 应用图论及算法 [M]. 成都: 电子科技大学出版社, 1995.

[4] VILLAMIZAR C, GOVINDAN R. BGP route flap damping, RFC 2439 [EB/OL]. (2006-11) [2009-04]. http://www.cnpaf.net/Class/Rfcen/200502/3502_2.html.

[5] KRISHNAMURTHY B, WANG J. Topology modeling via cluster graphs [C]//Proc. of ACM SIG COMM '2001. San Diego: [s.n.], 2001: 19-23.

[6] BAO Guangbin, YUAN Zhanting. A novel algorithm to optimize QoS multicast routing [C]//HUANG Deshuang, LI Kang, IRWIN G W. Intelligent Control and Automation: International Conference on Intelligent Computing. New York: Springer-Verlag, 2006: 150-157.

[7] LABOVITZ C, MALAN G R. Internet routing instability [J]. IEEE/ACM Transactionson Networking, 1998, 6(5): 515-528.

[8] 李昌兵, 胡 华, 杜茂康, 等. 基于免疫遗传算法的多播 QoS 路由算法 [J]. 兰州理工大学学报, 2008, 34(5): 105-109.