域间路由不稳定性的分析方法研究*

王旸旸[1,2]，翟 鹏[2,3]，毕经平[2]

(1. 首都师范大学 计算机科学联合研究院，北京 100037；2. 中国科学院 计算技术研究所，北京 100080；3. 山东科技大学，山东 青岛 266510)

摘 要：从分析 <对等体，前缀> 二元组的 BGP 更新消息序列出发，从四个不同层面提出了域间路由不稳定性的分析指标和方法。在此基础上开发出工具 FlapView，可以全面有效地监测域间路由的不稳定性。

关键词：BGP 路由；域间路由；路由不稳定性；网络测量

中图法分类号：TP393.04 文献标识码：A 文章编号：1001-3695(2006)04-0240-03

Analysis Method of Inter-domain Routing Instability

WANG Yang-yang[1,2], ZHAI Peng[2,3], BI Jing-ping[2]

(1. *Joint Faculty of Computer Scientific Research, Capital Normal University, Beijing* 100037, *China*; 2. *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080, *China*; 3. *Shandong University of Science & Technology, Qingdao Shandong* 266510, *China*)

Abstract: This paper starts with analysising the BGP update message array specified by the given < peer, prefix > tuple, then makes comprehensive evaluation from four different different levels. Some useful measurement metrics and methodologies have been defined, based on which the tool named FlapView was developed for monitoring inter-domain routing instability.

Key words: BGP Route; Inter-domain Route; Routing Instability; Network Measurement

今天的 Internet 和较大的 ISP 的网络被分成大量的自治域(Autonomous System, AS)，自治域定义了管理控制的区域和作用于自治域范围的路由策略。在自治域的边界路由器上，通过运行外部网关协议，与其他自治域的边界路由器交换路由信息。目前外部网关协议的事实标准是 BGP-4。

BGP 路由更新消息有两种类型，即路由宣告和路由撤销。路由宣告消息表明到某目的前缀有新的或者更优的路由；路由撤销说明到某目的前缀的路由被撤销，不可达。撤销可以分为两种方式，即显式撤销和隐式撤销。显式撤销明确发出撤销消息；隐式撤销是指存在的路由没有用明确的撤销消息撤销掉，而直接被新宣告的路由代替。在一个 BGP 更新消息报文中，可以包含多条路由宣告和撤销。

路由不稳定性主要表现为到某个目的前缀的路由在短期内快速地反复撤销和重现，这也称为路由震荡、路由波动。在路由协议上表现为 BGP 路由器发出一个路由宣告，然后又很快撤销它。一个接收更新和撤销消息的路由器会把这些消息传播给它的对等体。如果这种行为继续传播下去，路由系统性能将会受到严重影响。严重的不稳定性会增加路由收敛延迟、数据的传输时间、丢包率、路由器 CPU 和内存等资源的开销。

有很多因素造成路由不稳定性，文献[4]中作了详细讨论。通常主要有路由器配置错误，短暂的物理或数据链路故障、路由协议软件缺陷、路由器故障、链路严重拥塞等。

目前降低路由不稳定性的主要方法有路由抑制算法(Route Dampening)和路由聚类(Route Aggregation)。其中路由聚类是把一些较小的 IP 前缀聚合，产生一个较为不明确的路由宣告，这些方法在一定程度上可以减少路由的不稳定因素。然而，随着 Internet 上 ISP 和骨干交换网提供商的网络变得越来越庞大和复杂，路由聚类和路由抑制算法等方法的效果也变得越来越弱。因此在对路由系统的监控中，路由不稳定性的监测就至关重要。

收稿日期：2005-04-14；修返日期：2005-05-26

基金项目：国家自然科学基金资助项目(90104006)；国家自然科学基金 QoS 资助项目(60273021)

1 相关研究

文献[1]从美国 Routing Arbiter 项目采集五个核心公共交换节点的 BGP 更新消息，对九个月内的域间路由不稳定性进行了分析。文献[2]也对路由不稳定性作了分析。这些研究侧重探寻路由不稳定性的某些特性和原因。文献[3]讨论了在 AS 级上定位不稳定性来源的方法。另外 ASExploer[5]，BGP-Player[4] 这些工具，通过表现网络拓扑和路由的变化，在一定程度上也可以反映 AS 级路由的不稳定现象。然而，上述工作在不稳定性的分析指标和方法上非常有限，无法实时详细地监测和分析域间路由的不稳定性状况。

2 实验数据

我们实验的数据是从 RIPE NCC 的 RIS 项目网站上[6] 获得的。RIS 是 RIPE NCC 组织 1999 年的一个项目，目的是在 Internet 上一些主要交换节点基本实时地采集 BGP 路由更新信息，并存储在数据库中，以便网络运营商分析和发现路由存在的问题。另外，这些数据压缩后在 Internet 上公开，用于研究。现在有 12 个数据收集点，多于 300 个 IPv4 和 IPv6 对等路由器，分布在欧洲、日本和北美。图 1 是 RIS 的路由信息采集系统的基本结构。

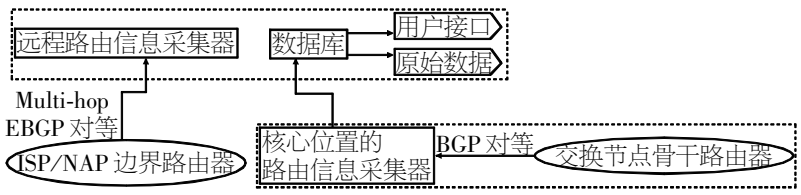

图 1　RIS (Routing Information Service)的基本结构图

在图 1 中, 运行 GNU-Zebra 路由软件的路由信息收集器, 与 ISP 或交换节点的骨干路由器建立 BGP 对等关系, 收集来自对等体路由器的 BGP 更新消息, 并把它们封装成 MRT 标准格式, 保存在数据库中, 供用户分析。用户可以使用 Multi-thread Routing Toolkit(MRT) 的相关文档和工具解析原始数据中路由信息的内容。

3　路由不稳定性分析指标和方法

为了实时全面地监测域间路由不稳定状况, 通常与多个路由器建立对等关系。这些路由器也因此称为对等路由器或者对等体。如何选择它们是个重要问题, 但本文不予讨论。

我们周期性地分析 T 内的不稳定状况。选取的周期 T 应该可以使得大部分的前缀能够出现至少两次路由的变化, 过短和过长的周期 T 都不利于对不稳定状况的监测和分析。在文献[7] 中选取 30 分钟时间作为短期路由表波动的上界, 因此, 我们这里选择 1 个小时作为 T。实验统计发现, 有 98% 的前缀可以出现至少两次路由变化。

3.1　事件定义

BGP 是路径向量路由协议, 域间路由不稳定性主要体现在 AS 级的路由变化, 我们根据文献[1] 以及能否反映 AS 级路由变化定义了如下三类事件:

(1) 重复宣告。它是两个连续且重复的路由宣告, 这里重复的路由宣告是指: 后面路由宣告的 AS_PATH 属性值和前面宣告并被隐式撤销的 AS 级路由相同。这表明一条路由被隐式撤销且被重复的路由代替。该事件不反映 AS 级路由的变化, 但它可能反映了病态行为或路由策略的改变。这种重复的路由宣告过多, 会加重路由器的负载, 降低路由系统的工作稳定性。

(2) 冗余撤销。它是两个连续关于同一不可达前缀的 BGP 撤销消息。到某一前缀的路由被显式撤销之后, 不应该再发出对该不可达前缀的撤销消息, 所以这种撤销是多余的。如果这种多余的撤销消息过多, 会加重路由器的负载, 降低路由系统的工作稳定性。

(3) 路由波动。这有两种情况: ①一条路由在它不可达时被显式或隐式地撤销, 后来它又被另一条到达相同目的的不同路由所代替。这对应于 $A_1W + A_2$ 和 A_1A_2 消息序列, 反映了 AS 级路由变化。其中 A_1 和 A_2 表示具有不同 AS-PATH 属性值的前缀 P 的路由宣告, "W + "表示一个或多个对前缀 P 的路由撤销。② 一个路由在不可达时被显式地撤销, 然后又被宣布可达。这对应 $A_1W + A_1$ 消息序列, 反映了短暂的拓扑故障(链路或是路由器) 。我们可以把这两种情况看成两类事件。

3.2　特定 < 对等体, 前缀 > 二元组的路由不稳定状况

该层面的分析是最基本的, 它展示了特定对等体 Peer 到特定前缀 Prefix 的路由不稳定状况。我们把 BGP 更新消息按 < 对等体, 前缀 > 二元组分组。对每一组消息, 按时间先后顺序排列。选取对应于 < Peer, Prefix > 的 BGP 消息序列组, 我们考查如下指标:

(1) 路由持续时间和有效时间。BGP 宣告消息的 AS_PATH 属性中的 AS 序列代表从 Peer 到 Prefix 的一条 AS 级路由 R, 路由 R 从宣告到被显式或隐式撤销的时间可以称为该路由的一次生存期。只要这次生存期与时间 T 有重叠, 我们就称该路由 R 在时间 T 内出现一次, 其重叠部分的时间就是在 T 内该路由一次出现的持续时间。设该路由在 T 内出现并持续了 n 次, n 次持续时间的总和称为该路由的有效时间, 记为 $A(R)$。< Peer, Prefix > 的路由可能有若干条, 其中有效时间最长的路由称为主路由。$A(R)/n$ 定义为路由 R 在 T 内的持续时间, 记为 $A_{avg}(R)$。持续时间说明路由出现后持久的程度, 持续时间和有效时间越长, 说明该路由越稳定。

(2) 前缀可达时间和可达持续时间。在分析时间 T 内, 所有出现的路由的有效时间之和记为 RT, 定义为 < Peer, Prefix > 的前缀可达时间。如果所有路由在 T 内总共出现 N 次, 则 RT/N 定义为 < Peer, Prefix > 的前缀可达持续时间, 记为 RT_{avg}。前缀可达时间长, 表明该前缀可达性良好; 较差的可达性反映了路由较差的稳定性。前缀可达持续时间说明了前缀可达状态持久的程度, 持续时间越长, 到该前缀的路由越稳定。

(3) 平均路径长度。设 $AS(R)$ 表示路由 R 经过的 AS 数量, 对 T 内出现的所有路由, $\sum \frac{A(R)}{T} AS(R)$ 定义为 < Peer, Prefix > 的平均路径长度。虽然也有一些特殊情况, 但是一般来说路径越长, 不稳定的概率越大[2]。

(4) 事件频度。单位时间内事件发生的数量定义为事件频度。事件频度越高, 表明路由的不稳定性越大。我们可以统计 T 内发生的事件数量来计算 < Peer, Prefix > 的事件频度, 也可以用(T 内发生事件的次数/ T) 计算更小单位时间的事件频度。

3.3　到指定前缀的路由不稳定状况

第 3.2 节前缀的数量可达十几万或更多, 我们不必要也不可能对它们逐一考察。本节在第 3.2 节的基础上, 分析所有对等体到指定前缀的不稳定状况, 展示出被监测网络中到指定前缀的整体不稳定状况。当发现它不稳定时, 再通过第 3.2 节进行更具体的考查。我们考查如下指标:

(1) 前缀可达时间和可达持续时间。我们使用均值来概括整体状况。设集合 PEER = { Peer |Peer 是一个对等体, 指定前缀 Prefix 是 Peer 的可达前缀} , C(PEER) 表示 PEER 中对等体的个数, $\frac{6}{Peer \in PEER} RT_{avg}(Peer, Prefix)/C(PEER)$ 定义为被监测网络中指定前缀 Prefix 的可达时间。类似地, $\frac{6}{Peer \in PEER} RT(Peer, Prefix)/C(PEER)$ 定义为被监测网络中 Prefix 的可达持续时间。其中 RT(Peer, Prefix) 和 RT_{avg} (Peer, Prefix) 是对应于 < Peer, Prefix > 的前缀可达时间和可达持续时间。

(2) 事件频度。同前面类似, 如果 Event(Peer, Prefix) 是对应于 < Peer, Prefix > 的某类事件的频度, 则 $\sum_{Peer \in PEER} Event(Peer, Prefix)/C(PEER)$ 为任一 Peer 到 Prefix 的某类事件平均频度, 定义为 Prefix 的事件频度。

3.4　网络局部的路由不稳定状况

我们不妨进行如下分析:

(1) 局部前缀可达时间和可达持续时间。设时间 T 内, 指定对等体 Peer 的所有可达前缀的集合是 P, Prefix 是 P 中的一

个前缀, $C(P)$ 表示 P 的前缀个数, 则 $\sum_{Prefix\in P} RT(Peer, Prefix)/C(P)$ 为 Peer 到 P 中所有可达前缀的平均可达时间, 定义为局部前缀可达时间。$\sum_{Peer\in P} RT_{avg}(Peer, Prefix)/C(P)$ 为 Peer 到 P 中所有可达前缀的平均可达持续时间, 定义为局部前缀可达持续时间。其中 RT(Peer, Prefix) 和 RT_{avg}(Peer, Prefix) 分别表示 < Peer, Prefix > 的前缀可达时间和可达持续时间。

(2) 局部事件频度。同上面的方法类似, 若 Event(Peer, Prefix) 表示 < Peer, Prefix > 的某类事件频度, 则 $\sum_{Prefix\in P}$ Event(Peer, Prefix) / $C(P)$ 为 Peer 到其任一可达前缀的某类事件平均频度, 定义为 Peer 的局部事件频度。

3.5 网络整体的路由不稳定状况

有时候我们关心整个被监测网络的不稳定状况, 第 3.3 节和第 3.4 节的分析无法给出。下面在第 3.3 节的基础上, 分析从所有对等体看到的, 网络到任一可达前缀的路由不稳定状况, 反映出被监测网络整体的稳定状况。我们作如下分析:

(1) 前缀可达时间和可达持续时间。一个对等体的可达前缀较多, 其观察到的局部范围则会较广, 因此它所观察到的局部不稳定状况越能代表整体状况。基于这种思想, 如果用 PEER 表示所有对等体的集合, Peer 是其中的一个对等体, CP(Peer) 表示 Peer 的所有可达前缀的数量, RT(Peer) 和 RT_{avg}(Peer) 分别表示第 3.4 节的前缀可达时间和可达持续时间, 则 $\frac{\sum_{Peer\in PEER} RT(Peer)\,CP(Peer)}{\sum_{Peer\in PEER} CP(Peer)}$ 定义为被监测网络的前缀可达时间。把表达式中的 RT(Peer) 换成 RT_{avg}(Peer), 得到的表达式定义为被监测网络的前缀可达持续时间。

(2) 事件频度。与前面一样, 如果 Event(Peer) 表示第 3.4 节的局部于 Peer 的事件频度, 把上面表达式中的 RT(Peer) 换成 Event(Peer), 则得到的表达式定义为被监测网络整体的事件频度。

4 分析工具和实验结果

我们使用上面的指标和方法开发了相应的分析工具 FlapView。该工具分析指定时间长度内的 MRT 格式封装的原始 BGP 更新消息, 使用灵活、通用性好。可以用于实时监测域间路由的不稳定状况。我们采集第 1 节描述的环境中的数据; 路由信息采集器 RCC0 和 14 个核心网络的骨干路由器对等, 我们选择 RCC0 上面 2004 年 11 月 24 日从 0 点开始的 24 小时的数据, 用 FlapView 进行分析。限于篇幅, 这里选取了其中一条最长的 BGP 消息序列的分析结果。这条序列对应于 AS3549 中的对等体 208.51.134.248 到前缀 202.64.49.0/24, 共 978 个宣告消息, 3 126 个撤销消息, 分析周期是 1 个小时。部分重要指标的结果如图 1 ~图 3 所示。

从图 1 可以看到, 在每个分析周期内, 前缀可达时间在 94% 以上, 说明从这个对等体观察, 该前缀的可达性良好。从图 2 可以看到, 重复宣告事件比路由波动事件少得多, 没有冗余撤销事件发生。在路由波动事件中, 绝大部分是在不同路由之间的波动, 极少量的波动是相同路由的撤销和重现, 这说明从这个对等体来看, 链路或路由器问题引起的拓扑瞬时故障很少, 路由波动主要是 AS 级路由的变化。从图 3 看出, 从该对等体到该前缀的可达持续时间在 30s 左右, 其大小与图 1 的可达时间和图 2 反映出的路由波动事件是一致的, 基本上等于可达时间与路由波动事件数量的比值。图 1 和图 3 没有出现很低的数值, 图 2 也没有出现很大的值, 这都说明分析时间内该对等体到该前缀的域间路由比较稳定。

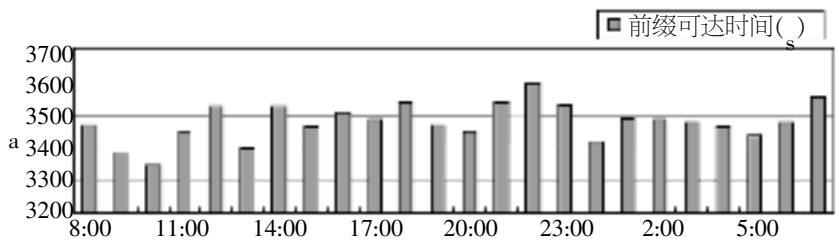

图 1 前缀可达时间

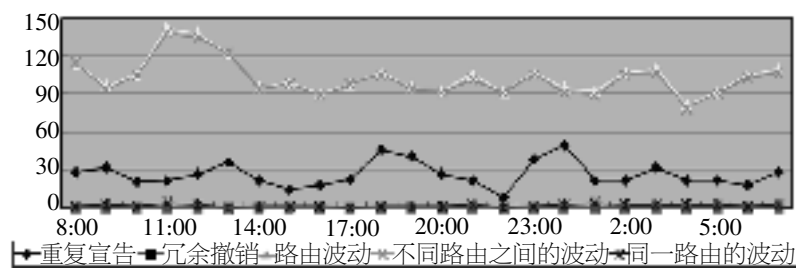

图 2 事件频度

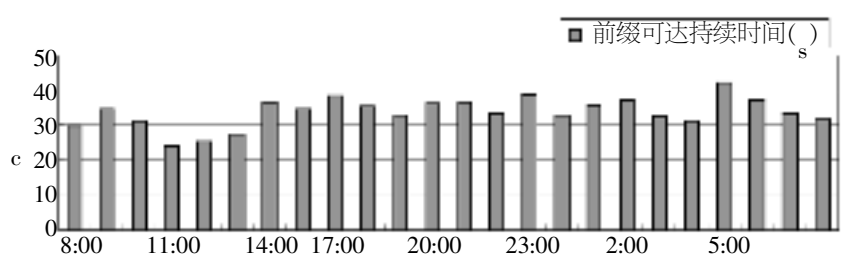

图 3 前缀可达持续时间

5 结束语

本文主要讨论了域间路由不稳定性的分析方法, 它从四个不同层面对域间路由的不稳定性进行分析, 提出相应的分析指标和方法, 能较全面而详细地反映域间路由的不稳定性状况。以此开发的工具 FlapView 实际分析了 RIPE 项目采集到的 BGP 更新消息数据, 分析结果较详细地反映出域间路由的不稳定性状况。本文所研究的分析方法和工具, 可用于对网络路由有效地监测和维护。

参考文献:

[1] Labovitz, G R Malan, F Jahanian. Internet Routing Instability[J]. IEEE/ACM Transactions on Networking, 1998, 6(5) : 515-528.

[2] Ramesh Govindan, Anoop Reddy. An Analysis of Internet Inter-Domain Topology and Route Stability[C] . INFOCOM, 1997. 850-857.

[3] Anja Feldmann, Olaf Maennel, *et al*. Locating Internet Routing Instabilities[J] . SIGCOMM, 2004. 205-218.

[4] Craig Labovitz, G Robert Malan, Farnam Jahanian. Origins of Internet Routing Instability[C] . INFOCOM, 1999. 218-226.

[5] http: //bgplay. routeviews. org/bgplay/, 2004-04[EB/OL] .

[6] http: //www. merit. edu/ ~ ipma/docs/Manual/asexplorer. html, 2004-04[EB/OL] .

[7] http: //www. ripe. net/projects/ris/index. html, 2004-04[EB/OL] .

[8] C Labovitz, A Ahuja, *et al*. Delayed Internet Routing Convergence [J] . IEEE/ACM Transactions on Networking, 2001, 9(3) : 293-306.

作者简介:

王旸旸(1979-) , 男, 安徽淮北人, 硕士研究生, 研究方向为网络测量; 翟鹏(1978-) , 男, 山东邹城人, 硕士研究生, 研究方向为网络安全; 毕经平(1974-) , 女, 副研究员, 硕士生导师, 博士研究生, 研究方向为网络测量。