文章编号:1001 - 2486(2006)03 - 0054 - 05

ATS:一种基于通告时间戳的 BGP 收敛性改进机制*

赵 锋,苏金树,王宝生
(国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:研究 BGP 收敛性改进问题,考虑网络运行状况,提出一种新的机制,称为通告时间戳机制 ATS (advertisement time stamp)。在该机制中,BGP 路由器在向对等体发送路由通告报文时,将相应的路由前缀打上时间戳。当通告报文准备好时,通过查看相应前缀的时间戳来判断报文是否可以立即发送还是应该等待。该机制充分融合了每对等体每目的网络速率限制定时器和每对等体速率限制定时器各自的优点,克服各自缺点。分析表明,该机制可以较大地减少 Tup、Tlong 及 Tshort 事件的收敛延迟。

关键词:边界网关协议;收敛性;速率限制定时器;时间戳
中图分类号:TP393 **文献标识码**:A

Improving BGP Convergence through Advertisement Time Stamp

ZHAO Feng, SU Jin-shu, WANG Bao-sheng
(College of Computer, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: This paper presents a new mechanism called BGP(border gateway protocol) with ATS (advertisement time stamp), which can greatly reduce routing convergence delay for BGP greatly. In the approach presented, when a BGP speaker sends a peer an advertisement packet, it marks the corresponding routing prefix/prefixes with the system time. When the next advertisement packet is ready, the BGP speaker decides whether the packet should be sent immediately or not via the corresponding time stamp checking. It shows that ATS takes advantage of the per-peer MRAI(min-route-advertisement-interval) Timer and per-peer per-destination MRAI Timer while overcoming their disadvantages. Theoretical analysis demonstrates that ATS can achieve a dramatic improvement in routing convergence for Tup, Tshort and Tlong events in comparison with BGP and other existing solutions.

Key words: BGP; convergence; MRAI; time stamp

实时的 IP 应用,比如声音、视频及交互式的网络游戏,要求 IP 网络稳定可靠。G.Hudson Gilmer[1] 指出,提高网络稳定性、可靠性的一个重要步骤就是要提高路由表的收敛性。导致网络拓扑发生变化的网络事件产生时,路由器的转发表应该快速收敛到最新的路由信息。假设每天发生一次协议收敛事件,每次持续 10 ~ 20s 的话,那么当前一年因协议收敛所损耗的代价估计可达上千万美元,而随着实时应用和用户数的增长,协议收敛事件所导致的损耗代价会更大。因此,为了减少收敛事件对实时服务直接和间接的影响,应该尽可能减少协议收敛时间。

Internet 的域间路由协议 BGP 解决了传统的距离矢量协议计数到无穷问题,因而人们曾认为它能在网络拓扑或者网络可达性发生变化后快速收敛(有文章认为该延迟是 30s 或更小[2])。然而 Craig Labovitz[3] 的研究结果表明,实际的恢复和重路由过程非常慢,平均需要 3min,路由表的波动有时会持续 15min。

Craig Labovitz 把 Internet 的路由事件划分为四类:Tup:原先不可用的路由被通告可以使用了,这代表路由恢复;Tdown:原先可用的路由被撤销了,这代表路由故障;Tshort:一条活跃的有比较长的自治系统路径的路由被一条具有更短自治系统路径的新路由代替,这代表路由恢复及路径切换;Tlong:一条活跃的有比较短的自治系统路径的路由被一条具有更长自治系统路径的新路由代替,这代表路由故障及路径切换。

* 收稿日期:2005 - 12 - 01
基金项目:国家重点基础研究发展规划 973 资助项目(2003CB314802);国家自然科学基金资助项目(90104001,90204005,90412011)
作者简介:赵锋(1980—),男,博士生。

在发现路由收敛过程比较慢之后,许多研究人员侧重于研究如何减少 BGP 的收敛延迟,特别是 Tdown 和 Tlong 事件的收敛延迟。BGP 收敛时间比较慢的一个原因是当故障产生时要探查大量的备份路径,而许多备份路径可能已经失效。为了减少探查的路径数,降低传播的消息数目,BGP-Assertion[4] 使用 AS 路径信息来检查路由一致性,确定不可行路由,在其网络测试环境中,该方法可以将 Tdown 事件的 BGP 收敛时间从 30.3s 减少到 0.3s,在 60 个自治系统组成的网络拓扑模拟实验之中,该方法可以将 Tdown 故障撤销收敛时间从 337s 减少到 19.5s。BGP-RCO[5] 给 BGP 引入一个新的过渡可选属性 RCO (route change origin),讨论了由于源撤销引起的 Tdown 事件的收敛属性,认为其可以得到较大改进。BGP-GF[6] 通过轻微修改 BGP 规则,使得如果一条路由切换到次优路由并且定时器还没超时,那么就发送路由撤销消息。Dan Pei[7] 引入一个 BGP-RCN(BGP with root cause notification)机制:每一个路由更新消息都携带有关触发路由更新的具体原因的信息,在 110 个自治系统组成的网络拓扑模拟实验中,当一个目的网络不可达时,该方法可以从 715.3s 减少到 1.3s。BGP-FESN (forwarding edge sequence number)[8] 使用转发边系列号代替 BGP-RCN 结点系列号,其 Tdown 和 Tlong 的收敛延迟和 RCN 类似。

上述改进收敛性的方法都大大地减少了 Tdown 事件的收敛延迟,但其对于 Tlong 事件只能作部分改进,并且这些方法并不改进 Tup 和 Tshort 事件的收敛性。因此本文提出了一种基于通告时间戳的方法 BGP-ATS(advertisement time stamp),可以较大改进 Tlong、Tup 和 Tshort 事件的平均收敛时间。该方法可以和其它方法结合使用,从整体上提高 BGP 的收敛性。

1　相关工作

1.1　BGP 收敛时间的上界

RFC1771[9] 指出,为了控制路由流量开销,BGP 使用最小路由通告间隔 MRAI(min route advertisement interval)参数限定同一个 BGP 发言者的针对同一个目的网络的两次路由通告所间隔的最小时间。

MRAI 定时器是影响 BGP 收敛性的一个重要因素。Davor Obradovic[10] 从理论上证明采用最短路径策略的 TSPP 实例可以在 $D\cdot w$ 时间内收敛,其中 D 为最长的可允许路由长度,w 为最大的边延迟。层次性的 TSPP 实例可以在 $2M\cdot n$ 时间内收敛,其中 M 为最大的边延迟,n 为路由器的数目。在前面阐述的改进 Tdown 和 Tlong 事件收敛时间的方法之中,BGP-Assertion 和 BGP-GF 并没有给出 Tlong 收敛时间的上界,BGP-RCN 给出假如采用每邻居每对等体速率限制定时器,那么 Tlong 收敛时间的上界为 $d\cdot(2u+M)$,其中 d 为网络直径,u 为最大的结点延迟,表示消息经过一跳 AS 的时间,包括处理延迟和传播延迟,M 为 MRAI 定时器的值;假如采用每对等体速率限制定时器,那么 Tlong 时间的上界可能会增加到 $d\cdot(2u+60)$。

因此 BGP 收敛时间的上界受 MRAI 定时器的限制,要减少 BGP 收敛时间的上界,需要减少 MRAI 定时器的值。但 Timothy Griffin[11] 指出如果没有 MRAI 定时器或者定时器值过小,那么就可能产生大量的路由通告报文,导致路由收敛时间变得非常长,因而不能单靠减少 MRAI 的值来提高收敛性。

1.2　BGP 平均收敛时间

对 BGP 平均收敛时间的分析涉及到 BGP 的速率限制定时器的具体实现。虽然 MRAI 针对每一个 BGP 对等体而设置,但速率限制过程是针对每一个目的网络的。对于当前的 Internet,假设路由前缀数目为 100K,那么一个有 k 个对等体的 BGP 发言者可能需要维护 $k\cdot 100K$ 个独立的定时器,因此为每一个对等体及每一个目的网络维护一个单独的定时器是不切实际的[12]。所以大多数 BGP 实现对每一个对等体只保持一个定时器,将其应用到所有的目的网络。在发送一个新的通告到一个对等体之前,检查定时器是否有其它通告在前 MRAI 秒内发送到那个对等体,如果没有,则发送,并启动定时器,否则要等定时器超时才能发送。对于 Internet 来说,几乎每一个给定的时刻系统都有变化发生,因而基于每个对等体的 MRAI 定时器一到期后往往被立即重置。Brian J. Premore[12] 指出,实际的 BGP 测量数据表明,在商业化的 BGP 实现中,不管有无通告发送,MRAI 定时器一到期就被简单重启。免费路由软件 Zebra 实现也是如此。对于连续重启的每对等体定时器实现而言,虽然路由器要向对等体发送一个通告报文时应该等待多长时间比较难以精确估计,但测量表明实际经历的平均等待时间约为 MRAI 的一半[13],假

设目的网络 X 起源于自治系统 A,那么一个结点 u 学习到到达 X 的活跃路由的时间 $t(P)$ 为

$$t(P) \approx \text{MRAI}/2 \cdot |P| \tag{1}$$

其中,P 是从 A 到 u 的最短路径。

假设 Internet 自治系统级网络拓扑的直径为 d,那么对于网络核心发生的 Tup 事件来说,其平均收敛时间 t 约等于 MRAI/2·d/2,而对于网络边缘发生的 Tup 事件来说,其平均收敛时间 t 约为 MRAI/2·d,因而对于任意的 Tup 事件来说,其平均收敛时间由式(2)所表征:

$$\text{MRAI}/2 \cdot d/2 \Leftarrow t \Leftarrow \text{MRAI}/2 \cdot d \tag{2}$$

Internet 自治系统级拓扑研究[14]表明,自治系统级网络拓扑的直径大约为 10。因此假设 d = 10,MRAI 按其缺省值 30s 来算,那么估计 Tup 的平均收敛时间在 75 ~ 150s 之间。

因此,只要实现时的定时器是基于每一个对等体的并且发送通告报文的过程保持不变,那么 Tup 的平均收敛时间必然受到式(2)的约束。而 Tlong 和 Tshort 事件影响的任何一个节点实际经历的平均等待时间也约为 MRAI 的一半,因此前面阐述的任何一种方法只能将 Tlong 和 Tshort 事件的收敛性改进到一定程度。

2 时间戳机制

2.1 时间戳机制的引入

前面的阐述说明,只要实现时的定时器是基于每一个对等体的并且发送通告报文的过程保持不变,那么前面提到的任何一种改进方法都不能将发送通告报文时的平均等待时间减少,因而其对 Tup、Tshort 及 Tlong 的平均收敛时间的改进受到固有的约束。因此,如果要改进这几种事件的收敛时间,那么就需要降低发送通告报文的平均等待时间。

目前的 BGP 实现之所以只对每一个对等体设置一个定时器,并将其应用到所有的目的网络,主要是因为对每一个对等体每一个目的网络设置一个定时器不切实际。但模拟实验说明采用每对等体每目的网络定时器可以很好地改进 BGP 的收敛性[12]。因此我们的想法就是能否采用某种可行机制减少定时器的数目但却拥有每对等体每目的网络定时器的优点。定时器规定系统在什么时间该做什么事情,和定时器相关的最重要的信息就是时间信息。因此我们采用替代方法,不使用定时器,只记录一些时间信息,根据这些时间信息来进行事件处理,使其达到和定时器类似的效果。

因此我们对 BGP 决策的路由通告过程做更改,BGP 路由器在向对等体发送路由通告报文后,就将相应的路由前缀打上时间戳,要发送通告报文时,通过查看相应前缀的时间戳来判断报文是否可以立即发送还是应该等待定时器到期再发送。如果报文不能被立即发送则将其相应信息放入通告摘要列表 advertisementsBrief。另外引入一个秒级定时器 oneSecondTimer,在其到期时检查通告摘要列表,只要报文等待时间大于等于 MRAI 就将其发送出去,检查完后重置定时器。引入 ATS 后 BGP 决策过程伪代码如图 1 所示。

2.2 时间戳机制的存储开销

BGP-ATS 要求对于每一个对等体每一个目的网络都存储相应的时间戳信息,并且等待发送的报文的摘要信息也要保存。因此假设一个路由器支持的对等体数目为 k,所支持的前缀数目为 n,那么该机制需要 $O(k \cdot n)$ 的存储开销。由于路由器控制平面使用的内存相对比较廉价,而且当前核心路由器实现时所支持的内存容量可以达到数百兆,甚至千兆。因此对于数百 K 的前缀数目及数十个对等体甚至数百个对等体,ATS 机制所耗费的内存空间都是可以接受的。该机制以部分内存代价换来较好的 BGP 收敛性。

```
BGP_Decision_Process(dst)                    /* Modified BGP pseudo code for ATS */
 {
  ...
  For(each peer pr)
   {
    If(currentSystemTime >= ATS(pr,dst) + the MRAI value for peer pr)
      SendAnnouncement(dst,pr);
    Else
      Insert < dst,pr,ATS(pr,dst) > into list advertisementsBrief;
   }
 }
 SendAnnouncement(dst,pr)
  {
   send message(announcement,ASpathdst,dst)to peer pr;
   ATS(pr,dst) = currentSystemTime;
  }
 OnExpirationOf oneSecondTimer()
  {
   CheckAdvertisementsBrief();
   Reset oneSecondTimer;
  }
 CheckAdvertisementsBrief()
  {
   For(each brief r in advertisementsBrief)
    {
     If(r.ATS <= currentSystemTime-mrai(r.pr))
      {
       SendAnnouncement(r.dst,r.pr);
       Delete r from advertisementsBrief;
      }
    }
  }
```

图 1　引入 ATS 后的 BGP 决策微码

Fig.1　Modified/new BGP pseudo code

3　BGP-ATS 收敛时间分析

我们采用和文献[13]类似的 BGP 收敛性模型,把 Internet 建模为一个有向图 G,其结点集合为 V,边集为链路(u,v)的集合 E,u 和 v 属于集合 V,结点代表自治系统集合,边集代表 EBGP 连接,采用最短路径策略。由于任何 Tshort、Tup 和 Tlong 事件都和某个目的网络相关,都起源于某个自治系统,因而用 X 表示该目的网络,用 s 表示该自治系统。边(u,v)存在当且仅当 u 会向 v 通告到目的网络 X 的路由。简单路径 P 的定义不变,但是对路径时间 $t(P)$的定义我们简单修改为路径终点学习到目的网络 X 的新路由并写入路由表的时间。

3.1　Tup 事件的收敛时间

在修改的 $t(P)$定义下,文献[13]的 VI.3 定理依然成立,即如果一个结点到起源 s 的最短路径为 P,那么该结点学习到目的网络为 X 的路由项的时间为 $t(P)$。根据 $t(P)$定义,边(u,v)时间 $t(u,v)$为从 u 往路由表中写入目的网络为 X 的新路由项的时间到 v 写入相应的新路由项的时间,把其称为边延迟时间。那么根据文献[13]的 VI.3 定理,可以得出如下推论。

推论 1　假设 h 为以图 G 和源自治系统结点 s 构造的组播最短路径树的高度,e 为平均边延迟时间,那么 Tup 事件的平均收敛延迟为 $h\cdot e$。

边延迟时间分为两大部分 边等待时间和边传播处理时间。(u,v)等待时间为 u 向 v 发送目的网络为 X 的通告报文时由于 MRAI 定时器的速率限制而导致的时间延迟。等待时间在 0 到 MRAI 之间。为了计算平均等待时间,我们引入一参数 p,表示目的网络为 X 的两个连续的路由通告报文发送时间间隔为 MRAI 的概率,也即当要发送一个通告报文时需要等待的概率。因此一个通告报文的平均等待时间为$(1-p)\cdot 0 + p\cdot \mathrm{MRAI}/2$ 即 $p\cdot \mathrm{MRAI}/2$。(u,v)传播处理延迟包括 u 发送通告报文的时间、该报文的传播时间、结点 v 接收该报文的时间及结点 v 处理该报文进行路由决策写入相应的路由项的时间。假设平均边传播处理时间为 k,那么 Tup 平均收敛时间 C_u 可以由下式描述。

$$C_u = h\cdot(p\cdot \mathrm{MRAI}/2 + k) \tag{3}$$

假设图 G 的直径为 d，由 $d/2 \Leftarrow h \Leftarrow d$ 可以得出平均收敛时间范围：

$$d/2 \cdot (p \cdot \mathrm{MRAI}/2 + k) \Leftarrow C_u \Leftarrow d \cdot (p \cdot \mathrm{MRAI}/2 + k) \tag{4}$$

3.2 Tlong 和 Tshort 事件的收敛时间

对于传统的 BGP、BGP-AS 及 BGP-GF，发生 Tlong 事件时，它们并不能阻止所有失效路径的传播。Tlong 事件发生时所切换的路径长度、网络中所有结点的备份路径的长短以及数目使得分析传统 BGP 的收敛事件的平均收敛时间非常困难。但 BGP-RCN 和 BGP-FESN 使得路由器不会选择失效路由，不会通告失效路径。因此我们只对路由器不会通告失效路径这类情况进行分析。

Tup 和 Tdown 事件影响图 G 中的所有结点，但是对于 Tlong 和 Tshort 事件而言，可能只有一部分结点收到更长/更短路径的通告报文并进行路由决策，而其它结点不受影响。因而我们考虑子图 S，所有发出或者收到通告报文的结点组成结点集 Vs，边集 Es 为 E 的子集，每条边的结点都在 Vs 中。根据文献[7]对于 Tlong 事件的收敛上界的证明可知，在路由器不会通告失效路由的情况下，对于子图 S，Tlong 和 Tshort 事件类似于 Tup 事件，推论 1 和式(4)的结论都成立。而子图的直径 $ds \Leftarrow d$，因此 Tlong 和 Tshort 事件的平均收敛时间 C_1 可以由式(5)表述：

$$C_1 \Leftarrow C_u \tag{5}$$

3.3 等待概率分析

ATS 对于收敛性的改进效果依赖于等待概率 p 的大小。路由器在网络中所处的位置以及网络变化的频繁程度对于等待概率 p 的大小都有影响。我们对一些顶级 ISP 和欧洲的几个主要 ISP 的一些路由器在 2004 年 2 月所产生的更新报文进行分析，得出表 1 结果。从表中可以看出，p 值往往比较小。因此如果实现 ATS 的话，收敛时间可以得到较大提高。

表 1 几个大的 AS 的等待概率

Tab.1 The waiting probability of some large ASes

AS number	AS513	AS1103	AS3333	AS4608	AS4777	AS7018	AS9177	AS13129
Waiting probability	0.23228	0.37677	0.35132	0.26568	0.32042	0.30579	0.29281	0.18241

4 结论

本文引入了时间戳机制 ATS。ATS 可以将通告报文的平均等待时间从传统实现方法下的(MRAI/2 + k)减少为($p \cdot$MRAI/2 + k)，相应地也减少了 Tlong 和 Tshort 事件的收敛时间。针对 Internet 自治系统的 BGP 更新数据分析表明，概率 p 较小，因此 ATS 机制能很好地提高 BGP 的收敛性。

参考文献：

[1] Hudson G. The Real-time IP Network [EB/OL]. http://www.avici.com/technology/whitepapers/The_Realtime_Network.pdf,2003.

[2] Kaat M. Overview of 1999 IAB Network Layer Workshop[S]. RFC2956,Nov.1999.

[3] Labovitz C, Ahuja A, Bose A,et al. Delayed Internet Routing Convergence [A]. Proceedings of ACM Sigcomm[C], August 2000.

[4] Pei D, Zhao X, Wang L, et al. Improving BGP Convergence through Ssertions Approach[A]. Proceedings of the IEEE INFOCOM[C], June 2002.

[5] Luo J, Xie J, Hao R, et al. An Approach to Accelerate Convergence for Path Vector Protocol [A]. Proceedings of IEEE Globecom[C], Nov. 2002.

[6] Bremler A, Afek Y, Schwarz S. Improved BGP Convergence via Ghost Flushing [A]. Proceedings of the IEEE INFOCOM[C], April 2003.

[7] Pei D, Azuma M, Nguyen N, et al. BGP-RCN: Improving BGP Convergence through Root Cause Notification [J]. Computer Networks, 2005, 48 (2): 175 - 194.

[8] Chandrashekar J, Duan Z, Zhang L, et al. Limiting Path Exploration in Path Vector Protocols [A]. Proceedings of the IEEE INFOCOM[C], March 2005.

[9] Rekhter Y, Li T. Border Gateway Protocol 4, RFC 1771 [S]. SRI Network Information Center, July 1995.

[10] Obradovic D. Real-time Model and Convergence Time of BGP [A]. Proceedings of the IEEE INFOCOM[C], June 2002.

[11] Griffin T, Premore B. An Experimental Analysis of BGP Convergence Time [A]. Proceedings of ICNP[C], November 2001.

[12] Premore B. An Analysis of Convergence Properties of the Border Gateway Protocol Using Discrete Event Simulation [D]. Dartmouth College Department of Computer Science, May 2003.

[13] Labovitz C, Wattenhofer R, Venkatachary S, et al. The Impact of Internet Policy and Topology on Delayed Routing Convergence [A]. Proceedings of the IEEE INFOCOM[C], April 2001.

[14] Magoni D, Pansiot J J. Analysis of the Autonomous System Network Topology [J]. ACM SIGCOMM Computer Communication Review, 2001,31 (3): 26 - 37.