

BGP 收敛性质改善的 MRAI 时钟设置方案

王文化, 沈庆国, 王 滨

(解放军理工大学通信工程学院, 南京 210007)

摘 要: 针对当前边界网关协议(BGP)路由存在慢收敛会引起网络数据转发层服务质量下降问题, 基于一个简化的 BGP 路由模型和核心网络拓扑结构, 提出一个新的 MRAI 时钟设置方案。该方案需要根据已知网络条件先计算后设置。通过使用 ssfnet 仿真软件测试表明, 与 RFC1771 中时钟抖动方案相比, 该方案能够减少 BGP 平均网络收敛延时和更新消息交互数量。

关键词: 边界网关协议; 收敛延时; MRAI 时钟; 网络拓扑

MRAI Timer Setup Scheme for BGP Convergence Properties Improving

WANG Wen-hua, SHEN Qing-guo, WANG Bin

(Institute of Communications Engineering, PLA University of Science and Technology, Nanjing 210007)

【Abstract】 Aiming at the problem that the quality of service in data forwarding plane may suffer heavily caused by slow convergence in Border Gateway Protocol(BGP) routing, based on a simplified BGP routing model and core network topology, this paper puts forward a new MRAI timer setup scheme. The scheme first calculates, then setup according to known network condition. Though simulation software ssfnet test, results show that compared with MRAI timer jitter setup in RFC1771, BGP convergence properties can be improved greatly by the proposed scheme.

【Key words】 Border Gateway Protocol(BGP); convergence delay; MRAI timer; network topology

1 概述

边界网关协议(Border Gateway Protocol, BGP)是当前主要使用的域间路由协议, 完成不同自治系统间相互发现和建立路由表的功能。在路由节点或链路发生故障后, 大量的路由更新消息在网络中传播, 通告路由发生的变化。BGP 属于路径矢量协议, 在路由更新消息中携带了到目的前缀的完整路径信息。由于 BGP 缺乏检测路径间关联性的机制, 导致路由收敛过程持续很长时间。实验表明, 在故障事件发生后, BGP 路由收敛过程延时大约为 15 min^[1]。在路由慢收敛过程中, 数据转发层会遭受路由黑洞和路由环路问题的影响, 分组端到端转发服务质量变差, 从而影响网络顶层业务的服务质量。加快 BGP 路由收敛过程是 BGP 协议研究的重要内容^[2-4]。描述路由收敛过程主要参数有收敛延时和交互路由消息数量。本文基于一个简化的 BGP 路由模型和核心网络拓扑结构, 提出一个新的 MRAI 时钟设置方案以改善 BGP 路由收敛性质。

2 路由模型

为了便于 BGP 协议分析, 本文对协议作如下简化:

- (1) 每个自治系统只有一个边界路由器运行 BGP 协议。
- (2) IBGP、复杂路由策略和 AS 间的多 BGP 会话(multi-BGP session)不做考虑, 路由策略简化为最短路径优先, 且路径长度相同时下一跳路由节点 id 号大的优选。
- (3) 更新消息的转发延时对于不同路由器是一样的。BGP 路由模型可分解为 2 个子模型: 消息传播模型和节点消息处理模型。

传播模型描述了路由更新消息在相邻节点间传递的规则。路由器节点关于特定目的前缀的最优路径从空变为非空

或者从非空值变为不同的非空值时, 发送路由通告消息到相邻节点。路由器节点特定目的前缀最优路径从非空变为空时, 发送路由撤销消息到相邻节点。路由器节点特定目的前缀最优路径取值未发生变化, 则不发送任何路由更新消息。节点每次等待 MRAI 时钟超时后, 才将路由更新消息向相邻节点传递。MRAI 时钟有 2 种实现方法 per-peer 和 per-destination。

节点消息处理模型描述了路由更新消息在路由器节点的逻辑处理过程, 该过程可以分为 3 步: 从相邻对端接收路由更新消息, BGP 路径选择过程, 向相邻对端发送路由更新消息。一个 BGP 路由器从相邻 BGP 对端路由器收到路由更新消息后, 根据入路由策略(in-policy)进行过滤, 通过过滤的路由更新消息将记入表 rib-in。rib-in 表项表示特定对端到特定目的前缀的 AS 路径。路由更新消息使 rib-in 更新, 启动路径选择过程。根据路由策略, 路由器计算到特定目的前缀 AS 路径的 preference 值, 值最大的作为最优路径, 存入路由器的路由表 local-rib 中。如没有任何可行路径, 则路由器认为到目的前缀不可达。按照出路由策略(out-policy)对 local-rib 表中的路径信息过滤, 将符合策略要求的更新路径放入表 rib-out 中, 然后封入路由更新消息发送到 BGP 对端路由器。

BGP 路由器节点收敛延时为 $t_c - t_s$, 其中, t_s 表示源头路由器发送路由更新消息时刻; t_c 表示在该时刻后该路由器节点始终有到目的前缀的可达路径或者发现前缀不可达。路由器节点收敛延时越大, 数据转发层因该点引起的服务质量下

基金项目: 国家自然科学基金资助项目(60472050)

作者简介: 王文化(1981-), 男, 博士研究生, 主研方向: 网络路由协议, 网络服务质量; 沈庆国, 教授; 王 滨, 博士研究生

收稿日期: 2010-01-10 **E-mail:** wenhuawang1981@gmail.com

降持续时间越长。BGP 网络收敛延时为所有路由器节点收敛延时中最大值。BGP 平均网络收敛延时为各个路由器节点收敛延时平均值。与网络收敛延时相比，平均网络收敛延时更能清晰地刻画 BGP 路由收敛过程对数据转发层服务质量的影响程度。链路故障引起目的前缀不可达而发生的路由收敛过程称为 p_{down} 。与其他事件引起的收敛过程相比， p_{down} 的收敛性质最差，本文针对全互联拓扑(称为 clique)下 p_{down} 过程，通过合理设置 MRAI 时钟取值，改善路由收敛性质。

3 MRAI 时钟设计方案

在 RFC1771 中 MRAI 时钟抖动设置要求如下：时钟默认值乘以 $[0.75, 1.00]$ 范围内服从均匀分布的随机数为路由节点时钟取值。RFC 中对抖动的取值范围和分布特征未做进一步说明。对 N 个节点的 clique 拓扑，节点 1 因故障向相邻节点发送特定目的前缀的撤销消息，引起路由收敛过程。在 ssfnet 软件中改变时钟抖动取值范围和分布特征，统计收敛延时和交互消息数量，发现收敛性质有明显改善的仿真测试中，不同节点的时钟触发时刻有着共同的特征，如图 1 所示。

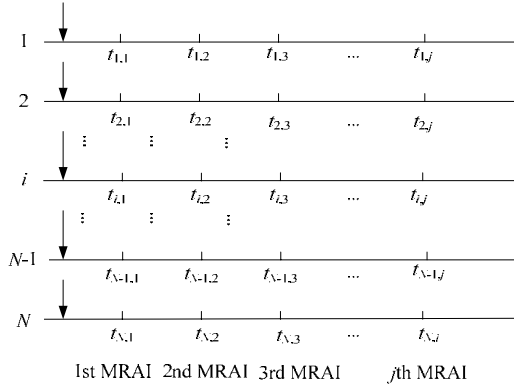


图 1 MRAI 时钟触发时刻时序关系

图 1 表示 N 个节点的全互联拓扑收敛性质显著改善时的时钟触发时刻时序关系。在每轮 MRAI 内，随着节点序号递增，MRAI 时钟触发时刻递增。能够实现该时序关系的时钟设置方法称为递增量设置法。该方法需要已知网络拓扑节点数目，节点 CPU 处理延时分布特征，收敛过程所含 MRAI 时钟轮数和 MRAI 时钟基值。假设节点 1 的 MRAI 时钟值 M_0 为默认基值，从节点 2 到节点 N 时钟值依次递增 ΔM ，路由收敛过程经历 k 轮时钟，节点 i 的 CPU 处理延时为随机变量 $d_{i,CPU}$ ，链路传播延时为常数 d_{link} ，起始状态下每个节点 MRAI 时钟设置时刻为 t_0 。节点 i 的 MRAI 时钟 M_i 为

$$M_i = M_0 + (i-1) \cdot \Delta M \quad (1)$$

节点 i 第 j 轮时钟触发时刻为

$$t_{i,j} = t_0 + j \cdot M_0 + (i-1) \cdot j \cdot \Delta M \quad (2)$$

时钟递增量 ΔM 必须满足如下 2 个条件才能保证图 1 时序关系：(1)任意节点 i 在发送更新消息后，其任意相邻对端在本地 MRAI 触发前必须有足够的时间接收和处理该消息。(2)随着时钟轮数递增，任意 2 个节点的 MRAI 触发时刻间隔时间递增。为了简化时钟设置和收敛过程分析，触发时刻间隔时间应当小于节点 MRAI 时钟值。

条件(1)对应的不等式组为

$$\begin{cases} d_{link} + E(d_{n,CPU}) < t_{n,j+1} - t_{i,j} & n=1,2,\dots,i-1 \\ d_{link} + E(d_{n,CPU}) < (n-i)j\Delta M & n=i+1,i+2,\dots,N \end{cases} \quad (3)$$

条件(2)对应的不等式组为

$$t_{i,j} - t_{n,j} < M_n, n = 1, 2, \dots, i-1 \quad (4)$$

若 $d_{n,CPU}$ 为独立同分布的随机变量，从式(3)和式(4)可以推出 ΔM 的上界和下界。

$$\frac{d_{link} + E(d_{n,CPU})}{j} \leq \Delta M \leq \frac{M_0}{(N-1)j} \quad j = 1, 2, \dots, k \quad (5)$$

由式(5)可以看出， ΔM 的下确界为常数 $d_{link} + E(d_{n,CPU})$ ，上确界为 $\frac{M_0}{(N-1)k}$ 。 ΔM 上确界和网络大小 N 呈反比。当 N 足

够大时， ΔM 上确界可能小于下确界。对时钟递增量设置法做如下扩展：将路由收敛过程中的 MRAI 轮数分为若干周期，约束 ΔM 的 2 个条件仅在每个周期内满足；在每个周期末尾各个节点的时钟触发时刻调整为相同取值； ΔM 上界中参数 j 为周期长度，初始值设置为 k 的一半，若 ΔM 的上界仍小于下界，则将周期长度 j 减半直到矛盾消除为止。以 clique30 为例， $N=30$ ， $M_0=6$ s， $k=N-1=29$ ， $d_{link}=0.01$ s， $E[d_{n,CPU}]=0.01$ s，通过 2 次迭代，得到周期长度为 8 轮时钟， ΔM 的取值范围是 $[0.02, 0.026]$ s，节点 i 的 MRAI 按照式(6)设置：

$$M_i = \begin{cases} M_0 + (i-1) \cdot \Delta M, & j \% 8 \neq 0 \\ M_0 + (239 - 7j) \cdot \Delta M, & j \% 8 = 0 \end{cases} \quad j = 1, 2, \dots, k \quad (6)$$

收敛性质显著改善原因可以通过收敛过程中不同节点路由表变化加以解释。以 clique5 拓扑为例，表 1 表示递增量法收敛过程对应路由表的变化，表 2 表示时钟随机抖动最差情况下收敛过程对应路由表变化。表项记录了某节点在某轮时钟触发时刻路由表内容。表中*标记的路径为最佳路径。

表 1 clique5 时钟递增量方案下收敛过程路由表

MRAI 轮数	2	3	4	5
1st MRAI	3,1,0 4,1,0 5,1,0*	2,5,1,0 4,1,0 5,1,0*	2,5,1,0 3,5,1,0 5,1,0*	Null Null Null
2nd MRAI	3,5,1,0 4,5,1,0*	2,4,5,1,0 4,5,1,0*	Null Null Null	Null Null Null
3rd MRAI	3,4,5,1,0* Null Null	Null Null Null	Null Null Null	Null Null Null
4th MRAI	Null Null Null	Null Null Null	Null Null Null	Null Null Null

表 2 clique5 时钟随机抖动最差情况下收敛过程路由表

MRAI 轮数	2	3	4	5
1st MRAI	Null 4,3,1,0 5,4,1,0*	2,1,0* Null 5,4,1,0	2,1,0 3,1,0* Null	2,1,0 3,1,0 4,1,0*
2nd MRAI	Null Null 5,4,3,1,0*	2,5,4,1,0* Null Null	Null 3,2,1,0* Null	Null 3,2,1,0 4,3,1,0*
3rd MRAI	Null Null Null	Null Null Null	Null Null Null	Null Null 4,3,2,1,0*
4th MRAI	Null Null Null	Null Null Null	Null Null Null	Null Null Null

在表 1 中，节点 5 在首轮时钟触发时最先发现目的前缀不可达，向相邻节点发送撤销消息，随后节点 4 在第 2 轮时钟触发时发现目的前缀不可达，发送撤销消息，直到节点 2 在最后一轮时钟触发时发现目的前缀不可达，此时所有节点均发现目的前缀不可达，路由收敛。在表 2 中，节点 2 到节点 4 在第 3 轮时钟触发时才发现前缀不可达，向相邻节点发送撤销消息，节点 5 在第 4 轮时钟触发时发现目的前缀不可达，此时所有节点发现目的前缀不可达，路由收敛。表 1 所描述收敛过程的平均网络收敛延时和消息交互数量明显小于表 2 收敛过程。对 N 个节点全互联拓扑，利用上述路由表分析法和归纳法可得：使用递增量方案下，网络收敛延时为

kM_2 , 平均网络收敛延时为 $\frac{N-1}{2}$, 交互消息数量为 $\frac{N(N-1)^2}{2}$; 在随机抖动方案最差情况下, 网络收敛延时为 kM_N , 平均网络收敛延时为 $\frac{(N-1)^2}{N}$, 最大交互消息数量为 $(N-1)^3$ 。可以看出, 两者的网络收敛延时相近; 当 N 较大时, 前者的平均网络收敛延时和交互消息数量趋于后者的一半。

4 仿真条件设置和仿真结果说明

对 clique 拓扑, 通过 ssfnct 仿真, 可以对比 RFC 中均匀分布随机抖动时钟设置方案和本文所提递增量设置方案对平均网络收敛延时和交互消息数量的影响。仿真参数设置如表 3 所示。在不同方案下, 平均网络收敛延时和网络节点数关系如图 2 所示, 更新消息数量和网络节点数关系如图 3 所示。从图 2 中可以看出, 时钟递增量方案与均匀分布随机抖动设置方案下收敛性质最差情况相比, 平均网络收敛延时和交互更新消息数量均减少约 50%; 与时钟随机抖动设置方案下某一具体时钟取值情况相比, 平均网络收敛延时和交互更新消息数量均减少约 [20%, 30%]。

表 3 仿真参数设置

参数	取值方式
WRATE	true
SSLD	false
SH	false
MRAI 时钟设置	Per-peer, 连续不间断设置
路由节点 CPU 处理延时	服从区间[0.001, 0.020] s 均匀分布
默认 MRAI 时钟基数值/s	6

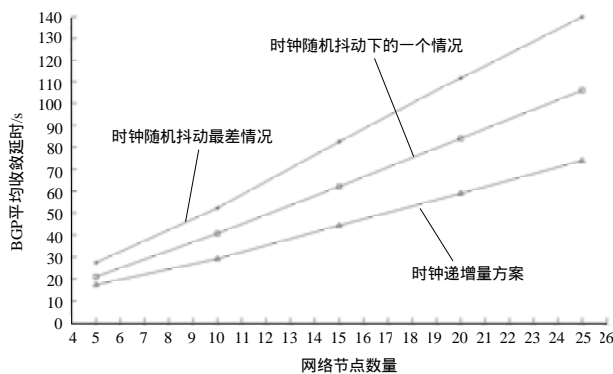


图 2 不同时钟设置方案下平均网络收敛延时比较

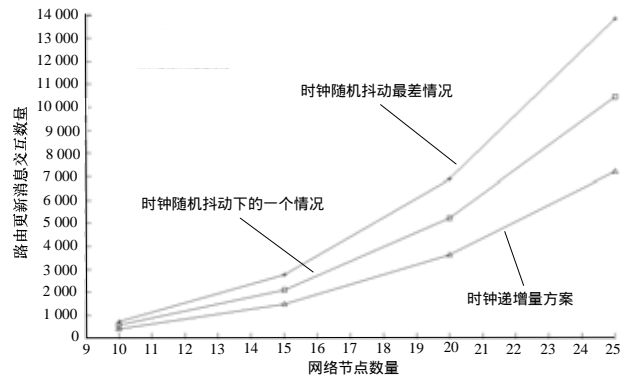


图 3 不同时钟设置方案下交互消息数量比较

5 结束语

本文针对核心网的全互联拓扑模型 clique, 提出一种新的时钟设置方案。该方案不同于 RFC 规范中的随机抖动设置, 需要根据已知网络条件先计算后设置。模型分析和仿真验证表明, 相对于已有方案, 该方案在保证边界路由器不会突发收到大量路由更新消息前提下, 能够明显减少平均网络收敛延时和更新消息交互数量。在任意拓扑下, 改善 BGP 收敛性质的 MRAI 时钟设置方案设计将是下一步工作。

参考文献

- [1] Labovitz C, Malan G R, Jahanian F. Origins of Internet Routing Instability[C]//Proc. of the 18th Annual Joint Conference of IEEE Computer and Communications Societies. New York, USA: [s. n.], 1999: 218-226.
- [2] Pei Dan, Zhao Xiaoliang, Wang Lan, et al. Improving BGP Convergence Through Assertions Approach[C]//Proc. of the 21st Annual Joint Conference of IEEE Computer and Communications Societies. New York, USA: [s. n.], 2002: 902- 911.
- [3] Bremner Barr A, Afek Y, Schwarz S. Improved BGP Convergence via Ghost Flushing[C]//Proc. of the 22nd Annual Joint Conference of IEEE Computer and Communications Societies. San Francisco, California, USA: [s. n.], 2003: 927- 937.
- [4] Sun Wei, Mao Zhuoqiang, Shin K G. Differentiated BGP Update Processing for Improved Routing Convergence[C]//Proc. of IEEE International Conference on Network Protocols. Santa Barbara, CA, USA: [s. n.], 2006: 280-289.

编辑 索书志

(上接第 18 页)

参考文献

- [1] 孙利民, 李建中, 陈 渝, 等. 无线传感器网络[M]. 北京: 清华大学出版社, 2005.
- [2] 崔 莉, 鞠海玲, 苗 勇, 等. 无线传感器网络研究进展[J]. 计算机研究与发展, 2005, 42(1): 163-174.
- [3] 韩月敏, 刘非平, 王永峰. 模拟对抗演习探测识别信息量化及其效能评估[J]. 系统仿真学报, 2009, 21(3): 757-760.
- [4] 庄 伟, 宋光明, 宋爱国. 用于未知环境的混杂传感器网络交互策略[J]. 通信学报, 2008, 29(11): 121-127.
- [5] 袁凌云, 朱云龙. 基于 NS2 的无线传感器交通监控网络仿真[J].

系统仿真学报, 2007, 19(3): 660-664.

- [6] 孟旭东, 王建安, 陆 凯. 家庭网络中的 BAN 和远程健康监护[J]. 中兴通讯技术, 2006, 12(4): 26-30.
- [7] 杨 军, 黄建民. 智能家居生活支援系统[J]. 机器人技术与应用, 2007, (6): 10-13.
- [8] 刘 航, 廖桂平, 杨 帆. 无线传感器网络在农业生产中的应用[J]. 农业网络信息, 2008, (11): 16-21.
- [9] 张媛媛, 杨祝红, 文高飞, 等. 我国饮用水水质标准研究进展及新增项目检测[J]. 中国公共卫生, 2007, (3): 275-276.

编辑 索书志