ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.11, November 2008, pp.2979−2989
DOI: 10.3724/SP.J.1001.2008.02979

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

改善BGP路由收敛的时间窗口机制*

王立军+, 吴建平

(清华大学 计算机科学与技术系,北京 100084)

Time Window Mechanism to Improve BGP Routing Convergence

WANG Li-Jun+, WU Jian-Ping

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: E-mail: mail2wlj@gmail.com

Wang LJ, Wu JP. Time window mechanism to improve BGP routing convergence. *Journal of Software*, 2008, 19(11):2979−2989. http://www.jos.org.cn/1000-9825/19/2979.htm

Abstract: In this paper, a time window mechanism based on the penalty value of route flap damping is designed to improve routing convergence. This mechanism judges the route stability in the network from BGP (border gateway protocol) routes correlation by observing multiple routes received from different peers jointly. Then BGP speaker can find instable routes earlier and makes stable routes get the chance to be selected earlier in route selection, thus curtail path exploration. Simulation results prove that with proper parameters this method can reduce convergence delay and communication overhead obviously. Furthermore, without addition information in BGP Update message, time window mechanism is a practical method to be deployed in the Internet.

Key words: border gateway protocol; time window mechanism; inter-domain routing; route flap damping

摘 要: 提出了一种时间窗口机制,能够基于路由抖动抑制中路由惩罚值的变化改善 BGP(border gateway protocol)路由收敛.这种新机制把来自不同邻居的路由变化情况结合起来,利用 BGP 路由传播过程形成的路由相关性判断路由在网络中的稳定情况.时间窗口机制使BGP路由器能够更早地发现不稳定路由,优先将稳定路由选择为最优路由,终止路径搜索过程.模拟实验的结果表明,通过选择适当的参数,时间窗口机制能够大大缩短 BGP 路由收敛延时,减小收敛过程中的通信开销.而且,这种方法不需要在 BGP 的路由消息中增加额外的信息,因此容易在实际网络中逐步部署.

关键词: 边界网关协议;时间窗口机制;域间路由;路由抖动抑制

中图法分类号: TP393 **文献标识码**: A

边界网关协议 BGP4[1]是互联网域间路由协议的事实标准,BGP(border gateway protocol)路由的稳定性对互联网上应用的端到端性能有很大的影响.Labovitz 等人[2]发现,实际网络中 BGP 路由的稳定性比预想的要差.互联网上的实验[3]表明,路由失效(fail-down)和路由切换(fail-over)事件导致的BGP路由收敛延时甚至会长达十几分钟.路由收敛过程不仅会增加路由器的通信和处理开销,而且会加大分组在网络中的传输延时和丢包率[4].

* Supported by the National Natural Science Foundation of China under Grant No.60473082 (国家自然科学基金); the National Basic Research Program of China under Grant No.2003CB314801 (国家重点基础研究发展计划(973))

Received 2006-11-08; Accepted 2007-05-31

BGP 协议规范和实现中引入了多种改善路由稳定性的机制.路由抖动抑制 RFD[5]使用惩罚值(penalty)记录一条路由过去变化的剧烈程度,并基于此判断路由未来可能的变化情况.当惩罚值超过一定阈值时,路由将不能参加 BGP 路由选择.在 Cisco 路由器的 BGP 实现中[6],优先选择更加稳定的路由以提高路由稳定性.如果两条路由的其他情况相同,则更早收到的路由会被优先选择.

作为一种路径向量协议,BGP 使用“存储-添加”模式传播路由.边界路由器存储邻居发送来的路由,根据路由策略从中选择一条最优路由,在其中的 AS(autonomous system)路径前端加入本地的自治系统号码后,将路由进一步发送给邻居自治系统的边界路由器.这种传播模式使互联网中到达同一目的网络的路由具有相关性.网络中的一个事件,比如链路失效,会同时引起多条相关路由的变化.

路由相关性和 BGP 路径搜索有着紧密的关系.链路或者设备故障会使网络中一些路由成为无效路由.这些无效路由是同一事件引起的相关路由,不会立即从网络中消失.路由器不断地在无效路由中搜索最优路由并传播,直到稳定的可靠路由被选择为止.路径搜索过程大大增加了 BGP 路由收敛延时.更为严重的是,相对稳定的路由,比如发生故障马上恢复的路由,可能会在路径搜索和 RFD 的联合作用下被错误地抑制较长时间[7].

本文提出了一种时间窗口机制,利用路由相关性使稳定路由得到更多的机会被选为最优路由,缩短路径搜索过程以改善 BGP 路由收敛.时间窗口机制的基本思想是,根据多条路由的变化特点判断它们之间的相关性,具体方法是利用相关路由 RFD 惩罚值的变化在时间上大致同步的特点来判断相关路由.路由的第 1 次变化打开时间窗口,时间窗口会持续固定的一段时间.窗口打开期间,到达同一目的网络路由的变化,使路由的 RFD 惩罚值增加并超过预定的阈值,就会被记录在时间窗口中.时间窗口关闭时,如果其中的路由数目超过一定数量,则 BGP 重新执行路由选择,并忽略时间窗口中的路由,尽管这些路由没有被 RFD 抑制.于是,其他与窗口中路由相关性小的稳定路由得到被选为最优路由的机会,减小了路由进一步变化的可能性.

本文第 1 节介绍 BGP 路由传播和收敛的特点,包括 BGP 路由传播过程的“存储-添加”模式、由此形成的路由相关性和 BGP 路由收敛中的路径搜索过程.第 2 节综述相关的研究工作.第 3 节使用同步网络模型介绍时间窗口机制的原理,并给出时间窗口关闭算法.第 4 节通过模拟实验的结果评价时间窗口机制的有效性.最后总结全文.

1 域间路由的传播和收敛

1.1 “存储-添加”模式

目前的互联网由超过 2 万个处于各种机构控制下的自治系统(AS)互连构成.每个自治系统由唯一的 16 位自治系统号码(autonomous system number,简称 ASN)标识,内部使用域内路由协议,如 RIP 和 OSPF,自治系统之间通过 BGP 传递网络可达性信息.BGP 传递路由信息的路由更新消息(update)包括两种类型:路由声明(announcement)和路由取消(withdrawal).BGP 发现新路由后,给邻居发送路由声明消息,其中包括网络层可达信息(network layer reachability information,简称 NLRI)和描述路由特征的路由属性.其中,AS_PATH 属性包含路由所经过自治系统的 ASN 序列,根据这个序列可以避免产生路由环路.BGP 发送路由取消通知邻居以前发送的路由不再可用,分为显式取消(explicit withdrawal)和隐式取消(implicit withdrawal)两种,前者直接取消前面发送的路由,后者通过发送一条新的可用路由代替以前发送的路由.BGP 发送到达某个目的网络路由的最小间隔为 MRAI(minimum route advertisement interval),间隔中间如果路由发生变化,只有最后一个路由被发送给邻居,但是为了避免把分组转发给错误的下一跳,路由取消消息不受 MRAI 的限制.

BGP 是一种增量协议.当 BGP 会话建立时,整个路由表被传递给对方,之后,只有路由发生变化时,边界路由器才会发送一个路由更新消息.这种机制有利于提高 BGP 扩展性,但是要求路由器记录邻居发送来的全部路由.BGP 协议给每个邻居设定一个 *Loc-RIB-In*,存储该邻居发送来的路由,图 1 描述了有两个邻居的边界路由器的 BGP 路由表结构.收到邻居发送来的路由后,输入策略首先会过滤掉一些路由,改变路由的某些属性.之后,路由判决过程从多个 *Loc-RIB-In* 中选择出到达目的网络的最优路由,存储在 *Loc-RIB* 中.最优路由被发送给邻居前需要经过输出策略,本自治系统的 ASN 被加入路由的 AS_PATH 属性中,并把路由记录在与该邻居对应的

Loc-RIB-Out 中.由此,BGP 处理路由的过程可以抽象为“存储-添加”模式,“存储”是指存储从邻居收到的路由,“添加”是指路由被发送前在 ASN 序列中加入本自治系统的 ASN.

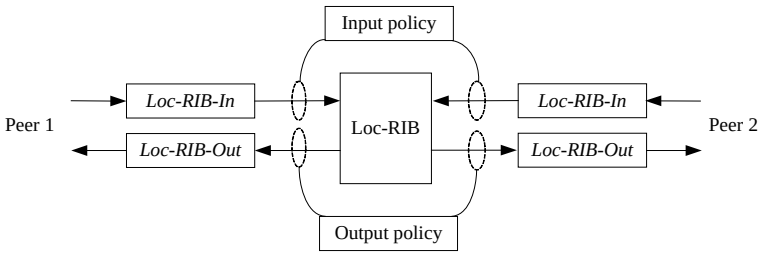


Fig.1 BGP route information base of BGP speaker with two neighbors

图 1 有两个邻居的边界路由器的 BGP 路由信息存储结构

1.2 路由相关性

“存储-添加”的路由传播模式会使互联网中到达同一目的网络的 BGP 路由间形成一定的相关性.同一自治系统中的边界路由器具有一致的路由策略,因此,本文的讨论中把自治系统抽象成一个点,节点间的边表示自治系统间的 BGP 邻居关系.这样,互联网能够表示为无向图 $G=(V,E)$,其中,V 和 E 分别是节点和边的集合.首先给出路由相关性的定义.

定义 1(路径依赖性). $P_1=\langle e_1^1,e_2^1,...,e_m^1\rangle$ 和 $P_2=\langle e_1^2,e_2^2,...,e_n^2\rangle$ 是无向图 $G=(V,E)$上的两条路径,如果 $m>n$ 并且 $\langle e_1^2,e_2^2,...,e_n^2\rangle$ 是 $\langle e_1^1,e_2^1,...,e_m^1\rangle$ 的一部分,那么 P_1 依赖于 P_2.

定义 2(路由依赖性). r_d^1 和 r_d^2 是到网络 d 的路由,如果 r_d^1 的 AS_PATH 依赖于 r_d^2 的 AS_PATH,那么, r_d^1 依赖于 r_d^2.

定义 3(路由相关性). r_d^1 和 r_d^2 是边界路由器从不同邻居收到的到达目的网络 d 的两条路由,如果 r_d^1 和 r_d^2 的 AS_PATH 都依赖于某条路径 P,那么, r_d^1 和 r_d^2 是相关路由.

尽管路由相关性由路由消息经过的ASN序列决定,但是,根据路由中静态的ASN序列不能准确推断路由变化间的相关性.在图 2 所示的简单拓扑中,节点到达 d 的路由记为 r_d.节点 H 从邻居共收到 4 条到 d 的路由:来自 D 的[*H-D-C-B-A*]、来自 C 的[*H-C-B-A*]、来自 B 的[*H-B-A*]以及来自 G 的[*H-G-F-E-A*].如果从节点 C 和 D 收到取消[*H-D-C-B-A*]和[*H-C-B-A*]的路由消息,节点 H 就不能区分路由变化是源自 e_1 还是 e_2,也就不能判断网络中 r_d 的变化是否影响到[*H-B-A*].

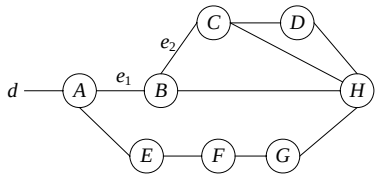


Fig.2 A simple AS connection topology

图 2 简单的自治系统连接拓扑

定理 1. 一条路径上的链路状态发生变化时,依赖该路径的相关路由的变化在时间上具有相关性.

证明:链路因为故障发生变化后,BGP 作为动态路由协议,会及时地将路由变化通知给每个 AS.AS 从邻居收到的依赖该路径的相关路由,在相近的时间发生变化.链路被修复后,BGP 也会将路由变化通知给每个 AS.AS 收到的相关路由的变化在时间上也是接近的.因此,依赖该路径的相关路由的变化在时间上具有相关性. □

由定理 1 可知,相关路由的特点是,当依赖路径发生故障时,相关路由的变化在时间上具有相关性.因此,当网络中路由发生变化时,路由的变化模式是否一致可以作为判断相关路由的依据.

相关路由的变化对 BGP 路由稳定性有很大的负面作用.例如在图 2 中,节点 H 收到的 r_d 中,来自节点 B,C 和 D 的路由是依赖于路径$\langle e_1\rangle$的相关路由.假设节点优先选择路径最短的路由,并且路由消息在每条边传输的延时相同.在 e_1 失效后,节点 H 首先收到来自节点 B 的路由取消,经过路由选择,来自节点 C 的路由被选为新的最优路由.之后,节点 H 先后收到来自节点 C 和 D 的路由取消,最终选择来自节点 G 的路由.如果边 e_1 不断反复"故障-修复"的变化,那么尽管只有来自节点 G 的路由是稳定的,节点 H 仍然会循环选择来自 B,C,D,G 的路由.

1.3 路径搜索过程

"存储-添加"模式对 BGP 路由收敛影响的具体体现是路径搜索过程.链路发生故障后,存储在网络中依赖故障链路的路由并不会立刻被取消.节点取消当前的路由后,选择的下一条最优路由可能由于链路故障也不再可用,继而又被发送给邻居,在网络中传播.节点在 *Loc-RIB-In* 的路由中不断搜索最优路由,直到选择一条不依赖于失效链路的稳定路由.

定义 4(无效路由). 无效路由是指沿着路由的 AS 路径已经不能将分组转发到达目的网络,但仍然作为最优路由或者有机会被选为最优路由的 BGP 路由.

定理 2. BGP 路径搜索过程中产生的无效路由是相关路由.

证明:在链路故障导致的路径搜索过程中产生的无效路由必然经过故障链路,否则,该路由就是一条可用路由.也就是说,无效路由都依赖于故障链路,所以,所有的无效路由都是相关路由.网络设备发生故障的情况与链路故障相同,相当于与故障设备相连的链路同时失效. □

如图 3(a)所示的由 5 个 AS 组成全连接拓扑(clique5),AS0 的网络设备发生故障后,到达 dst 的路由会在 AS1,AS2,AS3 和 AS4 中经历路径搜索过程.为了简化分析,本文采用文献[8]中的同步网络模型描述到达 dst 的路由的收敛过程.同步网络的含义是指:

(1) 任意 AS 间的传输延时都是固定值,不妨假设为 1 秒;

(2) 网络中的 AS 同时收到邻居在上一轮发送来的路由消息,选择新的最优路由并传播出去;

(3) AS 优先选择路径长度最短的路由,对于相同长度的路由,选择来自 ASN 较小的 AS 发送来的路由.

在图 3 中,*ASpath* 表示 AS 当前到达 dst 的最优路由,{}表示没有到达 AS0 的路由,实线方框中的路由表示该 AS 本轮发送给邻居的路由,没有使用实线方框围起来的表示上一轮发送给邻居的路由,*w* 表示发送的是路由取消.

AS0 失效后,在 t=1 时刻,AS3 删除来自 AS0 的所有路由,在 AS1,AS2,AS4 对应的 *Loc-RIB-In* 中仍然存有到达 dst 的路由,路径分别是[10],[20],[40].尽管这 3 条路径由于 AS0 的故障已经成为无效路由,但是 AS3 无法知道这一点,于是从中选择[10]为新路由,并发送[310]给邻居.同时,AS1,AS2,AS4 的操作与 AS3 类似,并没有取消到达 dst 的路由,而是在存储在 *Loc-RIB-In* 中的无效路径中搜索一条最优的,传播给其他 AS.

在 t=2 时刻,所有 AS 同时收到邻居 t=1 时刻发送来的路由声明,更新对应的 *Loc-RIB-In* 并重新选择路由.由于收到的来自 AS2,AS3,AS4 的路由形成路由环路(路径中都包含 AS1),AS1 不再有到达 dst 的路由,向邻居发送路由取消.收到来自 AS1 的[120]后,AS2 选择[310]为最优路由,AS3 和 AS4 选择[120]为最优路由.AS2,AS3,AS4 新选择的路由受 MRAI 时钟的限制,不会马上发送给邻居.路由取消不受 MRAI 的限制,AS1 发送的路由取消在 t=3 时刻到达 AS2,AS3 和 AS4.AS3,AS4 重新选择[210]为最优路由并继续受到 MRAI 的限制.

t=31 时刻,AS2,AS3,AS4 的 MRAI 时钟超时,发送各自的最优路由给邻居.t=32 时刻,AS2 收到的来自 AS3 的路由[3210]和来自 AS4 的路由[4210]包含路由环路,因此不再有到达 dst 的路由,而向邻居发送路由取消.AS3 收到来自 AS2 的路由[2310]也包含路由环路,重新选择[4210],但是发送新的路由声明再次受到 MRAI 的限制.AS3 和 AS4 在 t=33 时刻收到来自 AS2 的路由取消,AS4 选择来自 AS3 的[3210]为最优路由,新路由的发送仍然受到 MRAI 的限制.

直到 t=61 时刻 MRAI 时钟超时,AS3 和 AS4 互相发送路由声明.t=62 时刻,AS3,AS4 发现对方发送来的路由存在路由环路,*Loc-RIB-In* 中不再有到达 dst 的路由,路由收敛过程结束.可见,在 BGP 路由收敛的过程中,每个 AS 会在所有邻居的 *Loc-RIB-In* 中不断搜索、选择和传播无效路由,直到所有 AS 的 *Loc-RIB-In* 中都不再存在

到达 dst 的路由.

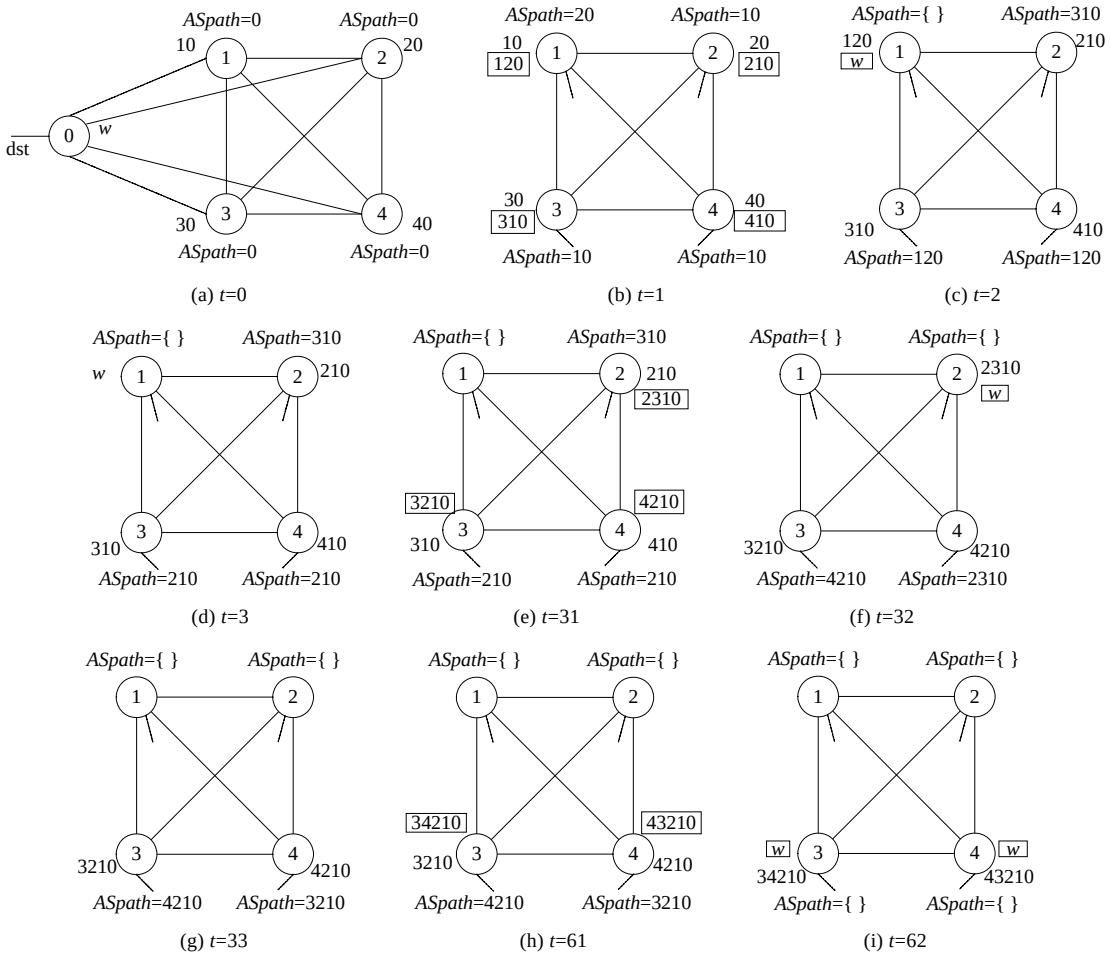


Fig.3 BGP path exploration procedure after the failure of node 0 in a clique with 5 nodes

图 3 5 个 AS 的全连接拓扑上,AS0 发生故障后,标准 BGP 的路径搜索过程

2 相关工作

根据 Labovitz 等人[3]在互联网上所做的实验,路由失效和路由切换事件后的路由平均收敛延时是 3 分钟,最大收敛延时甚至达到 15 分钟.理论分析表明,在 n 个节点的网络中,最坏情况下,路由收敛过程中路径搜索的复杂度为 $O(n!)$.MRAI 对路径搜索过程中的路由更新消息起到打包作用,如果网络中自治系统路径最长为 N,那么最大收敛延时为 $\tau_{MRAI}\cdot N$,其中,τ_{MRAI} 为 MRAI 的值.

Pei 等人[9]提出,在每个最初生成的路由更新消息中增加一种新的路由属性,根源原因通知 RCN,内容包括引起路由消息的链路、该链路当前的状态和一个与该链路相关序号.根据路由更新消息中的 RCN,边界路由器能够判断引起路由变化的位置和原因:RCN 中的链路提供位置信息,链路状态提供路由变化的原因,序号则可以区分同一链路的多次变化.对于同一事件引起的多个路由更新消息,根据 RCN 提供的链路信息,路由器判断当前所有 *Loc-RIB-In* 中的路由哪些会由于该事件而成为无效路由,避免选择和传播无效路由,从而缩短路径搜索过程.但是,RCN 增加了 BGP 通信开销,而且要求每个边界路由器都能探测到相连链路的状态,支持发送和识别 RCN,这不利于 RCN 在网络中逐步部署.

Bremler-Bar 等人[8]提出的 Ghost Flushing 对 BGP 作了很小的改进,但是能够大大改善 BGP 路由收敛.其主要思想是,把存储在所有路由器 *Loc-RIB-In* 中的无效路由看作游荡在网络中的"幽灵(ghost)",在路由收敛之前,首先将这些"幽灵"从网络中驱逐(flush)出去.具体方法是,在到达 dst 的路由被更换为一条比较差(比如 AS 路径更长)的路由并且 MRAI 时钟没有超时的情况下,边界路由器给所有邻居发送路由取消,清除网络中存在的无效路由.传递"坏消息"的隐式路由取消在时间上(先发送显式取消,后发送路由声明)和物理上(两个路由更新消息)分割为一个显式路由取消和一个新路由的路由声明.显式路由取消不受 MRAI 限制,能够迅速清除网络中的无效路由,新路由在网络中的收敛时间远远小于路由失效和路由切换的收敛时间[3].

Ghost Flushing 将路由失效和路由切换后的收敛延时由平均 3 分钟缩短为 10s~15s.然而,这种方法改变了标准 BGP 的语义,标准 BGP 中,显式路由取消表示没有到达目的网络的路由;而 Ghost Flushing 中,显式路由取消被用来通知邻居以前发送的路由不再有效.在某些情况下,Ghost Flushing 会使 BGP 变得不安全,额外发送的路由取消导致某些 AS 没有到达目的网络的路由,尽管邻居 AS 有到达目的网络的可用路由.传递可达性信息是 BGP 的主要功能,因此,提高 BGP 的收敛性不能以损害路由可达性为代价.

3 时间窗口机制

本文提出的时间窗口机制的基本思想是,把相关路由的稳定性加入到路由选择中,使稳定路由有更大的机会被选为最优路由.路由是否参加路由选择,不仅根据该路由的历史变化情况,而且也取决于相关路由的稳定性.

3.1 设计原理

根据相关路由的特点,这里采用根据路由的变化行为来识别相关路由的方法:如果几条到达 dst 的路由在某段时间的变化几乎同步,就认为这些路由有很大的相关性,同时受到了网络中的某个事件的影响.识别相关路由的时间窗口机制的主要数据结构是时间窗口.当第 1 次有来自某个邻居的到达 dst 的路由发生变化时,时间窗口被打开,窗口打开期间来自所有邻居的到达 dst 的路由变化都被记录在时间窗口中.根据前面路径搜索过程的分析可知,如果发生变化的相关路由数目较少,它们对路由稳定性的负面影响不大;如果时间窗口中的路由超过一定数目,边界路由器就可以推断网络中的某个事件导致在互联网中或者互联网的一部分到 dst 的路由处于不稳定状态.由此,当时间窗口关闭时,路由器忽略时间窗口中的相关路由,在时间窗口外的路由中选择最优路由.因为这条路由与时间窗口中同步变化的路由没有相关性,因此可以预期,以后会有较好的稳定性.

时间窗口的生存周期记为 T_{tw},窗口关闭时,只能判断其中的路由在同一段时间发生变化,然而,仅仅由此并不能推断路由间的相关性.但是,如果几条路由多次都出现在相继的时间窗口中,路由在同步变化的可能性就大为增加.但是,后面的时间窗口不能获知其中的路由是否在之前的时间窗口中出现过.因此,除了判断相关路由在同一窗口中出现以外,还需要提供路由同步变化的依据.

路由抖动抑制机制为能够为判断 BGP 路由过去的稳定性提供依据.RFD 原理是,路由以后的稳定情况与其在最近一段时间的变化相关:如果路由最近一段时间变化频繁,以后该路由发生改变的可能性就大;如果路由最近一段时间保持稳定状态,以后该路由处于稳定状态的可能性就大.RFD 使用惩罚值记录路由过去变化的剧烈程度,每次路由发生变化,惩罚值就增加一个常量,隐式路由取消的惩罚值增量记为 P_{AC},显式路由取消的增量记为 P_W.路由变化得越剧烈,惩罚值增长得越快.如果惩罚值超过了代表路由变化最大忍受程度的阈值 P_{cutoff},该路由就被抑制而不能参加路由选择,以限制它在网络中的传播.当路由处于稳定状态时,惩罚值以 H 为半衰期,按指数规律衰减,直到小于另一个阈值 P_{reuse},路由才能重新参加路由选择.但是需要注意的是,RFD 仅仅根据该路由本身的变化历史来判断路由以后的稳定性.

将时间窗口和 RFD 惩罚值提供的历史变化信息结合起来,能够大大增加判断相关路由的准确性:如果路由最近一段时间不够稳定,而又出现在同一时间窗口中,那么,这些路由很可能在同步变化.因此,在时间窗口机制中设定阈值 $P_{timeWin}$,将变化路由加入时间窗口前检查该路由的 RFD 惩罚值,只有惩罚值大于 $P_{timeWin}$ 的路由才被记录在时间窗口中.时间窗口关闭时检查其中路由的数量,如果多于 K,则执行路由选择过程,忽略窗口内的路

由,只有窗口外的路由才能参加.时间窗口关闭算法的描述如图 4 所示.

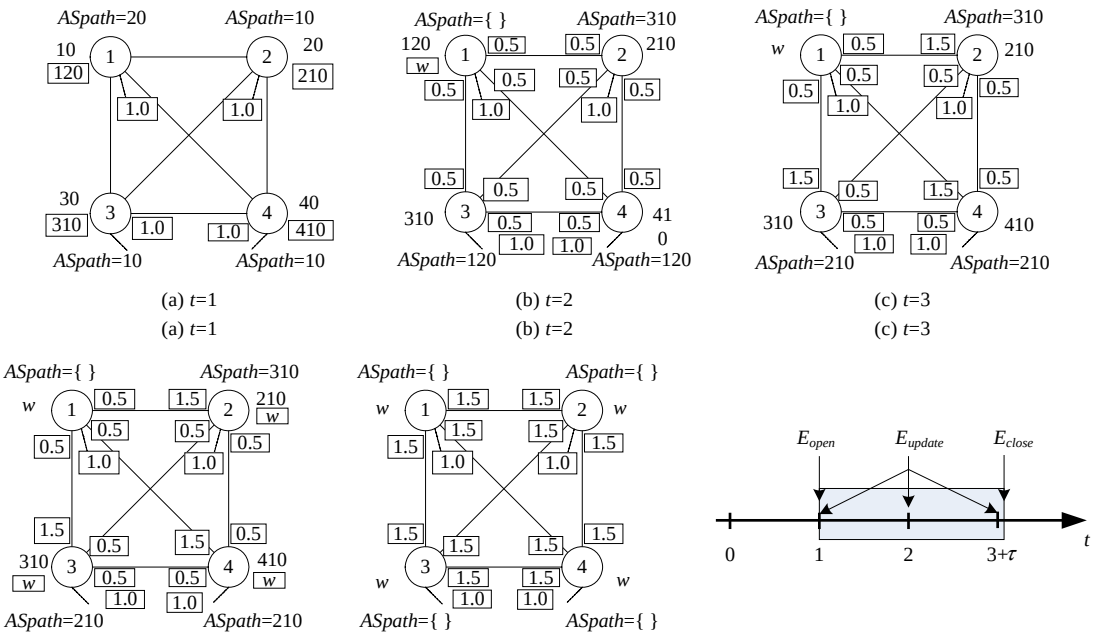
```
1: R_tw←route set enclosed in the time window
2: R←route set received from all neighbors
3: r*←the best route selected to dst
4: IF r*∈R_tw AND |R_tw|>K THEN
5:     r*=Decision process on R–R_tw
6:     IF r* is NULL THEN
7:         Send Withdrawal of dst to neighbors
8:     ELSE
9:         IF MRAI timer does not timeout THEN
10:            Stop MRAI timer
11:        ENDIF
12:        Send Announcement of r* to neighbors
13:    ENDIF
14: ENDIF
```

Fig.4 Algorithm on time window closure

图 4 时间窗口关闭算法

3.2 路由失效

这里以图 3(a)的同步网络模型路由失效的情况来说明时间窗口机制改善路由收敛的原理.t=0 时刻 AS0 失效,在时间窗口机制作用下,到达 dst 的路由收敛过程如图 5 所示.虚线包围的数值表示从该邻居获得的路由的 RFD 惩罚值.RFD 参数的设置与 Cisco 路由器的默认值成比例,P_W=1.0,P_{AC}=0.5,P_{cutoff}=2.0,P_{reuse}=0.75,H=900s.



(a) t=1 (b) t=2 (c) t=3

(a) t=1 (b) t=2 (c) t=3

(d) t=3+τ (e) t=4+τ (f) Time window

(d) t=3+τ (e) t=4+τ (f) 时间窗口

Fig.5 BGP convergence in clique of size 5 after fail-down event of node 0 with time window mechanism enabled

图 5 5 个 AS 的全连接拓扑上 AS0 失效时间后,使用时间窗口机制的 BGP 路由收敛过程

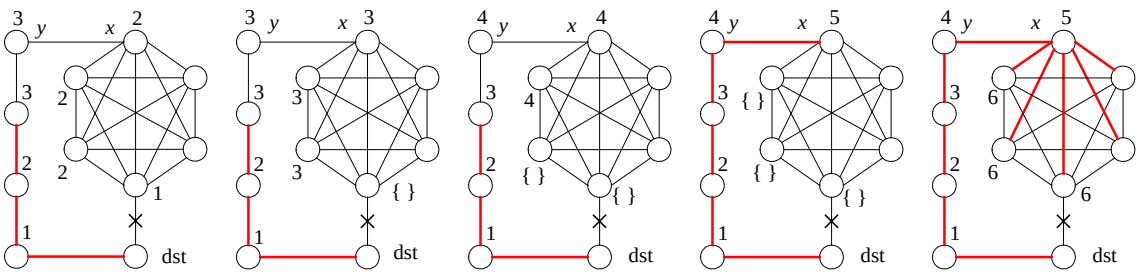
在 t=1 时刻,AS2 收到来自 AS0 的路由取消,打开时间窗口,在图 5(f)上表示为 E_{open} 事件.时间窗口的持续时间设为 T_{tw}=2+τ(秒),其中的τ表示很小的增量.在 t=3+τ时刻,AS2 关闭时间窗口,在图 5(f)上表示为 E_{close} 事件.在

时间窗口打开期间,来自 AS1,AS3,AS4 的路由也发生了变化,在图 5(f)中都用 E_{update} 表示,因此,这些路由的 RFD 惩罚值也相应增加.由于在同一段时间里,所有邻居 *Loc-RIB-In* 中的路由都发生了变化,AS2 判断网络中到达 dst 的路由不够稳定.因此,当时间窗口关闭时,AS2 重新选择到 dst 的最优路由,但是忽略时间窗口中记录的路由(图 4 中第 5 行).路由选择的结果是没有到达 dst 的路由,AS2 向邻居 AS1,AS3,AS4 发送路由取消(图 4 中第 7 行、第 8 行).同样的过程也在 AS1,AS3,AS4 上发生.在收到来自邻居的路由取消后,所有 AS 清除 *Loc-RIB-In* 中的无效路由,收敛时间缩短为 $4+\tau$(s).

3.3 路由切换

路由切换与路由失效的不同之处在于时间窗口关闭时在时间窗口外还有其他路由.这些路由相对稳定,与窗口内的变化路由不相关,时间窗口关闭算法会从中选出一条最优路由(图 4 中第 5 行).对于选出的最优路由,时间窗口关闭算法与标准 BGP 的操作有所不同(图 4 中第 10 行~第 13 行),新选出的路由不受 MRAI 限制,立即通过路由更新消息发送给邻居,目的是使网络中路由尽快达到一致的稳定状态.

图 6 是在时间窗口机制作用下,路由切换事件后的路由收敛过程.在初始状态,每个 AS 到目的网络 dst 的路由都处于稳定状态,AS 旁边的数字表示到 dst 的跳数,{}表示节点在该时刻没有到 dst 的路由.*x* 和 *y* 到 dst 的跳数分别是 3 和 2.在 t=0 时刻,dst 与 6 个 AS 的全连接拓扑间(clique6)的链路中断.在 t=1 时刻,*x* 收到取消到达 dst 路由的更新消息,时间窗口被打开,持续时间设为 $2+\tau$.当 *x* 的时间窗口在 t=3+τ时刻关闭时,*x* 忽略所有 clique6 中 AS 发送来的到 dst 的路由,因为时间窗口记录了这些路由的变化.来自 *y* 的路由位于时间窗口外,被选为最优路由并被发送给 clique6 中的其他 AS.在 t=4+τ时刻,所有 AS 的路由都切换到稳定的备用路由,整个路由收敛过程不超过 5s,标准 BGP 的收敛时间是 121s.



(a) t=0 (b) t=1 (c) t=2 (d) t=2+τ (e) t=4+τ

Fig.6 Routing convergence after fail-over event, with time window mechanism enabled

图 6 在使用时间窗口机制后,路由切换事件后的路由收敛过程

3.4 与Ghost Flushing的比较

Ghost Flushing 是消除路径搜索过程中无效路由、加速路由收敛的有效方法.对图 3(a)中的情况,Ghost Flushing 使收敛时间缩短到 3s,比使用时间窗口机制的效果还要好.但在某些情况下,Ghost Flushing 会导致存在可用路径而目的网络不可达的情况.

在图 7 中,假设在 *A* 和 *B* 之间的链路发生故障之前,*X* 选择[*KBA*]作为到达 dst 的路由,通过 *K* 转发到达 dst 分组.*A* 和 *B* 之间的链路发生故障之后,假设 *X* 分别在时刻 T,$T+\tau_1$ 和 $T+\tau_2$ 收到来自 *K*,*M*,*N* 的路由取消.如果使用标准 BGP,*X* 会依次选择[*MBA*],[*NBA*]和[*YZDCA*].*X* 选择[*MBA*]后,给 *S* 发送路由[*XMBA*],但此后给 *S* 发送[*XMBA*]和[*XYZDCA*]时,需要等待 MRAI 时钟超时.在时间间隔[$T+\tau_1$,$T+\tau_{MRAI}$]中,*X* 的路由先后是[*NBA*]和[*YZDCA*],*S* 的路由是[*XMBA*],可见标准 BGP 发生了 AS 间路由不一致的情况.但是,这没有妨碍网络的可达性和正确转发分组,*S* 到 dst 的路径不正确但是路由的下一跳正确,源自 *S* 到达 dst 的分组被转发给 *X*,再经过[*NBA*]或者[*YZDCA*]转发到 dst.直到 $T+\tau_{MRAI}$ 时刻,*X* 的 MRAI 时钟超时,发送[*XYZDCA*]给 *S*,各个 AS 的路由达到一致状态.

如果使用 Ghost Flushing,则 *X* 在收到来自 *K* 的路由取消、选择路由[*MBA*]后,发送额外的路由取消给 *S*.在时间间隔[$T+\tau_1$,$T+\tau_{MRAI}$]中,*S* 都没有到达 dst 的路由,直到 $T+\tau_{MRAI}$ 时刻,*S* 才收到[*XYZDCA*].可见,在图 7 的例子中,

Ghost Flushing 使备用路由的收敛变坏,破坏了网络的可达行,但是时间窗口机制却仍然能够改善路由的收敛情况.*X* 收到来自 *K* 的路由取消后,发送[*MBA*]给 *S* 并打开时间窗口.在窗口打开期间,*X* 相继收到来自 *M* 和 *N* 的路由取消.假设 $T_{tw}>\tau_2$,*X* 在 $T+T_{tw}$ 时刻关闭窗口时,选择[*YZDCA*]并立即发送给 *S*.可见,时间窗口机制的设计不仅加速了路由收敛,而且消除了标准 BGP 中节点路由不一致的情况.

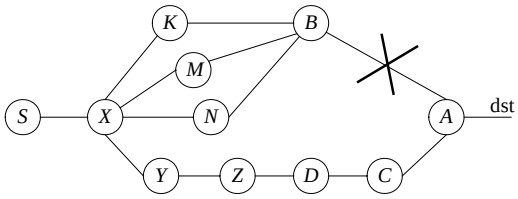


Fig.7 Scenario in which time window mechanism is superior to Ghost Flushing

图 7 时间窗口机制优于 Ghost Flushing 的情况

3.5 参数设置

时间窗口机制的参数包括窗口持续时间 T_{tw}、时间窗口中路由的惩罚值最小值 $P_{timeWin}$ 和触发时间窗口关闭算法的相关路由最小数目 *K*,这 3 个参数对准确识别全部相关路由、避免错误抑制不相关路由有很大作用.

无论 T_{tw} 设为何值,如果变化路由的惩罚值超过 $P_{timeWin}$,则会被记录在时间窗口中.因为在路由发生变化时,如果已经有时间窗口打开,则该路由被记录在该窗口中;如果没有时间窗口打开,该路由变化则打开一个新的时间窗口.但是,窗口持续时间 T_{tw} 的大小会影响对相关路由的判断:较小的 T_{tw} 使时间窗口对同步的路由变化快速作出反应,但是,如果 T_{tw} 过小,时间窗口不能捕捉到全部变化路由,或者窗口中路由数目小于 *K* 而不能触发窗口关闭算法;过大的 T_{tw} 会使来自某个邻居的路由在同一窗口中变化多次,也降低了时间窗口机制的有效性.BGP 给邻居发送到达同一目的网络路由的间隔为 MRAI,因此,T_{tw} 应该设为小于 MRAI 的值.在保证 T_{tw} 能够捕捉到所有同步变化路由的情况下,*K* 值一般不能过小:一方面路由样本数量过少,不足以判断网络中路由的稳定情况;另一方面,少量相关路由对路由收敛的影响也较小.

只有 RFD 惩罚值大于 $P_{timeWin}$ 的路由才能被加入时间窗口,时间窗口机制以此作为路由同步变化的依据.但是,这只是判断相关路由的必要条件,不是充分条件.如果多条相关路由同步变化,则路由惩罚值也同步增加,从而被某个时间窗口记录;反之,根据 RFD 惩罚值超过 $P_{timeWin}$ 来判断路由在过去同步变化,则会有误判的情况发生.落入时间窗口的偶然变化的路由,会被错误地判断为相关路由而在路由选择中被忽略,导致没有可用路由或者选出的路由不是最优的.最好的方法是把所有时间窗口中的内容都记录下来,在相继多个窗口中出现的路由可以认定为在同步变化,但是,这种方法开销较大.时间窗口机制中采用了比较路由RFD惩罚值和$P_{timeWin}$的方法,如果几条路由过去的变化超过一定程度(用 $P_{timeWin}$ 来描述),并在某段时间都发生了变化,则认为这些路由为同步变化的相关路由.$P_{timeWin}$ 值如果设置得过大,则时间窗口机制对相关路由的反应变慢;如果设置得过小,则不能为判定路由同步变化提供充分的依据,增大了误判的可能性.如果把最近两次变化时间非常接近作为判定相关路由的依据,则被记录在时间窗口中的路由 RFD 惩罚值应该大于 $P_{timeWin}=p_{-1}\times e^{-\lambda(\tau+T_{tw})}+p_0$,其中:$p_0$ 和 p_{-1} 分别是指路由当前变化和上一次变化引起的惩罚值增量,由于一般情况下,$P_{AC}<P_W$,p_0 和 p_{-1} 的值应该取较小的 P_{AC};$\lambda=\ln 2/H$ 是惩罚值的衰减速度;τ表示当前变化和上一次变化之间的时间间隔,反映了对相关路由变化周期的估计值.

4 模拟实验

4.1 实验环境和设置

模拟使用规模为 3~20 个自治系统的全连接拓扑来评价时间窗口机制改善路由收敛的作用.我们在模拟软件 SSFNet[10]的 BGP 源代码中增加了时间窗口机制的实现.每个自治系统只包括一个 BGP 路由器,路由器间链

路的传输延时设定为固定值 0.01s.模拟中的 MRAI 设为默认值 30s.为了避免路由器收到的路由消息过于集中,在 MRAI 的基础上增加了随机抖动.RFD 的参数与 Cisco 路由器的默认值成比例,P_W=1.0,P_{AC}=0.5,P_{cutoff}=2.0,P_{reuse}=0.75,H=900s.时间窗口的参数是 T_{tw} 设为 4s~6s,$P_{timeWin}$=0.75,K=3.初始状态,每个路由器都有到达目的网络 d 的稳定路由 r_d.在 T_0 时刻,源节点取消 r_d.实验记录了使用标准 BGP 和使用具有时间窗口机制的 BGP 两种情况的路由收敛过程,从中统计路由收敛时间和路由更新消息的数量.在不同规模拓扑上,每种参数设置的实验分别做了 10 次,每次实验中采用不同的生成随机数的种子,根据 10 次实验的结果得到收敛延时和路由更新消息数量的平均值.

4.2 实验结果

如图 8 所示,标准 BGP 的路由收敛延时随着拓扑规模呈线性规律增加.启用时间窗口机制后,收敛延时得到不同程度的改善.当 T_{tw} 设为 4s 时,收敛延时的大致变化规律与标准 BGP 相一致,但在某些规模的拓扑,如 8,11,14 个节点的拓扑上,收敛延时减小很多,多数其他规模拓扑上的收敛延时改善不大.当 T_{tw} 设为 5s 时,节点数量小于 13 的拓扑收敛延时大致为常数,约为 30s.在规模较小的拓扑上,路由 r_d 被源节点取消后,拓扑中的节点重新选择新路由并发送给邻居,路由的 RFD 惩罚值也相应地增加 P_{AC}.收到新路由后,经过 MRAI 时间,节点再次选择和发送路由消息,RFD 惩罚值又增加 P_{AC},累计的惩罚值超过 $P_{timeWin}$.时间窗口中的数量大于 K 触发路由选择,窗口中的路由都被忽略,网络中的无效路由很快被清除.当拓扑规模大于 13 个节点时,路由收敛分为 3 种情况: (1) 在 14,17,19,20 个节点拓扑上,收敛时间与使用标准 BGP 的收敛时间相当;(2) 在 15 个节点的拓扑上,收敛时间大约是标准 BGP 收敛时间的一半;(3) 在 16,18 个节点的拓扑上,收敛时间约为 30s,远远小于标准 BGP 的收敛时间.出现这种情况的原因在于,没有被记录在时间窗口中的无效路由会增加收敛时间,当拓扑中的节点较少时,无效路由落在时间窗口外的可能性小,收敛延时也较小.当 T_{tw} 增加到 6s 时,所有规模拓扑的收敛时间大致为常数 30s,与标准 BGP 相比大为提高.

比较图 8 与图 9,收敛过程中发送的路由消息数量与收敛延时有大致的对应关系,收敛时间短,收敛中的通信开销也相对较小.标准 BGP 收敛中的路由消息数量呈指数过滤增长.使用时间窗口机制的 3 种情况,路由收敛的通信开销都有不同程度的降低.当 T_{tw} 设为 6s 时,路由消息数量减少得最为明显.当 T_{tw} 设为 4s 时,路由消息的数量比标准 BGP 有所减少,但并不明显,多数拓扑上消息数量与标准 BGP 大致相当;当 T_{tw} 设为 5s 时,在某些规模的拓扑上,路由消息数量与使用标准 BGP 时大致相当,而其他拓扑上的路由消息数量与 T_{tw} 为 6s 时的数量大致相当.

实验结果说明,在模拟所使用的网络拓扑中,当 T_{tw} 设为 4s 时,时间窗口机制对 BGP 路由收敛的改善不大;当 T_{tw} 设为 5s 时,时间窗口的作用受网络中随机事件的影响,部分地发挥了作用;当 T_{tw} 设为 6s 时,时间窗口机制极大地改善了 BGP 路由收敛.

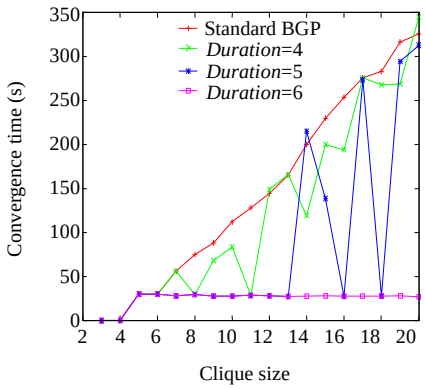


Fig.8 Convergence delay comparison

图 8 路由收敛延时的比较

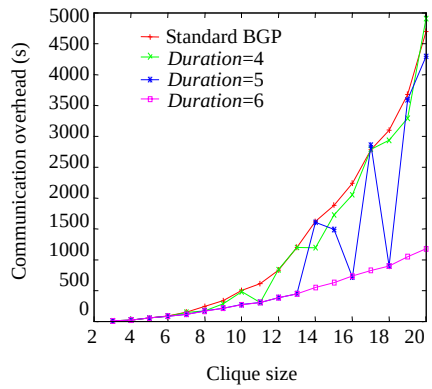


Fig.9 Communication overhead comparison

图 9 路由收敛过程中通信开销的比较

5 结束语

BGP 路由慢收敛问题对互联网端到端的性能有显著的影响.本文利用 BGP 路由传播过程中形成的相关性,设计了一种新的时间窗口机制,以改善 BGP 路由收敛.时间窗口机制利用路由的 RFD 惩罚,通过发现路由同步变化判断路由相关性.模型分析和模拟实验结果说明,时间窗口机制能够大大缩短 BGP 路由收敛时间,减少收敛过程中的 BGP 路由更新消息数量.时间窗口机制不需要在 BGP 路由消息中增加额外的信息,因此更容易在互联网中逐步地部署.

致谢 在此,我们向对本文工作给予建议的同行和老师表示感谢.

References:

[1] Rekhter Y, Li T, Hares S. A border gateway protocol 4 (BGP-4). RFC 4271, 2006.

[2] Labovitz C, Malan GR, Jahanian F. Internet routing instability. IEEE/ACM Trans. on Networking, 1998,6(5):15−527.

[3] Labovitz C, Ahuja A, Bose A, Jahanian F. Delayed Internet routing convergence. IEEE/ACM Trans. on Networking, 2001,9(3): 293−306.

[4] Zhang BC, Massey D, Zhang LX. Destination reachability and BGP convergence time. In: Proc. of the IEEE Global Telecommunications Conf., Vol.3. Los Angeles: IEEE, 2004. 1383−1389.

[5] Villamizar C, Chandra R, Govindan R. BGP route flap damping. RFC 2439, 1998.

[6] Bartell M, Zhang R. BGP Design and Implementation. Cisco Press, 2003.

[7] Mao ZM, Govindan R, Varghese G, Katz RH. Route flap damping exacerbates Internet routing convergence. In: Proc. of the ACM SIGCOMM, Vol.32. New York: ACM, 2002. 221−233.

[8] Afek Y, Bremler-Barr A, Schwarz S. Improved BGP convergence via ghost flushing. IEEE Journal on Selected Areas in Communications, 2004,22(10):1933−1948.

[9] Pei D, Azuma M, Massey D, Zhang LX. BGP-RCN: Improving BGP convergence through root cause notification. Computer Networks, 2005,48(2):175−194.

[10] The SSFnet project. http://www.ssfnet.org/homepage.html

王立军(1978－),男,河北唐山人,博士,主要研究领域为互联网域间路由协议.

吴建平(1953－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为互联网网络体系结构,IPv6 和下一代互联网.