ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.12, December 2007, pp.3048−3059
DOI: 10.1360/jos183048

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

支持域间分布式分组过滤的 BGP 扩展*

王立军+, 吴建平, 徐 恪

(清华大学 计算机科学与技术系,北京 100084)

BGP Extension to Support Inter-Domain Distributed Packets Filtering

WANG Li-Jun+, WU Jian-Ping, XU Ke

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

+ Corresponding author: Phn: +86-10-62785822, Fax: +86-10-62785819, E-mail: wlj@csnet1.cs.tsinghua.du.cn

Wang LJ, Wu JP, Xu K. BGP extension to support inter-domain distributed packets filtering. *Journal of Software*, 2007,18(12):3048−3059. http://www.jos.org.cn/1000-9825/18/3048.htm

Abstract: To be trustworthy is an important characteristic of the next generation Internet. The routing system of the present Internet forwards packets only according to the destination IP address. Forged packets with spoofed source IP address will also be forwarded to the destination, which impairs the security of receiver and conceals the real identity of the sender. The trustworthy Internet requires the routing system not only forward packets correctly, but also validate the packets from the real sender. Inter-domain distributed packet filtering is an effective method to filter out spoofed packets. This paper proposes to extend BGP with route selection notice to provide filtering criteria. With the support, border routers can validate incoming packets and filter the spoofed packets form false autonomous systems. Simulation result indicates BGP route selection notice does not impair the routing function of BGP, and both proper design acceptable bandwidth cost and fast convergence may be achieved simultaneously.

Key words: trustworthy Internet; border gateway protocol (BGP); inter-domain routing; distributed packets filtering

摘 要: 可信任是下一代互联网的重要特征.目前,互联网的路由系统只按照分组的目的 IP 地址转发分组,携带虚假源 IP 地址的伪造分组也会被传输到目的地,这会在威胁接收方安全的同时,隐藏发送方的真实身份.可信任互联网的路由系统不仅需要能够正确地转发分组,而且能够验证分组来自正确的发送方.基于路由的域间分布式分组过滤是过滤伪造分组的有效方法.提出了 BGP 的路由选择通知功能扩展,为域间分组过滤提供过滤标准.在扩展的支持下,边界路由器能够鉴别进入本自治系统的分组的真实性,过滤掉伪造其他自治系统地址的分组.模拟结果表明,路由选择通知不会对 BGP 正常的路由功能产生负面影响,选择合理的路由选择时钟参数,可以在同时取得较小带宽开销和较快收敛速度的情况下,为域间分布式分组过滤提供支持.

关键词: 可信任互联网;边界网关协议;域间路由;分布式分组过滤

中图法分类号: TP393 **文献标识码**: A

* 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 2006-07-15; Accepted 2006-10-10

目前的互联网提供尽力而为的传输服务,路由器在转发表中查找分组的目的 IP 地址,将分组发往对应的下一跳.当接收方从网络收到一个分组时,只能通过源 IP 地址判断发送者.分组的发送者可以任意指定分组的源 IP 地址,以达到伪造和冒充的目的.携带不真实源 IP 地址的分组通常被称为伪造分组.接收方不能辨别分组中的源 IP 地址是真实的还是伪造的,无法确定分组是否来自真实的发送方.因此,目前的互联网是不可信的.多种网络攻击利用伪造源 IP 地址的方法,隐藏攻击者的真实身份以逃避惩罚.随着互联网技术的飞速发展和网络应用的普及,越来越多的对安全敏感的应用将被部署到网络中,互联网只有提供更可信任的网络服务才能满足未来的发展要求.保证分组源 IP 地址的真实性是实现可信任网络的核心问题.

可信任互联网的路由系统提供两个方面的服务:

(1) 尽力而为地将分组沿着可行的最优路径转发到接收方;

(2) 保证接收方收到的分组来自真实的发送方.

目前的路由系统只能为第 1 个方面提供支持,服务的第 2 方面需要提出新的方法来解决.Park[1]提出使用基于路由的分布式分组过滤(distributed packet filtering,简称 DPF)来解决伪造源 IP 地址的问题,路由器根据路由选择和网络拓扑信息使伪造分组到达接收者之前在网络中被过滤掉.基于 1997 年互联网拓扑的实验表明,如果 20%的自治系统部署 DPF,超过 80%的伪造分组可以被过滤掉.

路由系统主要分为数据平面和控制平面,控制平面计算转发表提供给数据平面以转发分组.分组真实性的检查是在数据平面完成,基于路由信息的检查规则也应该由控制平面提供.另一方面,互联网的路由包括两个层次:使用 OSPF,IS-IS,RIP 等协议的域内路由和以边界网关协议 BGP[2]为事实标准的域间路由.路由系统为真实 IP 地址提供支持应该结合路由系统的层次结构:域间路由和域内路由分别为粗粒度和细粒度的真实 IP 地址服务提供支持.BGP 只有在路由发生变化时才会发送路由信息,相对于周期性计算某种最优化量度的域内路由协议,其稳定性更好.尽管域间路由也存在稳定性的问题[3−5],但是已经提出的 BGP 配置原则[6]和多种改进措施[7−10]会在很大程度上提高 BGP 的稳定性.因此,基于路由的分布式分组过滤应用在域间作粗粒度的过滤比应用在域内更为可行.

基于 BGP 的域间分布式分组过滤的含义可以表述为:边界路由器根据 BGP 路由信息生成分组转发的路径约束,并对进入本自治系统的分组携带的地址信息进行检查,如果发现分组的转发路径与分组的地址信息不符,则判断该分组为伪造分组并丢弃,以确保转发的分组来源于真实的自治系统.目前,互联网使用的 BGP-4 无法提供验证分组真实性的信息.为了支持域间分布式分组过滤,我们采用扩展 BGP 的方法,增加了路由选择通知功能.基本思想是:自治系统 A 在选择了来自邻居自治系统 B 的路由 r_d 后,给 B 回送路由选择通知消息,其中包含被选择路由的 IP 前缀 p_d 和自治系统 A 的地址空间 s_A.根据路由选择通知消息,自治系统 B 可以知道路由 r_d 被 A 选择,A 中主机发送到 p_d 的分组会经过 B 转发.随着路由的传播,路由选择通知消息也需要沿着路由传播相反的方向传递,直到路由的源自治系统.设计中我们详细考虑了路由选择通知消息的通信开销和路由选择消息的收敛时间,设置路由选择时钟可以达到捎带更多路由选择通知消息的目的,减少路由器的通信开销.模拟实验的结果表明,扩展 BGP 路由选择通知功能不会对 BGP 的路由功能产生负面影响.

本文第 1 节介绍问题的研究背景并总结相关研究工作.第 2 节阐述路由选择通知的设计原理.第 3 节具体介绍对 BGP 所作的扩展.第 4 节通过模拟结果说明路由选择通知的性能.第 5 节进一步讨论了相关问题.最后总结全文.

1 研究背景

1.1 边界网关协议BGP

互联网由超过 22 000 个互连的自治系统(autonomous system,简称 AS)组成.自治系统的特点包括:

(1) 由唯一的 16 比特 AS 号码标识;

(2) 具有 1 个或者多个 IP 地址前缀;

(3) 内部的边界路由器具有一致的域间路由策略.

自治系统内部使用 IGP、自治系统之间采用 BGP 传递网络可达性信息.为了保证路由信息的可靠性,BGP 基于 TCP 建立连接.同一自治系统内的边界路由器间建立的 BGP 会话称为 iBGP(internal BGP),相邻自治系统间建立的 BGP 会话称为 eBGP(external BGP).BGP 的路由消息包括 4 种类型:Open,KeepAlive,Update, Notification,其中最主要的是传递路由信息的 Update 消息,包括声明新路由的 Announcement 和取消路由的 Withdrawal.BGP 的每条路由包括目的网络前缀和一系列描述路由特征的路由属性.BGP 的路由属性有很多种,为路由策略提供灵活、丰富的支持.BGP 具有很强的扩展性,表现为:增加新的路由属性以扩展 BGP 的功能;通过能力声明[11]确定建立 BGP 会话的双方是否支持某种新功能.BGP-4 被提出后,已经有多种功能[8,12]加入到 BGP 中.

BGP 是基于策略的路由,自治系统间的连接关系决定了路由选择和路由输出策略.主要的自治系统间关系包括 provider-customer 和 peer-to-peer 两种.在 provider-customer 关系中,作为 provider 的自治系统为 customer 提供通常是有偿的互联网接入服务;在 peer-to-peer 关系中,建立连接的两个自治系统间是对等关系,向对方提供访问本网络的服务,一般不需要支付费用.provider 会把所有的路由传递给 customer,customer 只把源自本自治系统的路由以及来自下级 customer 的路由传递给 provider,这样保证了 customer 对整个互联网的访问.peer 通常只把本自治系统的路由以及来自下级 customer 的路由传递给 peer,因此,peer 之间只能互相访问对方以及对方 customer 的网络.

1.2 相关研究

IP 追踪[13,14]通过追踪伪造分组的转发路径寻找分组真实的发送方.基于分组标记的 IP 追踪方法把分组所经过的路由器信息记录在分组头部的某个字段中,但这种方法只能追踪大量分组构成的分组流;基于分组记录的追踪方法在路由器上记录一段时间内经过的分组的部分信息,尽管能够追踪单个分组,但是会给路由器带来巨大的计算和存储开销.需要注意的是,IP 追踪是在危害产生后的补救措施,无法提供预防性的保护.

分组过滤技术是另一种解决伪造分组问题的方法,试图在伪造分组到达接收方之前,在网络中将其过滤.分组过滤通常根据一定的规则检查分组中的某些字段,以判断分组是否来自真实的发送方.基于转发表的过滤方法[15]规定,分组的进入端口必须与源地址的转发端口一致.但是如果使用这种方法,路由不对称会导致很多真实分组被过滤掉.入口过滤[16]在边界路由器上部署流量过滤器,只允许源地址属于边界网络地址的分组通过.这种方法的问题是不仅额外开销导致网络性能下降,使自治系统没有激励部署,而且只有在互联网上得到广泛部署才会有效.

Bremler-Barr 等人[17]从保护目的网络的角度提出了伪造预防方法 SPM.部署 SPM 的每对源/目的自治系统,通过协商得到一定时间内有效的唯一密钥.每个从自治系统 S 发出,目的自治系统为 D 的分组中都附加密钥 $K(S;D)$.到达 D 的分组密钥被验证正确后,即把密钥从分组中删除;如果验证不正确,就将分组丢弃.目的自治系统 D 可以通过 SPM 保证源地址属于 S 的分组真实性,实现粗粒度的真实 IP 地址访问.它的优势在于参加 SPM 的自治系统可以保护本网络中的用户,因此,自治系统有部署 SPM 的激励.但是,SPM 密钥的管理、协商和同步难以控制.

针对 DDoS 攻击分组中使用伪造源 IP 地址的问题,Park 等人[1]系统地阐述了基于路由的分布式分组过滤的原理,自治系统根据路由系统提供源地址与路由器端口的映射信息判断分组的真实性.Park 等人通过模拟实验研究了分布式分组过滤的有效性,发现在符合幂率特性的互联网拓扑中,20%的自治系统部署 DPF 就能消除超过 80%的伪造分组.Park 等人通过比较不同节点集合执行分组过滤的效果发现,如果限定过滤节点的比例,集合覆盖(vertex cover)方法选出的节点集合比随机选择的节点集合有明显的性能优势.尽管在一个图上找到最小节点覆盖集合是 NP-C 问题,但是,有很多启发式算法[19]能够提供近似解.互联网拓扑具有幂率特性[20],在 1997 年,互联网拓扑上选出的节点覆盖大约只占自治系统总数的 18.9%,效果却可以使 80%伪造源地址的分组被过滤掉,而且使用最大过滤和半最大过滤的效果非常接近.

然而,文献[1]中没有指出路由系统如何为分组过滤提供分组真实性判断的标准.源地址验证实施(SAVE)[18]是为了支持分布式分组过滤而设计的一种新协议.路由器向转发表中的每个目的网络发送一个含有本地网络

地址的 SAVE 更新消息.SAVE 更新消息经过的每个路由器记录消息中的网络前缀,并为每个路由器端口构建一个与之对应的源地址表.当路由器收到一个分组时,根据其进入端口的源地址表判断分组的真实性. SAVE 没有考虑路由系统的层次结构,是一种在所有路由器(包括域内路由器和边界路由器)间使用的协议.一方面,变化的域内路由可能频繁产生 SAVE 更新消息,增大通信开销的同时,也使源地址表不准确;另一方面,新协议在网络中难以逐步部署.

2 路由选择通知扩展的设计原理

2.1 分组过滤方法

互联网可以抽象为一个无向图 $G(V,E)$,其中,V 是节点集合,每个节点表示一个自治系统;E 是边的集合,每条边表示自治系统间的 BGP 会话.$\{R(d,s)|d\in V,s\in V\}$是图 G 上的路由集合,$R(d,s)$表示从节点 s 到节点 d 的路由,在不产生歧义的情况下,$R(d,s)$也表示从 s 到 d 的路径,分组 $P(d,s)$将沿着 $R(d,s)$上的边转发.$F_e:V^2\to\{0,1\}$是定义在 $e=(u,v)\in E$ 上、表示分组过滤的函数,当节点 v 收到来自边 e 的分组 $P(d,s)$时,$F_e(d,s)=1$ 表示丢弃分组,$F_e(d,s)=0$ 表示转发分组.基于路由的分组过滤可以表述为:如果 $e\in R(d,s)$,那么 $F_e(d,s)=0$.可见,基于路由的分组过滤不会影响真实分组的转发.

根据路由信息形成的约束,路由器对分组地址信息的检查分为两种:基于源/目的地址联合检查的最大过滤和只基于源地址检查的半最大过滤.

定义 1. 当且仅当存在路径 $P(d,s)$使 $e\in R(d,s)$,基于路由的分组过滤 $F_e(d,s)=0$.执行这种分组过滤的节点为最大过滤器,最大过滤可以表示为 $\tilde{F}_e(d,s)=\begin{cases}0, e\in R(d,s)\\ 1, e\notin R(d,s)\end{cases}$.

定义 2. 如果存在路径 $R(t,s),t\in V$ 使 $e\in R(t,s)$,基于路由的分组过滤 $F_e(d,s)=0$.执行这种分组过滤的节点为半最大过滤器,半最大过滤可以表示为 $\hat{F}_e(d,s)=\begin{cases}0, \exists t\in V, e\in R(t,s)\\ 1, \text{Otherwise}\end{cases}$.

半最大过滤器过滤分组的能力明显不如最大过滤器,但是最大过滤需要同时检查分组的源和目的地址,对数据平面的查找速度有更高的要求.在实际应用中,使用何种过滤方法应由网络管理者的配置决定.控制平面应该能够根据不同的过滤方法生成不同的分组正确性验证标准,为过滤伪造分组提供支持.

2.2 路由选择通知

与 SAVE 设计新协议相比不同,我们设计的路由选择通常采用扩展 BGP 的方法,优点包括:

(1) BGP 提供了很强的扩展机制,如属性扩展、能力协商等,有利于新的功能在网络中逐步部署;

(2) BGP 建立连接,维护连接、错误检验、消息传输等机制保证 BGP 会话有很强的可靠性,扩展 BGP 的方法可以利用已有的机制,减少重复设计和降低设计复杂性.

在设计中,我们没有把目标局限于解决某种特定问题,比如 DDoS 攻击,而是从系统的角度考虑,把真实 IP 地址访问作为路由系统提供的一般服务.

我们通过图 1 所示的拓扑连接说明路由选择通知如何为分组正确性验证提供支持.图 1 中的 AS_1~AS_6 的圆圈表示自治系统号码分别是 1~6 的自治系统.为了简化问题,假设每个自治系统只有一个边界路由器,由字母 A~字母 F 表示,自治系统所有的 IP 地址块用 p_1~p_6 表示.圆圈间的连接表示自治系统间的 BGP 会话,有向边表示 provider-customer 关系,箭头由 provider 指向 customer,无向边表示 peer-to-peer 关系.

使用现有的 BGP 协议,边界路由器在把一条路由传播给邻居路由器后,不能确定邻居是否选择该路由,也不知道邻居进一步把路由传播到哪里.因此,当以某个地址为源 IP 地址的分组从邻居转发过来时,边界路由器无法判断分组中的源 IP 地址是否真实.我们的主要设计思想是:如果边界路由器选择了一条路由,它要给发送这条路由的邻居回送一个消息,通知该路由被选择,而且在消息中附带本自治系统的地址空间,这也说明正确的源 IP 地址应该属于地址空间.为了完成这个功能,我们给 BGP 增加了一种消息类型,定义为选择通知(selection notice)

消息,简称 SN 消息.为了支持域间分组真实性验证,边界路由器记录收到的所有 SN 消息中的信息,称为路由选择信息(route selection information),简称 RSI.

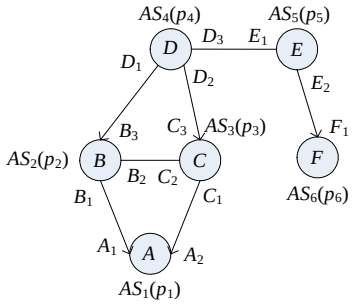


Fig.1 A simple AS connection topology

图 1 一个简单的自治系统连接拓扑

下面以 AS_2 中的路由器 B 为例,说明边界路由器获得路由选择信息的过程.在以下的讨论中,〈〉表示 BGP 消息(包括 SN 消息),Update 消息的格式是〈*p*,*nh*,[*asn*]〉,其中,*p* 表示路由目的网络的地址前缀,*nh* 表示路由的下一跳地址,[*asn*]表示路由的 *AS_PATH* 属性,是路由所经过的自治系统的 *AS* 号码组成的序列.*B* 收到的 Update 消息包括:AS_1 发送的〈p_1,A_1,[1]〉,AS_4 发送的〈p_1,D_1,[431]〉,〈p_3,D_1,[43]〉,〈p_4,D_1,[4]〉,〈p_5,D_1,[45]〉,〈p_6,D_1,[456]〉和 AS_3 发送的〈p_1,C_2,[31]〉,〈p_3,C_2,[3]〉.路由器 *B* 的路由判决过程选出到达每个目的网络的最优路由,存储在 BGP 路由表(BGP RIB)中,并向发送每条最优路由的邻居回送 SN 消息.SN 消息的格式是〈[*p*],[p_s]〉,其中,[*p*]表示目的网络地址前缀序列,[p_s]表示发送 SN 消息的自治系统的地址空间,序列中的每一项表示一个地址前缀.路由器 *B* 发送的 SN 消息包括:发送〈[p_1],[p_2]〉给 AS_1,发送〈[p_3],[p_2]〉给 AS_3,发送〈[p_4,p_5,p_5],[p_2]〉给 AS_4.

路由器 *B* 也会向邻居发送本网络的可达性信息,同时也会把选出的最优路由发送给邻居,但在发送前,会根据与邻居的关系类型应用路由输出策略.路由器 *B* 发送的 Update 消息包括:发送给 AS_1 的〈p_2,B_1,[2]〉,〈p_3,B_1,[23]〉,〈p_4,B_1,[24]〉,〈p_5,B_1,[245]〉,〈p_6,B_1,[2456]〉,发送给 AS_3 的〈p_1,B_2,[21]〉,发送给 AS_4 的〈p_1,B_3,[21]〉.如果邻居选择 *B* 发送的路由作为最优路由,会给 AS_2 发送 SN 消息.假设 AS_1 和 AS_4 都优先选择 AS_2 的路由,那么邻居发送给 AS_2 的 SN 消息包括:来自 AS_1 的〈[p_2,p_4,p_5,p_6],[p_1]〉,来自 AS_3 的〈[p_2],[p_3]〉,来自 AS_4 的〈[p_1,p_2],[p_4]〉.路由器 *B* 将这些 SN 消息的内容汇总成路由选择通知消息,其 BGP 路由表和路由选择信息见表 1,横列表项表示一条路由及其对应的路由选择信息,各项的含义是:*prefix* 表示路由的目的网络前缀;*next_hop* 表示路由的下一跳地址,在表中采用路由器接口表示,例如 A_1 表示路由器 *A* 接口 1 对应的 IP 地址;*as_path* 表示路由的 *AS_PATH* 属性;*RSI* 表示这条路由对应的路由选择信息.前缀 *p* 的路由选择信息包含 *asn*(p_s)表示路由器 *B* 会从 *AS* 号码为 *asn* 的邻居自治系统收到目的地址属于前缀 *p*,源地址属于前缀 p_s 的分组.例如,根据表 1 中的 *RSI*,如果路由器 *B* 收到的来自 AS_3 的分组的源地址不属于 p_3,就可以确定该分组是伪造的,因为 AS_3 只会向 *B* 发送目的地址属于 p_2、源地址属于 p_3 的分组.

Table 1 Route selection information on *B*

表 1 路由器 *B* 的路由选择信息

Prefix	*next_hop*	*as_path*	*RSI*	*RSI**
p_1	A_1	1	4(p_4)	4(p_4),4(p_5),4(p_6)
p_2		Local	1(p_1),3(p_3),4(p_4)	1(p_1),3(p_3),4(p_4),5(p_5),6(p_6)
p_3	C_2	3		
p_4	D_1	4	1(p_1)	1(p_1)
p_5	D_1	45	1(p_1)	1(p_1)
p_6	D_1	46	1(p_1)	1(p_1)

路由器 *D* 在选择了 *B* 发送来的路由后,会把路由进一步传播给 AS_5 和 AS_6.从 AS_5 和 AS_6 发出的目的地址属于 p_1 和 p_2,源地址属于 p_5 和 p_6 的分组经过路由器 *D* 转发给 *B*.如果 *B* 根据表 1 中的 *RSI* 执行分组真实性检查,

这些正常分组都会被过滤掉,原因在于 AS_5 在选择了路由后把 SN 消息发送给路由器 *D*,路由器 *B* 没有得到这些路由选择信息.可见,要保证路由器能够收集到所有路由选择信息,SN 消息需要在路由传播的同时一起反向传播.由此得到 SN 消息传播规律:自治系统 *x* 收到邻居发送来确认路由 *r* 的 SN 消息,如果路由 *r* 不是发源于本自治系统,那么,*x* 要将 SN 继续发送给路由 *r* 的下一跳自治系统.这里,下一跳自治系统是指发送 *r* 给 *x* 的自治系统.按照这个规律,路由器 *D* 把来自 AS_5 和 AS_6 的 SN 消息$\langle[p_1,p_2],p_5\rangle$,$\langle[p_1,p_2],p_6\rangle$发送给 AS_2.路由器 *B* 的路由选择信息在表 1 中 *RSI* 列的基础上进一步地扩展,*prefix* 为 p_1 的表项的路由选择信息扩展为 $4(p_4),4(p_5),4(p_6)$,*prefix* 为 p_2 的表项的路由选择信息扩展为 $1(p_1),3(p_3),4(p_4),4(p_5),4(p_6)$,最终结果为 RSI^*列.

2.3 有效性证明

本节证明根据 BGP 扩展提供的路由选择信息,路由器能够完成域间的最大过滤和半最大过滤.假设节点 *u* 是 *V* 中的任意一个节点,节点 *u* 的边集表示为 $E_u=\{e_u^1,e_u^2,...,e_u^k\}$,节点 *u* 的路由选择信息表示为 $\Omega_u=\bigcup_{i=1}^{k}\Omega_{e_u^i}$,其中,$\Omega_{e_u^i}$ 表示从边 e_u^i 收到的所有 SN 消息中的路由选择信息,不产生歧义的情况下也表示 SN 消息的集合.对任意的 SN 消息 $\omega\in\Omega_{e_u^i}$,s_ω表示发送 ω的节点,r_ω表示 s_ω节点选择的路由,d_ω表示发源路由 r_ω的节点.节点 *u* 根据Ω_u能够构造由节点对组成的过滤规则 $\tilde{C}_{e_u^i}=\bigcup_{\omega\in\Omega_{e_u^i}}\{(d_\omega,s_\omega)\}$.

假设节点 *u* 收到来自边 e_u^i 的分组 *P*(*d*,*s*),如果节点对 $(d,s)\in\tilde{C}_{e_u^i}$,那么 *P*(*d*,*s*)经过 e_u^i 符合路由选择形成的约束.这是因为节点对 $(d,s)\in\tilde{C}_{e_u^i}$ 表示存在 SN 消息 $\omega\in\Omega_{e_u^i}$,也就是说,节点 *s* 收到 1 条或者多条到达节点 *d* 的路由后,选择了其中经过边 e_u^i 的路由,即 $e_u^i\in R(d,s)$.因此,*P*(*d*,*s*)应被认为是真实地址的分组.

假设节点 *u* 收到来自边 e_u^i 的分组 *P*(*d*,*s*),如果节点对 $(d,s)\notin\tilde{C}_{e_u^i}$,那么 *P*(*d*,*s*)经过 e_u^i 不符合路由选择形成的约束.根据上一节介绍的路由选择通知机制和 SN 消息传播规律,如果节点选择了新路由将发送一个 SN 消息,如果节点收到一个 SN 消息并判断其响应的路由不是源自本节点,将转发该 SN 消息.因此,节点 *s* 选择经过 e_u^i 传播到达 *d* 的路由后发送的 SN 消息 ω一定会经过 e_u^i 传递给 *u*,于是总有 $\omega\in\Omega_{e_u^i}$.但是,$(d,s)\notin\tilde{C}_{e_u^i}$ 与 $\omega\in\Omega_{e_u^i}$ 相矛盾,由此可以判断 *P*(*d*,*s*)是使用伪造地址的分组.

综上所述,节点 *u* 能够根据 $\tilde{C}_{e_u^i}$ 实现最大过滤器 $\tilde{F}$:假设节点 *u* 从边 e_u^i 收到分组 *P*(*d*,*s*),如果节点对 $(d,s)\in\tilde{C}_{e_u^i}$,那么 *u* 可以判断 $e_u^i\in R(d,s)$ 并转发分组,否则,可以判断 *P*(*d*,*s*)是伪造分组并丢弃.根据 $\tilde{C}_{e_u^i}$ 还可以构造 $\hat{C}_{e_u^i}=\{s\mid(d,s)\in\tilde{C}_{e_u^i},d\in V,s\in V\}$,作为半最大过滤器 $\hat{F}$ 的过滤规则:假设节点 *u* 从 e_u^i 收到分组 *P*(*d*,*s*),如果节点 $s\in\hat{C}_{e_u^i}$,则转发分组,否则将分组丢弃.执行最大过滤器和半最大过滤器所需的最小信息分别是 $\tilde{C}_{e_u^i}$ 和 $\hat{C}_{e_u^i}$,由于 $\tilde{C}_{e_u^i}>>\hat{C}_{e_u^i}$,可见最大过滤器的存储要求也远大于半最大过滤器.

3 BGP 扩展设计

为了支持发送,处理 SN 消息和存储路由选择信息,需要扩展现有的 BGP.扩展设计的目标包括:不影响 BGP 交换路由信息和降低路由收敛速度;对 BGP 路由稳定性没有负面影响;尽量少的通信、存储和处理开销.

3.1 SN消息

SN 消息的格式如图 2 所示,包括 5 个数据域.Select 域长度为 1 字节,可取值有两个:*Select*=1 表示发送者选择了到达 Destination prefix 的路由;*Select*=0 表示发送者取消了到达 Destination prefix 的路由.Destination prefix length 域长度为 2 字节,表示 Destination prefix 域的字节长度.Destination prefix 是一个变长域,其中的内容是被选择或者被取消路由的地址前缀,每个前缀由二元组$\langle$*length*,*prefix*$\rangle$表示.Source prefix length 域长度为 2 字节,表示 Source prefix 域的字节长度.Source prefix 是变长域,其中的内容是发送者所在自治系统的地址空间,由一个

或者多个地址前缀组成,地址前缀由二元组⟨*length*,*prefix*⟩表示.

根据 SN 消息传播过滤可知,SN 消息沿着 Update 消息相反的路径传播.边界路由器 *A* 收到来自邻居 *B* 的 SN 消息后,首先检查是否发送过到达 Destination prefix 的路由给 *B*,如果 *A* 没有发送过到达 Destination prefix 的路由给 *B*,则可以判断出现错误,*A* 向 *B* 发送 Notification 消息通知 SN 消息错误.如果 *A* 向 *B* 发送过到达 Destination prefix 的路由,*A* 根据 Select 域的值将消息中的信息加入路由选择消息(Select 值为 1)或者从中删除(Select 值为 0),并根据更新的路由选择信息生成新的分组过滤规则.如果消息中的 Destination prefix 不是本地自治系统的地址空间,边界路由器要把该 SN 消息进一步发送给下一跳自治系统.

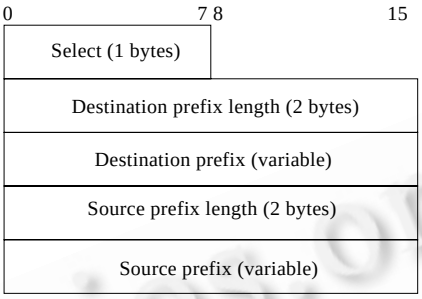


Fig.2 SN message Format

图 2 路由选择通知消息的格式

3.2 BGP能力协商

新的 BGP 功能在网络中部署是渐进的过程,路由选择通知在网络中部分部署的情况下,部署和未部署的自治系统间需要协调机制.另一方面,在某些情况下,自治系统间并不需要使用路由选择通知,例如只接入一个 ISP,地址空间由 ISP 分配的 Stub 自治系统,由于所有分组都从与 ISP 相连的链路上进入,没有执行分组过滤的必要,因此与 ISP 不需要发送 SN 消息给 Stub 自治系统.能力协商(capabilities negotiation)机制[8]为 BGP 增加新特性提供了很大的扩展能力.为了支持自治系统灵活配置路由选择通知功能,需要扩展 BGP 的能力协商功能,使 BGP 建立会话时决定是否向邻居发送 SN 消息.当属于两个自治系统的边界路由器 *A*,*B* 建立 BGP 会话时,能力协商的过程是:如果 *A* 需要 *B* 发送 SN 消息给自己,那么 *A* 发送给 *B* 的 Open 消息中加入标识路由选择通知的可选参数,否则,Open 消息中不加这个参数;如果 *B* 同意发送,则发送 KeepAlive 消息确认 *A* 的请求,否则发送 Notification 消息,其中的错误代码指出不支持路由选择通知.*B* 到 *A* 方向上的过程与此相同.所以,BGP 会话建立之后,发送 SN 消息的情况有 4 种:(1) *A*,*B* 互相发送;(2) *A* 向 *B* 发送,*B* 不向 *A* 发送;(3) *B* 向 *A* 发送,*A* 不向 *B* 发送;(4) *A* 和 *B* 互不发送.BGP 能力协商的支持,使自治系统在使用路由选择通知功能时有很大的灵活性,以适应自治系统间的不同关系.

3.3 体系结构

扩展了路由选择通知功能的 BGP 消息处理流程及其与数据平面的关系如图 3 所示.实线包围的部分是控制平面的处理过程,Update 消息和 SN 消息并行处理.Update 消息的处理过程与普通 BGP 基本相同,区别包括两个方面:(1) 选择新路由后,向邻居发送 Update 消息的同时给路由发送者回送 SN 消息;(2) 删除一条路由时,将该路由对应的路由选择信息删除.Announcement 和 Withdrawal 引起的操作不同:如果 Withdrawal 取消了 BGP 路由表中的路由后有新路由被选出,给新路由的发送者回送一个 SN 消息(Select 值为 1);如果 Announcement 中的新路由被选择,则给发送者回送一个 SN 消息(Select 值为 1);如果 Announcement 中的路由替换了 BGP 路由表中的路由,则给发送者回送一个 SN 消息(Select 值为 1),并给被替换路由的发送者发送 SN 消息(Select 值为 0).如果边界路由器执行域间分组过滤,则在生成分组转发表的同时,根据路由选择信息生成分组正确性检查规则.

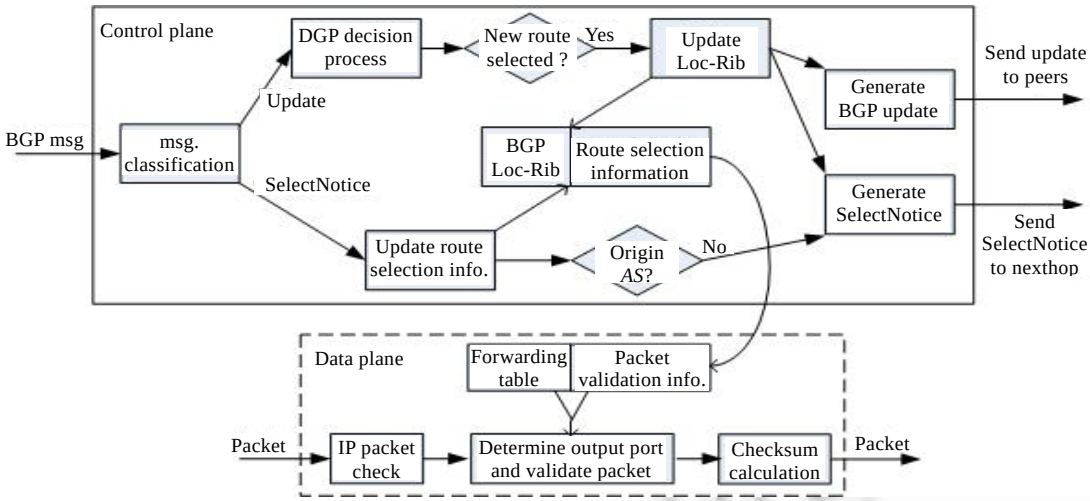


Fig.3 Processing on SN and Update message of BGP with route selection notice and its support to packet filtering

图 3 扩展路由选择通知功能的 BGP 对 SN 和 Update 消息处理过程以及对分组过滤的支持

3.4 路由选择时钟

根据 SN 消息传播规律,自治系统不仅要发送本地生成的 SN 消息给路由的发送者,还需要转发来自下游节点的 SN 消息,如果能够把来自下游自治系统的 SN 消息聚集为一个消息,就会减少上游自治系统需要传输和处理的 SN 消息的数量.例如在图 1 中,为了响应源自 AS_1 的路由,如果不做聚集,那么总共需要发送 11 个 SN 消息,而如果采用聚集的方法,那么最好情况是只需发送 5 个 SN 消息.为了聚集 SN 消息,减少通信开销,我们设计了路由选择时钟,其工作原理是:当节点选出新路由并向邻居传播后,并不马上给路由的发送者回送 SN 消息,而是启动一个时钟 T_s,我们称 T_s 为路由选择时钟,等待下游自治系统可能发送来的 SN 消息;如果收到来自下游自治系统的 SN 消息时,时钟 T_s 还没有超时,就将 SN 消息中的信息暂存起来,而不是根据 SN 消息传播规律向下一跳自治系统发送 SN 消息;直到时钟 T_s 超时或者收到了所有邻居发送来的 SN 消息,才向下一跳自治系统发送一个 SN 消息,其中聚集了所有收到 SN 消息中的信息.

时钟 T_s 值的设置主要基于以下两点考虑:

首先,到达某个目的前缀的路由可能会沿着多条路径传播到达某个节点,节点只需对从中选出的最优路由发送 SN 消息,因此,多条路由到达某个节点的先后顺序对发送的 SN 消息数量有一定影响.如果等待多条路由都到达节点,从中选出最佳路由后再发送 SN 消息,则只需要发送一个 SN 消息.因此,路由选择时钟的第 1 个作用是等待路由稳定后再发送 SN 消息.

其次,路由选择时钟使节点等待下游节点的 SN 消息完成路由选择信息的聚集功能.但存在的困难在于:上游节点不知道下游节点是否会选择路由以及如果下游节点选择了路由,如何预测 SN 消息传送回来的延时.为了能够捎带更多的 SN 消息,路由选择时钟应该依据可能最晚到达的 SN 消息来设置.这种设置方法不会影响路由选择信息的收敛速度,因为路由选择信息的收敛时间取决于最晚传送的 SN 消息.下一节的模拟部分会介绍一种具体的设置 T_s 的方法.实验结果表明,合理设计 T_s 能够有效地减少传递路由选择通知信息所需的通信开销.

4 模拟实验

我们在 SSFNet 模拟器[21]的 BGP 协议中实现了 BGP 路由选择通知的扩展,并做了模拟实验,目的是考察路由选择通知带来的通信开销和扩展对 BGP 本身功能的影响.同时也研究如何设置路由选择时钟 T_s 以平衡通信开销和路由选择信息的收敛时间.在模拟中,我们采用了两种规模的网络拓扑:29 个自治系统的 Net29 和 110 个

自治系统的 Net110.这两种网络拓扑都是基于互联网路由表生成[22].在两种拓扑中,每个自治系统都只由一个边界路由器构成.实验中,路由器对 BGP 消息(包括 SN 消息)的处理延时在 0.000 001s~0.000 005s 间均匀分布,路由器间的传输延时设为 0.000 1s,连续发送 Update 消息的时间间隔 MRAI 设为 30s.

4.1 对BGP收敛的影响

BGP 路由收敛延时对分组正确转发和互联网的通信质量有很重要的影响.我们比较了在扩展前后 BGP 路由的收敛延时和收敛过程中 Update 消息数量的变化,以此作为评价扩展路由选择通知对 BGP 收敛的影响.实验中比较了 3 种情况的数据结果:

(1) 没有经过扩展的标准 BGP;

(2) 扩展了路由选择通知功能但没有使用路由选择时钟 T_s 的 BGP;

(3) 扩展路由选择通知功能并且使用路由选择时钟 T_s 的 BGP.

模拟结果显示,3 种情况下路由的收敛延时及期间发送的 Update 消息数量完全相同,这表明 BGP 的路由选择通知扩展和路由选择时钟 T_s 对 BGP 的路由功能没有任何影响.从图 3 中可以看出,SN 消息的处理与 Update 消息的处理是并行的,对路由收敛没有影响,因此这个结果是合理的.需要指出的是,实际中,SN 消息的传输和处理以及路由选择时钟都会给路由器带来额外的处理器开销,负载的增加可能会对 BGP 其他方面的处理产生微小影响.

4.2 带宽开销

我们采用比较 SN 消息的数量和 Update 消息的数量来评价路由选择通知的通信开销,如果 SN 消息数量小于 Update 消息数量或者相当,那么扩展带来的额外通信开销是可以接受的.实验中记录了一个路由变化引起的 Update 消息和 SN 消息数量.

图 4 中前 3 个柱型和后 3 个柱型分别是 net29 和 net110 的实验结果,其中,SelNot1 表示不使用路由选择时钟情况下发送的 SN 消息的数量,SelNot2 表示使用路由选择时钟 T_s(T_1=0.025s,t=0.00055s)发送的 SN 消息数量,T_1 和 t 的含义见第 4.3 节.可以看出,SN 消息数量小于 Update 消息数量的水平,而且路由选择时钟进一步减少了 SN 的通信开销.SN 消息的结构比 Update 消息的结构简单,因此扩展路由选择通知不会带来显著增加的通信开销.

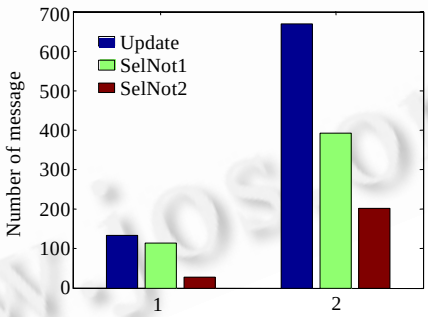


Fig.4 Communication overhead comparison between SN message and Update message

图 4 SN 消息与 Update 消息通信开销比较

4.3 路由选择时钟

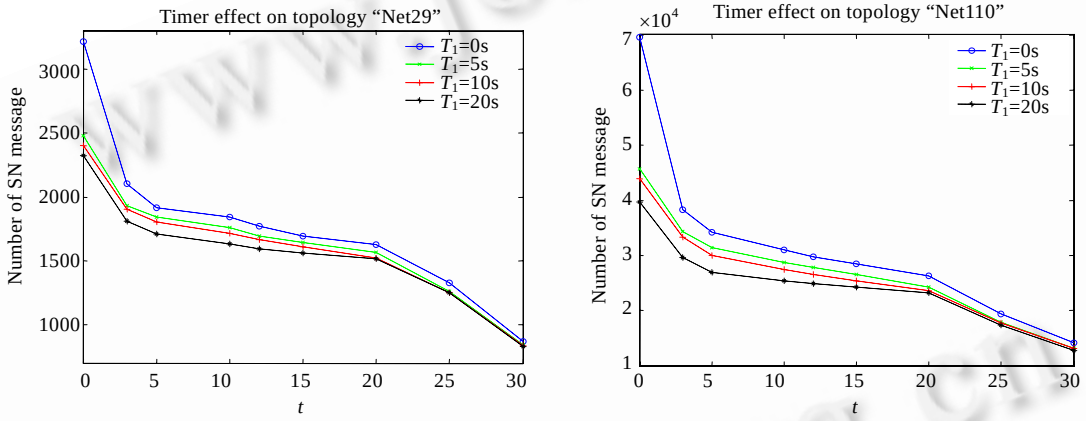
从上面的实验可以看出,路由选择时钟对减少路由选择通知扩展带来的通信开销有很大的作用.根据路由选择时钟的工作原理,可以有多种具体实现.我们在模拟实验中设计的路由选择时钟 $T_s=t\times(L_{\max}-L_{as\_path})$,其中,$L_{\max}$ 表示该节点 BGP 路由表中路由的最大路径长度,L_{as_path} 表示被响应的路由中 *AS_PATH* 属性的长度.为了控制域间流量,有些自治系统会在路由的 *AS_APTH* 属性中重复填充多个自治系统号码(AS prepending).这些重复

自治系统号码都不计入 $L_{\max}$ 和 L_{as_path}.从节点的角度看,$L_{\max}$ 可以作为一条路由在网络中最大传播跳数的估计值.由此,一条已经传递了 L_{as_path} 跳的路由被进一步传递的最大跳数就可以用($L_{\max}$–L_{as_path})估计.邻居自治系统间 Update 消息和 SN 消息的传输和处理延时之和是随机的,我们用 t 来估计.当不使用 MRAI 时,t 的估计比较简单,t^*=链路传输延时×2+BGP 消息处理延时的平均值×2,但在使用 MRAI 的情况下,几乎无法判断估计节点对发送的 Update 消息是否应用 MRAI.因此,MRAI 会使 t 的估计非常困难.在后面的实验结果中可以看到,取得良好聚集效果的 t 值往往大于 t^*,这就是由于 MRAI 所发挥的作用.

综上,发送 SN 消息的等待时间由两部分组成:T_1 等待节点路由稳定的延时,T_2=t×($L_{\max}$–L_{as_path})等待下游节点回送 SN 消息的延时.

路由选择时钟与 BGP 的 MRAI 时钟的作用相似,都是为了减小 BGP 消息的通信开销.但是,MRAI 时钟一般设为固定值,默认为 30s;而路由选择时钟为了捎带更多的信息,计算更加复杂.MRAI 会增大 BGP 路由的收敛延时,路由选择信息的收敛作为 BGP 路由收敛的逆过程,也会受 MRAI 的影响而增大.

路由选择时钟 T_s 对路由选择信息收敛的影响如图 5 所示,在实验中,MRAI 的值设为默认值 30s.可以看出,T_1 对减少通信开销的作用并不明显,只有在 t<5s 的情况下,T_1 才会起到作用.另一方面,在 t<5s 时,随着 t 的增大,SN 消息数量很快减少,而当 t>5s 时,随着 t 的增大,SN 消息数量变化缓慢.在 t=30 秒时,SN 的消息数量减小到最低水平,这时,路由选择时钟最大程度地发挥了捎带功能.因此,T_s 的设置主要在于参数 t 的选择,MRAI 应该是参数 t 的上限值.



(a) The simulation result on Net29
(a) Net29 上的模拟实验结果

(b) The simulation result on Net110
(b) Net110 上的模拟实验结果

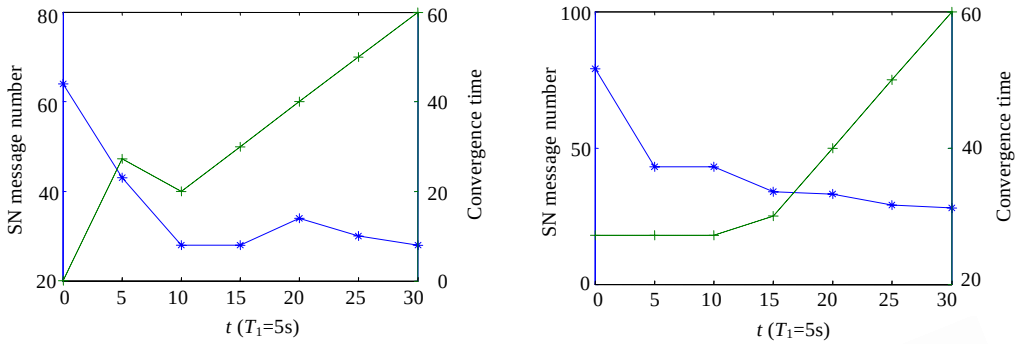
Fig.5 Effect of route selection timer on SN message communication overhead.

图 5 路由选择时钟对 SN 消息通信开销的影响

4.4 路由选择信息的收敛延时

T_s 减少了 SN 消息的数量,同时也带来了一些问题:给路由器带来了额外的存储和处理开销;使路由选择信息的收敛速度变慢.

通过模拟实验,我们观察了通信开销和收敛延时的关系.图 6 显示了 T_1=5s 和 T_1=10s 时,SN 消息数量和收敛时间随参数 t 的变化情况.在实验中,我们记录了 Net29 中一次路由变化引起的 SN 消息数量和路由选择信息的收敛时间.随着 t 的增加,收敛时间增大,SN 消息的数量减少.当 t<10s 时,收敛延时增加缓慢而 SN 消息的数量减少很快,而当 t>15s 时,收敛延时增加很快,而 SN 消息数量基本保持稳定.可见,将参数 t 的值设为 15s 左右能够达到收敛延时和通信开销综合最优的效果.



(a) T_1=5s (b) T_1=10s

Fig.6 Relationship between SN message communication overhead and RSI convergence delay

图 6 SN 消息的通信开销与路由选择信息收敛延时间的关系

5 结束语

可信任是基于 IPv6 的下一代互联网的重要特征,是容纳更多的客户、支持更多应用的需求.路由系统提供真实 IP 地址服务是可信任互联网的核心特征,在自治系统间过滤伪造源地址的分组是实现真实 IP 地址访问的重要一步.本文提出通过扩展 BGP 为域间分布式分组过滤提供支持,使路由系统不仅正确转发分组,而且保证分组来自真实的自治系统.扩展 BGP 的方法利用了 BGP 灵活的扩展机制,能够大幅度降低设计的复杂度.模拟实验表明,BGP 的路由选择通知扩展,不会影响 BGP 的路由功能,能够及时、有效地为域间分布式分组过滤提供支持.

路由选择通知与数据平面具体的分组过滤操作分离,边界路由器是否执行分组过滤由网络管理员决定.BGP 能力协商使路由选择通知扩展有很强的灵活性,能够适应复杂的网络连接和自治系统间关系.路由选择信息的传输和存储会增加边界路由器的存储和处理开销,基于 IPv6 的下一代互联网,BGP 路由能够高度聚合,将减少路由选择通知带来的开销.

随着互联网的进一步发展,更多的自治系统采用 peer-to-peer 和 multihoming 的连接方式.互联网连接拓扑特性的变化对域间分布式分组过滤效果以及部署方案都会产生一定的影响.域间的伪造分组过滤只能保证分组来自真实的自治系统,需要进一步提出解决保证域内 IP 地址真实性的方法,以形成完整的可信任网络体系结构.这两个方面将是我们下一步的工作方向.

致谢 在此,我们向对本文的工作给予支持的老师、对论文提出批评和建议的编辑同志和审稿老师表示感谢.

References:

[1] Park K, Lee H. On the effectiveness of route-based packet filtering for distributed DoS attack prevention in power-law Internets. Proc. of ACM SIGCOMM, 2001,31(4):15−26.

[2] Rekhter Y, Li T. A border gateway protocol 4 (BGP-4). RFC 1771, 1995.

[3] Labovitz C, Malan GR, Jahanian F. Internet routing instability. IEEE/ACM Trans. on Networking, 1998,6(5):515−527.

[4] Griffin TG, Shepherd FB, Wilfong G. The stable paths problem and interdomain routing. IEEE/ACM Trans. on Networking, 2002, 10(2):232−243.

[5] Labovitz C, Ahuja A, Bose A, Jahanian F. Delayed Internet routing convergence. IEEE/ACM Trans. on Networking, 2000,9(3): 293−306.

[6] Gao L, Rexford J. Stable Internet routing without global coordination. IEEE/ACM Trans. on Networking, 2001,9(6):681−692.

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

[8] Chen E. Route refresh capability for BGP-4. RFC 2918, 2000.

[9] 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.

[10] Chandrashekar J, Duan Z, Zhang ZL. Limiting path exploration in BGP. In: Proc. of the IEEE INFOCOM, Vol.4. IEEE Press, 2005. 2337−2348.

[11] Chandra R, Scudder J. Capability advertisement with BGP-4. RFC3392, 2002.

[12] Bates T, Rekhter Y, Chandra R, Katz D. Multiprotocol extensions for BGP-4. RFC 2858, 2000.

[13] Savage S, Wetherall D, Karlin A, Anderson T. Practical network support for IP traceback. Computer Communication Review, 2000, 30(4):295−306.

[14] Snoeren AC, Partridge C, Sanchez LA, Jones CE, Tchakountio F, Kent ST, Strayer WT. Single-Packet IP traceback. IEEE/ACM Trans. on Networking, 2002,10(6):721−734.

[15] Baker F. Requirements for IP version 4 routers. RFC 1812, 1995.

[16] Ferguson P, Senie D. Network ingress filtering: Defeating denial of service attacks which employ IP source address spoofing. RFC 2827, 1998.

[17] Bremler-Barr A, Levy H. Spoofing prevention method. In: Proc. of the IEEE INFOCOM. IEEE Press, 2005. 536−547.

[18] Minimum vertex cover. 2006. http://www.nada.kth.se/~viggo/wwwcompendium/node10.html

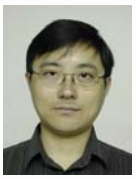
[19] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the Internet topology. Computer Communication Review, 1999,29(4):251−262.

[20] Li J, Mirkovic J, Wang M, Reiher M, Zhang L. SAVE: Source address validity enforcement protocol. In: Proc. of the IEEE INFOCOM, Vol.3. New York: IEEE Inc., 2002. 1557−1566.

[21] The SSFnet project. 2006. http://www.ssfnet.org/homepage.html

[22] Multi-AS topologies from BGP routing tables. 2006. http://www.ssfnet.org/Exchange/gallery/asgraph/index.html

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



徐恪(1974－),男,博士,副教授,CCF 高级会员,主要研究领域为路由器软件体系结构.

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体系结构,网络协议测试.