MFT²-BGP:基于多转发树的无中断域间路由协议

胡乔林[1)] 彭 伟[2)] 陈 新[1)] 苏金树[2)]

[1)](空军预警学院五系 武汉 430019)

[2)](国防科技大学计算机学院 长沙 410073)

摘 要 BGP通过触发全局、反应式收敛应对链路或节点失效引起的拓扑变化,然而BGP协议收敛时间长、收敛过程中的瞬时失效严重降低了数据平面转发性能,难以支持关键业务流量.该文提出了容忍失效的MFT²-BGP,通过利用路径标识符以较低的消息开销构造符合BGP策略的多转发树,使得每个AS获得多样性路径,当出现瞬时失效时,在不改变协议动态性的情况下,允许节点动态切换报文转发路径以实现无中断报文转发,通过嵌入"失效根源信息"以降低收敛时间,抑制瞬时失效以降低路由系统的扰动.通过在Internet-like拓扑上的大量实验表明,在链路失效场景中与其它协议相比,MFT²-BGP能有效改善收敛时间,降低转发中断时间,改善路由系统稳定性.

关键词 域间路由;瞬时失效;多转发树;多路径;快速恢复

中图法分类号 TP393 **DOI号**: 10.3724/SP.J.1016.2012.02023

MFT²-BGP: Achieving Disruption-Free Inter-Domain Routing Protocol Using Multiple Forwarding Trees

HU Qiao-Lin[1)] PENG Wei[2)] CHEN Xin[1)] SU Jin-Shu[2)]

[1)](*Department Five, Air Force Radar Academy, Wuhan* 430019)

[2)](*School of Computer, National University of Defense Technology, Changsha* 410073)

Abstract BGP triggers global and reactive routing convergence to dynamically adapt to network topology and policy changes such as link or node failures. Nevertheless, BGP faces the challenge of long convergence time, transient failure during the routing convergence, which affects the forwarding performance so as to hard to support mission critical traffic. This paper presents the protocol for achieving failure tolerant inter-domain routing protocol named MFT2-BGP, it constructs policy-compliant multiple forwarding trees using path identifier with low message and process overhead, which make every AS discovery diversity path, Every AS can switch traffic to backup path dynamically to achieve disruption-free forwarding without comprising the routing behavior in the presence of transient failure scenario. MFT²-BGP reduces the convergence time by carrying "root cause notification" in update messages and reduces the churn of routing system by suppressing non-necessary routing updates. Detailed experiments on internet-like topology suggest that MFT²-BGP reduces convergence time and forwarding disruption time in link failure scenarios, enhances the routing stability than other protocols.

Keywords inter-domain routing; transient failure; multiple forwarding trees; multiple path; fast recovery

收稿日期:2012-06-30;最终修改稿收到日期:2012-08-09.本课题得到国家自然科学基金(61070199)、国家"九七三"重点基础研究发展规划项目基金(2009CB320503)资助.**胡乔林**,男,1979年生,博士,讲师,主要研究方向为容迟容断网络、网络生存性和虚拟网络.E-mail:huqiaolin@nudt.edu.cn.**彭 伟**,男,1974年生,博士,副研究员,主要研究方向为路由协动、移动网络.**陈 新**,男,1981年生,博士,主要研究方向为网络安全.**苏金树**,男,1962年生,博士,教授,博士生导师,主要研究领域为计算机网络、信息安全.

1 引 言

Internet 已经发展成为重要的通信基础设施，承载了许多如 VoIP、在线游戏、远程医疗等对延迟及中断敏感的关键业务，然而 Internet 却容易遭到破坏，例如软硬件故障、误配置、攻击等因素都会造成网络失效. 通过对 Sprint 骨干网链路故障进行分析[1]，人们发现失效几乎每天都会发生，且 90％属于瞬时性故障. 理想的路由协议应保证只要底层拓扑连接就能提供持续通信. 然而 BGP 协议[2]缺乏生存性保证，在失效时仅触发全局、反应式收敛应对拓扑变化，相对于 IGP 秒级的收敛，BGP 由于大量“路径探索”[3]以及 MRAI 和 WRATE 定时器的限制，导致收敛时间甚至达到 30 分钟[4]，这进一步恶化了转发中断. Katz-Bassett 等人[5]发现即使多宿主 AS 并不总能实现容错，BGP 导致大量多宿主 AS 经历不可达事件，持续时间甚至长达数小时或天，其数量与持续时间都远超预期，严重降低数据平面的转发性能. Wang[6]通过对顶级 ISP 进行测量，发现 BGP 收敛期间的动态性造成大量路由器经历瞬时路由失效、环路，即使在失效切换和链路恢复时路由黑洞也会造成数十秒的报文丢失.

可达性是路由协议的顶级目标，研究者通过加速收敛以降低失效对转发中断的影响，包括 Ghost Flushing[7]、EPIC[8]等通过抑制或嵌入“失效根源”信息以快速作废非法路径，但 Ghost Flushing、EPIC 在 Internet-like 拓扑下，其加速收敛效果并不明显. 而且由于多数是瞬时失效，过快收敛反而给路由器增加负担，带来路由的不稳定性. 加速收敛并不能满足关键应用需求，也不能消除收敛过程对转发中断的影响. 一些研究工作集中于扩展 BGP，比如 MIRO[9]、R-BGP[10]、Splicing[11]、Consensus[12]、YAMR[13]等采用预计算备份路径的方法，不同程度上暴露底层拓扑冗余路径，但这些协议 AS 之间的备份路径缺乏协调可能形成环路，也并不能保证每个 AS 获得多样性路径，且部分协议难以实现完全无中断转发.

针对抑制瞬时失效、无中断报文转发、低开销的需求，本文提出了通过构造多转发树容忍失效的域间路由协议(Multiple Forwarding Tree and Failure Tolerant BGP，MFT^2-BGP)以实现快速路由恢复，其主要面临如下挑战：(1)在标准 BGP 中，AS 节点无全局拓扑，使得每个 AS 难以获得、区分多样性路径；(2)在构造多转发树时要求降低收敛时间，并保证处理开销在合理范围内；(3)保证报文在多转发树之间切换时不会形成环路. 为此，MFT^2-BGP 设计主要包括 3 个关键部分，即控制平面的多转发树构造、加速收敛与瞬时失效抑制、数据平面无中断报文转发算法. 该协议的特点在于：(1)充分利用底层冗余拓扑，提供可保证的结构化多样性路径，避免了备份路径之间可能的环路；(2)抑制瞬时失效，避免失效全局可视化，增强了 BGP 系统的稳定性；(3)从数据平面保护流量，采取本地先验式的方法处理瞬时失效引起的报文丢失，支持关键业务的应用；(4)具有可扩展性、兼容性，便于增量部署.

本文第 2 节介绍相关工作；第 3 节描述问题模型；第 4 节设计 MFT^2-BGP 协议框架以及控制平面的多转发树构造、加速收敛、失效抑制以及报文转发算法；第 5 节对协议的正确性进行分析；第 6 节通过模拟验证 MFT^2-BGP 的收敛性以及无中断转发性能；最后对本文工作进行总结.

2 相关工作

当前大部分研究工作主要利用冗余路径以应对出现的失效，IETF RTGWG 工作组致力于 IGP 协议的 IPFRR(IP 快速重路由)，比如 Deflections[14]等，其主要思想是为可能出现的链路、节点失效预计算备份路径. Medard 等人[15]最早提出了冗余树的方法以容忍单链路失效，Kini 等人[16]通过构造多冗余树以容忍两条链路失效，然而在图论中计算冗余路径的经典算法都需要获得全局一致性拓扑信息，而域间路由系统中每个 AS 仅能获得局部拓扑信息，无法进行有效集中式计算；其次由于 Internet 的 AS 数量大，集中式计算复杂度非常高，即使是分布式方法也由于协议复杂度过高而难以实现；最后由于 BGP 协议需要保证策略实现，进一步限制了可达性，因此许多 IGP 中的方法难以直接应用于 BGP 中.

现有 BGP 协议缺乏对多路径的考虑，缺乏发现、使用多路径的能力，大量研究表明多路径可增强网络容量、可靠性. MIRO[9]通过复杂的协商机制获得多路径，依赖于配置通告备份路径的策略，是一种无结构路径的方法，MIRO 采用隧道机制区分那些需要在备用路径上传输的报文，并且没有考虑快速路由恢复. R-BGP[10]为 AS 通告一条与最佳路径不相交的 AS 级备份路径以提高可靠性，这依赖于大量手工配置通告备份路径的策略，是一种无结构路径的方法，信令机制较复杂，且并不能完全保证获得多样性路径，且需要特殊报文封装机制进行转发.

Consensus[12]路由协议借鉴了分布式系统中快照实现一致性的算法，将路由计算集中于顶级 AS，这却违反了 AS 自治性，当链路失效造成报文不能正常转发时进入瞬时模式，采用偏转(Deflection)、回退(Backtracking)、备份(Backup)机制发起本地重路由. Splicing[11]利用 BGP 路由器中已有的多条路径构造多路径转发. YAMR[13]为每个 AS 最佳路径的每一跳都获得冗余标签路径，其平均 RIB/FIB 存储平均至少增加 4.6 倍，且部分冗余路径对于快速恢复难以发挥作用. ACF[17]在路由不一致的情况下采用检测并恢复的方法，通过在报文头中添加路径踪迹域、黑名单域等辅助路由决策和转发，较大程度增加了报文头开销，可能形成违反 AS 策略的转发路径. Add-Path[18]通告多路径，需要全网路由器更新，且增加了路由震荡的风险.

上述针对 BGP 的多路径协议主要通过修改策略使其在控制平面通告多路径，这需要复杂的信令、决策机制等，是一种无结构的方法，并没有对 BGP 协议进行系统化的设计，对于多链路失效的转发缺乏考虑；且这些协议对于路径多样性暴露程度并没有确定性的保证，且基本上没有考虑抑制瞬时失效的方法，对于多路径协议的收敛期间的动态行为缺乏考虑. 对于降低协议开销、保证无中断转发、抑制瞬时失效这 3 个相互影响的问题，已有工作并不能完全解决. MFT²-BGP 通过为目标前缀构造冗余树，每个 AS 自动获得结构化的多样性路径，而无需大量人工配置复杂策略，可以实现负载均衡、降低 BGP 扰动，面临失效时动态切换路径以实现快速路由恢复，在不影响报文转发的情况下抑制瞬时失效，增强 BGP 路由系统的稳定性.

3　网络模型和问题分析

3.1　网络模型

MFT²-BGP 做出如下设定(其中设定(1)是为了简化表述，这是当前研究普遍采用的设置，不影响协议的正确性，而(2)、(3)是研究 BGP[2]的基本设置)：

(1) 每个 ASv 仅由单边界网关路由器节点组成(不考虑 iBGP)，v 对每个$u\in Peers(v)$($Peers(v)$为v的邻居节点)的导入、导出路径分别保存在 rib_in($v\leftarrow u$)、rib_out($v\rightarrow u$)中，loc_rib(v)保存本地最佳路径；

(2) AS 的导入导出策略服从“Gao-Rexford”原则[19]，即仅出现“valley-free”的路径；

(3) AS 间建立 eBGP 会话，路由更新消息服从 SPVP[20]模型.

3.2　问题的提出与分析

在给定拓扑下网络容忍失效的性能主要取决于拓扑暴露程度、协议响应性两个因素. BGP 协议在这两个方面却存在不足. 虽然 BGP 将快速响应性置于一致性之上，但是却带来了大量环路以及路由黑洞，直接影响转发性能；同时，尽管 Internet 底层路径具有充分冗余[21]，AS 无可替代路径的不到 5%[9]，BGP 协议对于底层路径暴露程度不足，缺乏发现、区分和使用多路径的能力，多宿主 AS 也仅能控制报文出口，而仍然缺乏容错性[5]，这主要是由于 BGP 缺乏故障隔离能力、仅通告并使用单路径、协议的动态性造成的瞬时性不可达、环路(将其统称为瞬时路由失效)[22]，从而造成了网络可用性相对较低，而一些多路径 BGP 协议却增加了环路发生的可能性[18]，比如 R-BGP 为当前下一跳通告备份路径时，如图 1 中 AS2 为下一跳 AS1 通告备份路径(2-0)，当 AS1-AS0 和 AS2-AS0 多链路同时失效时，由于 AS1 和 AS2 独立决策，将会在造成备份路径之间形成环路，甚至某些单链路特殊场景下也可能造成环路. Splicing[11]仅利用已有路径，却没有考虑如何获得多样性路径，比如 AS0-AS3 断开时 Splicing 仍将产生失效. 合理发现、利用多样性路径是解决这些问题的关键. 基于以上观察，在设计 MFT²-BGP 时对 SPVP[20]定义进行扩展以发现、区分多路径；在多路径开销和性能之间折衷，确保协议的收敛性和正确性.

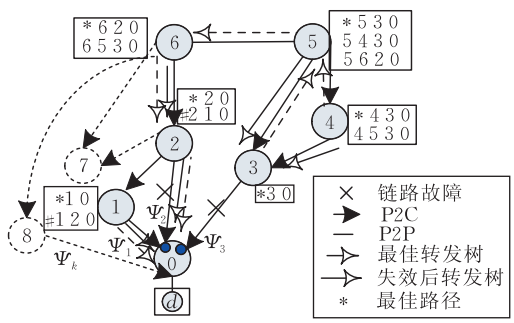


图 1　链路故障导致 BGP 瞬时失效

定义 1.　合法路径. BGP 中合法路径必须服从策略、能够到达 d(即不含失效链路/节点)；其次路径 $P=\pi(u,d)$ 是合法的必须满足：$\forall i,n\geq i\geq 1$，$\pi(u,d)=\pi(u,v_i)+\pi_u(u\in v_i,d)=\pi(u,v_i)+\pi(v_i,d)$，即 $\pi_u(u\in v_i,d)=\pi(v_i,d)$，即合法路径不会造成报文偏转. 其中路径 $P=\pi(u,d)=(u,v_1,v_2,\cdots,v_n,d)$，表示 u 分配的到达 d 的节点序列，$\pi_u(u\in v_i,d)$表示在 u 的路由记录中所看到的 v_i 到达 d 的路

径，而 $\pi(v_i,d)$ 表示 v_i 实际采用转发路径，$P \cdot Q$ 表示为路径 P、Q 拼接.

设非空合法路径 P、Q 完全不相交；或者在某点交叉，即 $\exists i,j$，使得 $v_i=w_j\neq d$，或者 P 与 Q 仅在 d 汇合，如果 $\pi_u(u\in v_i,d)=\pi_z(z\in w_j,d)$，那么 P、Q 之间是一致的，一致性路径不会造成转发环路.

定义 2. 隐藏路径与备份路径. 指底层拓扑存在到达 d 的合法路径，但 AS 却没有获取该路径，其具有相对性，比如图 1 中，AS3 可以通过(3-5-6-2-0)或(3-4-6-2-0)合法路径封装报文到达目标，却被 AS4、AS5 隐藏. 备份路径是指可到达 d 且没有被选作最佳路径的合法路径，备份路径也可能被隐藏.

定义 3. 路径不相交度. 对于给定路径 P、Q，如果 P 和 Q 具有不同下一跳，$Disjoint(P,Q)=1-|P\cap Q|/|P|$（计算 $|P\cap Q|$ 时不包含源、目的节点）；否则 $Disjoint(P,Q)=0$，这是由于具有相同下一跳的路径无法用于快速恢复. 对于节点 u 的合法路径 P、Q，如果满足 $Disjoint(P,Q)>0$，则称 u 具有多样性路径.

定义 4. 转发树. 对于给定目标前缀 d，存在稳定的路径分配使 BGP 收敛后，所有节点最佳路径最终形成以 d 为根节点的转发树[20]，转发树中所有节点转发路径具有一致性，即如果 $\pi(u,d)=(u,v,P)$，那么 $\pi(v,d)=(v,P)$. 由于 BGP 的策略限制，其转发树并不一定是最短路径树.

定义 5. 路径标识(Path Identifier，PID). 当存在多条到达目标前缀的路径时，为了区分、处理每条合法路径，节点需要同时根据前缀 $prefix$ 和特定标识符 ψ 区分路径，即利用 $(prefix,\psi)$ 路径标识符唯一标识一条路径 P.

定义 6. 秩函数 $\lambda^v:P^v\rightarrow N$，对于 $P\in P^v$，$P^v=\{P|P=v\cdot\pi(v,d),v\in Peers(u)\}$，$P^v$ 表示从 v 到达 d 的合法路径集合，如果 P_1、$P_2\in P^v$，$\lambda^v(P_1)>\lambda^v(P_2)$，那么 P_1 具有更高优先级.

定义 7. 瞬时失效节点. 指出现失效后，节点 v 的 $rib\_in(v)$ 中无可到达 d 的合法路径，此时节点可行路径集为空或包含非法路径，此时会造成报文丢失或者环路[22]. 如图 1 所示，目标前缀 d 位于 AS0，AS 旁矩形框为根据 Valley-Free 得到的路由表，比如 AS5 具有 3 条到达 AS0 的路径，AS5 根据“Prefer Customer”优选(5-3-0)作为最佳路径. 假设链路 AS0-AS3 断开，AS3 发送撤销消息到 AS4 和 AS5，AS5 将根据其偏好策略优先选择(5-4-3-0)，而 AS4 将选择(4-5-3-0)路径作为最佳路由，WRATE(撤销速率限制定时器)使得 AS2 和 AS3 延迟通告新的可用路径，此时 AS5、AS4 之间形成环路，AS3 无到达 AS0 的路径. 即 AS3、AS4、AS5 经历瞬时失效.

定义 8. 转发中断. 转发路径是指在时刻 t，报文从 u 发送到 d 所实际经过的节点序列，表示为 $forward(u,d)$，如果 $\exists v\in forward(u,d)$，$v$ 的转发平面中的下一跳为空或陷入环路，则称之为转发中断(本文忽略失效检测时间，BFD 机制可将失效检测时间降至 15 ms 以内，Cisco 路由器已经普遍实现 BFD). 需要注意的是，节点瞬时失效仅是转发中断的必要条件，比如在图 1 中，如果 AS2-AS0 链路失效，虽然 AS6 采用(6-2-0)不是合法路径，但是由于 AS2 具有冗余路径，在检测到失效以后能够迅速切换(2-1-0)，此时并无节点经历转发中断. 而在一些加速收敛协议中，如 Ghost Flushing 将使 AS2 快速撤销路径(2-0)，如果 AS6 无备份路径，将导致 AS6 也经历转发中断，数据平面的转发中断相对于 BGP 收敛时间更准确度量协议性能[23]，文献[5]也指出控制平面的可达与数据平面的可达并不完全相关，本文也将用瞬时失效率、转发中断作为衡量协议性能的重要指标.

4 MFT²-BGP 协议设计

4.1 MFT²-BGP 基本框架

BGP 路由系统表示为 $S=\langle G,Policy(G),s\rangle$，其中 AS 图表示为 $G=(V,E)$，$V=\{0,1,\cdots,n\}$，$Policy$ 为所有节点的策略集合. BGP 系统状态表示为元组 $\langle r_0,r_1,\cdots,r_n\rangle$，其中 r_i 表示 AS i 的路由表中所包含的到达 d 的路径集合，$p\in r_i$，s^0 表示 d 发起通告时的初始状态，$s^0=\langle r_0^0,r_1^0,\cdots,r_n^0\rangle$. 当处于收敛状态 s^{final} 时，所有节点的转发路径将形成以 d 为根的最佳转发树 $Best\_Tree(d,\langle r_0^{final},r_1^{final},\cdots,r_n^{final}\rangle)=G(V',E')$，且有 $V'=\{n_j\in V|\exists p\in r_j\}$，

$$E'=\bigcup_{n_j\in V'}\{(u,v)\in E\mid \exists p\in r_j\wedge p=(v_1,v_2,\cdots,v_m)\wedge((u=n_j\wedge v=v_1)\vee(u=v_k\wedge v=v_{k+1}))\} \quad (1)$$

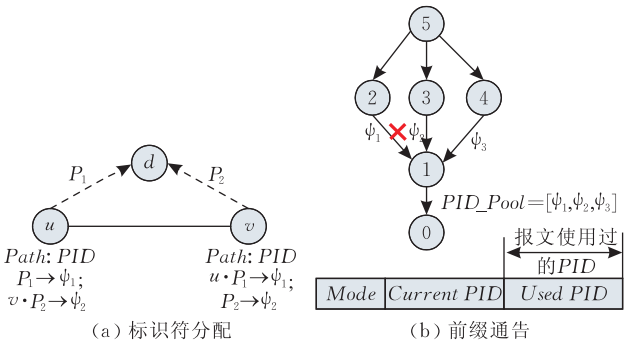
当链路失效后 BGP 重新收敛形成新的以 d 为根的转发树. 如图 1 中，AS0-AS3 断开后收敛最终得到失效后转发树. 基于以上观察，如果各个节点在失效前能够发现、预计算出备份转发树，就可以保证每个 AS 在面临链路/节点失效的时候，仍然存在合法备份路径，从而将流量切换到备份转发树中，且备

份转发树路径上具有完全一致性，其结构化性质阻止了环路形成. 然而，由于 BGP 并不能获得全局拓扑进行有效集中式计算链路或者节点不相交树[15-16]，而注入失效链路并强制其收敛获得失效后转发树，其收敛时间过长且会中断报文转发. 为此，MFT²-BGP 对协议通告过程进行修改，添加路径标识符以区分、获得结构化的多样性路径. 节点为了能够利用多条路径进行无环转发，也需要为多路径分配路径标识符，任意节点 v 必须实现如下映射：

$$\begin{cases} PathID_v(path,d):\ Path \rightarrow PID \\ Forward_v(PID,d):\ PID \rightarrow Path \end{cases} \tag{2}$$

路径标识函数 $PathID_v$，即节点 v 需要为每条合法路径从路径标识集合 PID 中选择合适标识符 ψ_v^i，其中 ψ_v^i 表示节点 v 的第 i 条路径标识符，而转发函数 $Forward_v$ 则根据 $\psi(\forall\psi\in PID)$ 进行报文转发，将 ψ 映射到对应的路径 $P=(v_1,v_2,\cdots,v_n)$，节点 v_k 和 v_{k-1} 需正确解释 ψ，且 v_k 能够将报文转发到 v_{k-1}，即 $Forward_{v_k}(\psi,d)=v_{k-1}$，$1\le i\le n$. 其中路径标识符可以是全局一致标识或本地分配的标识符，而本地分配标识符更加灵活，但 v_k、v_{k-1} 之间需对本地标识符 $\psi_{v_k}^i$ 以及 $\psi_{v_{k-1}}^i$ 按照式(2)的条件协商后进行映射转换，即对于标识符 ψ 路径 P 上相邻节点，如图 2(a)所示，$map(\psi_{v_k}^j)=map(\psi_{v_{k-1}}^i)=\psi$ 以保证转发一致性，标识符映射可采用与 Deflection[14] 类似的方法，为表述方便本文采用全局一致性标识.

在不影响理解的情况下，本文中将网络节点等同于 AS 节点. 对于标识符 ψ 和其对应转发树 ψ 也不进行区分，对于转发树 ψ(非最佳转发树)中节点对应 ψ 的路径，也称为该节点的备份路径.



(a) 标识符分配　　(b) 前缀通告

图　2

4.2　控制平面多转发树构造与更新

MFT²-BGP 对路由通告、路由决策进行了相应的修改，如图 1 中，AS0 根据本地策略向 3 个邻居节点 AS1、AS2、AS3 通告前缀时分别分配标识符 ψ_0^1、ψ_0^2、ψ_0^3(简记为 ψ_1、ψ_2、ψ_3). 各节点(不含 AS7、AS8)最终获得多路径如图 3 所示，其中对更新消息处理顺序可能导致中间状态不同，但最终收敛结果一致，形成 3 个不同转发树，简记为 ψ_1_Tree、ψ_2_Tree、ψ_3_Tree. 其中对于AS7仅能通过上行路径(uphill)[19]到达目标前缀，AS6 和 AS2 会导出所有本地合法路径到 AS7，AS7 可根据本地策略决定各个路径标识上的最佳路径. 图 3 表现了协议运行过程，暂时并没有考虑将具有相同下一跳的路径合并以降低开销以及加速收敛等优化操作. 其中，图 3 中“⇒”表示发送更新消息，下划线“—”表示该节点 rib_in 发生变化，路由表中斜体部分表示可以合并的路径，设置最佳路径标识符“*Best*”是为了与当前协议兼容，并且避免由于标识符数量过少而导致部分节点使用次优路径(如果不考虑“*Best*”，更容易实现，但是却违反了标准 BGP 语义).

更新消息中含: PID,AS_Path		AS1	AS2	AS3	AS4	AS5	AS6
0⇒1: ψ_1,0	0⇒3: ψ_3,0						
0⇒2: ψ_2,0		ψ_1:1-0	ψ_2:2-0	ψ_3:3-0			
1⇒2: ψ_1,1-0	3⇒4: ψ_3,3-0						
	3⇒5: ψ_3,3-0		ψ_1:2-1-0		ψ_3:4-3-0	ψ_3:5-3-0	
2⇒1: ψ_2,2-0	2⇒6: ψ_1,2-1-0						ψ_1:6-2-1-0
4⇒5: ψ_3,3-0	ψ_2, 2-0	ψ_2:1-2-0				ψ_1:5-3-0	ψ_2:6-2-0
6⇒5:ψ_1:6-2-1-0	5⇒6:ψ_3:5-3-0					ψ_1:5-6-2-1-0	
ψ_2: 6-2-0						ψ_2:5-6-2-0	ψ_3:6-5-3-0
6⇒2:ψ_3:6-5-3-0	5⇒3/4:ψ_1:5-6-2-1-0 ψ_2:5-6-2-0			ψ_1:3-5-6-2-1-0 ψ_2:3-5-6-2-0	ψ_1:4-5-6-2-1-0 ψ_2:4-5-6-2-0		
	2⇒1:ψ_3,2-6-5-3-0		ψ_3:2-6-5-3-0				
最终稳定状态		*Best*: 1-0 ψ_1:1-0 ψ_2:1-2-0 ψ_3:1-2-6-5-3-0	*Best*: 2-0 ψ_1:2-1-0 ψ_2:2-0 ψ_3:2-6-5-3-0	*Best*: 3-0 ψ_1:3-5-6-2-1-0 ψ_2:3-5-6-2-0 ψ_3:3-0	*Best*: 4-3-0 ψ_1:4-5-6-2-1-0 ψ_2:4-5-6-2-0 ψ_3:4-3-0	*Best*: 5-3-0 ψ_1:5-6-2-1-0 ψ_2:5-6-2-0 ψ_3:5-3-0	*Best*: 6-2-0 ψ_1:6-2-1-0 ψ_2:6-2-0 ψ_3:6-5-3-0

图 3　MFT²-BGP 更新消息处理过程示例

当节点检测到最佳路径的下一跳失效后，如果该节点具有不相交度大于 0 的备份路径，通过修改报文头以动态切换到具有不同标识符上的转发树，实现无中断转发.

4.2.1　前缀通告

在 MFT²-BGP 协议中，ASv 发起前缀通告时，针对不同的节点对$\langle v,u\rangle$为每条路径分配标识符，v 的 loc_rib 发生改变并在通过导出策略后，需将对应标识符(ann_pid)的路径信息发送到邻居节点 u. 当 v 的邻居节点数大于可用路径标识符数量，将其余的邻居 AS 路径标识符设置为隐藏标识符($Hide\_PID$)，表示该路径仅可能被部分 AS 选作最佳路

径,但是并不会传播到全网,从而在暴露的路径数量以及开销之间进行折衷.如果发起前缀通过的 AS 仅有单条路径,如图 2(b)所示,此时 AS0 可以将路径标识符池[ψ_1,ψ_2,ψ_3]通告给 AS1,然后 AS1 从标识符池中取出标识符并通告给不同邻居,或者 AS1 自主对不同邻居分配路径标识符.

通过分配路径标识符使得路由器能够区分、利用多路径,v 存储的 rib_in($v \leftarrow u$)、rib_out($v \rightarrow u$)、loc_rib(v)以及更新报文 *update* 数据结构修改为映射类型,比如某个完整的更新报文将包括($pefix$,PID,AS_Path)三个域,分别表示前缀、标识符以及 AS 路径.其通告过程如过程 1 所示,其中省略了 MRAI 设置等语句.过程 1 中第 1～3 步表示构造原始的通告路径.第 4～10 步表示依据路由策略从路径标识符池中选取路径标识符,并构造对应标识符的更新消息.第 11～13 步则表示将包含标识符的更新消息发送到邻居节点中.

过程 1. 路由器 v 目标前缀通告过程.

/* v 表示发起通告的路由器,$prefix$ 为目标前缀,$Peers(v)$ 表示其邻居列表,PID_Set 表示可用路径标识符集合,*update* 为路由更新消息 */

1. $P_{origin} := path\_init(prefix)$ //构造通告路径
2. $v.origin\_rib(prefix) := P_{origin}$
3. IF $v.Path\_Selection(prefix) ==$ True //检测 loc_rib 是否改变
4. FOR EACH $w \in peers(v)$ DO
5. IF $v.export\_policy(w, prefix, origin\_path) ==$ True && $v.get\_link(w) \neq$ down //检测到达对应邻居节点的路径无失效且能通过导出策略
6. IF $PID\_Set == \varnothing$
7. $ann\_pid := Hide\_PID$
8. ELSE
9. $ann\_pid := PID\_Set.pop()$ //获得路径标识符
10. ENDIF
11. $update = v.loc\_rib[ann\_pid]$//根据标识符构造更新消息
12. $v.export\_action(ann\_peer, prefix, update)$
13. $v.send\_update(w, ann\_pid, update)$//发送更新报文到 w
14. ENDIF
15. ENDFOR
16. ENDIF

4.2.2 路由更新

节点接收到路由更新消息后,将启动路由更新发送的过程,其服从 SPVP 模型[20].如过程 2 所示,其中第 1～4 步表示 v 接收到 u 更新报文以后,v 根据导入策略对更新报文进行过滤.第 5～9 步中,v 从更新报文中提取所有路径标识符对应的路径信息,v 更新所保存的 rib_in($v \leftarrow u$).第 10～13 步中 v 调用本地路径选择过程,使用秩函数 λ^v_{best} 从所有可行路径中选择最佳路径,然后为每个路径标识符计算新的最佳路径,并最终下载到 FIB 中.其中在过程 2 中,v 中具有标识符 ψ_i 的路径集合表示为 $P^v_{\psi_i}$,β_{ψ_i} 表示对应标识符的最佳路径,令 $\lambda^v_{\psi_i}$ 表示对应路径标识符的秩函数,要求选择最大不相交多样性路径.

过程 2. 路由器 v 路径选择过程.

1. IF $loop\_detection(update) ==$ True //环路检测
2. RETURN
3. $v.import\_policy(u, update)$ //根据导入策略对 *update* 报文进行过滤并修改路径属性
4. $v.import\_action(u, update)$
5. FOR EACH $(PID, Path) \in update[prefix]$ DO //遍历 *update* 中所有标识符对应的路径信息
6. IF $Path == \varnothing$
7. Del $v.peers[u]$.rib_in$[prefix][PID]$
8. ELSE
9. $v.peers[u]$.rib_in$[prefix][PID] := Path$ //v 更新对应的 rib_in($v \leftarrow u$)
10. $old\_rib = Copy(v.loc\_rib[prefix])$ //复制原有 loc_rib
11. $\beta_{best} := \max(\lambda^v_{best}(\bigcup P^v_{\psi_i}))$ //从所有路径标识符中选择最佳路径
12. $v.loc\_rib[prefix][best] := \beta_{best}$
13. $install\_FIB(\beta_{best})$
14. FOR EACH $\psi_i \in PID\_Set$ DO
15. $\beta_{\psi_i} := \max(\lambda^v_{\psi_i}(Path^v_{\psi_i}))$//为每个路径标识符选择对应的最佳路径,且满足路径多样性要求
16. $v.loc\_rib[prefix][\psi_i] := \beta_{\psi_i}$
17. $install\_FIB(\beta_{\psi_i})$
18. IF $old\_rib \neq v.loc\_rib[prefix]$ //如果 loc_rib 改变,则将向每个 *peer* 发送更新消息
19. RETURN True, THEN $send\_update(peers(v))$
20. ELSE
21. RETURN False

4.3 多转发树下的加速收敛

与标准 BGP 协议构造单个以 d 为根的转发树相比,MFT^2-BGP 在构造多转发树时,为了传播隐藏路径也相应增加了一定的开销.MFT^2-BGP 则将 EPIC[8] 的思想扩展到多路径场景以加速协议收敛(简称 MFT^2-EPIC).MFT^2-EPIC 主要针对各标识符上的路径之间存在的依赖性,在每条标识符路径上添加不同的序列号($fesn$)进行区别,即 $fesn = (\langle x, y\rangle : N)$表示链路两边节点有序对与序列号的组合.即完整的更新消息包含[$pefix$,PID]+[AS_Path]+{$fesnList$},$fesnList$ 包含对应AS_Path中的 $fesn$ 列表.当出现链路失效(或更改策略)撤销[$pefix$,ψ]相关路径时,u 发送的撤销消息中包含 $fesnList_1$ 信息,令 $fesnList_1 = [(\langle a_1, b_1\rangle, n_1), (\langle a_2, b_2\rangle, n_2), \cdots, (\langle a_k, b_k\rangle, n_k)]$,$v$ 接收撤销消息后,v 可

利用撤销消息中的 $fesnList$ 信息作废其它邻居节点通告的所有标识符上的非法路径，设 v 存储的某条路径 P 对应的 $fesnList_2=[\langle a'_1,b'_1\rangle,n_1),\langle a'_2,b'_2\rangle,n_2),\cdots,\langle a'_l,b'_l\rangle,n'_l)]$，且 $k\leq l$. 如果存在 $(a_i==a'_i)\wedge(b_i==b'_i)\wedge(n_i==n'_i))$，$\forall i=1,2,\cdots,k$，即 $fesnList_2$ 与撤销路径 $fesnList_1$ 相关，此时可以快速作废非法路径 P，从而加速收敛并消除收敛过程中可能的环路.

4.4　瞬时失效抑制

在不影响转发连续的情况下，可以利用多路径压制瞬时失效以实现故障隔离能力，避免频繁收敛带来的路由震荡，从而降低控制平面的扰动. 基本思想是当最佳路径失效后，如果备份路径具有不同的下一跳，则并不发送更新消息，并且设置合理定时器（绝大多数 BGP 故障将会在 10 min 内恢复[24]，可据此设置默认定时器）. 如果定时器超时则需要发送路由更新消息，触发路由重新收敛. 比如 u 检测到路径标识符 ψ 对应的路径失效，如果 u 仍然具有其它可行转发树对应的路径，在不影响持续转发的情况下，u 可对与 ψ 相关的失效消息进行抑制，将该策略称为 SUP（Suppress all UPdate message）.

SUP 策略将会使得部分节点在压制失效后可能使用非最佳路径，比如图 1 中链路 AS0-AS3 失效后，将导致 AS3、AS4、AS5 的最佳路径发生变化，如果 AS3 压制失效信息，在消息压制期间，将会使得受到失效影响的节点 AS3、AS4、AS5 到达目标的跳数增加. 为此，设置新的路由策略 SNUM（Suppress Non-Best Update Message），即 u 在检测到失效后，如果其 u 的最佳路径改变，u 在利用备份路径转发的同时也发送更新消息，如图 1 中 AS0-AS3 失效后，AS3 将发送与 ψ_3 相关的路由更新消息，但是由于 ψ_3 并不改变图 1 中 AS1、AS2、AS6 的最佳路径，因此，AS5 可以安全压制更新路由消息，不再进一步向 AS6 发送消息，并避免持续使用次优路径.

4.5　报文转发过程

为了保证转发一致性，需要修改报文头格式并增加路径标识符 ψ，通常可以在 IP 报文头中添加额外的 shim 报文头[14]，或者利用现有 IP 报文头中 TOS 或 DSCP 域，通过报文头中标识符 ψ 指示报文当前所采用的转发树.

在链路失效时，失效检测节点具有特殊的角色，失效节点在查找备份可用转发树的同时，还需要将报文头中的路径标识符修改为所切换的转发树标识符. 在单链路失效时，由于转发树中对应的所有节点转发视图具有一致性，在节点间无需协调的情况下，检测失效的节点仅需要发起本地重路由即可保证无环转发.

当出现节点失效、多链路失效时，以及采用瞬时失效抑制，将会造成各个 AS 节点之间转发视图不一致，单失效下的转发机制并不能充分利用已经获得的冗余路径，而且还可能出现转发环路，因此还需要信令机制以避免环路. 如图 1 中，如果 AS2-AS0 和 AS1-AS0 同时断开，如果 AS2 检测到失效后采用图 3 中标识符为 ψ_1 的路径，修改报文头标识符为 ψ_1，并将报文转发到 AS2；同时，由于 AS2 发现 AS2-AS0 断开，由于 AS1 与 AS2 之间无协调，如果 AS1 选择 ψ_2 路径，将导致 AS1 和 AS2 之间形成环路，在这种多失效场景下必须利用辅助信息才能无环转发.

为了充分利用已获得的多转发树，MFT²-BGP 在报文中添加“报文经历信息”$Used\_PID$，表示报文已经使用过的转发树. 如图 2(b)中所示，节点在接收到报文后，提取报文头中 $Used\_PID$ 信息以避免重复选择同一转发树，该信息存储实际不超过 2 bits.

具体报文转发如过程 3 所示，其中 $Best\_Tree$ 表示最佳转发树对应的路径，其下一跳表示为 $best\_nhop(d,Best\_Tree)$，而根据路径标识符 ψ 形成的转发树为 ψ_Tree，其下一跳表示为 $best\_nhop(d,\psi\_Tree)$，$forward\_packet(nhop,tree)$ 表示根据对应转发树转发报文，$Change\_header(tree,transient)$ 表示修改报文头中的路径标识符为特定的转发树，并且设置报文进入失效转发模式.

过程 3.　报文转发过程.

```
1.  nhop := Best_nhop(d, packet.current_PID)//按指定标识符转发
2.  IF d≠nhop
3.    IF v.get_link(nhop)==UP
4.      forward_packet(nhop, packet.current_PID)
5.    ELSE
6.      FOR EACH(PID, AS_Path)∈v.loc_rib[prefix] DO
7.        IF PID not in packet.Used_PID and PID≠Best_Tree
8.  IF AS_Path.valid==False//如果 PID 对应的 AS_Path 是非法路径,需要将其加入报文 Used_PID 中
9.          packet.Add_Used_PID(PID)
10.       CONTINUE
11.     ENDIF
12.     (newhop, new_tree) := find_Best_PID_nhop(d, PID)
                  //根据策略从所有可行标识符中选择路径
13.   ENDIF
14.     IF newhop≠∅ and v.get_link(newhop)≠Down
15.       Change_header(new_tree, transient)
```

```
16.      packet.Add_Used_PID(new_tree)
17.      forward_packet(newhop,new_tree)
18.    ELSE
19.      IF v.loc_rib[prefix][Best_Tree].valid==True
                //在无对应路径标识符时,如果Best_Tree
                  可行则通过Best_Tree转发
20.        nhop := Best_nhop(d,packet.current_PID)
21.        forward_packet(nhop,packet.current_PID)
22.      ELSE
23.        Drop packet
24.      ENDIF
25. ELSE
26.   forward_success()
27. ENDIF
```

过程 3 中的第 1～4 步,报文依据对应标识符转发.第 6～12 步在未使用过的转发树中选择备份路径(瞬时失效中优先选择非 *Best_Tree* 以降低协议动态性的影响),需要说明的是,当节点具有多个备份转发树时,节点可以根据策略选择最短路径树 *STF*(或者固定优先选择某个特定标识的转发树 *SPT*[16],并且将已失效转发树标识加入报文中的 *Used_PID* 域中).第 14～17 步中,节点将根据修改后的转发树标识进行转发.第 19～21 步中,当节点在无可用转发树时,则尝试使用 *Best_Tree* 转发.

该转发过程可以有效解决多链路失效下的转发环路,比如在图 1 中,当节点失效引起 AS2-AS0 以及 AS1-AS0 断开后,报文到达 AS2 后,其 *Best*、ψ_2 皆失效,如果 AS2 选择了 ψ_1 将报文转发到 AS1,AS1 将通过报文中 *Used_PID* 以及当前链路状况,仅能选择标识符 ψ_3,最后通过 ψ_3 成功转发报文.当路径标识符数量为 k 时,为了表示报文头的状态,其报文头开销的最大长度约为 $1+2\lceil\lg(k)\rceil$ bits.

5 协议属性分析

5.1 协议收敛属性

定理 1. 对于给定的 SPVP 实例 Z,如果它不含 Dispute Whee[20],那么 MFT^2-BGP 最终收敛;且在路由收敛期间不会出现瞬时环路.

证明. Dispute Wheel[20] 的定义略,本文将 SPVP 扩展到多路径路由.与标准 BGP 不同,v 具有多条到达目标 AS0 的路径,v 采用路径标识符 ψ 进行区分,路由更新消息中包含 ψ 以及 AS_Path 路径信息.v 的最佳路径标识符 β_{best} 将从 v 的可行路径集合 P^v,即从 rib_in($v\leftarrow peers(v)$)中选择,因此在同样的拓扑、策略下,如果 SPVP 在无 Dispute Wheel 的情况下收敛,那么 MFT^2-BGP 中最佳路径 β_{best} 上也一定收敛.

其次,对于路径标识符 ψ_i,v 在路由决策过程中计算 ψ_i(除最佳路径 β_{best})路径时与 ψ_j($i\neq j$)的路径隔离,即在 ψ_i 上其可行路径集合即仅在 $G_{\psi_i}\subset G$ 中选择路径,显然,如果 SPVP 在拓扑 G 中无 Dispute Wheel,G_{ψ_i} 对 G 删除了部分边,G_{ψ_i} 也不会含有 Dispute Wheel,那么 MFT^2-BGP 中标识符为 ψ_i 的路径 β_{ψ_i} 在 G_{ψ_i} 上也一定收敛. 证毕.

5.2 报文转发属性

引理 1. 对于任意节点 u,如果在链路 e 失效前,其最佳路径上包含 e,在 BGP 协议中最终收敛能够得到不含 e 的路径,那么 MFT^2-BGP 保证 u 存在某条标识符 ψ 的路径不通过 e.

证明. 首先,显然当 MFT^2-BGP 使用充分数量的路径标识符时,总能获得底层拓扑中服从策略多样性路径,并形成多转发树.设最佳转发树(*Best_Tree*)中失效链路 e 的失效根节点为 q,设 e 失效之前,$\pi_{Best}(u,d)=P_{best}=(u,v_1,v_2,\cdots,v_n)\cdot\pi(q,d)$,且 $\pi(q,d)=(q,w_1,w_2,\cdots,w_m,d)$;在 e 失效之后如果 BGP 协议收敛能够使 u 获得到达 d 的路径为 $P_{backup}=(u,z_1,z_2,\cdots,z_t,d)$,即 u 在失效前、后的最后一跳分别为 w_m 和 z_t.下面证明 MFT^2-BGP 能够获得多样性路径.

(1) 如果 $w_m\neq z_t$,根据前缀通告过程,AS0 发起前缀通告时将会针对每个邻居节点 w_m、z_t 分配不同的路径标识符 ψ_i、ψ_j($i\neq j$)进行区分;节点 z_t 到达 d 的路径必不包含失效链路 e.因此,z_t 需要将 ψ_j 的路径通告给 z_{t-1},依次类推,可知 u 将会存在 ψ_i、ψ_j 上的多样性路径.同样,通过归纳法证明,u 将 ψ_j 路径通告给 v_1,任意 v_i($n>i\geq 1$)中存在某条标识符 ψ_j 的路径不通过 e,且最终 w 也将会获得不包含 e 的路径.

(2) 如果 $w_m=z_t$,说明 P_{best} 与 P_{backup} 路径重叠,AS0 仅有一个邻居节点.如果 w_m 为每个邻居节点分配标识符,设失效前 P_{best} 路径的标识符为 ψ_i,z_{t-1} 到达 d 的路径不包含 e,其标识符为 ψ_j.依次类推,可得 u 将会存在不经过 e 的标识符为 ψ_j 的路径.对于 v_i($n>i\geq 1$),其证明与上面相同. 证毕.

引理 2. 在无失效场景下,报文总在最佳路径上转发;当出现单或多链路失效后引起路由收敛期间,如果协议获得的多样性路径中存在可行的转发路径,那么报文不会被丢弃,且不会陷入环路.

证明. 分 3 种情况证明,即无失效、单链路失效、多链路失效场景.显然,在无失效场景下 MFT^2-BGP 处于收敛状态,v 的最佳路径是对所有可行路

径集合 P^v 计算得到一致性路径，报文在最佳路径上安全转发.

其次，当出现单链路 $e=(u,v_1)$ 失效以后，设失效检测节点 u 的最佳路径为 $\pi_{Best}(u,d)=(u,v_1,v_2,\cdots,v_n,d)$，当 e 失效后，u 切换到路径 $\pi_{\psi_i}(u,d)=(u,w_1,w_2,\cdots,w_k,d)$，如果 $Disjoint(\pi_{Best}(u,d),\pi_{\psi_i}(u,d))=1$，报文将在转发树 ψ_i_Tree 中无环转发. 如果 $0<Disjoint(\pi_{Best}(u,d),\pi_{\psi_i}(u,d))<1$，即 $\pi_{Best}(u,d)$ 与 $\pi_{\psi_i}(u,d)$ 部分重叠，由 MFT²-BGP 报文转发算法可知必有 $v_1\neq w_1$，设 $\pi_{Best}(u,d)$ 与 $\pi_{\psi_i}(u,d)$ 在某点 $v_i=w_j$ 相交，其中 $v_i\in\pi_{Best}(u,d)$，$w_j\in\pi_{\psi_i}(u,d)$，$1<i\leq n$，$1<j\leq k$. 如果 $\pi_{Best}(v_i,d)=\pi_{\psi_i}(w_j,d)$，即 v_i 中标识符为 *Best* 和 ψ_i 具有同样的路径，由于其任意转发树 ψ 上转发路径具有一致性，即报文不会回退或环路；而当 $\pi_{Best}(v_i,d)\neq\pi_{\psi_i}(w_j,d)$，即 v_i 中标识符为 *Best* 和 ψ_i 具有不同的路径，报文 p 在 ψ_i 可无环转发.

出现多链路失效时，根据报文转发过程，具有如下属性：(1) 报文 p 不会 2 次遇到同一个失效链路，该属性主要是由于检测失效的节点会将其失效转发树的标识符 ψ_i 加入到报文头中，下行节点选择转发树时将避免 ψ_i，这使得 p 不会两次遇到同一个失效链路；(2) 当出现第 $k(k\geq 2)$ 个链路失效后，当前节点 u 将根据 $\psi_k(\psi_k\neq\psi_i)$ 转发报文 p，如果在 ψ_k 不能成功转发，说明出现了新的失效链路，u 将避免使用 $p.Used\_PID$，并选择新路径，此时 p 要么被成功转发，或由于无可行路径而丢弃.　　证毕.

MFT²-BGP 获得多转发树的同时也一定程度上增加了计算、存储开销，在不采用路径合并的时候，其最大存储开销线性增加. 根据文献[25]的结论，硬件的发展仍然满足 Moore's 定律，其多路径计算、存储带来的开销是在可承受范围之内，且如果将同一前缀的多标识符路径进行合并到单个转发表，其存储开销将会进一步减低.

由于互联网 AS 拓扑具有幂率特性，其平均 AS 跳数为 4，可证 MFT²-BGP 增加转发路径的长度非常小，在本文后实验中也证明了这一点.

6　性能评价

为验证 MFT²-BGP 性能，本文对轻量级 SimBGP①进行扩展，实现了所有 BGP 相关的重要特征. 根据 CAIDA② 的数据分析可知当前 AS 数量为 31212，链路数量为 60052，多宿主 AS 占总数的 53.4%，由于其规模过大，采用 Dimitropoulos[26] 根据 CAIDA 数据标注推断的 AS 级 Internet-like 拓扑图，该拓扑具有 AS 之间的商业关系以及 Internet 的复杂结构特征，Splicing[11] 也使用了该拓扑. 本文采用的 800 节点的拓扑如图 4 所示，链路数量为 1582，协议模拟参数如表 1 所示. 实验对 MFT²-BGP 在各种场景下的收敛时间、消息数进行比较；为验证 MFT²-BGP 的无中断转发的性能，还从平均 AS 瞬时失效率、平均转发中断时间、路径伸展度等方面进行比较.

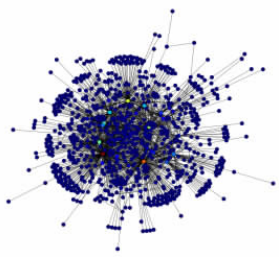


图 4　800 AS 节点拓扑

表 1　实验参数设置

参数	默认值
MRAI	30 s (Peer-based)
SSLD, WRATE	True, False
导入导出策略	Gao-Rexford
其它设置	基于标准 BGP
MRAI 抖动	[0.75,1] * MRAI
链路队列延迟	[0.01,0.1] ms
FIB 处理更新延迟	[0.001,0.01] ms
默认 PID 数量	2

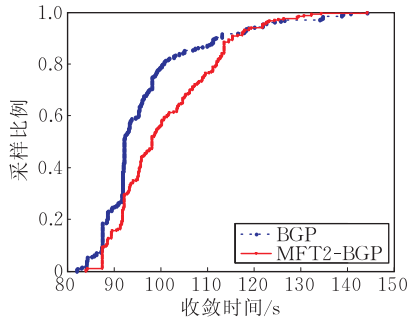
6.1　多转发树构造开销

该场景中主要针对多宿主 Stub AS 发起前缀通告，测试其收敛时间以及消息数量. 在 BGP 实现中 MRAI 通常设置为 Peer-Based. 由于源节点在发起通告时，所有节点并没有对其邻居设置 MRAI，为保证模拟的真实性，降低 Peer-Based 的影响，因此设置发送消息等待时间为[0，MRAI]随机分布，以此作为背景 MRAI.

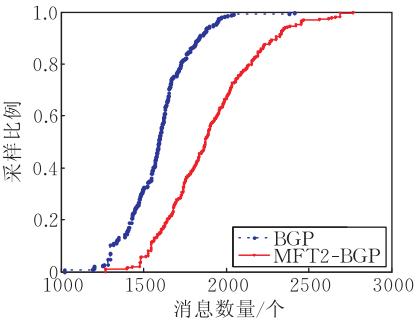
(1) 当使用路径标识符数量为 2 时，随机选择 150 个多宿主($Provider\geq 2$)的 AS 共测试 648 次. 收敛时间、消息数量累积分布概率如图 5(a)、图 5(b)所示，与 BGP 相比，收敛时间增加仅约 4%，总体消息数量增加约 20%，以合理的开销获得了多样性路径. MFT²-BGP 在前缀通告时消息数量有所增加，这是由于为了将隐藏合法路径暴露给其它 AS，而收敛时间与 BGP 相比几乎无增加，这是由于不同标识符上的收敛具有相对隔离性质，而 BGP 收敛时间主要与网络直径、MRAI 设置等因素相关[27]，理论上收敛时间为 $\max\{\psi_1\_ann,\psi_2\_ann,Best\}$，其中 ψ_i_ann 为对应标识符上的收敛时间. 此外 MFT²-BGP 中 MRAI 采用 Peer-Based 设置，由于 ψ_1、ψ_2 会相互影响，也一定程度上会限制更新消息传播速度，稍微增加收敛时间.

① simBGP: A lightweight event-driven BGP simulator. http://www.bgpvista.com/simbgp.php

② CAIDA AS relationships(20090429). http://www.caida.org/data/request_user_info_forms/as_relationships.xml



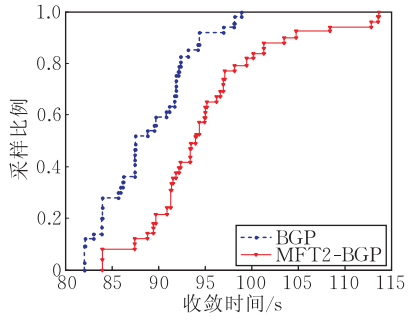
(a) 收敛时间累积分布(≠*PID*=2)



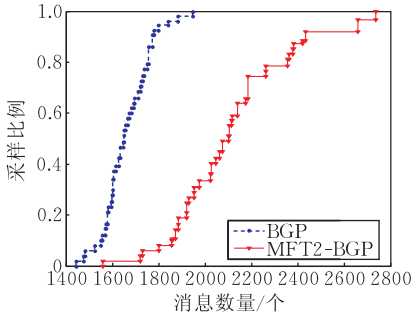
(b) 消息数量累积分布(≠*PID*=2)

图 5 标识符数量为 2 时的构造开销

(2) 当使用路径标识符数量为 3 时,随机选择 71 个多宿主(*Provider*≥3)测试 214 次.收敛时间、消息数量累积分布概率如图 6(a)、图 6(b)所示,其收敛时间仅增加 10%,总体消息数量增加 40%,但考虑到每个节点获得的路径数量将会达到 3,此开销仍然在合理范围内.



(a) 收敛时间累积分布(≠*PID*=3)



(b) 消息数量累积分布(≠*PID*=3)

图 6 标识符数量为 3 时的构造开销

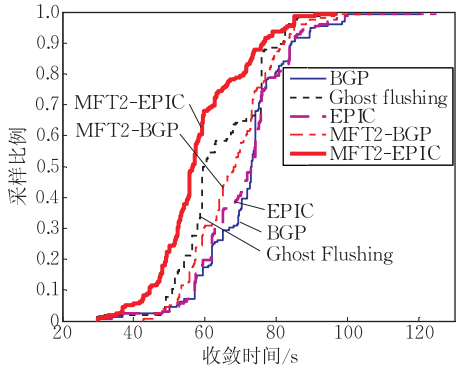
MFT²-BGP 在初始化时,带来了一定的消息开销,但这种开销也仅出现在多转发树的初始构造中.

6.2 单失效下协议性能

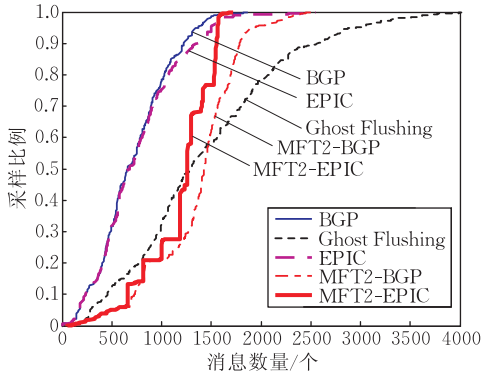
针对单链路失效,测试了边缘链路、核心链路失效两种场景.

(1) 在边缘链路失效中,随机选择 300 个多宿主 AS(*Provider*≥2),断开多宿主 AS 与其中的一个 *Provider* 的链路(这是最极端情况),共进行了 650 次实验,分别测试了 BGP、Ghost Flushing[7]、EPIC[8]、MFT²-BGP 以及加速收敛的 MFT²-EPIC.

从图 7(a)中 CDF 图中可以看出,Ghost Flushing、EPIC 协议在 Internet-like 拓扑中 Fail-over 收敛时间与 BGP 相比改善效果并不明显[27].MFT²-BGP 协议收敛时间与 EPIC、BGP 相比有所降低,总体上 MFT²-BGP 比 BGP 收敛时间降低了 8%以上.尤其值得注意的是,MFT²-EPIC 收敛时间均优于其它所有协议的收敛时间,与 BGP 相比降低了 30%以上,且在绝大多数实验中,收敛时间都降低 50%以上.发生这种现象的原因主要是 BGP、Ghost Flushing、EPIC 等协议在 Fail-over 场景中,要传播撤销消息以及通告新的路径,从而增加了收敛时间,收敛时间主要受到失效位置与失效后的路径到达目标的距离两个因素的影响[27];而 MFT²-BGP 在前



(a) 收敛时间累积分布



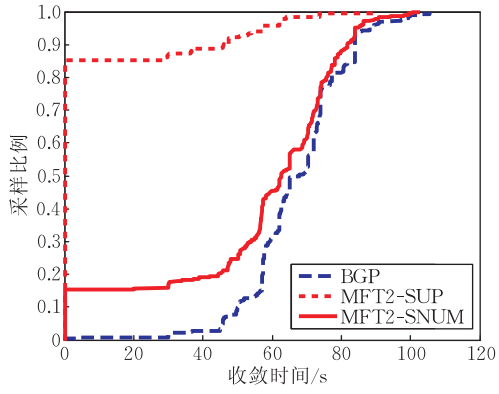
(b) 收敛时间累积分布

图 7 单失效下控制平面开销

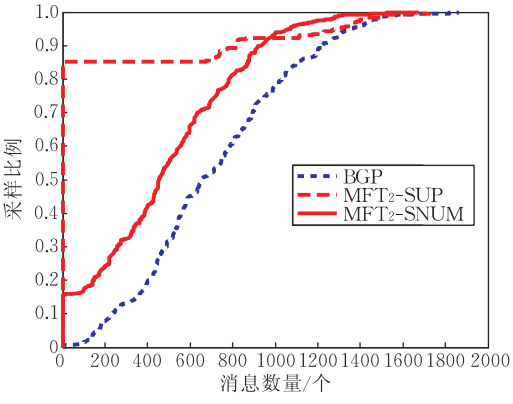
缀通告的时候已经比较充分地获得了隐藏路径，当出现链路失效时，MFT²-BGP、MFT²-EPIC 主要传播撤销消息，且 MFT²-EPIC 降低了路径探索，因此收敛时间大为改善. 此外在实验中也发现，在设置 WRATE 时，MFT²-BGP、MFT²-EPIC 收敛时间与 BGP 相比将会进一步降低，其改善效果更加明显.

从图 7(b)中 CDF 图中可以看出，Ghost Flushing 以增加消息数量为代价降低了收敛时间. MFT²-BGP 和 MFT²-EPIC 由于需要将特定路径标识符上的撤销消息传播到全网，略微增加了消息数量. 但是 MFT²-EPIC 与 BGP 相比，其消息数量的最大值与 BGP 相比反而降低了.

由于 MFT²-BGP 具有多条合法路径，仅当失效检测节点存在备份路径时可以抑制瞬时失效. 如图 8(a)、图 8(b)所示，当采用 SUP 策略时，其收敛时间和消息数量均降低了一个数量级；而采用 SNUM 策略时也有较大程度的改善. 在不影响持续转发的同时，SUP 和 SNUM 策略均极大地降低了网络的抖动.



(a) 失效抑制收敛时间

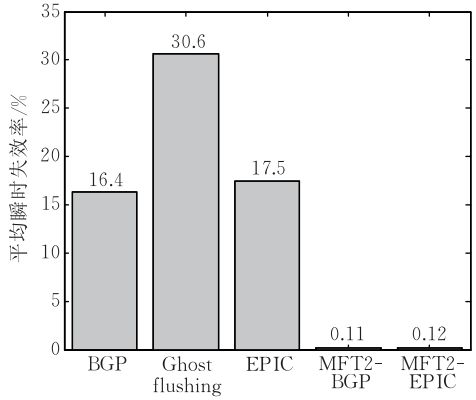


(b) 失效抑制消息数量

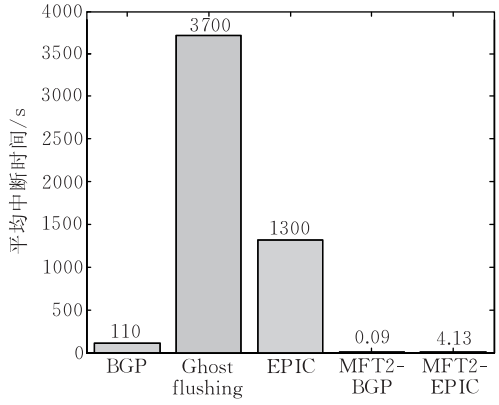
图 8　失效抑制策略下控制平面开销

当链路断开时，测试各个协议经历路由瞬时失效的节点数目(其中对永久失效节点不计入内)如图 9(a)所示. 在协议动态收敛过程中，MFT²-BGP、MFT²-EPIC 降低平均瞬时失效率至 0.11% 和 0.12%，经过分析发现，MFT²-BGP 中瞬时失效主要是由于部分核心节点采用“Gao-Rexford”导入导出原则，使其仅能获得单条路径，当该路径被撤销以后，从而引起其它节点转发中断. 而 Ghost Flushing、EPIC 发送过多撤销消息使得大量节点缺乏可用路径，从而增加了瞬时失效节点的数目.

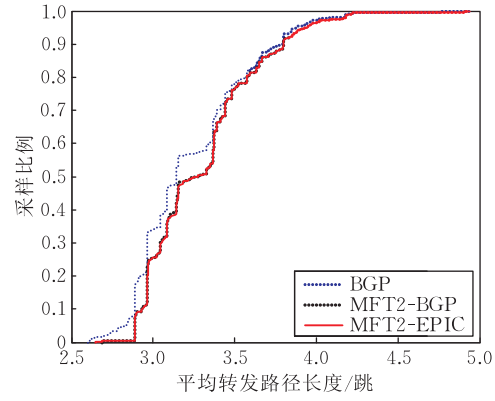
在收敛过程中通过对每个节点模拟转发报文，计算平均中断转发时间如图 9(b)所示(其中对永久失效节点不计入内). MFT²-BGP、MFT²-EPIC 中节点可以动态切换转发路径，从而降低平均转发中断



(a) 平均瞬时失效率



(b) 平均中断时间



(c) 平均转发路径长度

图 9　单失效下协议转发性能

时间至 0.09 s 和 0.13 s，而 EPIC 和 Ghost Flushing 由于快速传播路由撤销消息，部分 AS 节点又缺乏可用的备份路径，从而造成转发中断时间大量增加.

如图 9(c)所示，MFT²-BGP、MFT²-EPIC 实际的平均转发路径长度几乎没有增加，与失效后的最佳路径长度相比增加了不到 1%. 这主要是因为当 AS 节点检测到失效时，节点动态切换的路径在绝大多数情况下实际就是最终收敛后的最佳路径.

(2)在核心链路失效中，模拟中选择 43 条核心链路断开进行测试(不计由于核心链路密集而无失效的场景). 从图 10 中可以发现，所有协议的平均瞬时失效节点都有所降低，这是因为在 BGP 协议中，核心节点中通常具有多条可行路径，降低了路径探索时间.

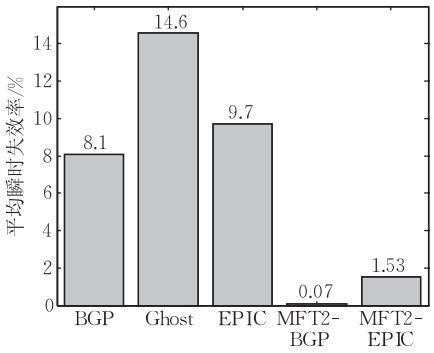


图 10 核心链路失效时平均瞬时失效率

6.3 多失效下协议性能

在域间路由发生多并发失效的情况是非常少见的，绝大多数出现在灾难场景中. 本文主要模拟了双链路失效情景.

(1)双链路失效场景 1. 当使用路径标识符数为 2 时，模拟中采用双链路失效的最极端场景，即断开多宿主 AS 与其 Provider 的一条链路，并随机选择备份路径标识符上的一条链路失效(第 2 条失效链路不能为多宿主 AS 与其 Provider 的链路，否则可能彻底断开)，共随机测试 248 次. 从图 11 中可以看出，与单链路失效场景相比，虽然 MFT²-BGP、MFT²-EPIC 经历瞬时失效的比例有所增加，但是仍然比 BGP 等协议瞬时失效比例大幅降低.

(2)双链路失效场景 2. 为了验证 MFT²-BGP 和 MFT²-EPIC 的可扩展性，使用 3 个路径标识符时，随机选取了 70 个多宿主 AS($Provider \geq 3$)共测试 156 次，测试场景仍然为最极端的情况，即将目标 AS 与 Provider 连接其中两条链路同时断开. 从图 12(a)中 CDF 图中得知，MFT²-BGP、MFT²-EPIC 的收敛时间与 BGP 相比均有所降低，且 MFT²-EPIC 总体收敛时间降低了 19%.

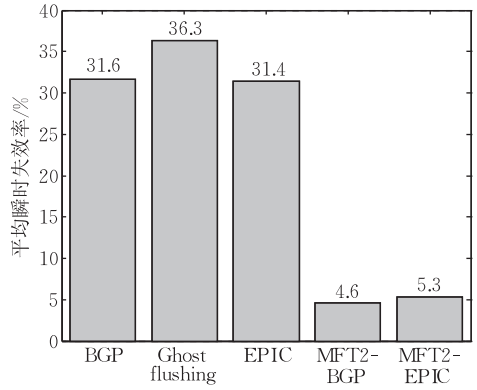
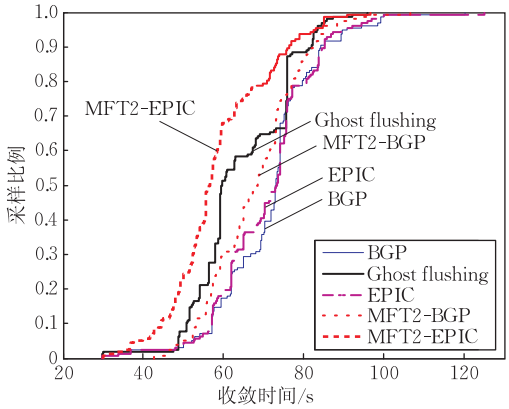
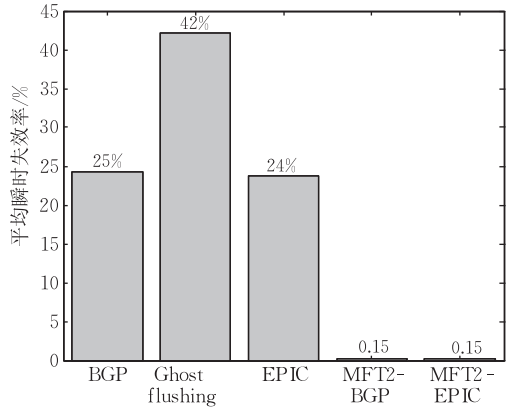


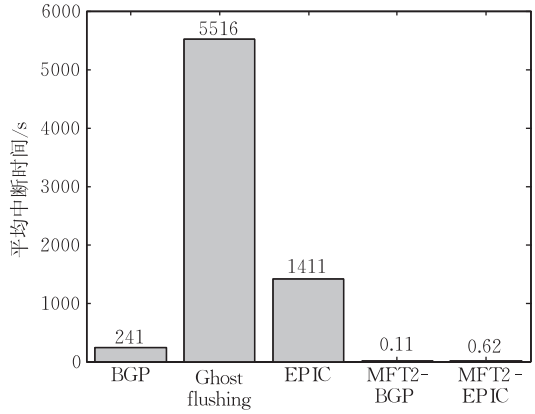
图 11 双链路失效下平均瞬时失效率(♯PID=2)



(a) 双链路失效收敛时间(♯PID=3)



(b) 双链路失效下平均瞬时失效率(♯PID=3)



(c) 双链路失效下平均中断时间(♯PID=3)

图 12 多失效下协议转发性能

从图 12(b)、图 12(c)中可知，在这种最为极端的失效情场景下，增加路径标识符数量可以获得更充分的多样性路径，容忍多失效的能力得到了进一步的提高和验证，从图中可知 MFT²-BGP、MFT²-EPIC 平均失效率仅为 0.15%，中断时间仅分别为 0.11 s 和 0.62 s，出现失效的原因也是由于“Gao-Rexford”策略过滤导致部分节点并没有获得多样性路径.

以上实验表明，MFT²-BGP 以相对合理的消息开销构造了结构化的多转发树，且在构造过程中对收敛时间几乎无影响. 当出现链路失效时，MFT²-BGP、MFT²-EPIC 不仅有效降低了收敛时间，而且通过报文在多转发树之间的切换，降低了失效引起的转发中断；且由于构造的多转发树的结构化特性，更易于实现对多失效的处理.

7　结论与未来工作

提高域间路由协议的生存性、实现无中断的转发是网络体系结构研究的重点之一，也是网络作为基础设施、新兴应用发展的需求. 与 BGP 协议以及一些加速收敛的协议相比，MFT²-BGP、MFT²-EPIC 以相对较低的消息开销获得了多样性路径，给了用户更大的路径选择权利[14]以满足 QoS 需求，可以针对特定标识符的路径实现细粒度的策略控制，避免了由于修改路由策略而造成的环路以及报文丢失. MFT²-BGP 也具有兼容性，遗留路由器忽略路径标识符即可进行正常的协议收敛和转发. 当出现链路失效后，MFT²-BGP、MFT²-EPIC 利用获得的多路径保证无中断转发，在不限制协议的动态性时进一步降低了收敛时间，当对瞬时失效进行抑制时，网络的稳定性得到增强. 下一步的工作需要将该模型进行扩展，将多路径的快速恢复协议扩展到 iBGP 协议中，并进一步考虑路由恢复所带来的流量转移可能造成的拥塞，以优化流量负载均衡分布.

参　考　文　献

[1] Markopoulou A, Iannaccone G, Bhattacharyya S et al. Characterization of Failures in an Operational IP Backbone Network. IEEE/ACM Transactions on Networing, 2008, 16(4): 749-762

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

[3] Oliveira R, Zhang B, Pei D et al. Quantifying path exploration in the Internet. IEEE/ACM Transactions on Networking, 2009, 17(2): 445-458

[4] Labovitz C, Ahuja A, Bose A et al. Delayed internet routing convergence. IEEE/ACM Transactions on Networking, 2001, 9(3): 293-306.

[5] Katz-Bassett E, Madhyastha H V, John J P et al. Studying black holes in the Internet with hubble//Proceedings of the USENIX Symposium on Networked Systems Design and Implementation. San Francisco, USA, 2008: 247-262

[6] Wang F, Mao Z M, Wang J et al. A measurement study on the impact of routing events on end-to-end internet path performance//Proceedings of ACM SIGCOMM. Pisa, Italy, 2006: 375-386

[7] 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

[8] Chandrashekar J, Duan Z, Zhang Z et al. Limiting path exploration in BGP//Proceedings of the IEEE INFOCOM. Miami Florida, USA, 2005: 2337-2348

[9] Xu W, Rexford J. MIRO: Multi-path interdomain routing//Proceedings of the ACM SIGCOMM. New York, USA, 2006: 171-182

[10] Kushman N, Kandula S, Katabi D et al. R-BGP: Staying connected in a connected world//Proceedings of the USENIX Symposium on Networked Systems Design and Implementation. Cambridge, USA, 2007: 341-354

[11] Motiwala M, Elmore M, Feamster N et al. Path splicing//Proceedings of the ACM SIGCOMM. Seattle, USA, 2008: 27-38

[12] John J P, Bassett E K, Krishnamurthy A et al. Consensus routing: The Internet as a distributed system//Proceedings of the USENIX Symposium on Networked Systems Design and Implementation. Berkeley, USA, 2008: 351-364

[13] Ganichev I, Dai B, Godfrey P B et al. Yamr: Yet another multipath routing protocol. ACM SIGCOMM Computer Communication Review, 2010, 40(5): 13-19

[14] Xiaowei Y, David W. Source selectable path diversity via routing deflections//Proceedings of the ACM SIGCOMM. Pisa, Italy, 2006: 159-170

[15] Edard M M, Finn S G, Barry R A et al. Redundant trees for preplanned recovery in arbitrary vertex-redundant or edge-redundant graphs. IEEE/ACM Transactions on Networing, 1999, 7(5): 641-652

[16] Kini S, Ramasubramanian S, Kvalbein A et al. Fast recovery from dual link failures in IP networks//Proceedings of the IEEE INFOCOM. Rio de Janeiro, Brazil, 2009: 1368-1376

[17] Ermolinskiy A, Shenker S. Reducing transient disconnectivity using anomaly-cognizant forwarding//Proceedings of the ACM SIGCOMM workshop on Hot Topics in Networks. Calgary, Canada, 2008: 91-96

[18] Van den Schrieck V, Bonaventure O. Routing oscillations using BGP multiple paths advertisement. Internet IETF draft, draft-vandenschrieck-bgp-add-paths-oscillations-00.txt, 2007

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

[20] Griffin T G, Shepherd F B, Wilfong G. The stable paths problem and interdomain routing. IEEE/ACM Transactions on Networking, 2002, 10(2): 232-243

[21] Mülbauer W, Maennel O, Uhlig S. Building an as-topology model that captures route diversity//Proceedings of the ACM SIGCOMM. Pisa, Italy, 2006: 195-206

[22] Wang F, Qiu J, Gao L et al. On understanding transient interdomain routing failures. IEEE/ACM Transactions on Networing, 2009, 17(3): 740-751

[23] Sahoo A, Kant K, Mohapatra P. Improving packet delivery performance of BGP during large-scale failures//Proceedings of the Global Communications Conference. Washington, USA, 2007: 1850-1854

[24] Mahajan R, Wetherall D, Anderson T. Understanding BGP misconfiguration//Proceedings of the ACM SIGCOMM. Pittsburgh, Pennsylvania, USA, 2002: 3-16

[25] Fall K, Iannaccone G, Ratnasamy S et al. Routing tables: Is smaller really much better?//Proceedings of ACM SIGCOMM Workshop on Hot Topics in Networks. New York, USA, 2009: 1-6

[26] Dimitropoulos X, Krioukov D, Vahdat A et al. Graph annotations in modeling complex network topologies. ACM Transactions on Modeling and Computer Simulation, 2009, 19(4): 1-29

[27] Pei D, Zhang B, Massey D et al. An analysis of convergence delay in path vector routing protocols. Computer Networks, 2006, 50(3): 398-421

HU Qiao-Lin, born in 1979, Ph. D.. His current research interests include delay/disruption tolerant network, network survivability and network virtualization.

PENG Wei, born in 1979, Ph. D., associate professor. His research interests include mobile network and routing protocol.

CHEN Xin, born in 1981, Ph. D.. His research interests focus on network security.

SU Jin-Shu, born in 1962, professor, Ph. D. supervisor. His current research interests include computer network and information securtiy.

Background

Recent widely deployed business applications such as VoIP, online games, video conferencing, VPN, tele-medicine, are very sensitive to packet loss and network delay. Unfortunately, Many measurements on internet have shown that network failures are very common and have serious impact on these applications. However, BGP, the current de facto inter-domain routing protocol are well known to suffer slow convergence when network topology and policy changes. Nevertheless, BGP ofen suffers noticeable transient routing loop and loss of connectivity during periods of convergence, which degrading the forwarding performance of data plane. Many researches have sought to reducing forwarding disruption. However, Some researches focous on speeding up the BGP routing convergence faces inherent limitations given the large scale of network and stringent demands of application. Others precompute backup paths for ensuring continuous connectivity at the extra cost of storage and computing multipath. These protocol can't gurantee every AS discovery diversity path. Espeically, because of lacking of consistency among these precompute backup paths, which may result forwarding loop in the condition of multiple links failure.

In this paper, We design a new interdomain protocol, MFT^2-BGP. It constructs policy-compliant multiple forwarding trees using path identifiers with low message processing overhead, which makes every AS discovery diversity path, AS to freely switch forwarding path to achieve disruption-free forwarding without comprising the routing dynamic behavior in the presence of transient failure scenario. MFT^2-BGP reduces the convergence time by carrying "root cause notification" in update message and reduces the churn of routing system by suppress non-necessary routing updates. MFT^2-BGP fully exploits the redundancy of underlying topology to improve internet reliability and reduce the forwarding disruption effectively.

Our work is supported by the National Natural Science Foundation of China under Grant No. 61070199, the National Grand Fundamental Research 973 Program of China under Grant No. 2009CB320503.