ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.9, September 2006, pp.1922−1932
DOI: 10.1360/jos171922

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

域间路由系统自组织特性*

卢锡城, 赵金晶+, 朱培栋, 董 攀

(国防科学技术大学 计算机学院,湖南 长沙 410073)

Self-Organization of Inter-Domain Routing System

LU Xi-Cheng, ZHAO Jin-Jing+, ZHU Pei-Dong, DONG Pan

(School of Computer, National University of Defense Technology, Changsha 410073, China)

+ Corresponding author: Phn: +86-731-4574606, E-mail: misszhaojinjing@hotmail.com, http://www.nudt.edu.cn

Lu XC, Zhao JJ, Zhu PD, Dong P. Self-Organization of inter-domain routing system. *Journal of Software*, 2006,17(9):1922−1932. http://www.jos.org.cn/1000-9825/17/1922.htm

Abstract: The inter-domain routing system is a complex macrosystem just like the Internet, and the self-organization theory is the efficient utility for studying complex system. This paper analyzes the intrinsic rules and behavioral exhibitions of inter-domain routing system from the view of self-organization, and evaluates the mending methods to BGP protocol for improving the scalability, convergence, stability and security of inter-domain routing system, in order to extract good experience and find out the deficiency. Based on the development forecast of BGP, the rules and techniques of using the self-organization character to solve the inter-domain system problems are presented.

Key words: self-organization; inter-domain routing system; BGP; scale-free; small world

摘 要: 域间路由系统与 Internet 一样是一个复杂巨系统.自组织理论是当前对于复杂性系统研究的重要成果,是研究复杂系统的有效工具.所以,从自组织特性的角度分析了域间路由系统的内在规律和外在表现,并且评价了为了改善域间路由系统的扩展性、收敛性、稳定性和安全性而对 BGP 协议进行改进的各种方法.在对 BGP 的发展趋势进行预测的基础上,给出了利用自组织特性解决域间路由系统问题的指导原则和几种可行的方法.

关键词: 自组织;域间路由系统;BGP;无尺度;小世界特性

中图法分类号: TP393 文献标识码: A

基于 BGP (border gateway protocol)的域间路由系统作为 Internet 的核心设施,不但是传递网络可达信息的基本机制、自治系统互连的纽带和 ISP(Internet service provider)实现策略控制的主要手段,而且对 Internet 的演化起着关键的作用.目前,域间路由系统在扩展性、收敛性、稳定性、健壮性和安全性等方面存在着诸多问题,对 Internet 的性能和安全造成不良影响,并会制约下一代互联网的健康发展.

已有的研究工作基于传统路由系统的严格层次模型,采用静态的基于图论的方法,没有很好地把握域间路

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2005AA121570 (国家高技术研究发展计划(863)); the National Grand Fundamental Research 973 of China under Grant No.2005CB321801 (国家重点基础研究发展规划(973))

Received 2005-10-30; Accepted 2006-04-20

由系统的拓扑规律和动态行为模型,域间路由系统的许多问题没有得到彻底的解决.随着 Internet 的规模扩展和商业化进程的加速,域间路由系统表现出了开放复杂巨系统的特性.本文基于复杂系统理论,从自治系统互连的无尺度(scale-free)特性和 ISP 之间交互的自组织规律出发,系统地分析了已有的改善域间路由系统性能和安全的方法与机制,吸取经验,分析不足;并试图从域间路由系统的基本规律入手解决 BGP 存在的各种问题,为构造安全、可信、可控、可管的下一代互联网、促进下一代网络的持续健康发展做出贡献.

本文第 1 节介绍自组织特性的基本理论,分析 Internet 的自组织特性.第 2 节从实际运行和理论研究两方面总结域间路由系统自组织特性的外在表现和发展走向.在第 3 节中,从性能和安全等几个方面研究 BGP 协议暴露出的问题和目前的一些解决方案,并指明域间路由系统的发展趋势.根据现有的解决方案,在第 4 节中,从域间路由系统自组织性和实际应用的角度出发,给出利用自组织特性进行域间路由系统研究的基本方法.最后总结全文.

1 自组织的概念和 Internet 自组织特性分析

1.1 自组织理论研究

自组织理论是当前对于复杂性系统研究的重要成果,是研究复杂系统的有效工具.它出现在科学领域的各个方面,然而却没有一个普遍接受的定义.从系统论的观点来说,自组织是指一个系统在内在机制的驱动下,自行从简单向复杂、从粗糙向细致方向发展,不断地提高自身的复杂度和精细度的过程.在通信网络领域,Prehofer 和 Bettstetter 在 2005 年 7 月的 IEEE Communication Magazine 上给出了比较确切的定义[1]:系统由多个实体组成,如果这个系统是自组织的,那么它具有一定的结构和功能.结构指实体以特定的方式组织,并且彼此之间以某种方式沟通;功能是指整个系统完成一个特定的目标.

自组织系统的基本运行机制遵守以下规则[2]:

- 信息共享:系统中每一个单元都掌握全套的“游戏规则”和行为准则;
- 单元自律:自组织系统中的组成单元具有独立决策的能力;
- 短程通信:每个单元在决定自己的对策和行为时,除了根据其自身的状态以外,往往还要了解与它临近的单元的状态,单元之间通信的距离比起系统的宏观特征尺度来要小得多,而所得到的信息往往也是不完整和非良态的;
- 微观决策:每个单元作出的决策只关乎自己的行为,所有单元各自行为的总和,决定整个系统的宏观行为;
- 并行操作:系统中各个单元的决策与行动是并行的,并不需要按什么标准来排队以决定其决策与行动顺序.

应用自组织理论可以具备更强的驾驭复杂性的能力.非常复杂的行为模式可以由按自律原则组织起来的、大量相互作用的、相对简单的单元来实现.由于自组织系统的运行和演化是基于大量单元各自的微观决策,而少量单元的决策失误或者损毁并无关宏旨,因而这类系统具有更强的鲁棒性和适应环境扰动的能力.自组织系统一旦开始运行,就具有“自提升”的功能,在内部机制的作用下不断地优化其组织结构,完善其运行模式.

1.2 Internet的自组织特性分析

计算机网络只是一类人造大型网络,很多特性与生物网络和社会网络类似.借用物理学和系统科学的理论和方法揭示计算机网络演化规律和行为特征,近年来出现了很好的成果.例如,在 Nature 和 Science 上有多篇这方面的论文.

Internet 等大型计算机网络表现出了开放复杂巨系统的特性.从控制平面上看,拓扑结构上呈现出无尺度特征,协议实体之间的关系上表现出自组织规律;从数据平面上看,流量行为表现出自相似特点,多个路由器节点之间的流量的控制与管理可以采用网络演算(network calculus)理论进行刻画与分析.

认为 Internet 是一个复杂巨系统,可以从以下几个方面论证.“巨”,指的是子系统数目巨多.从数据平面上

讲,Internet 主机和用户多达数亿;从控制平面上看,自治系统数目 2001 年有 11 000 个,到 2005 年 9 月达到 20 500 个;核心 BGP 路由表 2001 年有 120 000 项,到 2005 年 6 月达到 205 600 项.“复杂”,指的是系统与子系统之间存在高度非线性的内部关系结构.规模扩展的同时,Internet 拓扑的复杂性也有非常突出的表现.新 ISP 不断涌现,ISP 多种对等关系和多宿主连接的建立,使网络拓扑变得更加密集,结构不再表现为简单的层次性,而是有明显的扁平化趋势.另外,协议体系在垂直方向上呈现多样化的层次结构,水平方向上以地域和功能为标准进一步形成分布和多级的结构;在业务性质上表现为多种业务的集成与综合,业务量突发性明显,不同业务的 QoS 要求不同;网络节点间、节点与数据分组间由于协议而产生的非线性作用以及用户之间的合作与竞争,使网络行为呈现出相当的复杂性并且难以预测.互联网的“开放”不但体现在采用统一的协议任意节点可以联入系统,更体现在系统与环境的交互.Internet 是人在其中的与社会系统紧密耦合的复杂巨系统.

Internet 表现出非常强的自组织性,有不少度量都接近理论的最佳值.1999 年,Faloutsos[3]兄弟三人对 BGP 数据和实时测量数据进行分析,发现 Internet 拓扑存在着幂率(power-law)特征.相对负指数分布,Internet 自治系统互联度数的概率分布曲线表现出拖尾(heavy-tailed)性.现在对 Internet 拓扑幂率特性的研究[3−6]如火如荼,因为这是分析 Internet 特性及设计协议的基础.

“小世界(small-world)”是各种网络中普遍存在的现象.社会网络也存在“六度分离”的现象,即每个人只需要很少的中间人(平均 6 个)就可以与全世界的人建立起联系.在 AS 级和路由器级,因特网的特征路径长度都接近或小于随机图,聚集系数比随机图大 3~5 个量级.这表明因特网有明显的 Small-World 特性.

Internet 是一个表现出简单特性的复杂巨系统,其原始结构特性跨越了多个科学领域.例如:网络传输特性可以看作时间序列;网络流量的调节用到开环和闭环控制的方法;各个层次和各个侧面的网络连接涉及到图论的知识;用户和协议实体的行为是形式规则的约定;单元之间的竞争涉及到博弈论的知识;ISP 的组织行为涉及到社会科学的范畴.与其他复杂巨系统比较,Internet 比较透明,这可能是因为其进化历史只有几十年的缘故.Internet 的发展经历着从层次结构(hierarchy)到无序互连(anarchy)、从事先规划(planned)向自组织(self-organizing)、从受控(controlled)到自我管理(self-governing)的变化历程.观察和思考作为开放复杂巨系统的计算机网络,必须把握并利用它的整体性规律.

2 域间路由系统的自组织特性

2.1 域间路由系统自组织特性的外在表现

2.1.1 幂律特性——Power-Law

幂律网络是动态演化和自组织的结果.网络从小到大成长,遵循偏好连接规律.1999 年,Faloutsos[3]兄弟三人对 BGP 数据和实时测量数据进行分析,发现 Internet 拓扑存在着幂律特征,即 $P(k)\sim k^{-\gamma}$,$2\le\gamma\le2.5$.相对负指数分布,Internet 自治系统互联度数的概率分布曲线下降得比较缓慢,表现出拖尾(heavy-tailed)特点.尽管大量节点具有较小的度,但是具有很高度数的节点不是特别少.目前,平均 AS 的互连度为 5~6 之间,但有些 AS 的度达到数百乃至几千.图 1 为 CAIDA[7]在 2004 年 5 月统计的 AS 的节点度分布图,其中最大的节点度数为 1 071,而平均节点度数只有 6.34 左右.图 2 为 ISP 之间连接的二维图形,它与相同规模的随机图的连接性很不相同,整个网络的节点连接度分布遵守幂率特性.

2.1.2 小世界特性——Small World

Small World 的两个主要参数是集群系数和平均路径长度.根据 2005 年 3 月 30 日的统计数据表明:AS 级的平均路径长度值为 2.853 1,集群系数达到 0.452 4.而在 2004 年 1 月 1 日,这两个参数值分别为 3.102 4 和 0.268 5.图 3 给出了在这段时间内 Internet 的 AS 和链路数目的增长曲线.

节点的近似增长函数为 $y=ax+b$,其中 $a=10.9$,$b=16700$;链路的近似增长函数为 $y=ax+b$,其中 $a=82.3$,$b=46700$.可以看出,它们都以近线性的速率增长.然而,在节点和链路都大量增长的情况下,整个系统的平均路径长度却急剧变小,集群系数急剧变大,小世界特性更加明显.

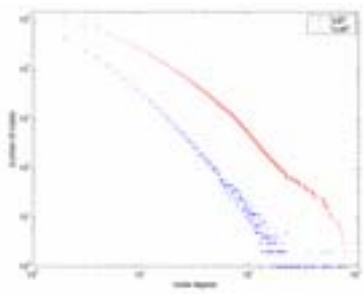

Fig.1 Degree distribution of Internet

图 1 Internet 节点度分布曲线

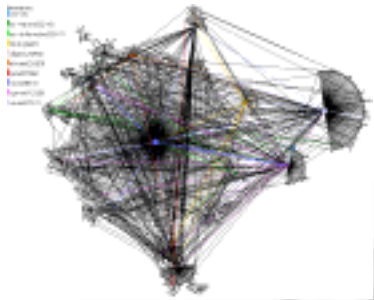

Fig.2 Conjunction graph of ISPs

图 2 ISP 之间的连接图

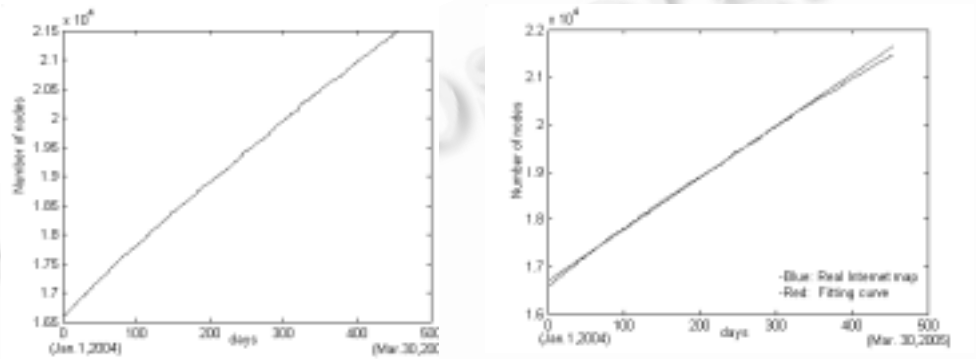

Fig.3 Increase of the ASs number and links number in the Internet from 2004.1.1~2005.3.30

图 3 2004.1.1~2005.3.30 间 Internet 上的节点增长曲线和链路增长曲线

2.1.3 AS 间的商业关系

一些研究人员将 ISP 之间的关系分为 4 种[8]:customer-to-provider(客户-提供商),provider-to-customer(提供商-客户),peer-to-peer(对等-对等)或者 sibling-to-sibling(同胞-同胞).域间路由系统是以 AS 为基本节点构成的自组织系统.单个 AS 体现出 ISP 的意志,但整个域间路由系统的运行却没有统一的管理.节点是高度自治的,各个 ISP 完全独立决策,并可在所辖 AS 内实施策略.AS 层次的自组织性对域间路由系统的构造、性能和安全部署具有重要的影响.Internet 所表现出的无尺度结构特性主要体现在 AS 的互连上,是 AS 局部利益极大化决策的结果,选择连附的上层 AS 的概率正比于提供商能够免费到达的网络或提供商互连的度数.多宿主是网络演化基本过程的结果,选择另外一个提供商的概率是呈指数递减的.ISP 互联的主要动力是减少转发代价.

2.2 域间路由系统自组织性的走向

2.2.1 好的方面

a) 结构扁平化

Kleinrock 在 1977 年提出层次路由的基本模型(简称 KK 模式)[9],并由此奠定了 Internet 路由的基础.KK 模式要求保证严格的层次性、严格的层次拓扑和严格层次性的寻址结构.Internet 发展的早期,网络结构比较稀疏,KK 模式还是比较有效的.但随着网络规模的扩大,域间模型无法严格地遵循 KK 模型.1988 年,Waxman 提出 Waxman 模型[10],即节点是随机放置在尺度为 L 的二维空间内.在当时,节点数还不足 10 000 个时,这种模型虽然简单却也反映了当时网络的状况.1997 年,由 Zegura 提出了一种传输-末端模型[11,12],这也是一种路由器级模型.他认为网络上的设备要么属于传输域,要么属于末端域.传输域只负责转发报文,是由路由器互联成的一个二维空间.末端域只负责发送和接收报文,连接在传输域上.传输-末端模型是一种层次模型结构,如图 4 所示.

从层次模型到幂率模型的转变,是 Internet 规模逐渐变大、复杂性逐渐加深的结果,同时也是 Internet 商业

化作用的产物.各个 ISP 由于商业关系的不同,所处地位的不同,在网络中有着多重身份,使得网络连接也不再是以往单一的层次式结构,而是使得整个域间路由系统成为大量 mesh 的聚集体,呈现出一种扁平结构.图 5 是现在域间路由系统的一种抽象表示.

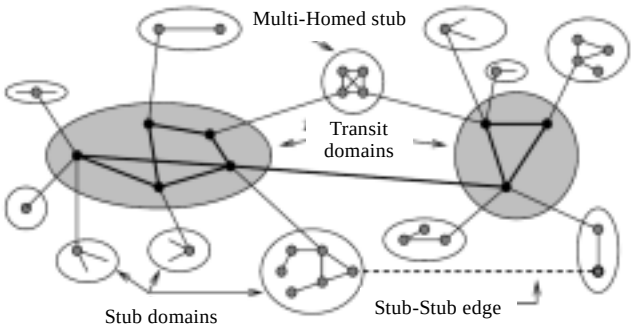

Fig.4 Transit-Stub model of ASs

图 4 过渡-分支模型结构

Fig.5 Current structure of the inter-domain system

图 5 现在的域间路由系统结构模型

基于层次网络划分的路由结构缺乏现实的扩展性[13].文献[14]中指出:采用 KK 模式会导致路径变长(path inflation),可能是最短路径的 15 倍.域间路由系统扁平结构的一个显著优点是网络的健壮性.根据统计,每时每刻大约有 0.3%的路由器失效,而域间路由系统仍然能够正确地工作.Albert 和 Barabasi 在 2000 年时研究表明:即使移走 60%的 Internet 节点,也不会对平均路径长度产生任何影响.即使从 Internet 路由器中随机选择的失效节点比例高达 80%,剩余的路由器还是能组成一个完整的集群并保证任意两个节点间存在通路.而在随机网络中,若有较大部分的节点被除去,网络必然溃散成彼此无法通信的小型孤岛.总的来说,无尺度网络对意外故障具有惊人的强韧性.

b) 信息扩散

域间路由系统的小世界特性主要体现在网络的平均路径短,集群系数比随机网络要大很多,使得网络的连接性比随机网络或者层次网络要好几十倍.这样有利于信息的传播,使得在很短时间内经过很少的跳数,就能够使信息从一个 AS 传递到另一个 AS.所以,我们在进行路由的时候要尽量选择最短路径,这样才能够充分利用网络的小世界特性.还可以利用小世界特性,使得 AS 之间自主地形成一个个信任联盟,进行安全决策.

c) 安全部署

由于域间路由系统自身的无尺度特性,使得系统的对于随机错误和失败的健壮性很好,面临的最主要的威胁是来自对关键节点的攻击和错误.所以在进行安全部署的时候,可以着重保护这些关键节点.

并且,对于域间路由的攻击,可以在只占少数的关键节点上部署过滤监测机制.这样可以在投入很低、时空开销很少的情况下,充分利用网络本身的特点,顺应网络发展的趋势,达到事半功倍的效果.

d) 商业关系

AS 之间在利益的驱动下,自动地形成了 customer-to-provider,provider-to-customer,peering-to-peering,sibling-to-sibling 这 4 种关系,并在相互协商的基础上,为了达到彼此的利益最大化,制定双方都遵守的规则.这样,在网络经济的驱动下,各个 AS 自主合作,使得整个域间路由系统能够健康、稳定地发展.

2.2.2 坏的方面

a) 不合作问题和信息隐藏

域间路由系统的各个 AS 之间不合作是由于自组织系统的自身特点,如短程通信、局部决策、自私性等特性所决定的.导致 ISP 之间不合作的原因可以从主观和客观两个方面进行分析:从主观上讲,由于各个 ISP 之间存在着商业关系,那么它们也应该遵循经济学的规律.从博弈论的思想来讲,各个 ISP 都是从自身的利益出发,目的都是为了获得利益最大化.所以在这种原则之下,在没有规则调和约束的情况下,无法获得全局利益的最大

化;从客观上来讲,由于自组织网络本身信息隐藏和局部决策的特性,使得各个 ISP 无法获得全局的视图,在决策的时候也无法考虑其他 ISP 的情况,所以也难免会出现这种被动的不合作问题.

由于 AS 的自私性和信息隐藏,影响了域间路由系统整体效能的优化及其为数据网络提供的服务质量,使端到端的数据传输路径没有实现全局最优,甚至出现路由收敛慢和不收敛的情况[15−19].

b) 抗毁性

由于域间路由系统的幂率特性,使得存在少量的集散节点,它们具有很高的连接度,在整个系统中有着举足轻重的作用.对集散节点的依赖,使得少数消息灵通的黑客只要攻击一些集散节点,就足以搞垮整个域间路由系统.无尺度网络的这一致命缺陷,引发了这样一个问题:到底有多少集散节点是必不可少的?研究表明:总的来说,只要有 5%~10%的集散节点同时失效,就足以搞垮系统.实验显示:一次有组织的协同攻击,只要去除掉若干个集散节点(先去除最大的,再去除次大的,依此类推),就足以造成重大破坏.因此,为了避免因恶意攻击带来网络的大规模破坏,最有效的办法就是保护好集散节点.

c) 病毒传播

过去数十年间,无论是流行病学家还是市场营销专家都在大力研究扩散理论.研究结果指出:一种传染病要在人群中传播开来,必须要跨越某一临界值.任何病毒、疾病或时尚的感染力一旦低于这个临界值,将不可避免地自行消亡;而一旦超过临界值,就会呈指数增长,最终传遍整个系统.然而在无尺度网络里,不存在上面所说的临界值.这就意味着所有病毒都可在网络中传播和长期存在,即便是那些传染力很低的病毒也是如此.因为集散节点会连接到很多其他节点,所以任何一个遭受病毒入侵的节点,都将连带感染至少一个集散节点.而一旦有集散节点被感染,它就会把病毒传播给众多的其他节点,当中也包括其他的集散节点,这就导致了病毒在整个网络里的传播.

3 域间路由系统面临的问题及发展趋势

3.1 域间路由系统面临的问题及研究现状

随着 Internet 规模的扩展和商业化进程的加速,域间路由系统的规模也在膨胀,AS 号和核心网络 BGP 转发表项的增长都非常显著.规模扩展的同时,Internet 拓扑的复杂性在路由系统上也有非常突出的表现.新的 ISP 不断涌现,ISP 多种对等关系和多宿主(multi-homing)连接的建立,使网络拓扑变得更加密集,结构不再表现为简单的层次性,而是有明显的扁平化趋势.2004 年 7 月,BGP 路由表有 227 750 项,转发表有172 820项,比值为1.317 9;到 2005 年 6 月,比值为 2.787 1,ISP 之间有更多的可达路径.多宿主技术使小型网络的前缀无法有效聚合,不但加剧了路由表增大的趋势,而且会增加整个路由系统管理的复杂性.2005 年 6 月,24 位前缀已占 44.05%,超过 24 位的达 20.2%之多[20].

域间路由系统的复杂性还表现在 ISP 互联策略的多样性上.BGP 是基于策略的路由(PBR),路由策略不但影响本自治系统的路由决策,还控制路由信息在 ISP 之间的传播.BGP 不是基于 AS-Path 进行简单的最短路径选择,而是综合利用 LP(local preference),MED(multi-exit discriminator),Origin,Router ID 等多种属性,出于商业利益、流量工程、安全等多个方面的要求进行路由控制.另外,IBGP 的互连往往采用联邦(confederation)、反射器(reflector)结构,使 BGP 的配置和 ISP 之间的交互更加复杂.

域间路由系统规模的扩展、结构的密集化、互联关系的复杂化和路由策略的多样性,带来了严重的扩展性、收敛性、稳定性、安全性问题,并给 Internet 的性能、安全和健康发展带来不利影响.下面我们分别对这些问题出现的原因以及研究现状进行分析.

3.1.1 扩展性

衡量路由系统的扩展性有两个标准:路由器和链路的资源消耗.一般来说,BGP 路由器只是增量式地更新消息,所以它的链路消耗并不是很巨大.BGP 路由器的资源消耗主要分为两大部分:1) BGP 会话建立、路由选择、路由信息处理以及路由更新处理的 CPU 处理消耗;2) 存储路由信息和多路径的存储消耗.这样来看,直接对 BGP 收敛性造成影响的两个方面就是 BGP 会话的数目和路由策略的复杂性.

现在,对 BGP 扩展性问题的研究主要集中在限制路由表数目的增长、AS 号的耗尽等方面.但由于多宿主技术的广泛采用,取得的效果并不太明显;另一个研究方面就是在 IBGP 的全互连结构如何减少通告数目和路由抖动的影响范围;还有就是如何在保证路由质量的前提下,减少 BGP 路由器维护的会话数目.

3.1.2 收敛性

近年来的工作表明:BGP 的收敛问题也是现今的主要研究方向.现在收敛性方面的研究主要从两个方面进行:能否收敛问题以及慢收敛问题.

保证收敛的解决方案分为 3 类:1) 预先检查路由策略中是否含有冲突[15],这需要向 ISP 中心路由注册机构注册路由策略,但是本地策略一般不会共享,并且 Griffin 已经证明检测路由策略的收敛性是 NP 完全问题;2) 约束路由策略或者约束路径选择.Gao[21]提出了根据 ISP 商业层次关系选择策略的准则来保证收敛性,但 ISP 都希望能够自由选择策略;另外一种解决方法[22]不限制选择策略,但是每个节点需要通过极其复杂的计算实时地约束路径的选择,不具有可行性;3) 携带路径历史信息,这会增加消息和存储开销,并且路径以前是环路并不表示现在还是,所以对链路失败引起的不收敛问题不适用.

对于收敛时间问题,2000 年 Labovitz 等人通过测试和理论分析[17]得到了 BGP 失败导致的路由重新收敛的时间上下界,平均时间大约是 3 分钟,但由于路由表振荡,也可能会达到十几分钟.在文献[23]中给出了 BGP 一个更真实的模型,考虑了 MRAI 定时器、不同的拓扑、策略设置等因素,认为 BGP 的收敛时间由 MRAI 定时器的设置、源与目的之间被选路径的最长长度以及所有可选路径所决定.

将现有的改进方法分为两类:本地通告不可用路径 LNPU 以及远程本地通告不可用路径 RNPU.前者包括 SSLD,WRATE,GF[24],CA[25]等;后者有 RCN[26],RON,FESN[27]等.Griffin 等人认为:SSLD(send side loop detection)只能很有限地减少收敛时间,并且无法保证每一种特定的网络拓扑都存在一个最优的 MinRouteAdver 值来减少收敛时间;在 GF(ghost flushing)中,尽管节点收到了一些关于路由无效的隐式消息,但不能保证无效的路由一定不会被节点所使用;CA(consistency assertions)没有定义慢收敛的类型,因此无法比较多跳之外节点的不一致性.无论是 RON(route change origin)还是 RCN(BGP with root cause notification),都不能在节点的多个邻居 fail-stop 的情况下保证无效路由不被使用.一致性断言和 RCN 在 AS 之间传播 entry-router-ids,而这是 AS 级的属性,所以会使本地 AS 内的变化传播到其他的 AS 中,有时会引起其他的不稳定性.FESN(forwarding edge sequence number)考虑了 IBGP 对域间收敛性的影响,使用转发边系列号代替 BGP-RCN 结点系列号,其 T_{down} 和 T_{long} 的收敛延迟与 RCN 类似.

3.1.3 稳定性

路由不稳定性也就是“路由抖动”,是由网络可达性和拓扑信息的快速变化而引起的.网络不稳定会有 3 种主要影响:1) 增加报文丢失率;2) 增加网络收敛延时;3) 增加网络的额外开销(内存、CPU 等).

当路由器收到一个拓扑变化时,会向它所有的 peer 发送撤销消息,无论它们之前是否向这个 peer 发送过这个路由的通告,发送的数目达到 $O(N\times U)$.U 是更新报文的数目,N 是对等体的个数.根据文献[28]中提供的数据,无缺省路由表一般包含 45 000 个前缀.然而在 Internet 的核心,每天观测到的路由前缀更新有 300 万~600 万个,每个不稳定和冗余的更新周期大概是 30 秒~60 秒,这对网络资源和处理效率都是巨大的消耗.

解决路由振荡问题的传统方法基本上分为静态和动态两类:静态方法包括路由策略的自动分析和 AS 之间的相互协调;动态方法主要是指扩展 BGP,使其可以动态检测并抑制基于路由策略的振荡.

美国自然基金会发起的 Routing Arbiter 项目设计了全局路由策略的自动分析系统[29],但没有产生很好的效果,路由的不稳定性仍然有不断增加的趋势.文献[30]中定义了 BGP 收敛或发散的若干条件,静态分析 BGP 收敛问题的时间复杂性为 NP-Complete 或 NP-Hard. Griffin[30]用 SPVP 模型描述了 BGP 协议的振荡行为,并针对该模型引入图论知识给出了路由收敛的充分条件.尽管 Griffin 使用了一个非常新颖的方法刻画了 BGP 振荡问题,但其结论本身意义不大.如果使用该结论来处理 BGP 振荡,分析 BGP 收敛问题的时间复杂性也为 NP-Complete 或 NP-Hard. Gao[21,31]利用了自治系统之间的逻辑关系,对处于各种逻辑关系下的自治系统配置策略给出了几个限制,然后利用 Griffin 的结论证明了满足这些限制条件的系统是收敛的.这些限制是针对各个自治系统的,是

一种局部措施,相对于其他方法而言具有较好的可行性.

对于动态方式,主要的研究是能够自动监测可能的路由振荡并自动进行抑制.路由抖动算法 RFD[32]的目标是:1) 减少不稳定性引起的路由器的处理负载;2) 阻止长久的路由抖动;3) 不影响正确运行的路由器的收敛时间.但有时可能会错误地惩罚一些稳定的路由器.路由器只是偶尔失败,很快就会恢复,而当它的惩罚计数器值没有低于重用门槛值,这条路径会一直被抑制.还有一些对它的改进方案,如 SRFD[33],RFD+[34]算法等,都有其无法避免的弱点,因为对于动态方式有两个固有的缺点:1) 不能够彻底消除 BGP 协议的振荡,只能使振荡以“慢动作”运行;2) 抖动事件不会通告给网络管理者,也就无法定位抖动的源.由策略冲突引起的路由抖动会和不可达引起的抖动相同处理,而这是不应该的.

3.1.4 安全性

BGP 作为 ISP 互连的基本手段,面临多种恶意攻击的威胁.例如,黑客组织 L0pht 曾宣称,能在很短的时间内利用 BGP 路由协议的安全问题来搞垮整个 Internet.

有大量的工作从理论的角度试图增强域间路由系统的安全性,然而却很难得以实施.以 BBN 公司提出的 S-BGP 为例,它通过路由更新报文来携带地址证书(AA),验证路由的所有者,通过路由证书(RA)验证使用该路由的邻居 AS 是否授权,从而对 BGP 路由前缀和路径信息进行全面的保护.从理论上分析,它是非常完美的,然而要广泛实施,就要建立一个大范围的域间公共密钥结构,以及在路由信息注册时执行中心路由广播机制或者其他类似的机制,而这是非常困难的.这几年的研究包括 soBGP 扩展[35]、S-BGP 协议扩展[36]以及路由过滤等机制.soBGP 是 CISCO 提出的对 BGP 的扩展,它通过对路由前缀来源(origin)的合法性进行验证,确保转发的前缀来自授权的 AS.路由过滤由对自治系统进行过滤的 Filter-List 和对路由前缀进行过滤的 Distribute-List 组成,并且具有路由聚合功能.这些 BGP 安全机制大都是基于中心的层次性方案,导致实现复杂、代价高昂,无法在实际中部署实施.

3.2 域间路由系统的发展趋势

下一代互联网是基于 IPv6 的新型网络.IPv6 提供了更大的地址空间,地球上每个人可以平摊 5×10^{28} 个.通过分析目前 Internet 域间路由系统面临的问题可以发现:这些问题与 IPv4 的地址空间没有必然联系.域间路由系统的问题主要源于 3 个方面:BGP 协议自身的特殊机制、路径向量路由算法自身的问题、可扩展策略路由的固有特性.这些问题随着 Internet 和路由系统规模的扩展和互连关系的复杂化,才得以充分表现并变得日益突出.所以,下一代互联网的寿命能够有多长,并不完全取决于 IPv6 的地址空间,在很大程度上将依赖于域间路由系统的持久健康发展.

IPv6 针对 IPv4 在地址格式、寻址方式等方面有较大变化,路由的层次性得到一定程度的改善;但是,IPv6 路由表增长的趋势是明确的.研究表明[14,37]:基于 BGP 路由协议现在的工作模式,在网络规模确定的情况下,路由表大小和路由的质量是相互矛盾的.路由节点中保存的路由数量少,得到的路由往往会比最短路由长.所以从这个意义上讲,将来 IPv6 路由表同样会经历快速增长的阶段.就域间路由系统而言,由于它所依附的网络的本质属性没有真正改变,如果不对 BGP 协议进行改进,目前所表现出来的问题在下一代互联网中仍然会非常突出.另外,由于 IPv6 网络普遍实行多地址接口,地址的指派有生命周期,将会使域间路由系统的配置和管理更加复杂.IPv6 环境下,网络 Multi-homing 互连的增加,使宣告到 Internet 核心网络并通过顶层 ISP 传播的路由增加,域间路由系统的收敛延迟加大,影响了网络的伸缩性,降低了路由系统的稳定性.

4 利用自组织特性解决 BGP 存在的问题

我们需要基于系统复杂性理论和 ISP 互连的自组织性,着眼于系统的整体性规律和自组织系统的基本机制来系统解决域间路由系统的扩展性、收敛性、稳定性和安全性问题.衡量一个解决方案是否合理地利用和遵循域间路由系统的自组织特性,首先要先明确系统中主体和外力的概念.域间路由系统的主体是指 AS 的集合以及它们之间的关系;而外力是指为了某种目的,人为施加在主体上的驱动力.我们认为:一个解决方案利用了自组织特性,是指外力顺着 AS 之间关系的趋势施加,其结果是使自组织特性得以体现.而一个解决方案违背了自

组织特性,是指外力的施加没有考虑到 AS 之间关系的方向.

观察和思考作为开放复杂巨系统的域间路由系统,必须把握和利用它的整体性规律,而不能单纯地靠还原论的方法把组件分解,分别分析.例如,对域间路由系统健壮性的研究,既然路由节点被攻击乃至被入侵不可避免,就应该着眼于提高网络整体的健壮性和可生存能力.与 Internet 的安全问题类似,复杂性使域间路由系统网络对抗的非对称性加强,因为防守一方要在巨量的脆弱点上处处设防,而攻击者可以攻其一点.因此,域间路由系统的健壮性同样需要基于整个网络的谋略.对域间路由系统的稳定性,更要从整体考虑.域间路由系统是复杂巨系统,个别事件的触发会影响全局的稳定.随着域间路由系统规模的扩展和互连密度的增加,发展中的复杂性和不可预知的敏感性交错上升,对域间路由系统攻击的扩散效应非常有必要建立系统模型.

利用结构的 Power-Law 和 Small-World 特性,是化解域间路由系统复杂性的可行途径.例如:基于处于第一层的枢纽 AS 建立路由传递的信任关系,可以缩短信任链,提高 BGP 路由安全机制的扩展性;基于一些活跃的枢纽节点,保存足够多的信息,对域间路由系统进行监管;设计基于多个枢纽节点的域间路由信息并行扩散模型.研究复杂巨系统的相变规律,用来探索域间路由系统运行的基本约束,确定无尺度网络上的性能上限.

采用经典的层次路由模型对网络进行多级划分,每级的网络直径都是网络规模的幂律分布函数,这与 Internet 的实际结构特征不符合.研究发现:采用新的路由模式, 对一般结构的图无法实现路由表的大小和所学到的路径长度同时减小;但是,当网络结构变化到域间路由系统这样的 Scale-Free 网络时,出现了相位的迁移(phase transition),可以在一个点上使两者同时出现最佳值.

图论是基本的网络理论,单纯图论的方法,一方面无法刻画动态演化特性;另一方面,由于图的优化往往是 NP 完全问题,往往需要全局信息,因而无法投入实际应用.但是,如果限定 Scale-Free 网络,则有可能得到较简单的方法.同时,基于自组织理论,采用局部决策,一些 NP 完全问题可能转化为简单的比较运算.

采用“自下而上”的设计方法,强化单个 ISP 节点的设计.节点必须能够主动、适时地感知其周围节点的信息,或通过对历史通信的记录学习得到环境信息,增强节点的自主决策水平.MIT 的 Clark 指出:实体之间的争斗(tussle)是整个 Internet 发展的动力.就域间路由系统而言,ISP 之间的利益和行为的冲突同样是域间路由系统演化的动力.通过分析域间路由策略的异常交互表现,我们发现 ISP 之间的冲突源于两个方面:一是 ISP 的主观自私性;二是 ISP 作为分布式系统的单个节点对域间路由系统和整个 Internet 视图的局部性.要通过 ISP 独立的“微观决策”实现域间路由行为的“整体协调”,必须约束自私性、克服局部性.为了克服微观决策的自私性,采用博弈论(game theory)的方法,通过引入各个 ISP 之间的主动或者被动的合作机制来提高全局的最优性.

5 总 结

鉴于域间路由系统的重要性及其对 Internet 发展的基础性作用,国外一些大学对域间路由系统的行为、模型进行了全面研究,以期解决暴露出来的问题,改善运行,并得到了网络运营商和设备制造商的大力支持.学术界和产业界的密切合作,说明这一研究方向是理论创新的源泉.具有重要的应用价值,并且二者能够很好地结合.在 2005 年的 IEEE INFOCOM 和 ACM SIGCOMM 上仍然是研究的热点问题[24−28].以美国的 MIT,AT&T,CAIDA,英国剑桥大学 Griffin 和 APNIC 的 Huston 研究最为活跃.IRTF 成立了 RR(routing research)工作组,对 BGP 路由协议的设计进行全面考察,并成立了未来路由协议扩展性研究小组(RR-FS).

现有工作存在的主要弊端是缺乏理论支持,缺乏对域间路由行为模型的准确认识.

(1) 缺乏相应理论的支撑

域间路由系统基本问题的解决需要理论的支撑.例如,路由系统的扩展性和路由质量之间的关系;信息隐藏是提高路由系统扩展性、改善稳定性的主要机制,但是却延缓了路由系统的收敛过程.如何得到这一对矛盾的内在约束关系;如何得到 BGP 协议收敛的充要条件;协议的收敛性如何实现与配置无关等.如果在基本问题上没有确定的答案,对 BGP 协议所做的修补工作只会使域间路由系统变得更加复杂,行为更加难以确定.即使得到部署,也只能在短期内缓解问题,并没有真正触及根本原因.

(2) 缺乏对域间路由行为模型的准确认识

对域间路由系统的基本问题能否找到有效而彻底的解决方法,在很大程度上取决于对域间路由系统行为模型的准确刻画.现有的域间路由系统的模型往往基于图论的方法来刻画,而图论主要用来处理静态确定的或随机平衡的网络,与域间路由系统的动态演化及自组织特性不符.所以单纯采用数学方法是远远不够的,还需要我们全面地把握域间路由系统的复杂行为来解决问题.

对域间路由系统,不能再当作思维的产物(the mind),而应该像考察其他自然现象一样探索这一人造自然(the nature)的演化规律;不能再使用由简单的还原论(reductionism)方法,而要系统考察其自然发生过程(emergence);自上而下(top-down)的方法要向由底向上(bottom-up)的方法转变;单纯数学的方法(mathematics)要向物理学的方法(physics)变化.这是所有网络技术研究人员都应该深刻感觉到的改变.

References:

[1] Prehofer C, Bettstetter C. Self-Organization in communication networks: Principles and design paradigms. IEEE Communications Magazine, 2005,43(7):78−85.

[2] Alderson D, Willinger W. A contrasting look at self-organization in the Internet and next-generation communication networks. IEEE Communications Magazine, 2005,43(7):94−100.

[3] Siganos G, Faloutsos M, Faloutsos P, Faloutsos C. Power-Laws and the AS-level Internet topology. IEEE/ACM Trans. on Networking, 2003,11:514−524.

[4] Zegura, E Calvert K, Donahoo M. A quantitative comparison of graph-based models for Internet topology. IEEE/ACM Trans. on Networking, 1997,5(6):770−783.

[5] Subramanian L, Agarwal S, Rexford J, Katz R.H. Characterizing the Internet Hierarchy from multiple vantage points. In: Proc. of the IEEE INFOCOMM 2002. 2002. 618−627.

[6] Tauro SL, Palmer C, Siganos G, Faloutsos M. A simple conceptual model for the Internet topology. In: Proc. of the IEEE GLOBECOM 2001. 2001. 1667−1671.

[7] 2005. http://www.caida.org/analysis/topology/

[8] Huston G. Interconnection, peering, and settlements. Internet Protocol Journal, 1999,45(3):136−152.

[9] Kleinrock L, Kamoun F. Hierarchical routing for large networks: Performance evaluation and optimization. Computer Networks, 1977,1:155−174.

[10] Waxman B. Routing of multipoint connections. IEEE Journal on Selected Areas in Communications, 1988,6(9):1617−1622.

[11] Calvert K, Doar M, Zegura E. Modeling Internet topology. IEEE Communications Magazine, 1997,35(6):160−163.

[12] Zegura, E Calvert K, Donahoo M. A quantitative comparison of graph-based models for Internet topology. IEEE/ACM Trans. on Networking, 1997,5(6):770−783.

[13] Eilam T, Gavoille C, Peleg D. Compact routing schemes with low stretch factor. Journal of Algorithms, 2003,46(3):97−114.

[14] Krioukov D, Fall K, Yang X. Compact routing on Internet-like graph. In: Proc. of the IEEE INFOCOM 2004. 2004. 219−220.

[15] Labovitz C, Ahuja A, Wattenhofer R, Venkatachary S. The impact of Internet policy and topology on delayed routing convergence. In: Proc. of the IEEE INFOCOMM 2001. 2001. 537−546.

[16] Yu H, Alaettinoglu C, Jacobson V. Towards Milli-second IGP convergence. IETF Internet Draft: Draft-Alaettinoglu-ISIS-Convergence-00. 2000. http://www.rtg.ietf.org/~fenner/ietf/xml/bibxml3/reference.I-D.alaettinoglu-isis-convergence.xml

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

[18] Varadhan K, Govindan R, Estrin D, Persistent route oscillations in inter-domain routing. Computer Networks, 2000,32(1):1−16.

[19] Wang L, Zhao X, Pei D, Bush R, Massey D, Mankin A, Wu SF, Zhang LX. Observation and analysis of BGP behavior under stress. In: Proc. of the ACM SIGCOMM Internet Measurement Workshop (IMW). 2002. 183−195.

[20] 2005. http://bgp.potaroo.net/as1221/bgp-active.html

[21] Gao L, Griffin T, Rexford J. Inherently safe backup routing with BGP. In: Proc. of the INFOCOM 2001. 2001. 547−556.

[22] McPherson D, Gill V, Walton D, Retana A. BGP persistent route oscillation condition. IETF Internet Draft: Draft-ietf-idr-route-oscillation-00.txt, Work in Progress. 2001. http://citeseer.ist.psu.edu/mcpherson02border.html

[23] Pei D, Zhang BC, Massey D, Zhang LX. An analysis of path-vector routing protocol. Technical Report, TR040009, 2004.

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

[25] Pei D, Zhao X, Wang L, Massey D, Mankin A, Wu FS, Zhang LX. Improving BGP convergence through assertions approach. In: Proc. of the IEEE INFOCOM 2002. 2002. 902−911.

[26] Pei D, Azuma M, Nguyen N, Chen J, Massey D, Zhang LX. BGP-RCN: Improving BGP convergence through root cause notification. Technical Report, TR-030047, UCLA CSD, 2003. http://www.cs.ucla.edu/ peidan/bgp-rcn-tr.pdf

[27] Chandrashekar J, Duan ZH, Zhang ZL, Krasky J. Limiting path exploration in BGP. In: Proc. of the IEEE INFOCOM 2005. 2005. 2337−2348.

[28] Shaikh A, Kalampoukas L, Dube R, Varma A. Routing stability in congested networks: Experimentation and analysis. In: Proc. of the ACM SIGCOMM 2000. 2000. 163−174.

[29] Kumar S, Lee WS. An architecture for stable, analyzable Internet routing. IEEE Network, 1999,13(1):29−35.

[30] Griffin T, Wilfong G. An analysis of BGP convergence properties. In: Proc. of the ACM SIGCOMM 1999. 1999. 277−288.

[31] Gao L, Rexford J. Stable internet routing without global coordination. In: Proc. of the ACM SIGMETRICS. 2000. 307−317.

[32] Cowie J, Ogielski A, Premore B, Yuan Y. Global routing instabilities during code red II and Nimda worm propagation. 2001. http://www.renesys.com/projects/bgp instability

[33] Feldmann A, Maennel O, Mao ZM. Locating Internet routing instabilities. In: Proc. of the ACM SIGCOMM 2004. 2004. 205−218.

[34] Labovitz C, Malan R, Jahanian F. Internet routing instability. In: Proc. of the IEEE INFOCOM 1999. 1999. 218−226.

[35] White R. Securing BGP through secure origin BGP (soBGP). The Internet Protocol Journal, 2003,6(3):15−22.

[36] Kent S, Lynn C, Mikkelson J, Seo K. Secure border gateway protocol (S-BGP). IEEE Journal on Selected Areas in Communication, 2000,14(4):114−117.

[37] Eilam T, Gavoille C, Peleg D. Compact routing schemes with low stretch factor. Journal of Algorithms, 2003,46:97−114.

卢锡城(1946－),男,教授,博士生导师,中国工程院院士,CCF 高级会员,主要研究领域为计算机网络,并行与分布处理,高性能计算.

赵金晶(1981－),女,博士生,主要研究领域为域间路由技术,复杂巨系统理论.

朱培栋(1971－),男,博士,副教授,主要研究领域为路由技术,移动网络,网络安全.

董攀(1978－),男,博士生,主要研究领域为信息安全,MANET.