

基于BGP的域间二维路由方案

A Design of Inter-Domain Destination and Source Routing Based on BGP

中图分类号: TN929.5 文献标志码: A 文章编号: 1009-6868 (2018) 06-0060-005

摘要: 提出了基于边界网关协议(BGP)的域间二维路由方案。该方案在进行路由决策时同时考虑目的地址和源地址,实现灵活和细粒度的流量控制。给出了二维路由的部分应用场景,设计了二维路由的控制层,修改了部分报文格式用来传递二维路由信息,同时保证二维路由与传统路由的兼容;还给出了二维转发表的方案,解决二维转发匹配造成的三态内容寻址存储器(TCAM)空间爆炸的问题。

关键词: 域间二维路由;源地址;报文设计;二维转发表

Abstract: In this paper, an inter-domain destination and source routing scheme based on border gateway protocol (BGP) is proposed. This scheme makes routing decisions considering both destination address and source address, which achieves flexible and fine-grained flow control. Some application scenarios of the scheme are presented. The control layer of the scheme is designed. Some message formats are modified to transfer two-dimensional routing information, and they are compatible with the traditional routing protocol. A practical design of two-dimensional forwarding table is then given, and the problem of ternary content addressable memory (TCAM) space explosion caused by two-dimensional matching is solved.

Key words: inter-domain destination routing; source address; message field design; two-dimensional forwarding table

耿男/GENG Nan¹
金飞蔡/JIN Feicai²
徐明伟/XU Mingwei¹

(1. 清华大学, 北京 100084;
2. 中兴通讯股份有限公司, 广东 深圳 518057)
(1. Tsinghua University, Beijing 100084, China;
2. ZTE Corporation, Shenzhen 518057, China)

由于互联网发展的历史原因,以传输控制协议(TCP)/网际协议(IP)为核心的互联网模型将“尽力而为”的可达性作为网络的首要任务,这使得报文携带的目的地址在路由过程中成为了唯一的决定因素。这种仅依靠目的地址的路由方式,大大限制了对报文转发控制的灵活性。并且随着互联网规模的迅速增长和用户业务的多样化,传统路由协议越来越难以满足许多业务对服务质量

的要求。

陆续出现的多协议标签交换(MPLS)^[1]、软件定义网络(SDN)^[2]等技术,试图寻找除目的地址之外的新的路由计算维度。MPLS通过标签交换提前建立好一条专门的转发路径,指定的流量可以按照MPLS建立的路径进行传输,然而这种源路由方式存在的安全性差、复杂性高以及开销大等问题,使得网络服务提供商(ISP)部署MPLS的积极性不高。SDN彻底解放了路由计算维度的限制,最新的OpenFlow协议^[3]支持44个匹配域,能够实现对网络流量的精细控制;但是这也造成了SDN流表空间爆炸的问题,使得SDN扩展性不高,另外集中式控制方式降低了网络的鲁棒性,因

而广泛部署SDN仍是一个很漫长的过程^[4]。

域间路由是互联网路由体系结构的重要组成部分,自治系统(AS)之间构成的域间路由具有管控难度大、路由因素复杂等特点。由于可扩展性等原因,MPLS、SDN等方案在域间难以实际部署,当前广泛应用的域间路由协议仍然是边界网关协议(BGP)^[5],那么能否在传统BGP的基础上,实现更加灵活的转发控制?

本文设计了一种基于BGP的域间二维路由方案,其在进行路由决策的时候,不仅仅考虑目的地址,而且考虑了源地址,这一思想弥补了传统BGP中源地址语义缺失的问题,可以实现对网络流量更加细粒度的控制,为满足用户和ISP的多样化需求提供了新的解决方案。整个方案包括控制层和数据层,控制层主要是协议设计,基于现有的多协议BGP(MP-BGP)^[6]进行扩展,实现二维路由信息的传递、管理和使用,并保证二维路由协议与传统路由协议的兼容性;数据层面主要是给出了支持二维匹配域的转发表结构方案,并进行了存储

收稿日期: 2018-09-23

网络出版日期: 2018-11-13

基金项目: 国家自然科学基金(61625203)、国家重点研发计划(2017YFB0803202, 2016YFC0901605)、北京市科技计划项目(Z171100005217001)

空间压缩。

1 二维路由简介

1.1 二维路由的概念

二维路由是一种新型路由协议,它在进行路由决策的时候,不仅仅考虑目的地址,还要考虑源地址。传统路由协议中,去往相同目的地址的报文的下一跳往往是相同的(不考虑等价多路径);但是在二维路由中,目的地址相同、源地址不相同的报文的下一跳可能不同。二维路由的这种特性带来2方面的好处:一是流量控制的粒度变细,网络管理者可以更加灵活地管理网络,如进行流量调度、策略路由等;二是用户的多样化需求可以被更好地满足,例如:享受专门的转发通道等。

本文介绍的是基于BGP的域间二维路由设计方案,它可以利用路径属性(如本地优先级和多出口分辨器)实现二维路由策略。简单来说,域间二维路由为所有可达的目的地址前缀设置一套默认的路径属性,然后再为<目的地址前缀,源地址前缀>这样的前缀对儿设置专门的路径属性,这种设置只作用于对应的前缀对儿,这样就可以生成一条区别于默认方式的转发路径。

1.2 域间二维路由的应用场景

本节通过2个例子来展示文中提出的域间二维路由的应用场景,每个场景都用域间二维路由来解决传统网络遇到的难题。

图1所示是一个多宿主场景,一个站点同时与2个上游ISP相连接,分别用ISP₀和ISP₁表示。其中ISP₀为该站点分配地址前缀为P₀的地址,ISP₁为该站点分配地址前缀为P₁的地址,因而网络中的终端系统具有2个地址,其中地址A属于P₀,地址B属于P₁。在传统的路由协议中,对于去往同一个目的地址的报文,无论终端系统使用地址A或是B作为源地址,

其都将会被转发至同一个上游ISP,假设是ISP₀,此时,如果ISP₀实施了源地址过滤机制,则有可能将来自源地址B的报文过滤掉,从而造成丢包。使用域间二维路由,可以根据报文的源地址选择合适的出口路由,即来自源地址P₀的报文将会通过路由器E₀转发至ISP₀,来自源地址P₁的报文将会通过路由器E₁转发至ISP₁,从而保证了多宿主环境下正常的数据通信。

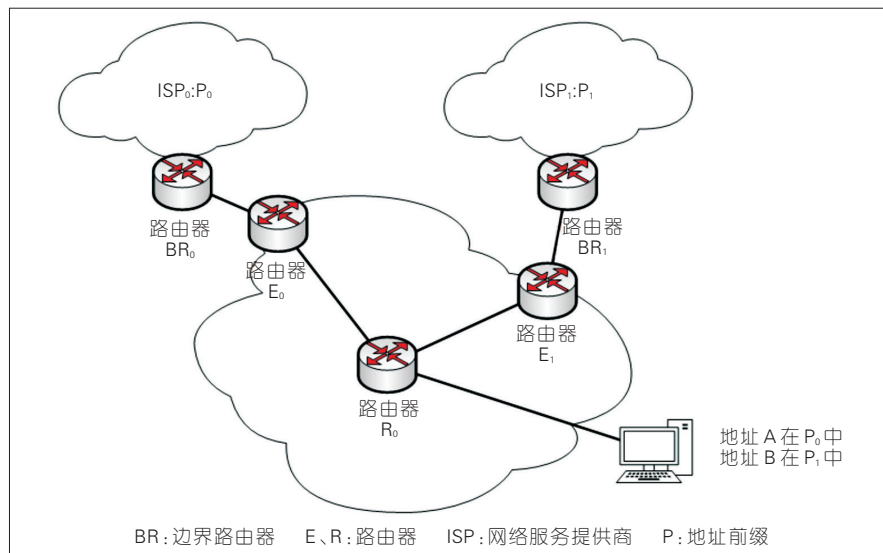
图2所示是一个负载均衡场景,网络中有2个源端AS,分别带有地址前缀P_{s1}和P_{s2},还有1个目的端AS,带有地址前缀P_d,源端AS和目的端AS通过中间的ISP相连。默认情况下,2个源端AS发出的报文可能选择了同一个ISP出口,如路由器E₂,这样可能

会造成与E₂相关联的设备或链路负载比较重,而E₃则负载很轻,出现了网络负载不均衡的情况。负载不均衡会造成网络应对突发问题的容忍性下降、资源利用率降低等问题。使用域间二维路由,可以区分报文的源地址,为不同的源端AS选择不同的出口路由,比如源前缀为P_{s1}的报文通过E₂到达目的端AS,源前缀为P_{s2}的报文通过E₃到达目的端AS,从而实现负载均衡。

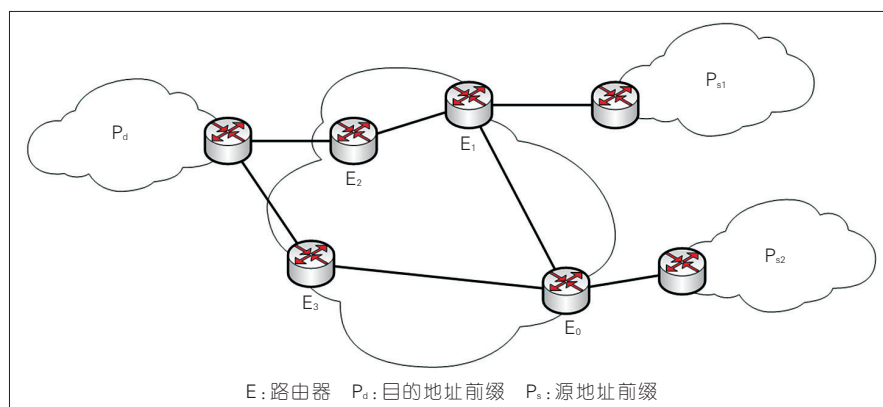
2 域间二维路由协议设计

2.1 基于BGP的二维路由动态学习过程

传统的BGP是通过4种报文完成



▲ 图1 多宿主环境下的域间二维路由应用



▲ 图2 负载均衡场景下的域间二维路由应用

路由信息交互的,分别为:OPEN、KEEPALIVE、UPDATE 和 NOTIFICATION。假设路由器 A 向路由器 B 已经建立了 TCP 链接,此时路由器 A 向路由器 B 发送 OPEN 报文,携带自己的协议版本等信息,如果路由器 A 在约定时间内(计时器计时)收到路由器 B 返回的 OPEN 报文,路由器 A 就会向路由器 B 发送 KEEPALIVE 报文保持刚刚建立的连接。如果在另一约定时间内路由器 A 收到对端发送的 KEEPALIVE 报文,则进入 ESTABLISHED 状态,在此状态下两个路由器可以任意传送 UPDATE 报文来更新路由或者撤销路由。整个过程中,路由器可以通过 NOTIFICATION 报文报告可能出现的错误信息。

基于 BGP 的二维路由协议与传统 BGP 的信息交互过程一样,只不过需要对信息交互报文进行必要的扩展,使其能够支持二维路由相关信息的传递。另外为了能够支持二维路由的增量部署,在设计 BGP 二维路由协议的时候,需要保证二维路由协议与传统路由协议的兼容。在本文中,我们是基于 MP-BGP 实现的域间二维路由协议。

域间二维路由器建立协议连接时,会向对端路由器其发送携带二维路由标识的 OPEN 报文。如果对端路由器支持域间二维路由则回复携带有相同标识的 OPEN 报文,如果对端路由器不支持二维路由即无法识别标识信息,则返回常规的 OPEN 报文。发起连接的域间二维路由器会根据返回的 OPEN 报文是否携带特定标识来识别和记录对端路由器是否支持二维路由,将来二维路由相关信息只会发给建立连接的二维路由器,而不会发给传统路由器。

进入 ESTABLISHED 状态后,域间二维路由器间使用 UPDATE 报文传递网络层可达信息(NLRI),该信息位于 MP-BGP 的可选属性 MP_REACH_NLRI 和

MP_UNREACH_NLRI 的 NLRI 字段,域间二维路由扩展了该字段使其除了包含目的地址信息外,还包含源地址信息。与传统 MP-BGP 类似,MP_REACH_NLRI 负责携带二维路由更新信息,MP_UNREACH_NLRI 负责携带二维路由撤回信息。

需要说明的是:域间二维路由在引入地址信息的时候,默认引入传统路由协议中的所有一维网络层可达信息,即目的地址信息,来保证最基本的可达性,在此基础上可以通过路由器指令等方式引入二维的网络层可达信息,实现进一步的二维路由策略。域间二维路由协议的选路规则同传统 BGP 相同,传统 BGP 的属性例如本地优先级等在域间二维路由协议中同样适用。此外,当需要使用 AGGREGATOR 等属性进行路由聚合的时候,只聚合目的地址而不用不聚合源地址^[7]。

2.2 基于 BGP 的二维路由的报文设计

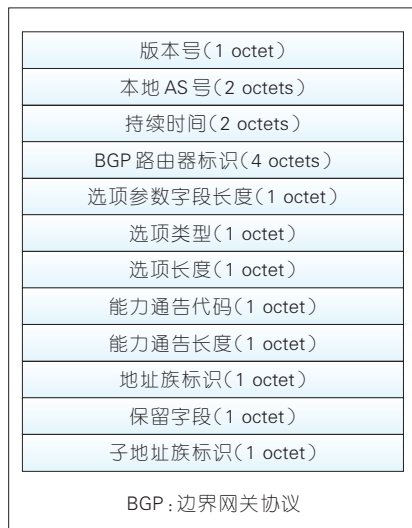
域间二维路由协议的报文设计主要是涉及 OPEN 报文和 UPDATE 报文,剩余的 KEEPALIVE 报文和 NOTIFICATION 报文不做修改,保持与传统路由协议一致。

2.2.1 域间二维路由协议 OPEN 报文

域间二维路由协议的 OPEN 报文跟原始的 OPEN 报文结构基本一致,不同点在于前者需要通过可选参数字段携带二维路由标识。本文中我们通过定义新的子地址族标识来作为二维路由标识,子地址族标识位于报文可选参数部分 Capability 属性包含的多协议扩展字段,具体携带方式如图 3 所示。该标识的具体数值需要向互联网数字分配机构(IANA)进行申请。

2.2.2 域间二维路由协议 UPDATE 报文

域间二维路由协议的 UPDATE 报文用于通知对端路由器有新的路由

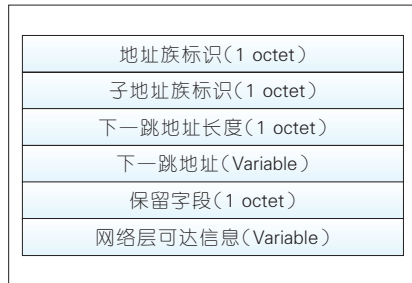


▲图 3 域间二维路由的 OPEN 报文结构图

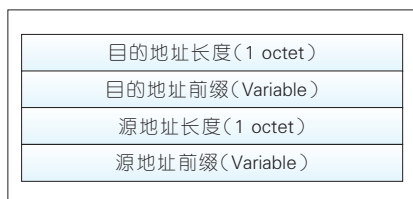
条目需要更新和撤销,协议格式同 MP-BGP 类似。域间二维路由协议使用可选属性 MP_REACH_NLRI 更新路由,使用可选属性 MP_UNREACH_NLRI 撤销路由,其中可选属性 MP_REACH_NLRI 字段结构如图 4 所示,其子地址族标识需要使用同 OPEN 报文一样的二维路由标识,另外其包含的 NLRI 字段(即图 4 中网络层可达信息字段)需要进行扩展。原始的 NLRI 字段只存放目的地址长度和目的地址前缀信息,域间二维路由协议对该字段扩展后,还存放了源地址长度和源地址前缀信息,扩展后的 NLRI 字段结构如图 5 所示。MP_UNREACH_NLRI 字段结构与 MP_REACH_NLRI 字段一致,不在此赘述。

3 域间二维路由转发方案

域间二维路由器收到报文时,不



▲图 4 MP_REACH_NLRI 字段结构图



▲图5 域间二维路由网络层可达信息字段的结构图

仅需要匹配目的地址,还要匹配源地址,因而转发表需要有2个地址匹配域。实际上,只有通过二维路由配置生成的转发表项才需要二维的匹配,其他由传统路由方式(仅基于目的)生成的转发表项只需要匹配目的地址。一个很直接的转发层设计方案为:路由器分配2个转发表,一个是二维转发表,另一个是传统的一维转发表。当路由器收到报文时,需先在二维转发表中进行最长前缀匹配,二维的最长前缀匹配规则是:先进行目的地址的最长前缀匹配,然后在匹配结果中进行源地址的最长前缀匹配,若找到匹配的表项,则终止查找,若无匹配表项,则寻找下一个最长前缀匹配的目的地址,然后再进行源地址的最长前缀匹配,这样重复下去直到找到匹配结果或查完所有表项^[8]。如果在二维转发表中找到匹配表项,就按照其下一跳转发报文,否则就查找传统的一维转发表。一维转发表的查找只进行目的地址的最长前缀匹配。这样的转发层设计十分简洁,但存在一个问题:当二维路由策略很多时,就会生成很多的二维转发表项,消耗大量的转发表空间。转发表一般使用三态内容寻址存储器(TCAM)实现高速匹配,但是TCAM成本高、能耗高,大大限制了二维转发表的空

间大小。

幸运的是:杨术等人提出的企业转发表结构(FISE)^[9]能够较好地解决该问题。FISE结合TCAM和静态随机存取存储器(SRAM)压缩了二维转发表所需的TCAM空间,同时保证报文的线速处理。

如图6所示,新型的转发表结构FISE有2张TCAM表和2张SRAM表。一张TCAM表存储目的表,每条表项是一个目的前缀到目的索引号的映射;另一张TCAM表存储源表,每条表项是一个源前缀到源索引号的映射。一张SRAM表存储所有转发规则的下一跳的索引号,该表被称为二维(TD)表,表里的每个单元格被称为TD单元。通过一个目的索引号和一个源索引号,可以定位到一个TD单元并获得一个下一跳索引号。另一张SRAM表中存储了下一跳索引号和下一跳接口信息之间的映射关系,该表被称为映射表,它能缓解

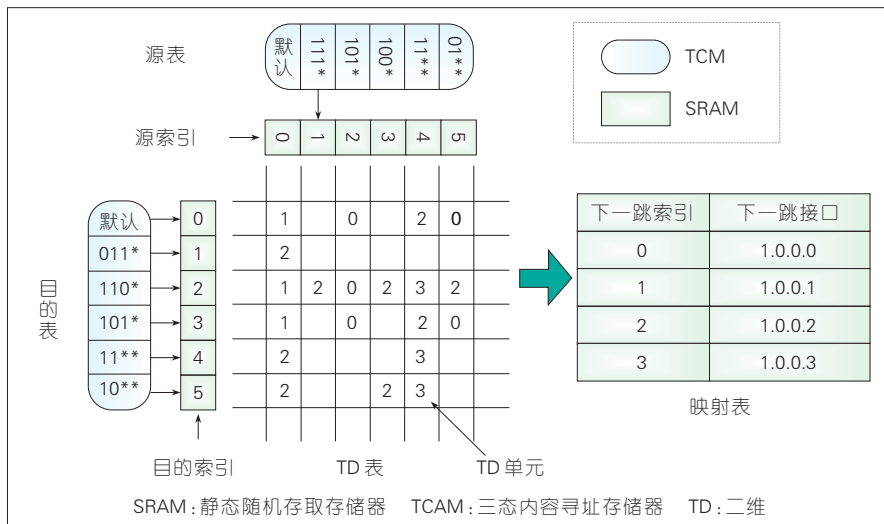
下一跳信息占用存储空间过大、冗余信息过多的问题。

可以看到:FISE将二维转发表的大部分开销从TCAM中移到了SRAM中。这是因为SRAM要比TCAM便宜得多(TCAM成本大概是SRAM的10~100倍),能耗也相对低得多(TCAM能耗大概是SRAM的几十倍到100倍)。FISE借助TCAM的高速匹配和SRAM的空间压缩大大降低了二维转发表的开销并保证了匹配性能。

图7展示了FISE的查找流程。当一个报文到达时,FISE首先在TCAM中分别匹配目的前缀和源前缀,通过目的表和源表可得指向TD单元的索引号,最后利用TD单元存储的下一跳索引号,在映射表中查找到下一跳的信息。

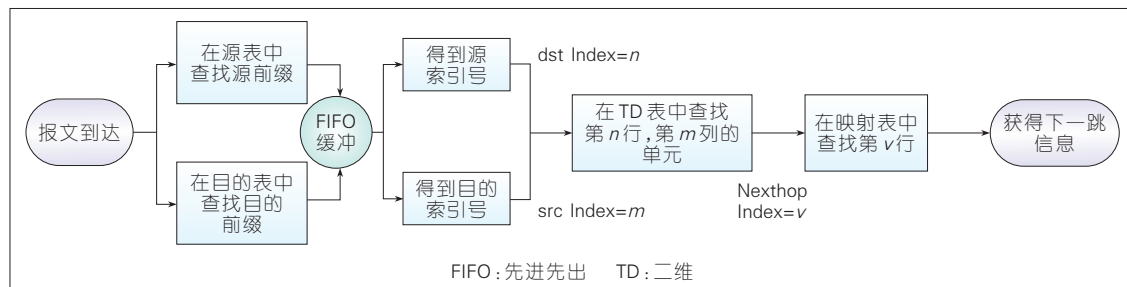
4 结束语

本文中我们提出了一个基于



▲图6 FISE的结构图

图7 FISE的查找流程图



BGP的域间二维路由的设计方案,域间二维路由在进行路由决策的时候同时考虑目的地址和源地址,实现了灵活的流量控制,能够满足用户的多样化需求。本文中我们设计了域间二维路由协议的控制层,使用MP-BGP的可选属性携带二维路由的配置信息,同时兼容传统路由协议,方便ISP进行增量部署。我们还给出了适用于域间二维路由的数据层设计,使用前人提出的FISE转发表结构,可以解决二维转发表造成的TCAM空间爆炸的问题。整体而言,域间二维路由是一种新型路由方式,灵活和细粒度的流量控制使得其具有很好的应用前景。

参考文献

- [1] Framework for Multi-Protocol Label Switching (MPLS)-Based Recovery: RFC 3469[S]. IETF, 2003

- [2] McKEOWN N, ANDERSON T, BALAKRISHNAN H, et al. OpenFlow[J]. ACM SIGCOMM Computer Communication Review, 2008, 38(2): 69-74. DOI:10.1145/1355734.1355746
- [3] RUCHANSKY N, PRPSEPIO D. A (not) NICE Way to Verify the Openflow Switch Specification [C]// ACM SIGCOMM 2013, ACM Special Interest Group on Data Communication 2013. China: ACM, 2013. DOI:10.1145/2486001.249171
- [4] VISSICCHIO S, TILMANS O, VANBEVER L, et al. Central Control over Distributed Routing [J]. ACM SIGCOMM CCR, 2015, 45(4): 43-56
- [5] A Border Gateway Protocol 4 (BGP-4): RFC1771[S]. IETF, 2005
- [6] Multiprotocol Extensions for BGP-4: RFC4760[S]. IETF, 2007
- [7] XIAO L, WANG J, NAHRSTEDT K. Optimizing IBGP route reflection network [C]// IEEE ICC 2003, IEEE International Conference on Communications 2003. USA: IEEE, 2003: 1765-1769. DOI: 10.1109/ICC.2003.1203903
- [8] SMIRNOV A. Draft-ietf-Rtggw-Dst-Src-Routing-06[R]. Internet Draft, 2017
- [9] YANG S, XU M W, WANG D, et al. Scalable forwarding Tables for Supporting Flexible Policies in Enterprise Networks[C]//IEEE INFOCOM 2014, IEEE Conference on Computer Communications, 2014. USA: IEEE, 2014: 208-216. DOI:10.1109/INFOCOM.2014.6847941

上接第55页

电力远远超过全球一些小国家(如约旦等)一年的用电量,相比于IoT设备的低耗能,区块链技术带来的安全防护可能远远小于它的资源浪费。

(4)跨链访问问题。可以预见,未来将会有更多的基于区块链的IoT平台出现,如基于区块链的车联网平台和基于区块链的智慧小区,而如何保证不同区块链平台的互通/互操作将会是促进区块链快速部署的重要条件。

(5)区块链自身的安全问题。区块链在为IoT提供安全防护的同时,自身也面临着一些安全威胁,如eclipse攻击、路由劫持攻击、51%攻击、智能合约漏洞等,这同样为2种技术的融合带来一定的安全隐患。

未来,区块链与IoT这2种技术将进一步融合,面向IoT安全的区块链技术也将逐渐兴起。共识效率更高、存储空间更小、绿色环保的安全区块链技术更能适应新型IoT技术的需求;而容错率高、鲁棒性强、去中心

化的IoT安全管控势必会推进区块链应用技术的发展。

5 结束语

基于区块链的新型IoT技术还处在“萌芽”阶段,对学术界与工业界来说既是机遇又是挑战。区块链与IoT两种技术的融合势必会推动IoT安全的迅猛发展,并由此带来技术的巨大升级。研究适合IoT特征的区块链安全防护技术是未来的一种趋势,但也存在各种挑战而任重道远。

参考文献

- [1] NAKAMOTO S. A Peer-to-Peer Electronic Cash System[EB/OL].(2008-10-31)[2018-10-16].<https://nakamotoinstitute.org/bitcoin/>
- [2] 徐恪,李沁.算法统治世界[M]. 北京:清华大学出版社, 2017
- [3] WU B, LI Q, XU K, et al. SmartRetro: Blockchain-Based Incentives for Distributed IoT Retrospective Detection[C]//IEEE MASS 2018. USA:IEEE, 2018
- [4] GAO F, ZHU L, SHEN M, et al. A Blockchain-Based Privacy-Preserving Payment Mechanism for Vehicle-to-Grid Networks [EB/OL].(2018-07-04)[2018-10-16]. <http://www.x-mol.com/paper/635775>

作者简介



耿男,清华大学计算机系在读博士生;主要研究方向为流量工程、二维路由、路由协议等。

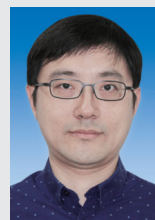


金飞蔡,电子科技大学计算机学院在读硕士生;现任中兴通讯股份有限公司承载网产品线平台系统部项目经理;主持研究了“863”计划项目,国家重点研发计划和省重点研究项目4项;发表专利10余篇。



徐明伟,清华大学教授、博士生导师;主要研究方向为互联网体系结构、互联网路由、高性能路由器、网络安全等。

作者简介



徐恪,清华大学教授、博士生导师、计算机系副主任,国家杰出青年科学基金获得者,中国计算机学会理事;主要从事计算机体系结构、网络空间安全和网络经济学等方面的研究;获国家技术发明二等奖1次、国家科技进步二等奖1次、中国电子学会电子信息科学技术奖一等奖1次,其他省部级一等奖3次等;发表论文20余篇,完成著作6本,获得中国发明专利20余项,美国发明专利授权8项。



吴波,清华大学在读博士研究生;主要研究方向为网络体系结构、网络安全和区块链等;发表论文6篇,获中国发明专利5项,参与制定中国通信行业标准3项。



沈蒙,北京理工大学副教授、硕士生导师,并担任多本国际期刊的审稿人,以及IEEE ICC、IEEE Globecom等国际会议的程序委员会委员;主要研究领域为区块链与数据隐私保护;已发表论文30余篇,获得中国发明专利授权12项,美国发明专利授权2项。