软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(Suppl.(2)):272–282

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

二维转发表的分解存储模型*

兰李晶, 陈文龙, 唐晓岚

(首都师范大学 信息工程学院,北京 100048)

通讯作者: 陈文龙, E-mail: chenwenlong@cnu.edu.cn

摘 要: 现有互联网基于目的IP地址实施报文转发,传输过程没有关注源IP地址,转发策略不够灵活.基于目的IP地址与源IP地址的二维路由,支持网络提供更灵活的转发服务.但是,源IP地址的引入会急剧增加转发表(forward information base,简称FIB)的存储空间,大幅增加硬件成本.提出了一种二维转发表分解存储模型(decomposition storage model of two-dimensional FIB,简称DSTF),基于目的IP前缀对转发表进行分解存储.它把归属于同一个簇头IP前缀(cluster-head IP prefix,简称CP)的二维转发项集合作为一个子二维块(sub two-dimensional block,简称STB),并根据*STB*所属*CP*的不同,将转发项存储到不同的线卡(line card,简称LC).报文转发时,系统可根据*CP*与*LC*间的对应关系,快速定位报文的宿主线卡(LC_{host}),并在宿主线卡中实施转发处理.实验结果表明,该模型能将二维转发表均衡存储在不同线卡上,有效地减少了二维转发表在路由系统中的存储空间.

关键词: 二维路由;转发表;分解存储;目的前缀;源前缀

中文引用格式: 兰李晶,陈文龙,唐晓岚.二维转发表的分解存储模型.软件学报,2016,27(Suppl.(2)):272–282. http://www.jos.org.cn/1000-9825/16041.htm

英文引用格式: Lan LJ, Chen WL, Tang XL. Decomposition storage model of two-dimensional FIB. Ruan Jian Xue Bao/Journal of Software, 2016,27(Suppl.(2)):272−282 (in Chinese). http://www.jos.org.cn/1000-9825/16041.htm

Decomposition Storage Model of Two-Dimensional FIB

LAN Li-Jing, CHEN Wen-Long, TANG Xiao-Lan

(College of Information Engineering, Capital Normal University, Beijing 100048, China)

Abstract: The existing network, which implements the packets forwarding based on the destination IP address, ignores the source IP address in the transmission process, hence its forwarding strategy is not flexible. Based on the destination IP address and the source IP address, the two-dimensional routing supports the network to provide a more flexible forwarding service. But the introduction of the source IP address dramatically increases the storage space of the Forward Information Base (FIB), in addition, it significantly increases the cost of the hardware. This paper presents a decomposition and storage model of two-dimensional FIB (DSTF). According to the destination IP prefixes, DSTF divides the two-dimensional FIB into blocks. And the collection of the two-dimensional routing entries (REs) that are included in the same Cluster-head IP Prefix (CP) is treated as a sub two-dimension block (STB). Then, depending on the different CPs that the STBs belong to, every STB is stored in the different line cards (LC). Meanwhile, in the process of the packets forwarding, according to the corresponding relationship between the CPs and the LCs, the forwarding mechanism can quickly locate the host *LC* (LC_{host}) of this packet, and implement the packets forwarding in the LC_{host}. Experimental results show that DSTF can decompose and store the two-dimensional FIB in different LCs averagely. Furthermore, it is effective to reduce the storage space of the two-dimensional FIB in the routing system.

Key words: two-dimensional routing; FIB; decomposition storage; destination prefix; source prefix

* 基金项目: 国家自然科学基金(61373161, 61502320); 北京市教委科技计划面上项目(KM201410028015); 北京市优秀人才培养资助青年骨干个人项目(2014000020124G133)

Foundation item: National Natural Science Foundation of China (61373161, 61502320); Science & Technology Project of Beijing Municipal Commission of Education under Grant (KM201410028015); Youth Backbone of Beijing Outstanding Talent Training Project under Grant (2014000020124G133)

收稿时间: 2016-06-05; 采用时间: 2016-10-18

现有网络只根据目的 IP 地址进行报文转发,依靠单条路径在可达性的基础上,实现尽力而为的传输服务.随着互联网的迅速膨胀,用户数量呈指数型增长,仅依靠目的 IP 地址提供单一尽力而为可达性服务已远远不能满足用户的需求.为了打破现有网络转发方式,增强互联网转发性能,实现报文传输过程中的区分服务,提出了基于目的 IP 地址与源 IP 地址的二维路由[1].二维路由通过增加源 IP 地址,丰富了路由信息,能够根据源 IP 地址区分不同报文,实现多宿主路由的灵活实施和更优的流量均衡,增加路由数据链路利用率.此外,使用二维路由可以实现多路径的连通性探测,在网络故障诊断中,利用源 IP 地址识别探测报文并将其转发至探测路径,保证多路径连通性探测不会受到网络工程的影响.二维路由还能够保证更好的可达性,实现高可靠传输.报文转发过程中,不仅需要进行目的 IP 地址最长前缀匹配,还需结合源 IP 地址实现报文的灵活转发.然而,二维路由[1]引入源 IP 地址,对传统路由体系产生了一定的影响,需要同时存储源 IP 地址与目的 IP 地址的二维转发表会急剧膨胀,极大地增加了转发表的存储空间,在一定程度上甚至会影响查找时间.

本文提出了一种二维转发表的分解存储模型(decomposition storage model of two-dimensional FIB,简称 DSTF),该模型依据目的 IP 前缀,实现二维转发表在不同线卡上的分解存储,不仅有效缩小了各线卡的存储空间,而且保证了各线卡的转发项存储均衡.DSTF 模型根据二维转发表中目的 IP 前缀之间的隶属关系,将隶属于同一个簇头 IP 前缀的二维转发项信息作为一个存储块(即子二维转发表(sub two-dimension block,简称 STB)).然后,依据不同的簇头 IP 前缀以及该 *STB* 所含转发项数目,分解存储一个或多个 *STB* 到不同线卡.*STB* 中每一个源 IP 前缀与目的 IP 前缀对都对应一个下一跳索引号,通过查询索引映射表,可以得到该下一跳索引号对应的具体下一跳以及相应的出接口.此外,分解存储过程中根据簇头 IP 前缀与所属线卡之间的对应关系,DSTF 模型生成一张前缀定位表,该定位表与映射表同时存储在每块线卡上.对于报文的转发,首先根据簇头 IP 前缀索引查询定位表确定宿主线卡,然后根据源 IP 前缀与目的 IP 前缀对所对应的下一跳索引号查询映射表确定下一跳和出接口,实施转发.DSTF 模型基于簇头 IP 前缀对二维转发表进行分解存储,有效缓解了各线卡的存储负担,在减少转发表存储空间和各线卡转发项负载均衡的情况下,实现了报文的快速查找与转发,并且,它还能有效地保证各线卡之间的协同工作.

与已有二维转发表存储结构相比较,DSTF 模型主要有以下优点:

(1) 根据簇头 IP 前缀对二维转发表进行分块存储,有效缩小了转发表的存储空间.

(2) 各线卡存储不同的子二维转发表,信息独立,无冗余存储现象.

(3) 各线卡分工合作,不仅保证了报文的准确转发,而且平衡了各线卡的工作负担,提高了线卡利用率.

本文第 1 节介绍相关工作.第 2 节详细介绍二维转发表的分解存储模型 DSTF.第 3 节进行性能评估和实验分析.第 4 节总结全文.

1 相关工作

近年来,研究工作者致力于转发表各方面的研究,包括转发表结构、转发协议以及转发表的分解存储等.

文献[1]通过分析现有网络的转发方法,指出单一目的 IP 地址转发的网络弊端.引入源 IP 地址,提出二维转发模型,并且设计了新型二维转发表结构 FIST.对于二维转发表的存储,文献[1]采用 TCAM 和 SRAM 相结合的方法,将部分 TCAM 存储内容转移到 SRAM 中进行存储.充分利用 SRAM 的灵活性,降低存储成本和查找成本.此外,文献[1]还设计了基于 FIST 的增量更新算法,保证了转发项快速更新的情况下访存次数最小.文献[2]提出的基于软件机制的源 IP 地址过滤器只通过检查源 IP 地址信息的某几位就可以保证恶意流量在网络中的过滤.该方法能够确保协同工作的各路由器流量负载均衡,此外,还有效地解决了路由的动态变化问题.

Bruni 在文献[3]中针对路由问题做出了相应的研究.将目前面临的路由问题定义为一个最优控制问题,用一个控制变量表示每个流经过可用路径(该可用路径为多路径路由)的百分比.此外,还利用一组成本函数实现流量均衡和瓶颈负载最小化的两个不同目标.该模型的主要问题集中在两个方面:(1) 对给定网络进行子网的划分.(2) 每一个划分的子网受一组独立的变量集所控制.根据变量与成本函数的计算以及子网的划分,文献[3]对强路由的最优控制问题进行了分解处理,保证最优控制问题能够在一个低维最优控制问题集合中得到简单、

快速的解决,节省了处理成本与时间.文献[4]主要研究了二维路由的增量部署.通过分析路由器的处理性能,提出了一种受约束的二维路由增量部署模型.根据部署节点个数和各路由器对二维路由的负荷量,定义 VIP 路径,该 VIP 路径是 VIP 流量流经的最佳路径.

文献[6]提出了一种规则的近似最优的转发表压缩方法.该方法基于文献[5]对转发表压缩进行了增量完善.对于文献[5]中转发表压缩生成树不断更新导致的路由信息错误,文献[6]通过恢复最近后裔前缀下一跳和回收非聚合前缀下一跳的存储空间的方法,保证了压缩存储后的转发表在转发项不断更新的情况下,仍可以保证转发项信息的正确转发.文献[7]引入了一种结合链路状态和路径矢量的混合路由协议,定义了下游节点和上游节点概念.通过有策略的洪泛让上游节点知道其下游节点的下游路径,可以有效地避免路由环路现象.此外,对于多宿主节点,依据某些节点链路的权限链表,不仅可以有效地避免环路现象,还能很好地进行策略路由.

文献[8]主要针对路由表构造的 Trie 树,提出了一种基于 SRAM 的多管道的并行查找模型,该模型通过对 Trie 树的划分、子树与管道之间的映射等操作进行路由信息的快速查找.文献[9]中的路由管理模型实现了路径压缩 Trie 树和 TCAM 硬件相配合的路由查找算法,解决了快速查找问题;采用广播更新,路由过滤等技术解决了主从路由表的同步问题.文献[10−12]分别研究了不同的软件与硬件对包的分类以及 TCAM 的分类,以此来提高分类速度,减少 TCAM 功耗.

以上研究主要集中在二维路由设计、转发项存储以及 TCAM 功耗等方面,都没有涉及二维转发表的分解存储问题.

2 二维转发表的分解存储模型

2.1 二维转发表结构

现有网络中,路由器进行报文转发时,根据报文目的 IP 地址进行最长前缀匹配(longest prefix matching,简称 LPM),然后根据匹配到的下一跳和出接口实现报文转发.FIB 中每条转发项信息都是一个三元组,即:〈目的 IP 前缀,下一跳,出接口〉.报文传输只依靠报文目的 IP 地址实现,会造成网络流量不均衡、可达性不稳定、网络资源浪费等问题.二维转发表将源端信息引入到转发项中,使得 FIB 不再只依据目的 IP 地址进行信息的转发.同时,根据源 IP 地址和目的 IP 地址将报文送至目的地,成功地将一维匹配转化为二维匹配,增强了网络灵活性.针对每条转发项,二维转发表需额外增加源 IP 前缀信息,在数据层中每条转发项信息增至为一个四元组,即:〈目的 IP 前缀,源 IP 前缀,下一跳,出接口〉.我们实施的报文转发策略为:当路由器转发报文时,首先需要进行目的 IP 地址最长前缀匹配,然后进行源 IP 地址最长前缀匹配,最后根据匹配的下一跳和出接口,完成报文的转发.

现有网络的 FIB 只需简单存储目的 IP 地址、下一跳和出接口,然而,二维转发表引入源 IP 地址会急剧扩大转发表的存储空间.对于每一个目的 IP 地址,可能存在多条与之对应的源 IP 地址,相对于现有的一维转发表,二维转发表会成倍地扩大.用 m 表示目的 IP 前缀存储空间,n 表示源 IP 前缀存储空间,如果每一条目的 IP 前缀都和全部源 IP 前缀匹配为一条二维转发项,那么 FIB 的存储空间将会从 m 直接扩大为 $m\times n$.这种存储空间数量级的增长会直接影响二维转发表的存储性能和查找速度.

2.2 DSTF 模型

文献[1]采用完全备份存储模式,二维转发表中所有转发项信息会同步存储在每一块线卡上,当网络流量集中流经其中一块线卡时,会造成该线卡持续工作而其他线卡处于空闲状态,无疑会浪费网络资源.此外,每块线卡存储相同的转发项信息,一旦二维转发表中表项信息出现更新,所有线卡上的转发项信息需要同步更新,在转发项信息频繁更新的情况下,会造成同步时间过长,降低了网络转发性能.除此之外,各线卡同步存储相同转发项信息也会造成存储空间的极大浪费.综上所述,我们提出了一种二维转发表的分解存储模型,通过将二维转发表划分成不同的子二维块(STB)存储到不同线卡,不仅减少了二维转发表的存储空间,而且保证了转发项均衡存储的情况下各线卡之间的协同工作.二维转发表在各线卡上的均衡存储就是将全部的二维转发项平均分摊

到每块线卡上,促使线卡共同完成转发工作.它能够提升网络的数据处理能力、增强网络的灵活性和线卡的可用性.

二维转发表中每条转发项为一个四元组信息,用 *DP* 表示目的 IP 前缀,*SP* 表示源 IP 前缀,*NH* 表示下一跳,*OI* 表示出接口,则四元组信息表示为 $FR\langle DP,SP,NH,OI\rangle$.DSTF 模型的分解存储方法只关注 *DP*(默认路由除外),依据 *DP* 之间的隶属关系将二维转发表划分为 *STB* 进行存储,每个 *STB* 都包含转发项的四元组信息.为了描述方便,以下只根据 *DP* 进行 *STB* 的划分与转发项说明.

对于 *DP* 之间的隶属关系,我们作如下说明:目的 IP 前缀 $address_1/PRElen_1$ 和 $address_2/PRElen_2$,若前缀掩码长度 $PRElen_1 \leqslant PRElen_2$,且目的 IP 地址 $address_1$ 和 $address_2$ 的前 $PRElen_1$ 位对应相同,则定义两个 *DP* 之间存在隶属关系,目的 IP 前缀 $address_2/PRElen_2$ 隶属于 $address_1/PRElen_1$.其中,*PRElen* 表示路由前缀掩码长度.

例如,目的 IP 前缀 198.64.0.0/16 与 198.64.192.0/20 之间存在隶属关系,且 198.64.192.0/20 隶属于 198.64.0.0/16.

为了更好地观察 *DP* 之间的隶属关系,方便、快捷地划分 *STB*,DSTF 模型首先依据前缀掩码长度对二维转发表中所有 *DP* 进行升序排序.

定义 1(簇头 IP 前缀 CP,cluster-head IP prefix). 根据 *DP* 之间的隶属关系,将集合 $\{DP_0,DP_1,\ldots,DP_n\}$ 划分为若干 *STB* 集合,若在 $STB_0=\{DP_0^0,DP_1^0,\ldots,DP_{k_0}^0\}$ 中存在且只存在一条目的 IP 前缀 DP_0^0,使得 $DP_j^0(2\leqslant j\leqslant k_0)\in DP_0^0$,并且任意 $DP_j^w(w\neq 0)\notin DP_0^0$,则称 DP_0^0 为 STB_0 的簇头 IP 前缀(CP).

STB_0 的簇头 IP 前缀 DP_0^0 满足:(1) 在 STB_0 中,DP_0^0 的前缀掩码长度最短.(2) 除 DP_0^0 外,STB_0 中其余 *DP* 均隶属于 DP_0^0.

将若干 *DP* 路由项集合根据簇头 IP 前缀 *CP* 划分为以下 *m* 个 *STB* 集合:

$$STB_0=\{DP_0^0,DP_1^0,\ldots,DP_{k_0}^0\},$$
$$STB_1=\{DP_0^1,DP_1^1,\ldots,DP_{k_1}^1\},$$
$$\vdots$$
$$STB_m=\{DP_0^m,DP_1^m,\ldots,DP_{k_m}^m\}.$$

集合内路由项按前缀长度升序排序,DP_0^0 是 $STB_0=\{DP_0^0,DP_1^0,\ldots,DP_{k_0}^0\}$ 的簇头 IP 前缀,DP_0^1 是 $STB_1=\{DP_0^1,DP_1^1,\ldots,DP_{k_0}^1\}$ 的簇头 IP 前缀,…,DP_0^m 是 $STB_m=\{DP_0^m,DP_1^m,\ldots,DP_{k_0}^m\}$ 的簇头 IP 前缀.

定义 2(子二维块 STB,sub two-dimension block). 如果 $STB_m=\{DP_0^m,DP_1^m,\ldots,DP_{k_m}^m\}$ 满足:(1) STB_m 中的任意 $DP_i^m(0\leqslant i\leqslant k_m)$ 均不隶属于 $DP_i^0(0\leqslant i\leqslant k_0)$,…, $DP_i^{m-1}(0\leqslant i\leqslant k_{m-1})$ 中的任意一项.(2) $DP_i^0(0\leqslant i\leqslant k_0)$,…, $DP_i^{m-1}(0\leqslant i\leqslant k_{m-1})$ 中任意一条 *DP* 均不属于 $DP_i^m(0\leqslant i\leqslant k_m)$,则称 STB_m 为最大子二维块.

STB 是 DSTF 模型分解存储过程中各线卡的一个最大存储单元.在 DSTF 模型的分解存储过程中,每个 *STB* 仅存储在一块线卡上,不同线卡根据自身存储容量的大小可以存储一个或多个 *STB*.此外,*CP* 与 *STB* 充分保证了分解存储过程中,各 *STB* 之间的信息独立,避免了各线卡上转发项的冗余存储.

转发表中 *DP* 的 Trie 树结构能够直观地描述 *DP* 之间的隶属关系,并能够简单、有效地确定 *CP* 及该 *CP* 所辖的所有转发项信息,该转发项集合即为 DSTF 模型中的一个 *STB*.通过对大量转发项前缀掩码长度的分析[13],我们得到前缀掩码长度小于 8 位的转发项不存在,也就是说,*CP* 的前缀掩码长度大于等于 8,故 Trie 树结构中 *CP* 所在层数大于等于 8.图 1 中,我们省略 Trie 树前 8 层中的部分层结构,从根节点对该转发表 Trie 树进行遍历,遍历得到的第 1 个真实节点即为 *CP*,其所辖子树中的真实节点的集合即为 DSTF 模型中的一个存储单元 *STB*.如图 1 所示,遍历到的真实节点 CP_i,CP_j 就是其所辖真实转发项集合的簇头 IP 前缀.虚线所辖范围就是 DSTF 模型中的一个 *STB*,其中,该 *STB* 的簇头 IP 前缀完全备份存储模式就是 CP_i.

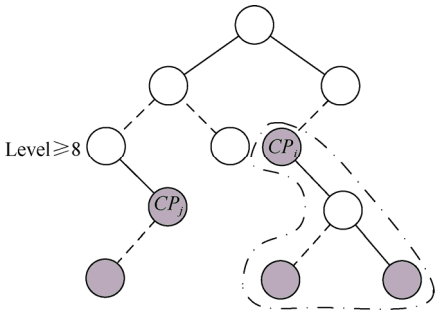

Fig.1 Trie tree of *DP* in DSTF

图 1 DSTF 中 *DP* 的 Trie 树

定义 3(压缩指数 CI,compression index). DSTF 模型下各线卡 *SP* 与 *DP* 总存储容量与原二维转发表下各线卡总存储容量的比值.见式(1):

$$CI=\frac{(DP+SP)_{\text{Full}}-(DP+SP)_{\text{DSTF}}}{(DP+SP)_{\text{Full}}} \tag{1}$$

对于 *SP* 为 50,*DP* 为 500 的二维转发表,原二维转发表在完全备份存储模式下,各线卡均存储 550 条前缀信息,而在 DSTF 模型中,4 块线卡的前缀存储容量分别为 176,176,171,177,对应 *CI* 分别为 0.680,0.680,0.689,0.678.

定义 4(二维转发规则). 路由器接收到目的 IP 地址为 *D*,源 IP 地址为 *S* 的 IP 报文后,首先进行 *DP* 的最长前缀匹配(LPM),得到多条与该 *DP* 对应的四元组信息.然后对这些四元组信息中的所有 *SP* 进行最长前缀匹配(LPM),最后根据 *DP* 与 *SP* 匹配对得到下一跳 *NH* 和出接口 *OI*,实施转发.

例如,IP 报文目的 IP 地址为 128.64.192.0,源 IP 地址为 192.128.64.0.路由器接收到该报文后,首先对目的 IP 地址进行 LPM,假设匹配到二维转发表中的 *DP*:128.64.128.0/17,该 *DP* 所对应的转发项有 3 条,即对应该 *DP* 的四元组信息有 3 条,分别为 FR_1⟨128.64.128.0/17,192.128.0.0/16,20.0.0.0,0⟩,FR_2⟨128.64.128.0/17,192.128.0.0/17,20.0.0.0,2⟩,FR_3⟨128.64.128.0/17,192.128.64.0/20,20.0.0.1,1⟩.然后根据源 IP 地址 LPM 可知,该 IP 报文最终匹配的转发项为 FR_3,则转发引擎根据下一跳信息 20.0.0.1 和出接口 1 实施报文转发.

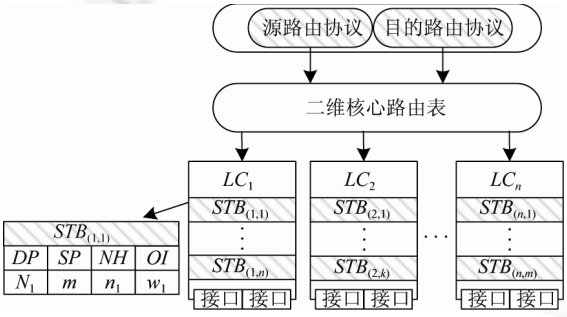

Fig.2 Two-Dimensional FIB of DSTF

图 2 DSTF 的二维转发表

如图 2 所示为 DSTF 模型中二维转发表结构.二维转发表引入了二维路由协议,该二维路由协议是通过对 OSPF 协议的扩展而实现的.二维路由协议中的源路由协议与目的路由协议的结合使得控制层的转发机制更加灵活.路由体系根据二维路由协议整合出报文转发所需的转发项信息集合,在 DSTF 模型的二维转发表结构中,我们根据 DSTF 方法分解存储该二维转发表到不同线卡.报文转发过程中,通过各线卡分摊处理报文,实施转发.DSTF 模型的分解存储原则是:(1) 首先根据 *DP* 之间的隶属关系将 *DP* 表分解为子 *DP* 表.(2) 每一个子 *DP* 表所包含的四元组 *FR* 为一个 *STB*.*STB* 根据 DSTF 算法均衡存储到各线卡.(3) 不同 *STB* 之间信息相互独立,不存在交叉、覆盖关系.(4) 每个 *STB* 仅存储在一块线卡上,然而,不同线卡可以存储一个或多个 *STB*.如图 2 所示,每一个 *STB* 只唯一存储在一块线卡上,不同线卡可以存储 *n* 个、*k* 个或 *m* 个 *STB*,除此之外,不同 *STB* 包含不同数目的二维转发项信息.例如,$STB_{(1,1)}$表示 LC_1 上的第 1 个子二维转发表,它包含 N_1 条目的 IP 前缀,*m* 条源 IP 前缀,n_1 条对应的下一跳,w_1 条出接口.虽然各线卡存储的 *STB* 子表数目不同,但是每块线卡上所存储的总二维转发项数目几乎是相同的.

2.3 DSTF 分解存储方法

根据二维转发表中 *DP* 之间的隶属关系以及二维转发规则,我们对二维转发表进行子表的划分,根据 *CP* 划分子表,将 *STB* 分解存储到不同线卡,具体分解存储过程如下.

(1) 首先根据 *DP* 的前缀掩码长度对二维转发表中的所有 *DP* 进行升序排序.然后依据 *DP* 之间的隶属关系,对 *DP* 表进行划分,归属于相同 *CP* 的 *DP* 集合为子 *DP* 表.

(2) 根据子 *DP* 表,将其对应的二维转发表划分为不同的 *STB*,每个 *STB* 包含与子 *DP* 表中各 *DP* 相对应的所有 *SP* 信息,以及 *DP* 与 *SP* 对所指向的下一跳索引号信息和对应的出接口.

(3) 根据其所含 *DP* 数目,对 *STB* 进行降序排序.统计各线卡所存 *DP* 数目,优先存放 *STB* 到总前缀数目(即 *DP* 与 *SP* 之和)最少的线卡.

(4) 线卡容量相同的情况下,则根据子 *DP* 表中的 *CP*,将其 *CP* 前 8 位相同的 *STB* 存储到同一块线卡上.

(5) 根据存储的 *STB*,各线卡存储相应的下一跳索引与下一跳的对应关系映射表,以及 *CP* 与 *LC* 的定位表.

如图 3 所示,根据 *DP* 之间的隶属关系,将 *DP* 集合划分为 3 个子 *DP* 表(DP table):DT_1\{00***,001***\},其 *CP* 为 00***.DT_2\{01***,0101*\},其 *CP* 为 01***.DT_3\{10***,101***,1001*\},其 *CP* 为 10***.将 DT_1 中 *DP* 所对应的所有 *SP* 信息以及 *DP* 与 *SP* 对所指向的下一跳索引号信息划分为一个转发项集合,该转发项集合就是 00***所对应的一个 *STB*,图 3 中虚线所含内容即为分解存储的一个 *STB*.

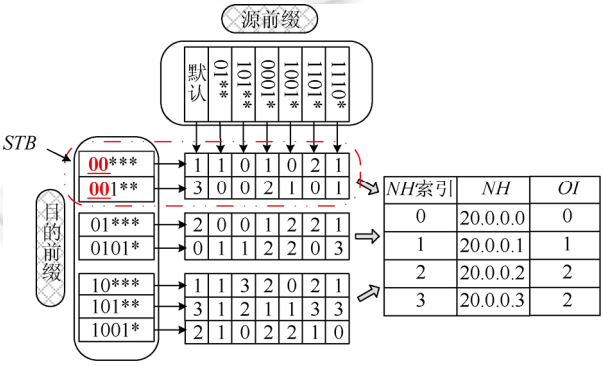

Fig.3 The decomposition storage structure of two-dimensional FIB

图 3 二维转发表分解存储结构

原二维转发表[1]在每块线卡上均存储所有表项信息,报文转发过程在一块线卡上就可以完成.DSTF 模型将二维转发表划分成不同的 *STB* 进行存储,如果接收该报文的线卡不能完成转发工作,则需要转发至其他线卡,协同实施报文转发过程.对于包含报文转发所需所有转发项信息的线卡,我们称为宿主线卡(LC_{host}).接收该报文的线卡,称为接收线卡(LC_{in}).将报文从适当出接口转发出去的线卡定义为发送线卡(LC_{out}).为了保证各线卡之间的协同工作,我们在分解存储 *STB* 到各线卡的过程中,根据 *CP* 与 *LC* 的对应关系生成一张 *DP* 定位表(location table,简称 LT),$LT\langle CP_k,LC_i\rangle$,$CP_k$ 表示子二维转发表的簇头 IP 前缀,LC_i 表示该 CP_k 所在的线卡号.该 *LT* 记录了 *STB* 中 *CP* 与 *LC* 之间的定位关系,当报文在 LC_{in} 上不能匹配到转发信息时,根据 *LT*,LC_{in} 会将该报文转发至 LC_{host},从而完成转发工作.

我们随机选取 10 条 *DP* 和 5 条 *SP*,构成 10×5 的二维转发表结构,通过分解存储算法分解存储到 4 块线卡:LC_0,LC_1,LC_2,LC_3.表 1 所示为源 IP 前缀表,表 2 为目的 IP 前缀表,目的 IP 前缀与源 IP 前缀匹配对对应的下一跳索引号见表 3.

Table 1 Source IP prefix

表 1 源 IP 前缀

No.	*SP*
①	默认路由
②	128.64.0.0/10
③	200.192.0.0/12
④	110.192.0.0/12
⑤	72.200.224.0/22

Table 2 Destination IP prefix
表 2 目的 IP 前缀

No.	*DP*	No.	*DP*
①	72.64.0.0/11	②	110.128.0.0/9
③	128.64.0.0/10	④	128.100.0.0/14
⑤	200.192.0.0/12	⑥	200.200.0.0/16
⑦	110.192.0.0/12	⑧	110.204.0.0/20
⑨	72.200.224.0/22	⑩	128.124.0.0/20

Table 3 Two-Dimensional index
表 3 二维索引

DP	*SP*				
	①	②	③	④	⑤
①	1	0	1	0	2
⑨	1	0	0	2	1
②	2	0	0	1	2
⑦	2	2	2	0	3
⑧	1	1	3	2	0
③	2	1	2	3	3
④	2	1	0	2	2
⑩	1	2	1	2	3
⑤	1	1	0	2	0
⑥	3	2	0	2	3

根据表 1 中的 *SP* 和表 2 中的 *DP* 构建二维转发表.假设表 2 中每条 *DP* 都可以和表 1 中所有的 *SP* 组成一条二维转发项,则该二维转发表一共有 50 条四元组转发项.根据第 3.2 节中 DSTF 模型分解存储的具体过程,我们首先分析 *DP* 表中各前缀之间的隶属关系,对 *DP* 表进行分块,由表 2 可知,①∈⑨,且①是 *CP*.{⑦,⑧}∈②,且②是其子 *DP* 块的 *CP*.同理可得,③是集合{③,④,⑩}的 *CP*,⑤是目的 IP 前缀⑤,⑥所属 *DP* 块的 *CP*.故表 2 共有 4 个 *CP*,也就是说,将表 2 中 *DP* 所对应的二维转发项信息按照所属 *CP* 的不同划分为 4 部分进行存储.

如表 3 所示,二维索引表记录了每一个 *DP* 与 *SP* 匹配对相对应的下一跳索引号信息,每一行表示每一个 *DP* 与不同 *SP* 相匹配所对应的下一跳索引号.我们根据 *DP* 之间的隶属关系,将二维索引表进行划分,①与⑨所含的两行二维索引信息作为一个子二维索①引表进行存储,②、⑦和⑧所含的 3 行二维索引信息同样作为一个子二维索引表进行存储,其余信息作相应处理.我们将属于同一个 *CP* 的 4 部分:源 IP 前缀表、子目的 IP 前缀表、子二维索引表以及子二维索引表中下一跳索引号对应的映射表作为一个 *STB* 存储在适合的线卡上.例如,经过上述分解存储后,①与⑨归属的 STB_1 包含的信息为:5 条 *SP*、2 条 *DP*、表 3 中①与⑨所对应的下一跳索引行信息以及这些下一跳索引号与下一跳出接口对应的映射表.为了方便描述,以下我们简单描述 STB_1 所含信息为①与⑨,表示为 STB_1={①,⑨}.故该 10×5 的二维转发表被分解为 4 个 *STB* 进行存储,分别为 STB_1={①,⑨},STB_2={②,⑦,⑧},STB_3={③,④,⑩},STB_4={⑤,⑥}.假设 4 块线卡 LC_0,LC_1,LC_2,LC_3 目前路由存储容量均为空,根据 4 个 *STB* 所含 *DP* 数目降序排序,我们首先将 STB_2 存储在线卡 LC_0 上,然后依次将其他 3 个 *STB* 存储在转发项数目最少的线卡上,那么 STB_3 将存储在线卡 LC_1 上,LC_2 存储 STB_1,LC_3 存储 STB_4.

2.4 报文转发过程

当线卡接收到一个 IP_j 报文后,首先根据 IP_j 目的 IP 地址在 STB 中进行目的 LPM,若匹配成功,则继续根据 IP_j 的源 IP 地址进行源 LPM,匹配成功,则根据 *DP* 与 *SP* 匹配对所指向的下一跳索引号查询映射表,根据下一跳和出接口实施转发.若目的 IP 地址 LPM 成功,源 IP 地址 LPM 没有成功,则匹配默认源 IP 前缀实施转发,此时即将二维匹配降低为一维匹配,根据目的 IP 地址 LPM 进行报文转发.若 LC_{in} 上目的 IP 地址 LPM 不成功,则根据 IP_j 目的 IP 地址查询定位表 *LT* 将其转发至 LC_{host} 进行 *DP* 与 *SP* 的最长前缀匹配.同样,若源 IP 地址 LPM 不成功,则直接根据目的 IP 地址 LPM 实施转发.如果报文 IP_j 在 LC_{host} 上没有与之匹配的转发项信息,则直接丢弃.

DSTF 模型采用非完全存储模式,每块线卡仅存储核心路由表的部分表项信息,根据 IP 报文的源 IP 地址与目的 IP 地址,判断该报文所需路由项所在的线卡,并将该 IP 报文转发至该线卡,很好地将报文转发工作分摊到各线卡并保证 LC_{in},LC_{host} 与 LC_{out} 之间的协同工作,有效提高了资源利用率,保证了更好的工作效率.

2.5 二维路由信息更新

随着转发表信息的不断变化,各线卡上所存转发项内容也会不断更新.对于二维路由的增加,首先根据 FR_i 的目的 IP 前缀 DP_i,确定各线卡是否存在其簇头 IP 前缀 CP_i.若存在,则直接添加到 CP_i 所在的线卡,若 CP_i 所在 *STB* 中不存在 FR_i 所含的源 IP 前缀 SP_i,则需在该 *STB* 中添加源 IP 前缀 SP_i.若不存在目的 IP 前缀 DP_i 的簇头

IP 前缀,则 FR_i 作为一条新的四元组转发项信息,根据第 3.3 节中的分解存储过程进行存储.

对于四元组转发项信息的删除,情况如下:(1) 所删四元组 FR_i 中 *DP* 非 *CP*,则直接删除该 *DP* 所包含的所有四元组信息即可.(2) 所删四元组转发项 FR_i 中 *DP* 是 *CP*,则需重新选取 *CP*.若所删 FR_i 所在的 *STB* 中其他 *DP* 均隶属于新的簇头 IP 前缀 CP_j,则删除 FR_i 后,更新定位表 *LT*.若所删四元组 FR_i 所在的 *STB* 中,无新的共同簇头 IP 前缀,则该 *STB* 信息需根据第 3.3 节的分解存储过程重新划分子二维表,更新定位表中 *CP* 与 *LC* 的对应关系.

3 性能评估和实验分析

DSTF 模型采用非完全存储模式,根据不同的 *CP* 划分二维转发表,通过将 *STB* 进行分解存储,大大缩小了二维转发表的存储空间,减轻了各线卡的存储负荷,在保证二维转发表最小存储空间的条件下,我们尽可能地保证了各线卡的转发项负载均衡.二维路由相对于一维路由,强化了源 IP 地址在报文转发中的作用,实现了报文传输过程中的区分服务.此外,相对于单一目的 IP 地址转发,二维路由保证了网络流量的均衡,避免了单条路径转发可能造成的网络拥塞问题.DSTF 模型还进一步保证了各线卡的流量负载均衡,将二维转发表的四元组转发项分摊存储到各线卡,避免了部分线卡空闲而部分线卡流量负载过大的问题,保证了各线卡的协同工作,提升了工作效率.网络中的转发项信息一直处于变化状态,分解存储有利于各线卡根据自身的需求自适应地进行转发项状态更新,使得各线卡转发项信息的更新速度快,更新周期短,在转发项信息更新的状态下,也能够保证报文的转发效率.二维转发表的分解存储使得每块线卡只存储转发表的一部分,为了实现报文的快速转发,需要引入 *CP* 与 *LC* 的定位表,该定位会占据额外的存储空间,所占最小存储空间为 446bytes.

为了分析 DSTF 模型的压缩存储性能以及验证各线卡在 DSTF 模型中的协同工作效率,实验系统利用 5 台 PC 机进行模拟,PC_0 模拟主控,PC_1~PC_4 模拟线卡 LC_0~LC_3,通过向主控中导入实验数据,实现二维转发表在 4 块线卡的分解存储.本文选取 9 个不同 ASs[13,14]下的 *SP* 和 *DP* 构建二维转发表,实验中 *SP* 表容量为 $Date_1$,*DP* 表容量为 $Date_2$.此外,为了验证各线卡在 DSTF 模型下的工作效率,我们选取“马上 6”软件的真实数据 $Date_3$ 进行实验分析.“马上 6”软件由清华大学开发,它基于 IPv6 隧道实现 IPv4 网络报文传输.

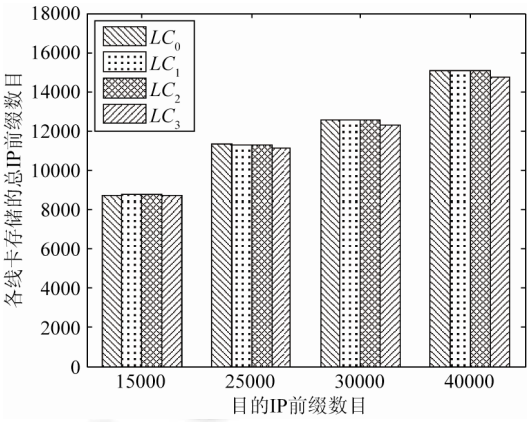

Fig.4 Decomposition and storage of DSTF in different data

图 4 不同数据下 DSTF 的分解存储

为了很好地验证 DSTF 模型对二维转发表的分解存储性能,我们选取 $Date_2$ 中不同数目的 4 组 *DP* 与 $Date_1$ 中的所有 *SP* 构成 4 张不同大小的二维转发表进行分解存储,实验结果如图 4 所示.相对于二维转发表的完全存储,该模型极大地压缩了存储空间,由图 4 可知,当 *DP* 数目为 40 000 时,完全存储下,每块线卡均需要存储的总前缀数目为 45 000,而 DSTF 模型中,线卡存储的最大总前缀数目为 15 100,最小总前缀数目仅为 14 975,存储空间压缩比例达到 66%以上.此外,DSTF 模型在不同数据下近乎都能达到各线卡的存储负载均衡,这不仅有效缓解了存储压力,同时也提高了路由器系统的可用性.

目前,二维转发表都是以完全备份存储模式进行存储,该模式下各线卡存储与核心路由表相同的表项信息.我们抽取 $Date_2$ 中的两组数据($Date_4$,$Date_5$)与 $Date_1$ 构成两张二维转发表,基于完全备份存储模式和 DSTF 模型进行存储比较,实验结果如图 5、图 6 所示.

由图 5 和图 6 可知,DSTF 模型相对于完全备份存储模式极大地减少了各线卡的总 IP 前缀存储数目,明显压缩了二维转发表的存储空间,并且,随着总 IP 前缀数目的增加,存储空间压缩效果更好.

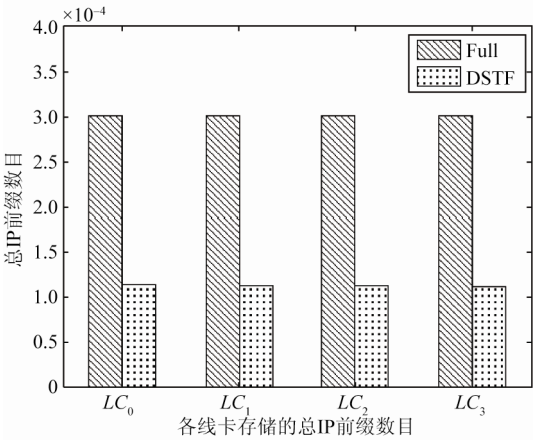

Fig.5 Storage based on $Date_4$ in different models

图 5 不同模型下基于 $Date_4$ 的存储

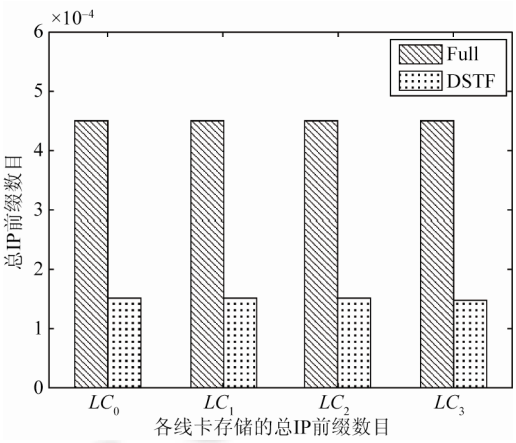

Fig.6 Storage based on $Date_5$ in different models

图 6 不同模型下基于 $Date_5$ 的存储

完全备份存储模式中,各线卡均存储与核心二维转发表相同的表项信息,故各线卡报文流量不均衡的情况下会导致部分线卡超负荷工作而部分线卡始终处于闲置状态.本文对 DSTF 模型中不同报文数目下各线卡的报文处理次数进行了实验分析,如图 7 所示.我们抽取 $Date_3$ 中 4 组不同报文数目在同一张 30000×5000(*DP* 数目为 30 000,*SP* 数目为 5 000)的二维转发表中进行报文处理次数的统计分析,其中,30000×5000 的二维转发表已经根据 DSTF 模型在各线卡进行了分解存储.图 7 显示,各线卡在同一张二维转发表中对不同数目报文的处理次数近乎相似,各线卡始终可以保持同一工作状态.由于转发表的分解存储会导致报文转发过程中线卡之间的板间通信,故总查询次数可能略大于总报文数目.板间通信量的增加可能引起转发时延和内部带宽消耗增大的问题,但是这些问题只需在目前高性能分布式路由器上作少许改进就可以得到解决.对于线卡协同工作而言,DSTF 模型有效地解决了报文转发过程中各线卡流量不均衡引起的线卡利用率不高的问题.

为了直观描述 DSTF 模型对二维转发表存储空间的压缩性能,我们定义了 *CI*,计算方法见公式(1).通过计算不同前缀数目下,每块线卡上的压缩指数 *CI*,以证明 DSTF 的总体压缩性能以及对比分解存储的转发项在不同线卡上的压缩性能.由图 8 可知,DSTF 模型压缩性能很好,最小压缩指数 *CI* 达到 56%以上,最大压缩指数 *CI* 甚至可以达到 67%以上.其中,LC_3 的转发项压缩指数略大,其他 3 块线卡的压缩指数呈现的 *CI* 折线图近乎重叠.图中折线斜率虽有减小趋势,但是该斜率仍表明,随着前缀数目的不断增加,*CI* 增长迅速.此外,图 8 还显著地说明了不同数据下 DSTF 模型在各线卡的压缩指数 *CI* 较稳定,受前缀数目变化的影响很小.

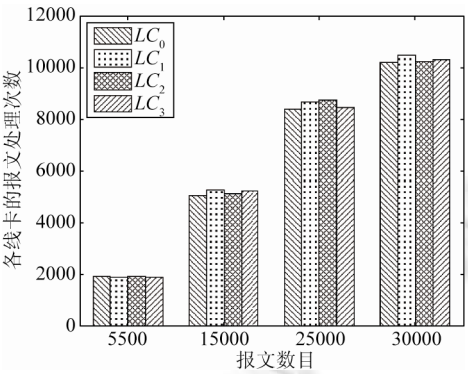

Fig.7 The number of processing in LCs of DSTF

图 7 DSTF 中各线卡的报文处理次数

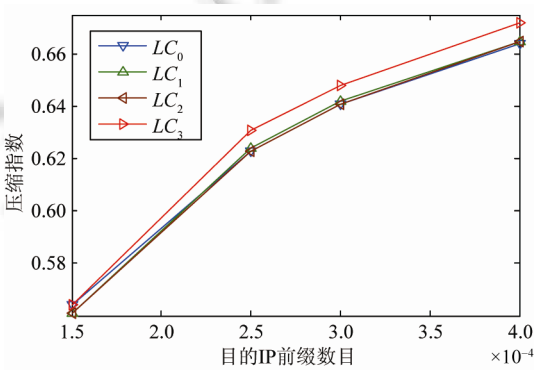

Fig.8 The *CI* of two-dimensional FIB of DSTF

图 8 DSTF 中二维转发表存储压缩指数

4 总 结

本文提出的二维转发表的分解存储模型充分利用目的 IP 前缀(DP)之间的隶属关系,实现二维转发表的分解存储.DSTF 模型在各线卡均存储源 IP 前缀信息,根据 *DP* 之间的隶属关系划分二维转发表,并将簇头 IP 前缀 *CP* 作为每一个子二维转发表 *STB* 的索引.然后根据 DSTF 模型的分解存储原则,实现二维转发表在各线卡的均衡存储.此外,各线卡均存储 *CP* 与 *LC* 之间对应关系的定位表,报文转发过程中,转发引擎通过定位表和最长前缀匹配,加速实施报文的转发.DSTF 模型极大地减少了二维转发表的存储空间,同时,也尽可能实现了各线卡的存储负载均衡,保证了各线卡之间的协同工作,节省了系统开销.与现有的二维转发表完全存储模型相比,DSTF 模型在存储空间和查找效率上有了进一步的提高.实验结果表明,DSTF 模型相对于完全存储模型在转发表存储空间上进行了很大的压缩,同时,在报文转发的过程中,提高了各线卡协同工作的效率.根据当前报文传输过程中区分服务的需求,二维路由将会是一个新的发展趋势,DSTF 模型在缩小转发表存储空间方面具有一定的实用价值,有很好的现实意义,有利于进一步应用到实际环境中.

References:

[1] Yang S. Algorithms and protocols for two dimensional-IP routing [Ph.D. Thesis]. Beijing: Tsinghua University, 2014 (in Chinese with English abstract).

[2] Yang S, Xu MW, Wang D. Source address filtering for large scale network: A cooperative software mechanism design. In: Proc. of the IEEE 21st Int'l Conf. on Computer Communications and Networks (ICCCN). Munich: IEEE Press, 2012. 1–7. [doi: 10.1109/ICCCN.2012.6289219]

[3] Radjenovic D, Hericko M, Torkar R, Zivkovic A. Software fault prediction metrics: A systematic literature review. Information and Software Technology, 2013,55(8):1397–1418. [doi: 10.1002/ett.2536]

[4] Yang S, Wang D, Xu MW. Efficient two dimensional-IP routing: An incremental deployment design. Computer Networks the Int'l Journal of Computer & Telecommunications Networking, 2014,59(3):227–243. [doi: 10.1016/j.bjp.2013.11.004]

[5] Draves RP, King C, Venkatachary S, Zill BD. Constructing optimal IP routing tables. In: Proc. of the IEEE Int'l Conf. 18th Annual Joint Conf. of the IEEE Computer and Communications Societies (INFOCOM). New York: IEEE Press, 1999. 88–97. [doi: 10.1109/INFCOM.1999.749256]

[6] Uzmi ZA, Nebel M, Tariq A, Jawad S, Chen R, Shaikh A, Wang J, Francis P. SMALTA: Practical and near-optimal FIB aggregation. In: Proc. of the Int'l Conf. ACM CoNEXT. Tokyo: ACM Press, 2011. 1–12. [doi: 10.1145/2079296.2079325]

[7] Zhang X, Adrian P, Zhang H. Centaur: A hybrid approach for reliable policy-based routing. In: Proc. of the IEEE Int'l Conf. 29th IEEE Int'l Conf. Distributed Computing Systems (ICDCS). Columbus: IEEE Press, 2009. 76–84. [doi: 10.1109/ICDCS.2009.77]

[8] Jiang W, Wang QB, Prasanna VK. Beyond TCAMs: An SRAM-based multi-pipeline architecture for Terabit IP lookup. In: Proc. of the IEEE Int'l Conf. IEEE INFOCOM. Phoenix: IEEE Press, 2008. 1786–1794. [doi: 10.1109/INFOCOM.2008.241]

[9] Liang ZY, Xu K, Wu JP, Xu MW. Routing management model in distributed routers. Journal of Tsinghua University (Sci. & Tech.), 2003,43(4):503–506 (in Chinese with English abstract).

[10] Sommers PB. On the prevalence and characteristics of MPLS deployment in the openInternet. In: Proc. of the Int'l Conf. ACM IMC. Berlin: ACM Press, 2011. 445–462. [doi: 10.1145/2068816.2068858]

[11] Meiners CR, Liu AX, Torng E. Split: Optimizing space, power, and throughput for TCAM-based classification. In: Proc. of the IEEE Int'l Conf. Architectures for Networking and Communications Systems (ANCS). Brooklyn: IEEE Press, 2011. 200–210. [doi: 10.1109/ANCS.2011.36]

[12] Zhao X, Liu Y, Wang L. On the aggregatability of router forwarding tables. In: Proc. of the IEEE Int'l Conf. IEEE INFOCOM. San Dieg: IEEE Press, 2010. 1–9. [doi: 10.1109/INFCOM.2010.5462137]

[13] BGP Routing Table Analysis Reports. http://bgp.potaroo.net

[14] BGP Routing Table. http://www.cidr-report.org/as2.0

附中文参考文献:

[1] 杨术.二维路由算法与协议研究[博士学位论文].北京:清华大学,2014.

[9] 梁志勇,徐恪,吴建平,徐明伟.分布式路由器的路由管理模型.清华大学学报(自然科学版),2003,43(3):503–506. [doi: 10.1109/INFOCOM.2008.241]

兰李晶(1992－),女,河南陕县人,学士,主要研究领域为网络体系结构,网络协议.

唐晓岚(1987－),女,博士,讲师,CCF 专业会员,主要研究领域为车载网络,无线传感器网络,城市计算.

陈文龙(1976－),男,博士,副教授,CCF 专业会员,主要研究领域为路由及交换技术,网络体系结构,下一代互联网及过渡技术.