DOI：10.7652/xjtuxb201706021

针对域内流量均衡的二维路由方案

赵成安[1]，王文东[1]，徐明伟[2]，陈文龙[3]
(1.北京邮电大学软件学院，100876，北京；2.清华大学计算机系，100084，北京；
3.首都师范大学信息工程学院，100048，北京)

摘要：针对域内流量均衡问题，提出了一种二维开放式最短路径优先(OSPF)路由方案TOL。在控制层面，通过对传统链路状态通告(LSA)的扩展，实现包括目的前缀和源前缀的二维路由信息的传递，路由器根据二维路由信息进行计算，生成二维路由表项。在数据转发层面，设计了一种基于传统一维转发表实现二维数据转发的方案，这种转发方案能够有效解决引入源前缀造成的转发表存储空间增长问题，兼容传统转发，为TOL提供保障。为了验证TOL的有效性和可行性，在商用路由器上实现了原型系统，测试和实验结果表明，TOL方案能够在传统IP网络结构和协议的基础上，有效实现流量均衡，减少链路拥塞，且不会带来较大的额外负荷。

关键词：二维路由协议；二维转发；流量均衡
中图分类号：TP393 **文献标志码**：A **文章编号**：0253-987X(2017)06-0129-09

Two-Dimensional Routing for Intra-Domain Load Balancing

ZHAO Cheng'an[1], WANG Wendong[1], XU Mingwei[2], CHEN Wenlong[3]
(1. School of Software Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
3. College of Information Engineering, Capital Normal University, Beijing 100048, China)

Abstract: Two-dimensional IP routing is proposed to solve the problem of load balancing in autonomous domain. For control layer, a two-dimensional OSPF protocol for load balancing, named as TOL, is proposed. TOL is able to realize path switching of flows, including two-dimensional routing information, by means of expanded LSA. For data layer, two-dimensional forwarding scheme is considered to achieve the aim of two-dimensional forwarding by the conventional one-dimensional forwarding tables. A prototype of TOL is realized on the commercial routers. Experimental results show that these schemes can be applied to the conventional IP network structures and protocols to effectively achieve network load balancing and relieve network congestion.

Keywords: two-dimensional routing; two-dimensional forwarding; load balancing

随着互联网的飞速发展，用户规模和应用数量不断增长，传统网络存在的问题也日趋明显。传统网络是基于IP模型的网络体系结构，强调的是信息的快速可达，采用的是基于最短路径进行报文转发的方式，对数据流提供同样的尽力而为的服务。这种方式容易造成有些链路利用率很高，甚至发生拥塞，而有些链路利用率低，甚至空闲。但是，由于路由协议和数据转发都是在目的地址这样一个维度上进行实现的，所以传统网络很难解决网络链路流量不均衡的问题。

收稿日期：2016-10-01。 作者简介：赵成安(1980—)，男，博士后；王文东(通信作者)，男，教授，博士生导师。 基金项目：国家"863计划"资助项目(2015AA015601，2015AA016101)；国家自然科学基金资助项目(61373161)。
网络出版时间：2017-03-15 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20170315.1643.002.html

针对这个问题，在网络发展过程中出现了不少解决方案，如等价多路径(ECMP)[1]、多协议标签交换(MPLS)[2-3]、策略路由等技术，但是这些技术都存在各自的问题。ECMP 能够实现多路径负载均衡，而且开放式最短路径优先(OSPF)中基本上都支持 ECMP 功能，但是在实际网络中，各路径间差异大时，效果并不理想。MPLS 是一种非常有效地解决方案，但是 MPLS 的封闭性给网络测量、管理和安全方面的工作带来了相当的不便，而且并不是所有网络都部署了 MPLS。策略路由需要在数据包经过的路径上进行配置，配置过程非常复杂。软件定义网络(SDN)[4]的出现能够很好地解决这个问题，但是 SDN 集中控制的属性有碍于大规模部署，而且目前 SDN 主要应用在新建的局域网或数据中心网络，因此与传统网络融合方面还存在问题。分段路由[5]也是一种不错的解决方案，但是分段路由的实现依赖于 MPLS 或源路由。

基于上述原因，本文将通过二维路由来解决网络链路流量均衡问题。文献[6]提出了基于源 IP 地址和目的 IP 地址进行二维路由转发，路由器对数据包进行路由转发时，不只考虑目的 IP 地址，还考虑源 IP 地址。文献[7]提出了地址驱动网络(ADN)体系结构，提出 IP 地址具有长度属性、逻辑属性、拓扑属性、空间属性、时间属性、所有者属性等多重属性，阐述了二维路由即利用了 IP 地址的拓扑属性和所有者属性，可以满足分级服务、流量工程、快速故障恢复等需求。

二维路由利用了 IP 自有的信息，即源地址信息，天然地与现有 IP 路由系统结合起来，另一方面，通过设计新的二维路由协议，以及改进现有的转发结构，能够使基于 IP 的网络满足各种网络需求。针对流量均衡问题，本文提出了一种二维 OSPF 路由方案 TOL。通过对传统链路状态通告的扩展，实现包括目的前缀和源前缀的二维路由信息的传递，路由器根据二维路由信息进行计算，生成二维路由表项。在数据转发层面，设计了一种基于传统一维转发表，实现二维数据转发的方案，这种转发方案能够有效解决引入源前缀造成的转发表存储空间增长问题。

1 方案概要

1.1 设计思想

在图 1 所示的网络拓扑中，链路上的数值就是 OSPF 链路 metric 值。传统 OSPF 中，节点 a、b、c 到节点 i 都要经过路径(d,h,i)，从而可能导致链路(h,i)的带宽利用率过高甚至产生拥塞，而链路(g,i)的带宽利用率较低。流量工程的目的是降低链路的带宽利用率，但是传统路由只能根据目的地址进行转发，难以实现链路流量均衡。因此，我们引入二维路由技术，利用源地址区分流量，使一部分流量能够切换到最短路径以外的其他路径。通过对链路状态通告(LSA)的扩展[8]，实现二维路由信息的传递。因为扩展的 LSA 内不仅包含目的前缀信息，而且包含源前缀信息，所以称为二维 LSA(TD-LSA)。以图 1 拓扑为例，假如链路(h,i)带宽利用率高于设定的阈值，在控制层面，可以通过手动配置或自动的方式，令节点 h 生成 TD-LSA 并洪泛。如果节点 h 发出 TD-LSA 包含[目的:i，源:b，metric:40]，节点 d 收到 TD-LSA 后，为源 b 生成一个虚拟拓扑，如图 2 所示。对于 b 来说，链路(h,i)的开销为 40，所以节点 d 在这个虚拟拓扑上运行最短路径优先(SPF)算法，得出到 i 的下一跳为节点 g，这样得到的一维路由表项[目的:i，下一跳:g]，再加入源 b，得到二维表项[目的:i，源:b，下一跳:g]。对于节点 a 和 c 来说，链路(h,i)的开销不变，所以还是经过原来的路径。数据层方面，转发表中的表项由二元组[目的前缀，下一跳]变为三元组，即[目的前缀，源前缀，下一跳]。当一个报文到达路由器的时候，路由器会查看报文的目的地址和源地址，得到下一跳信息。

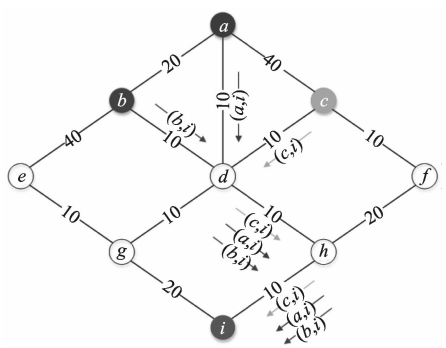


图 1 传统 OSPF 路径选择

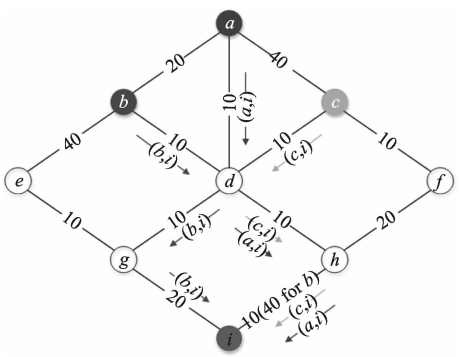


图 2 二维 OSPF 路径选择

1.2　相关模块

定义 1　宏流是由一个目的前缀和源前缀对来唯一标识的流量集合。

(1)流量监测。节点需要获取每个接口(一般为出接口)的带宽利用率，以及确定出接口有哪些宏流和宏流的流量大小。这个需求可用通过诸如 Netflow、IP Accounting 等功能来满足。通常，路由器支持流量监测功能。

(2) TD-LSA 处理。TD-LSA 处理包括 TD-LSA 生成和 TD-LSA 传播。TD-LSA 生成涉及到 metric 值确定和宏流选取两方面工作。在通常情况下，传统 LSA 多采用洪泛的方式传播。为了减少控制报文数量和计算负荷，针对 TD-LSA 提出反向路径和定向传输机制。

(3)路由计算。根据本文的设计思想，路由算法还是采用 SPF 算法，只不过针对不同源前缀，在不同的虚拟拓扑上分别运行 SPF 算法，这个算法称为 SP^2F 算法。

(4)报文转发。二维转发与传统一维转发相比，转发表中的表项由二元组[目的前缀，下一跳]变为三元组[目的前缀，源前缀，下一跳]。当一个报文到达路由器的时候，路由器会查看报文的目的地址和源地址，得到下一跳信息。

2　二维 OSPF 路由方案 TOL

本文基于传统 OSPF 协议提出了用于流量均衡的二维 OSPF 路由方案 TOL。TOL 方案引入源地址前缀对 LSA 进行扩展，形成一种新的 LSA 类型(TD-LSA)。路由器收到 TD-LSA 后，会将其存入扩展的链路状态数据库(TD-LSDB)，然后基于每个源前缀的虚拟拓扑进行 SPF 计算，从而得到二维转发表项(TD-ENTRY)，存储在二维转发表(TD-FIB)中。TOL 的实现涉及到 TD-LSA 生成、TD-LSA 传播、SP^2F 计算、二维转发等几个部分，下面将分别介绍。为了便于描述，表 1 给出了本节相关符号含义。

表 1　TOL 方案符号列表

符号	定义
U_S	期望的链路带宽利用率，如 SLA 中带宽利用率上限
$u_{i,j}$	链路(i,j)的实时带宽利用率
$b_{i,j}$	链路(i,j)的带宽容量
B	OSPF 计算 metric 的参考带宽
$m_{i,j}$	传统 LSA 中链路(i,j)的 metric 值
$m^t_{i,j}$	TD-LSA 中链路(i,j)的 metric 值
m^c_k	类型为 k 的链路 metric 值
p_s	TD-LSA 中源前缀
p_d	TD-LSA 中目的前缀
h_u	带宽利用率上限冗余
h_l	带宽利用率下限冗余
I	下一个 TD-LSA 的发送间隔
S_f	从流量监测模块得到的宏流集合
S_d	从 TD-LSA 中得到的宏流集合
f_m	宏流的标识信息，主要包括源前缀和目的前缀
o	出接口

2.1　TD-LSA 的生成

TD-LSA 可以通过两种方式产生。一是网络管理员配置的方式，网络管理员通过网络管理系统了解到某链路的带宽利用率较高，而另一链路的带宽利用率较低，只需在链路带宽利用率高的路由器上通过配置，令路由器生成 TD-LSA。一是节点自动生成 TD-LSA，路由器监测链路的实时带宽利用率 $u_{i,j}$，一旦 $u_{i,j}\geqslant U_S+h_u$ 满足，将启动 TOL 机制，生成 TD-LSA。手动配置方式下，网络管理员可以通过配置源前缀和目的前缀选择将哪些宏流切换路径，还可以设置 TD-LSA 中的 metric 值。下面，主要介绍自动方式下 TD-LSA 的生成过程。

2.1.1　*metric 值的确定*　传统 OSPF 中，链路(i,j)的 metric 值通常由下式计算得到

$$m_{i,j}=B/b_{i,j} \tag{1}$$

这样的 metric 值是静态的，路由算法在选择路径时，总是选择 metric 值小的链路，从而导致这些链路带宽利用率过大。

为了实现流量均衡，针对 TD-LSA 定义了新的公式确定 metric 值

$$m^t_{i,j}=\frac{B}{b_{i,j}(1-u_{i,j})\lambda_k} \tag{2}$$

式中：$\lambda_k=m^c_k/\sum_k m^c_k$。

式(2)考虑了实时带宽利用率，能够动态反映链路负载情况。路由器收到 TD-LSA，针对其中每个源前缀，计算最短路径时，对于带宽利用率超限的链路，以式(2)的结果作为链路开销，由于 $m^t_{i,j}$ 反映了链路可用带宽，所以能够减少超限链路的流量负载，实现流量均衡。

2.1.2　*宏流的选择*　图 3 给出了源前缀信息在 TD-LSA 中的结构[8-9]。TD-LSA 实际上是在传统 LSA(如 Intra-Area-Prefix-LSA、AS-External-LSA 等)后加入源前缀结构，可加入一个或多个源前缀。

类型		长度
前缀长度	前缀选项	0
地址前缀 ⋮		

图 3 TD-LSA 中源前缀信息结构

当路由器发现某条链路实时带宽利用率超限时，通过流量监测模块，得到有哪些宏流在此链路上，从中选择宏流，并将对应的前缀放入 TD-LSA 中。宏流的切换有可能导致目标链路产生拥塞，为了避免这种情况以及这种情况导致的流量频繁切换的现象，本文设计了最新生存期最小流量优先(LMF)选择算法。其主要过程：首先，查看 TD-LSAB，查找是否存在之前切换过来的宏流信息，如果存在，根据 TD-LSA 中的生存期，选择其中最新宏流的前缀信息放入 TD-LSA 中；否则，选择最小宏流的前缀信息放入 TD-LSA 中。

LMF 算法的目的是尽量保证宏流切换到链路不会由于流量的增长而过快的使带宽利用率超限，进而启动 TOL 机制。为了防止传输不必要的 TD-LSA，在上一个 TD-LSA 生成发送后，下一个 TD-LSA 生成发送之前，设定一个时间间隔 I。时间间隔过后，通过流量监测模块，检查该链路上是否存在上一个 TD-LSA 中的宏流，如果存在，则停止 TD-LSA 生成；否则，判断是否满足 $u_{i,j}>U_S$，如果满足，则运行 LMF 算法，否则停止。

算法 1 最新生存期最小流量优先算法

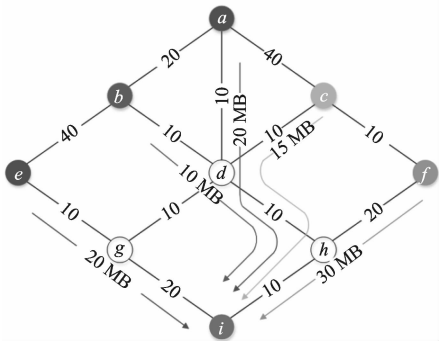
输入：S_d，S_f

输出：f_m

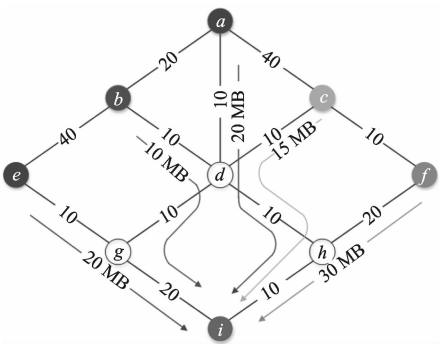
```
PROCEDURE MacroFlowSelect()
    IF S_d ≠ NULL THEN
        f_m ← GetMacroflowWithLastestAge(S_d)
        GOTO step 6
    f_m ← GetMacroflowWithMixRate(S_f)
    Insert(TD-LSA, f_m)
    Send(TD-LSA)
    Waiting I
    IF FindMacroflow(S_f, f_m)=NULL THEN
        IF u_{i,j} > U_S THEN
            GOTO step 2
END PROCEDURE
```

另外，链路(i,j)一旦满足 $u_{i,j}\leqslant U_S+h_l$，则路由器将生成 TD-LSA，其中设置 $m^t_{i,j}=m_{i,j}$，选择被切换的宏流加入 TD-LSA 并发出。其他节点收到这样的 TD-LSA 将删除与 TD-LSA 中宏流相关的路由机制和二维路由表项，使这些宏流按照传统 OSPF 进行转发。

图 4 展示了 TD-LSA 生成和宏流选择的过程。设 $U_S=60\%$，$h_u=10\%$，$B=100$ Mbit·s^{-1}。初始化状态如图 4a 所示，链路(h,i)带宽利用率 $u_{h,i}=75\%$，节点 h 生成 TD-LSA。按照 LMF 算法，首先会选择来自节点 b 的宏流进行路径切换。切换之后，$u_{h,i}=65\%$，仍然满足 $u_{h,i}\geqslant U_S$，这时选择来自节点 c 的宏流进行路径切换。但是，在节点 c 的宏流切换之后，$u_{g,i}=90\%$，节点 g 启动 TOL 机制，按照 LMF 算法，选择节点 c 的宏流进行路由切换。此时，在节点 d 上，$m^t_{h,i}=200$，$m^t_{g,i}=700$，所以节点 c 的宏流切换回原来的路线。最后的结果见图 4b。



(a)初始阶段



(b)运行 TOL 后的结果

图 4 宏流选择过程

2.2 TD-LSA 的传播及 SP²F 算法

如果 TD-LSA 也采用洪泛的方式传播，一方面浪费链路带宽；另一方面，收到 TD-LSA 的路由器都要进行处理，增加计算负荷。为了减少 TD-LSA 的传播数量，以及降低路由器计算负荷，本文提出反向路径定向(RPD)传输机制。在通常情况下，一条链路两个方向上的 metric 值相同，所以在 SPF 算法的作用下，网络中两个节点双向通信，所经过的路径相同。RPD 传输机制的主要思想：当路由器生成

TD-LSA 后，针对 TD-LSA 中所包含的源前缀，以此源前缀为目的地址查找 FIB，从得到的出接口发出；当路由器收到 TD-LSA 时，首先也进行与上面相同的处理，然后在运行 SP²F 算法后，如果二维路由表项中的出接口与原一维路由表项的出接口不同，则把 TD-LSA 从改变后的出接口发出。RPD 传输机制能够使 TD-LSA 通过最短路径从生产 TD-LSA 的路由器传播到拥有源前缀的路由器。沿途路由器会对 TD-LSA 进行处理，从而使包含在 TD-LSA 中的宏流实现路径切换。TD-LSA 也会沿切换后的路径传播，以保证数据层的正确转发。如图 5 所示，假设节点 h 针对节点 b、c 的宏流生成 TD-LSA，根据 RPD 传输机制，节点 h 将把 TD-LSA 沿链路(h,d)发出，节点 d 收到 TD-LSA 后，将分成两个 TD-LSA 分别沿链路(d,b)和链路(d,c)发送给节点 b 和 c，并会在运行 SP²F 算法后，将 TD-LSA 沿链路(d,g)发出。与洪泛相比，RPD 传输机制能够有效减少控制报文的数量，而且只有与被切换宏流相关的路由器才需要进行路由计算。

过程 1　反向路径和定向传播过程

Part Ⅰ　反向路径传输

```
1：PROCEDURE ReversePathTransmitting()
2：    FOREACH p_s in TD-LSA DO
3：        o ← Lookup(FIB, p_s)
4：        Send(TD-LSA with p_s only, o)
5：    END FOREACH
6：END PROCEDURE
```

Part Ⅱ　SP²F 算法

```
7：PROCEDURE TDRouteCalculating()
8：    FOREACH p_s in TD-LSA DO
9：        TD-ENTRY ← SPFCalculating(TD-LS-DB, p_s)
10：       Insert(TD-FIB, TD-ENTRY)
11：   END FOREACH
12：END PROCEDURE
```

Part Ⅲ　定向传输

```
13：PROCEDURE DirectionalTransmitting()
14：   FOREACH p_d in TD-LSA DO
15：       IF Lookup(FIB, p_d) ≠ o of TD-ENTRY
THEN
16：       Send(TD-LSA, o of TD-ENTRY)
17：   END FOREACH
18：END PROCEDURE
```

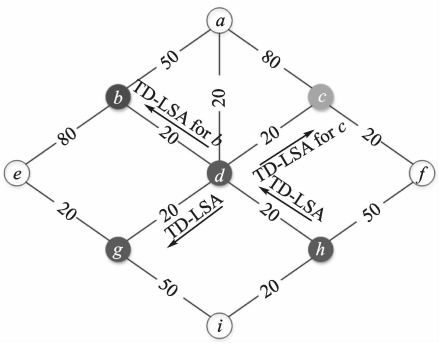


图 5　反向路径定向传输机制

2.3　二维转发

在高性能路由器中，为了保证转发性能，通常采用基于三元内容可寻址存储器（TCAM）的硬件实现。由于 TCAM 价格贵、功耗高、容量有限，通常情况下，采用 TCAM 加上静态随机存储器（SRAM）的方式实现。即将前缀信息存储在 TCAM 中，对应的下一跳信息存储在 SRAM 中，而逻辑控制模块可由现场可编程门阵列（FPGA）或特定用途集成电路（ASIC）来实现。针对二维转发，ACL-like 方案[10-12]的做法是将目的前缀和源前缀合成一条转发表项存储在 TCAM 中，TCAM 存储空间复杂度为 $O(MN)$，其中 M 为目的前缀数量，N 为源前缀数量。FIST 方案[13]将源前缀和目的前缀分别存储在不同的 TCAM 表中，从而使 TCAM 的存储空间变为 $O(N+M)$，将大量信息从 TCAM 移到成本和功耗更低的 SRAM 中。根据网络中数据包交互的过程可知，请求数据包的源地址即为相应的响应数据包的目的地址，所以对于路由器来说，传统路由的转发表中就包括源前缀，即源前缀转发表项在原有的目的前缀转发表中已存在。利用这个特点，本文基于 FIST 方案设计了 TOL 的转发方案 FIST＋，该方案进一步减少了 TCAM 存储二维转发表项的空间。二维路由由于引入了源地址，所以转发表中的前缀可以分为目的前缀和源前缀。

定义 2　设 P^{d} 为二维路由目的前缀的集合，P^{s} 为源前缀的集合，如果前缀 p 满足$(p\in P^{\mathrm{s}})\wedge(p\notin P^{\mathrm{d}})$，则称为 s-prefix；如果满足$(p\in P^{\mathrm{d}})\wedge(p\notin P^{\mathrm{s}})$，则称为 d-prefix；如果满足$(p\in P^{\mathrm{d}})\wedge(p\in P^{\mathrm{s}})$，则称为 b-prefix。

由定义 2 可知，b-prefix 即为源转发表和目的转发表的前缀重复表项。路由器中上述各类前缀的数量取决于路由器在网络中的位置和配置。通常在基于二维路由 TOL 的网络中，因为每个路由器都可以获取到所有路由前缀，所以所有前缀都可以认

为是 b-prefix。

根据定义 2,设计二维路由转发方案 FIST+。如图 6 所示,在 TCAM 中只有 1 张表 FIB,这张表由 d-prefix、b-prefix 和 s-prefix 构成,既作为目的前缀表,也作为源前缀表。在 SRAM 中,存储着 3 张表,第 1 张表存储着与 TCAM 中前缀表项对应的目的/源前缀指向的索引,称之为一级索引表(FITable);第 2 张表存储着所有转发规则的下一跳索引,称之为 TD 表(TDTable),通过一个目的索引和源索引,可以定位到一个 TD 表中的 TD 单元,TD 单元中存放着下一跳索引;第 3 张表存储着下一跳索引和下一跳之间的映射关系,称之为下一跳映射表(NMTable)。采用的二维匹配规则为:报文到达路由器时,提取目的地址和源地址,首先针对目的地址进行最长前缀匹配,得到目的索引,然后针对源地址进行最长前缀匹配,得到源索引,最后根据两个索引得到下一跳索引,进而获得下一跳。

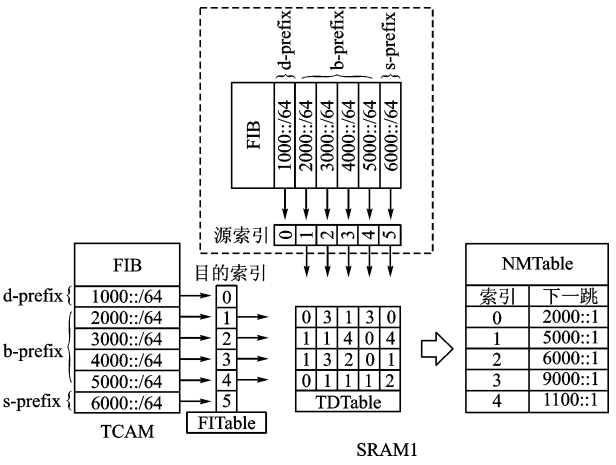


图 6 二维转发结构 FIST+

FIST+的 TCAM 存储空间复杂度为 $O(N+M-L)$,其中 L 为 b-prefix 的数量,通常 $L=M$。也就是说,通常情况下,FIST+的 TCAM 存储空间与一维路由情况相同。

3 实现与测试

3.1 实现架构

设计了 TOL 在路由器上实现的架构如图 7 所示,控制层除了传统的 OSPF 相关模块,还增加了一些功能模块。TD-LSDB 用于存放 TD-LSA,在实现上也可以通过对原有 LSDB 进行扩展,从而节省存储空间;SP^2F 算法模块以 TD-LSDB 为基础进行路由计算,生成的二维路由存储在二维路由表中。流量监控模块主要实现接口带宽利用率监测,以及接口上流量的采样(通过 Netflow 实现)。TD-LSA 生成模块根据流量监控结果生成 TD-LSA,通过 RPD 传输模块发送出去。在数据层方面,当收到报文后,通过查找 FIST+转发表确定出接口。

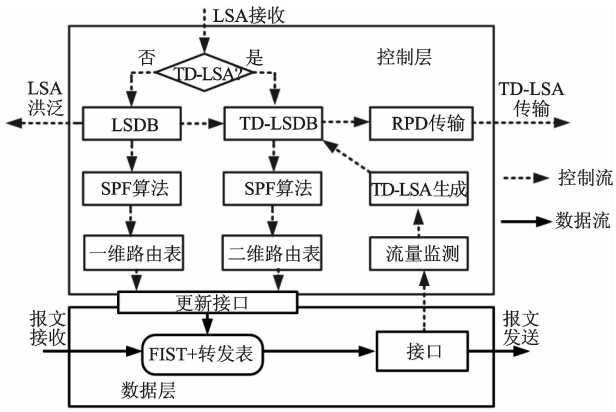


图 7 TOL 实现架构

3.2 二维转发测试

为了证实所提出的设计,以比威公司的商业路由器(Bit-Engine 12004s)为基础,开发了原型系统。Bit-Engine 12000 系列路由器此前已在 CERNET2 主干网上规模部署。Bit-Engine 12004s 具有一块主控卡和一块交换卡,可以同时支持 4 块线卡,其中每块线卡都由一块 CPU 板、两块 TCAM 芯片(IDT 75K62100)、一块 FPGA 芯片(Altera EP1S25-780),以及若干块级联的 SRAM 芯片(IDT 71T75602)构成,同时 FPGA 芯片内部也具有自带的 SRAM 存储。Bit-Engine 12004s 支持 Netflow v9,在网络协议方面也有较全面的支持。

3.2.1 转发性能 如图 8 所示,流量生成器(SmartBits 6000C)的两个接口通过光纤与路由设备两个接口相连接,路由设备从一个接口收到数据包,经过查询转发表,然后把数据包从另一个接口送回到流量生成器,流量生成器能够自动统计发送速率和接收速率。路由设备两个接口分别配置地址 1000::1/64 和 2000::1/64,流量生成器同时在两个接口以 1 Gbit·s^{-1}的速率发送 76 B(包括以太网帧头)的 IPv6 数据包,其中与 1000::1/64 相连接口发出的流量目的 IP 地址为 2000::2,源 IP 地址为 1000::2,与 2000::1/64 相连接口发出的流量目的 IP 地址为 1000::2,源 IP 地址为 2000::2,发送时间为 5 min。实验分别测试基于传统一维的转发和基于 FIST+的转发性能。

测试结果如图 9 所示,FIST+与传统路由一样都能达到线速转发。FIST+查找过程分为 4 个阶

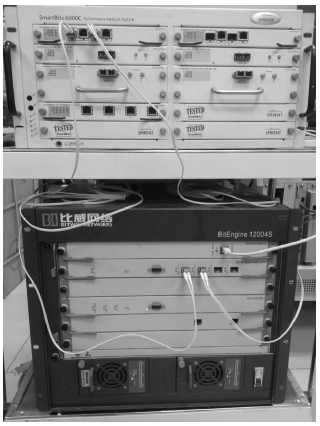


图 8　测试环境

段：TCAM 查找、一级索引表访问、TD 表访问、下一跳映射表访问。4 个阶段互相独立，不存在共享资源冲突，可以使用流水线技术提高查找速度。因为每个报文需要查找两次 TCAM，所以在采用流水线技术后，针对一个报文的查找需要 2 个时钟周期，比传统路由多了一个时钟周期。但是，在通常情况下，路由器并行处理位宽小于最小报文长度，传输处理一个报文的时钟周期大于 2，所以在一定的速率下，查找多出的 1 个时钟周期不会影响转发速率。

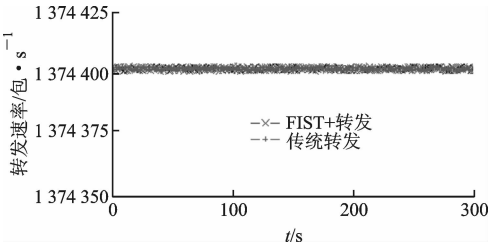


图 9　FIST＋的转发速率

3.2.2　*存储性能*　从 CERNET2 上收集了包括 6 000条 IPv6 地址前缀的 FIB 信息，将这些前缀信息分别以 ACL-like、FIST、FIST＋等转发结构下发到路由设备的 TCAM，并分别统计所占用的存储空间。首先将 6 000 条前缀作为目的前缀，然后从这 6 000条前缀中取出部分前缀作为源前缀，并与目的前缀两两生成对应的转发规则，逐渐增加源前缀的数量。TCAM 所用存储空间如图 10 所示，从图中可以看出，在 6 000 条前缀内，FIST＋的存储空间是保持不变的，而 FIST 呈线性增长，ACL-like 呈指数增长。FIST＋存储空间与一维转发表相同。

4　实验验证

4.1　CERNET2 存在的问题

图 11 给出了 CERNET2 的主干网拓扑结构。作为世界上规模最大的采用纯 IPv6 技术的下一代

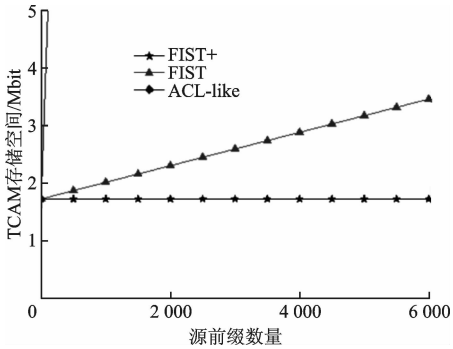


图 10　TCAM 存储空间

互联网主干网之一，CERNET2 同样面临着一些流量均衡问题。首先，CERNET2 有两个国际交换中心连接到全球互联网，北京的 CNGI-6IX 交换中心和上海的 CNGI-SHIX 交换中心。在网络运行的过程中发现，CNGI-6IX 中的路由器的流量过高，导致经常发生拥塞，同时 CNGI-SHIX 中的路由器却经常处于空闲状态。其次，由于用户数量的不断增长，CERNET2 主干网的主要链路都达到了带宽利用率较高的状态，但是也有一些链路带宽利用率很低。如果将一些节点去往国外的流量引入到 CNGI-SHIX，可能造成路径（wuhan，shanghai）带宽利用率过高。根据 CERNET2 的 SLA 参数的要求 $U_S<70\%$，所以如何利用现有空闲链路实现流量均衡是一件很有意义的工作。

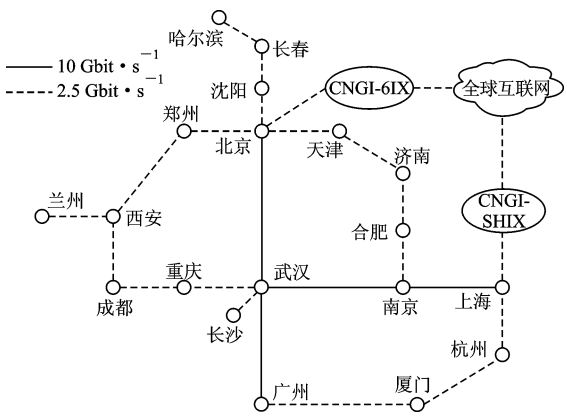


图 11　CERNET2 主干网拓扑结构

CERNET2 主干网各节点之间运行的路由协议是 OSPFv3，且不支持 MPLS，所以流量问题很难解决。特别是两个国际交换中心流量不均衡问题，由于 CNGI-6IX 和 CNGI-SHIX 分别属于不同的自治域，与 CERNET2 主干网通过边界网关协议（BGP）实现路由交互，所以 CERNET2 主干网在北京和上海的自治系统边界路由器（ASBR）需要把 BGP 路由重新分发到 OSPF 协议。在通常情况下，这些路

由在 OSPF 中的类型为 OE2(域外路由 2),特点是选路时不考虑域内开销,所以导致流量不均衡。如果通过在两地 ASBR 上配置修改域外路由的开销,或者采用 OE1(域外路由 1,累加域内开销),根据现有网络结构和 OSPF 选路特点,也会出现目前不平衡的现象。

4.2 实验安装

为了验证本文提出的方案在网络流量均衡方面的可行性,我们搭建模拟 CERNET2 主干网拓扑的真实环境。如图 12 所示,节点之间链路带宽均为 1 Gbit·s^{-1}。限于设备数量,只有需要数据层进行二维转发的节点和需要 4 个以上接口及 netflow 的节点采用 Bit-Engine 12004s,包括节点 B、W、G、N,其他节点采用 Netwire-4604s(比威公司商用过渡设备,具有路由器,支持 4 个千兆接口)。所有节点的控制层均支持本文所提方案。为了模拟 CERNET2 主干网,首先除了节点 F,其他各节点在同一域内配置 OSPFv3;节点 F 分别与节点 B 和 S 建立 BGP 邻居,即节点 F 与其他节点不在同一个自治域;其次,将拓扑中虚线表示链路的 metric 设为 4,实线表示链路的 metric 默认为 1,节点 B 向 OSPF 重分发 BGP 路由的 metric 设为 5,节点 S 向 OSPF 重分发 BGP 路由的 metric 设为 10;最后,通过灌注背景流量,设置链路的带宽利用率,其中链路(B,F)为 60%,链路(W,N)为 40%,其余链路均为 0。

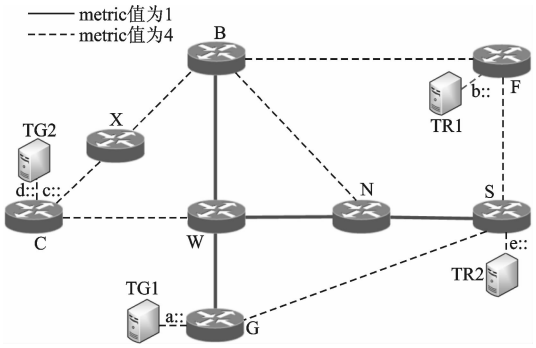


图 12 实验拓扑示意图

4.3 实验过程

设 $U_S=60\%$,$h_u=10\%$,$h_l=10\%$,$I=10$ s,支持手动和自动两种配置方式。在节点 B 上采用手动配置方式,而自动方式主要应用在节点 W 上。

首先,网络在背景流量下运行稳定后,流量发生器(TG1)向流量接收器(TR1)发送 200 Mbit·s^{-1} 的宏流,30 s 后,在节点 B 上配置命令启动流量均衡机制,并选择源地址前缀和目的前缀以及 metric 作为命令参数{src a::,dst b::,metric 50}。然后,过 30 s,TG2 向 TR2 发送两个宏流,一个源地址为 c::,流量为 150 Mbit·s^{-1},另一个源地址为d::,流量为 100 Mbit·s^{-1}。在这个过程中,将测量带宽利用率、控制报文数量、CPU 利用率等数据。

4.4 实验结果

网络运行 30 s 时,在节点 B 上启动流量均衡机制后,节点 B 生成 TD-AS-External-LSA,其中包含的路由信息为{src a::,dst b::,metric 50},并发送到节点 W。节点 W 收到后,通过 SP^{2}F 算法计算,生成二维路由表项[b::,a::,N],之后将收到的宏流(b::,a::)发送给节点 N。60 s 时,由于链路(W,N)带宽利用率超限,节点 W 自动启动流量均衡机制,生成 TD-INTRA-Area-Prefix-LSA,其中的路由信息为{src a::,dst b::,metric 44},并发送到节点 G。节点 G 收到后,通过 SP^{2}F 算法计算,生成二维路由表项[b::,a::,S],之后将宏流(b::,a::)发送给节点 S。由于链路(W,N)此时的实时带宽利用率大于 U_S,所以节点 W 生成 TD-INTRA-Area-Prefix-LSA,其中的路由信息为{src d::,dst e::,metric 15},通过 SP^{2}F 算法计算,生成二维路由表项[e::,d::,G],之后将宏流(e::.,d::)发送给节点 G。

图 13 给出了实验 1 中 4 条链路(B,F)、(W,N)、(G,S)、(S,F)的带宽利用率变化情况。由图中可以看出:30 s 时,链路(B,F)的带宽利用率下降到 60%,链路(W,N)和(S,F)相应增加 20%;90 s 时,链路(W,N)的带宽利用率先是增加到 85%,而后下降 20%,继而降到 55%,链路(G,S)由 0 增加到 30%。图 14 给出了在实验周期内,以 10 s 为时间间隔,测量的 OSPF 控制报文数量情况。图 15 给出了在实验周期内,所用路由器的平均 CPU 利用率的变化情况。从图中可以看出,在保证 TOL 实现流量均衡的前提下,相比于洪泛方式,RPD 传输机制可有效地减少控制报文的数量和路由器的计算负荷。

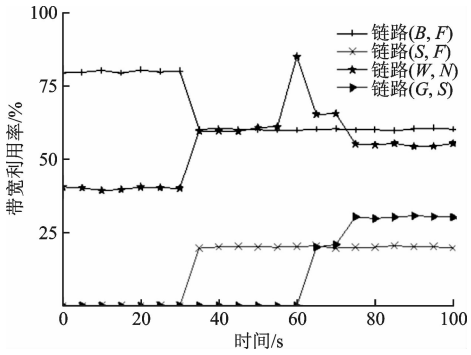


图 13 链路带宽利用率

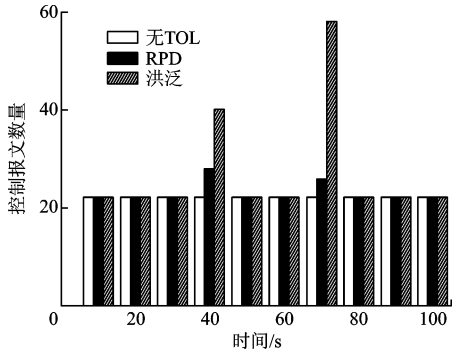


图 14　OSPF 控制报文数量情况

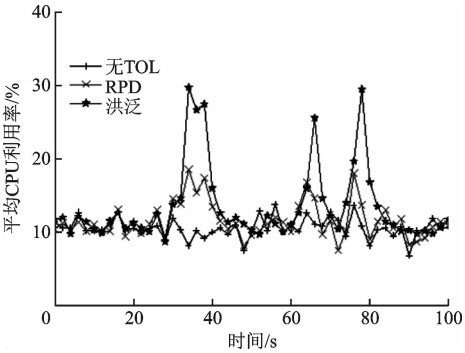


图 15　路由器平均 CPU 利用率

5　结　语

本文将利用二维路由来解决网络链路流量均衡问题，提出了一种二维 OSPF 路由方案 TOL，并在商用路由器上进行了实现。测试和实验结果表明，TOL 方案能够在传统 IP 网络结构和协议的基础上有效实现流量均衡，减少链路拥塞。基于二维路由的流量均衡方案没有 MPLS 的封闭性问题，也没有 SDN 方式与现有 IP 网络兼容问题，在网络发展演进过程中能够发挥承前启后的作用。二维路由主要包括二维路由协议和二维路由转发两方面。根据不同的网络需求，可以灵活地设计与传统路由兼容的二维路由协议。高性能二维转发的研究也会为今后的多维转发提供支持。

参考文献：

[1] FELDMANN A, GREENBERG A, LUND C, et al. NetScope: traffic engineering for IP networks [J]. IEEE Network, 2000, 14(2): 11-19.

[2] ROSEN E, VISWANATHAN A, CALLON R. Multiprotocol label switching architecture [EB/OL]. [2015-03-02]. http://www.ietf.org/rfc/rfc3031.txt.

[3] YASUKAWA S, FARREL A, KOMOLAFE O. An analysis of scaling issues in MPLS-TE core networks [EB/OL]. [2015-03-02]. http://www.ietf.org/rfc/rfc5439.txt.

[4] AGARWAL S, KODIALAM M, LAKSHMAN T. Traffic engineering in software defined networks [C]// Proceedings of the 32nd IEEE International Conference on Computer Communications. Piscataway, NJ, USA: IEEE, 2013: 2211-2219.

[5] BHATIA R, FANG H, KODIALAM M, et al. Optimized network traffic engineering using segment routing [C]// Proceedings of the 34th IEEE International Conference on Computer Communications. Piscataway, NJ, USA: IEEE, 2015: 657-665.

[6] XU M, YANG S, WANG D, et al. Two dimensional-IP routing [C]// Proceedings of the IEEE International Conference on Computing, Networking and Communications. Piscataway, NJ, USA: IEEE, 2013: 835-839.

[7] 吴建平，李丹，毕军，等. ADN：地址驱动的网络体系结构 [J]. 计算机学报，2015，38(6)：1-12.
WU Jianping, LI Dan, BI Jun, et al. ADN: address driven internet architecture [J]. Chinese Journal of Computers, 2015, 38(6): 1-12.

[8] BAKER F J. IPv6 source/destination routing using OSPFv3 [EB/OL]. [2015-05-10]. https://tools.ietf.org/html/draft-baker-ipv6-ospf-dst-src-routing-03.

[9] LINDEM A, MIRTORABI S, ROY A, et al. OSPFv3 LSA extendibility [EB/OL]. [2015-06-20]. https://www.ietf.org/archive/id/draft-ietf-ospf-ospfv3-lsa-extend-10.txt.

[10] MEINERS C R, LIU A X, TORNG E, et al. Split: optimizing space, power, and throughput for TCAM-based classification [C]// Proceedings of the 7th ACM/IEEE Symposium on Architectures for Networking & Communications Systems. Piscataway, NJ, USA: IEEE, 2011: 200-210.

[11] NORIGE E, LIU A X, TORNG E. A ternary unification framework for optimizing TCAM-based packet classification systems [C]// Proceedings of the 9th ACM/IEEE Symposium on Architectures for Networking & Communications Systems. Piscataway, NJ, USA: IEEE, 2013: 95-104.

[12] BLOCH G, RABENSTEIN I, MENES M, et al. Configurable access control lists using TCAM: US 8861347[P]. 2014-10-14.

[13] YANG S, XU M, WANG D, et al. Scalable forwarding tables for supporting flexible policies in enterprise networks [C]// Proceedings of the 33rd IEEE International Conference on Computer Communications. Piscataway, NJ, USA: IEEE, 2014: 208-216.

（编辑　杜秀杰）