doi:10.11959/j.issn.1000-436x.2016080

智慧协同网络中基于流量矩阵的负载均衡路由机制

贾濡[1]，郜帅[1]，罗洪斌[1]，张宏科[1]，万明[2]
(1. 北京交通大学电子信息工程学院下一代互联网互联设备国家工程实验室，北京 100044；
2. 中国科学院沈阳自动化研究所网络化控制系统重点实验室，辽宁 沈阳 110016)

摘 要：智慧协同网络具有能够实时准确测算流量矩阵的特点。将流量矩阵作为约束，对负载均衡路由优化问题进行建模，利用拉格朗日对偶方法，将原问题转化为优化目标易实现的对偶问题。为实现对偶问题优化目标，提出一种基于流量矩阵的负载均衡路由 (TM-LB，traffic matrix based load balancing) 算法，供控制层根据实时网络情况为后续流规划传输路径。利用 OMNET++仿真器在 NFSnet 拓扑结构上进行仿真实验，结果表明 TM-LB 相比传统路径规划机制能有效避免拥塞，实现负载均衡。最后，搭建原型系统对 TM-LB 算法的开销进行测试。
关键词：智慧协同网络；流量矩阵；路由优化；拉格朗日对偶；负载均衡
中图分类号：TP302 **文献标识码**：A

Traffic matrix-based load balancing routing in flow-based smart identifier network

JIA Ru[1], GAO Shuai[1], LUO Hong-bin[1], ZHANG Hong-ke[1], WAN Ming[2]
(1. National Engineering Laboratory for Next Generation Internet Interconnection Devices,
School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China;
2. Laboratory of Networked Control System, Shenyang Institute of Automation Chinese Academy of Sciences, Shenyang 110016, China)

Abstract: Smart identifier network (SINET) has an advantage of being aware of the accurate traffic matrix. Traffic matrix was use as the constraint, model on load balancing routing problem, and transform the primal problem into dual problem by using Lagrange function. Therefore, the optimization goal of the dual problem can be easily achieved. In order to achieve the dual problem optimization goal, a traffic matrix based load balancing routing algorithm (TM-LB) was proposed. By performing the TM-LB, the control plane can plan paths for subsequent traffic according to network fluctuation. OMNET++ was used to run the experiment simulation based on NSFnet topology. The results show that TM-LB can better reduce congestion and realize load balancing, compared with traditional routing mechanism. Finally, a proof-of-concept was built implementation and carry out experiments for testing the overhead of TM-LB algorithm.
Key words: smart identifier network, traffic matrix, routing optimization, Lagrange duality, load balance

1 引言

随着用户规模的增长、多媒体应用的增多，传统互联网已经逐渐暴露出各种弊端。因此，设计一种全新的未来网络体系架构[1]逐渐成为近年来研究的热点。

作为一种新型未来网络架构，智慧协同网络(SIENT, smart identifier network)相比传统互联网具有以下本质上的不同：其通信模式从传统的“以 IP 地址为中心”转变为“以服务内容为中心”；其控制逻辑从数据转发平面解耦出来，从而能够实现集中式的、可编程的、能够面向改革创新的智慧控制

收稿日期：2015-09-28；**修回日期**：2016-01-11
基金项目：国家重点基础研究发展计划（“973”计划）基金资助项目（No.2013CB329100）；国家自然科学基金资助项目（No.61232017, No.61271200, No.61501447）；国家高技术研究发展计划（“863”计划）基金资助项目（No.2015AA011906）
Foundation Items: The National Basic Research Program of China (973 Program)(No.2013CB329100), The National Natural Science Foundation of China (No.61232017, No.61271200, No.61501447), The National High Technology Research and Development Program of China (863 Program) (No.2015AA011906)

管理；SIENT 能够实时感知网络状态、准确测算流量矩阵，并动态分配网络资源。

流量矩阵（TM, traffic matrix）表达了网络中所有源和目的节点对（OD, origin-destination）之间的流量需求。准确及时的流量矩阵对于流量工程、网络服务供应、入侵检测来说是重要的输入。传统的流量矩阵测算方法[2~4]是通过 SNMP 等数据采集方式获取各链路流量值 $\boldsymbol{Y}$，及利用路由器配置信息或路由协议获得路由矩阵 $\boldsymbol{A}$，再将所有信息集中，根据 $\boldsymbol{Y=AX}$ 来估算流量矩阵 $\boldsymbol{X}$。这类方法准确性差，计算开销大，仅适用于小规模网络。

OpenFlow[5]技术的出现使网络中的路由器能够记录所有转发流的大小，OpenTM[6]利用这种固有特点，选取 OpenFlow 网络中的部分路由器来流信息来测算网络流量矩阵。OpenTM 直接以流为单位测算，而非通过链路总流量及路由矩阵估算，所以其准确度远远高于传统流量矩阵估算方法。但 OpenTM 的流量矩阵测算是由控制器集中完成，为控制器造成过大计算开销。

智慧协同网络转发组件也具备以流为单位实时记录转发流量的功能，还能够利用域间路由族群标识（PID, path identifier）和终端组件标识（NID, node identifier）进行分布式流量统计（具体介绍见第 2.2 节和第 2.3 节），在及时获得准确流量矩阵的同时，减小了资源管理器计算开销。文献[7]将智慧协同网络与 OpenTM 机制测算的流量矩阵进行对比，结果表明：智慧协同网络流量矩阵的测算机制比 OpenTM 具有更高的准确度。

本文以“将网络所有链路的最大链路利用率最小化”作为优化目标，以智慧协同网络测算出的实时流量矩阵作为流量守恒约束，对负载均衡路由优化问题进行数学建模，再通过拉格朗日对偶法将原问题转化为对偶问题。利用互补松弛定理，对原问题和对偶问题分析可知：对偶问题的优化目标可通过改进链路权重来实现，即可通过一个实际的网络路由机制来实现对偶问题优化目标，进而使原问题也达到最优化。因此，本文提出一种基于流量矩阵的负载均衡路由（TM-LB，traffic matrix based load balancing）算法，供 RM（resource manager）执行路径规划，然后 RM 给域内各个转发组件设定转发规则，从而实现避免拥塞，负载均衡。该路由机制不仅适用于智慧协同网络，也可应用于其他能够实现集中控制并且基于流的网络，比如软件定义网络(SDN)。在实验部分，本文利用 OMNET++仿真器在 NFSnet 拓扑结构上进行仿真实验，结果表明，TM-LB 相比传统路由机制能有效避免拥塞，实现负载均衡。最后，搭建原型系统对 TM-LB 算法的开销进行测试。

2　智慧协同网络体系架构

2.1　智慧协同网络基本概念

智慧协同网络[8~10]具有资源动态适配的“三层两域”体系结构。

如图 1 所示，智慧协同网络纵向分为三层：智慧服务层[11]，使用服务标识(SID, service identifier)和服务行为描述(SBD, service behavior description)来实现服务命名、查找及匹配；网络组件层[12]，由类似路由器的转发组件(FN, forwarding node)构成，使用组件标识 NID 和组件行为描述(NBD, node behavior description)来进行数据存储、转发及网路状态感知；资源适配层[13]，负责协调上层服务需求与下层网络状态，使用族群标识(FID, family identifier)和族群行为描述(FBD, family behavior description)构建网络族群，动态适配网络资源。族群的划分是基于特定的行为，如用域间路由族群（PID, path identifier）用来标识实现域间传输功能的路径。

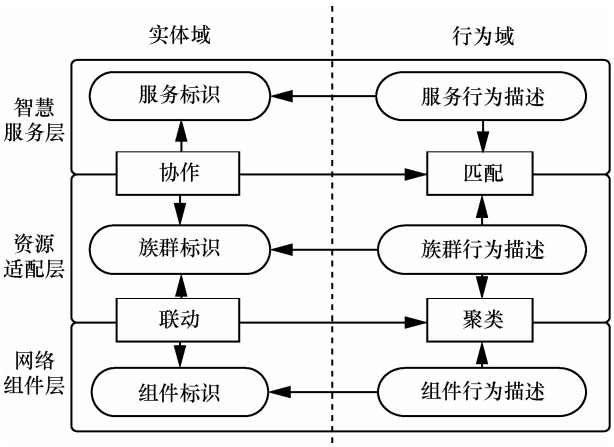

图 1　智慧协同网络三层两域模型[9]

2.2　智慧协同网络数据传输流程

智慧协同网络由自治域构成，分层级部署。其路由分为域内路由和域间路由。域内路由机制具有控制层面与数据层面分离，及以信息内容为中心的特点，每个自治域维护一或多个资源管理器 RM，作为逻辑集中的控制平面，用以维护服务的可达性信息、本域的网络资源并实现智慧资源适配。域间

路由则依靠域间协商的一或多条传输路径[14]，以域间路由族群 PID 来表示。如图 2 所示：域 D_2 与域 D_1 之间有一个域间路由族群 P_1，域 D_2 与域 D_3 之间有 2 个域间路由族群 P_2 和 P_3。

资源管理器通过服务注册获知以 SID 标识的各种服务的存储位置和可达性信息。每个边界转发组件 FN 维护一个域间流表(相当于路由表)，如图 2 中 R_4 维护的流表。所有连接在本域的终端可以视作一个小型的邻域，连接该终端与本域的转发组件也可是看作是一种边界转发组件。连接终端的转发组件维护的流表如图 2 中 R_8 流表。

如图 2 所示，当域 D_3 中以 NID_{r_1} 标识的请求方需要获取以 SID 标识的某个服务时，该请求方向其本地资源管理器 RM_3 发送服务请求消息。该服务请

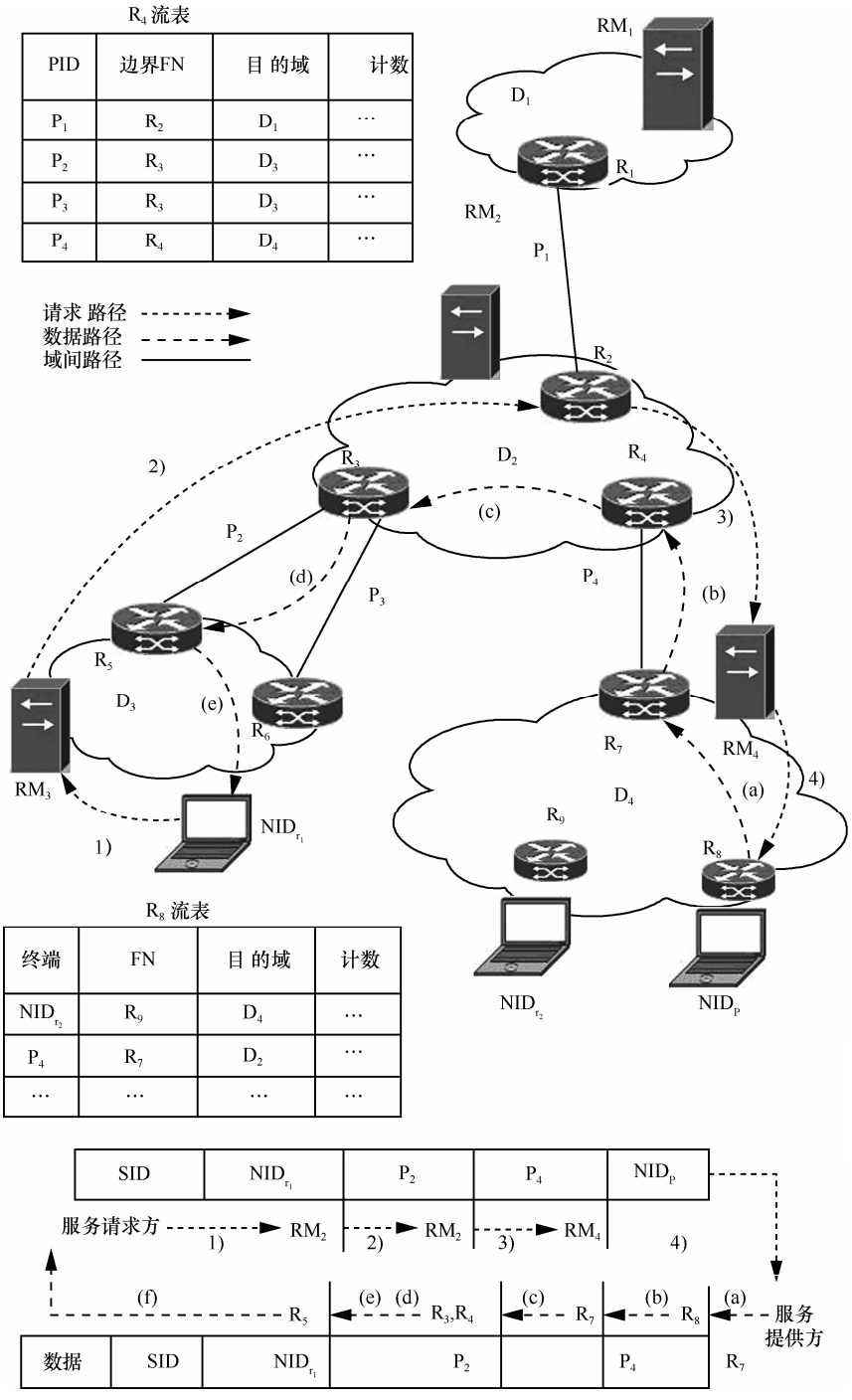

图 2　智慧协同网络数据传输流程

求消息包含该网络组件的组件身份标识、所需服务的服务标识等信息，如图 2 中消息格式 1）所示。后续转发处理过程可参考文献[8]，其数据分组格式变化如图 2 中消息格式 2)~4)。

当以 NID_p 标识的服务提供方收到服务请求后，将收到的域间路由族群(P_2、P_4)、被请求服务标识 SID、请求者的组件身份标识 NID_{r_1} 等放在数据头部，发送至 R_8，如图 2 中（a）所示。数据分组经过 R_7、R_4，最终由 R_5 发送给服务请求方 NID_{r_1}。数据分组格式如图 2 中消息格式(b)~(f)。

2.3　流量矩阵测算机制

如图 2 中 R_4 和 R_8，每个边界路由器都维护一个类似域间路由表的流表，该表记录了所有域间路由族群 PID/NID_r，各 PID/NID_r 连接的边界路由器和目的域，以及转发的流量值。

对于本域所有边界 FN 来说，为获得该点到其他 FN 的流量矩阵，需要将该节点流表中具有相同出口边界 FN 的流量累加，以图 2 所示流程为例，在 D_2 域中，R_4 到 R_3 的流量为流 P_1 和流 P_2 流量计数的累加。各边界 FN 仅需简单累加即可得到该点到本域所有其他点的流量矩阵/向量，资源管理器仅需获取本域所有 FN 节点计算出的流量向量，组成本域的全局流量矩阵。即通过本域所有边界 FN 进行分布式流量计算，进而完成全局流量矩阵的测算。由于智慧协同网络中组件标识是有分类的，即代表终端的组件标识（NID_{r_1}、NID_{r_2}）与代表转发组件的标识（R_7、R_8）是能够明显区分的，所以在计算本域流量矩阵时，仅选出 PID 和终端组件标识 NID 端条目来进行计算是容易实现的。

智慧协同网络中的流量矩阵测算，无需任何先验假设，直接由各边界 FN 计算该节点的流量矩阵，是一种分布式的并行计算模式，不会给 RM 增加过多负载，并且能够在数据转发的同时进行线速计算，具有低开销、实时、准确的优点。

智慧协同网络测算出的实时流量矩阵，可视作网络的流量需求，进而作为负载均衡路由优化问题的流量守恒约束。

此外，需要说明的是，一个自治域中所有 OD 对的流量矩阵由 2 部分构成：流经本域的域间流量，以及源端或目的端位于本域的域内流量。在智慧协同网络架构下，由于域间路由族群标识 PID 可以唯一确定本域连接邻域的某条路径，并且邻域间的路径数量也非常有限，所以本文可以通过统计边界 FN 中的 PID 流表项，简单直接地得到准确及时的域间流量矩阵，这也正是智慧协同网络流量矩阵测算的优势所在。而对于域内流量占多数的自治域，如数据中心网络和校园网络，虽然也可以将终端视作邻域，把终端 NID 视作 PID，用同样方法得到域内流量矩阵，但由于终端数量太过庞大，导致流量矩阵测算的及时性大幅下降，所以流量矩阵测算的优势便不复存在了。

因此本文提出的“利用及时准确的流量矩阵作为网络流量需求”针对的是域间流量占多数的网络类型，比如骨干传输网络。

3　负载均衡路由优化问题

3.1　负载均衡路由优化问题建模

给定一个网络拓扑图 $G(V,E)$，V 表示本域内所有转发节点的集合，$|V|=N$，该自治域有 N 个节点。E 表示所有链路的集合，$|E|=L$，该域有 L 条链路。

同时给定流量矩阵 $\boldsymbol{TM}$ 为一个 N 行 N 列的矩阵，其中，第 i 行 j 列（$i\in V, j\in V$）的值，代表以 i 为源点、j 为目的点的 OD 对之间的流量需求，仅考虑网络中 M 个非零流量需求时，$\boldsymbol{TM}=\{r_1,r_2,\cdots, r_M|r_{ij}\neq 0,\ i\in V,\ j\in V\}$，即对于一组 OD 对$(i_m,\ j_m)$, $m=1,\cdots,M$，用一个标量 $r_m(m=1,\cdots,M)$表示该 OD 对的流量需求，单位为 Mbit/s。

负载均衡路由优化问题以最小化最大链路利用率作为优化目标，用智慧协同网络测算出的实时流量矩阵作为流守恒约束，并且用链路的实际容量限制网络中的链路负载。负载均衡路由优化问题(P1)如下

$$
\begin{aligned}
&\text{(P1)} \qquad\qquad \min\ \theta \\
&\text{s.t.} \sum_{\{j|(i_m,j)\in E\}} x_{i_m j}(m) - \sum_{\{j|(j,i_m)\in E\}} x_{j i_m}(m) = r_m \\
&\qquad \sum_{\{i|(j_m,i)\in E\}} x_{j_m i}(m) - \sum_{\{i|(i,j_m)\in E\}} x_{i j_m}(m) = -r_m \\
&\qquad \sum_{\substack{\{j|(i,j)\in E,\\ i\neq i_m,j_m\}}} x_{ij}(m) \ - \sum_{\substack{\{j|(j,i)\in E,\\ i\neq i_m,j_m\}}} x_{ji}(m) \ = \ 0 \\
&\qquad \sum_{1\leqslant m\leqslant M} x_{ij}(m) \leqslant \theta u_{ij},\ (i,j)\in E \\
&\qquad x_{ij}(m)\geqslant 0,\ (i,j)\in E, m=1,\cdots,M
\end{aligned}
\tag{1}
$$

目标函数中的 θ 为最大链路利用率；约束条件 1 及约束条件 2 分别是在第 m 个流量需求 r_m 下流出源节点和流入目的节点的流量约束；约束条件

3 是除了源和目的节点之外的其他节点在第 m 个流量需求下的流入流量等于流出流量的约束；约束条件 4 是每条链路 l_{ij} 上在所有 M 个业务下的总流量小于等于该链路容量 u_{ij} 与最大链路利用率 θ 的乘积；约束 4 是在所有 m 个流量需求下每条边的流量必须大于等于 0。

因为（P1）模型中目标函数是关于变量 x_{ij} 的线性函数，而约束域为凸集，由凸规划理论[15, 16]可知（P1）模型为凸规划问题，所以存在唯一最优解，且其对偶问题也存在唯一最优解。

3.2 对偶问题转化

当实际网络规模较大时，该优化问题最优解的计算复杂度较高，因此，本文使用拉格朗日对偶法将原问题转化为对偶问题。

首先针对原问题（P1）的流量守恒约束设置对偶变量 p，针对（P1）的容量守恒约束设置对偶变量 q，并将其作为拉格朗日乘子，构造拉格朗日函数为

$$\begin{aligned}&L(x,p,q)=\theta+\\&\sum_{m=1}^{M}p_{i_m}(m)\left(r_m-\sum_j x_{i_m j}(m)+\sum_j x_{ji_m}(m)\right)+\\&\sum_{m=1}^{M}p_{j_m}(m)\left(-r_m-\sum_i x_{j_m i}(m)+\sum_i x_{ij_m}(m)\right)+\\&\sum_{m=1}^{M}\sum_{i\neq i_m,j_m}p_i(m)\left(-\sum_j x_{ij}(m)\ -\sum_j x_{ji}(m)\right)+\\&\sum_{(i,j)}q_{ij}\left(\theta u_{ij}-\sum_{m=1}^{M}x_{ij}(m)\right)\\&\qquad p\text{ free};\ q\leqslant 0\end{aligned}\tag{2}$$

根据对偶理论中互补松弛定理[15,16]可知对偶变量 p 为自由变量，q 不大于零。

对式（2）整理如下

$$\begin{aligned}&L(x,p,q)=\sum_{m=1}^{M}r_m\left(p_{i_m}(m)-p_{j_m}(m)\right)+\\&\left(1+\sum_{(i,j)}u_{ij}q_{ij}\right)\theta+\sum_{m=1}^{M}\sum_{(i,j)}\left(p_j(m)-p_i(m)-q_{ij}\right)x_{ij}\\&\qquad p\text{ free};\ q\leqslant 0\end{aligned}\tag{3}$$

拉格朗日函数 $L(x,p,q)$基于变量 x_{ij} 上的对偶目标函数 $D(p,q)$则可以定义为

$$D(p,q)=inf_x L(x,p,q)\tag{4}$$

可以证明原问题目标函数和对偶目标函数对于任意的可行解 x_{ij} 和(p,q)都满足

$$\begin{aligned}&f(x_{ij})\geqslant D(p,q),\\&f(x_{ij})\,|\,\min\theta,\theta=\max\left[\frac{x_{ij}}{u_{ij}}\right],(i,j)\in E\end{aligned}\tag{5}$$

那么当原问题在最优解处得到下限值 $f(x_{ij})$*的时候，其对偶问题将获得最大值。首先，式(3)右边第一部分为对偶问题的目标函数；然后，根据互补松弛定理[15,16]可知，原问题中 θ 为自由变量，若想对偶问题获得最大值，式(3)右边第二部分中系数项必须等于零，即

$$1+\sum_{(i,j)}u_{ij}q_{ij}=0\tag{6}$$

令 $w_{ij}=-q_{ij}$（代表链路权重），式(3)右边第二部分为

$$\sum_{(i,j)}u_{ij}w_{ij}=1\tag{7}$$

最后，由于原问题（P1）中 $x_{ij}\geqslant 0$，所以式(3)右边第 3 部分系数必须等于零。

通过以上分析得到原问题（P1）的对偶问题（P2）如下

$$\begin{aligned}&(\text{P2})\ \max\sum_{m=1}^{M}r_m\left(p_{i_m}(m)-p_{j_m}(m)\right)\\&\text{s.t.}\qquad\sum_{(i,j)}u_{ij}w_{ij}=1\\&\qquad p_j(m)-p_i(m)+\sum_{(\mathrm{i,j})}w_{ij}\geqslant 0,\forall m,\forall(i,j)\\&\qquad w_{ij}\geqslant 0,\ \forall(i,j)\end{aligned}\tag{8}$$

根据互补松弛定理[15,16]分析得：当原问题（P1）与对偶问题（P2）都取到最优解时，原问题中最优解 x_{ij} 取值大于零，其对应的对偶约束取值等于零；原问题中最优解 x_{ij} 取值等于零，其对应的对偶约束取值大于等于零。

结合网络中的实际意义，对原问题和对偶问题的最优解进行分析。

原问题最优解是实现负载均衡的流量分配

$$\begin{aligned}&f(x_{ij})^*=\{x(m)_1^*,x(m)_2^*,\cdots,x(m)_L^*\},\\&(i,j)\in E,\ |E|=L,m=1,\cdots,M\end{aligned}\tag{9}$$

$f(x_{ij})$*中有 k（$k\leqslant L$）个为非零值，即只有 k 条链路上有流量 r_m 经过（$x_{ij}>0$），其余链路上没有 r_m 流经。以图 3 网络拓扑为例，（为叙述方便，仅考虑流量矩阵 $\boldsymbol{TM}$ 中某个特定流量需求 r_m 的流量分配情况），链路 a、b、c 上有 r_m 流量经过，$x_a,x_b,x_c>0$。

链路 d 上没有 r_m 流量，x_d =0。

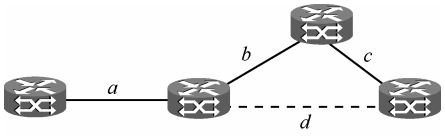

图 3　最优解流量分配示意

根据互补松弛定理分析：原问题和对偶问题都获得最优解时，原问题的解 x_a,x_b,x_c >0 对应的对偶约束取等号

$$p_j(m)-p_i(m)+\sum_{(i,j)}w_{ij}=0 \tag{10}$$

即

$$\sum_{(i,j)}w_{ij}=p_i(m)-p_j(m) \tag{11}$$

对路径 a−b，x_a, >0，x_d =0，所以对偶约束大于等于零

即

$$p_j(m)-p_i(m)+\sum_{(i,j)}w_{ij}\geqslant 0 \tag{12}$$

$$\sum_{(i,j)}w_{ij}\geqslant p_i(m)-p_j(m) \tag{13}$$

根据互补松弛定理，原问题取最优解时，对偶问题也取得最优解 w_l，l∈[1,L]。由式(11)和式(13)可知，实现负载均衡最优路由规划时，被分配流量的路径是链路权重和最小的路径。对偶问题的优化目标可通过改进链路权重来实现，即可通过一个实际的网络路由机制来实现对偶问题优化目标，进而使原问题也达到最优化。

4　基于流量矩阵的负载均衡算法 TM-LB

对偶问题（P2）优化目标

$$\max\sum_{m=1}^{M}r_m\left(p_{i_m}(m)-p_{j_m}(m)\right) \tag{14}$$

可以理解为优先给流量需求高的 OD 对分配链路权重和最小的路径。智慧协同网络能够实时提供准确的流量矩阵作为流量需求，所以该优化的关键在于如何设置链路权重。

现有互联网一般采用 OSPF (open shortest path first) 、IS-IS(intermediate system-to-intermediate system)等协议在域内决策路由，该种协议的本质是为每个节点选择最优下一跳来进行路由。导致网络中不同 OD 对间的路径趋于重合，使某些链路被多条传输路径占用，负载过大，造成局部网络拥塞[17]。

引入图论理论中边转发指数定义[18,19] 如下：在图 $G(V, E)$中，某路由选择 $\boldsymbol{\rho}$ 所确定的路径经过链路 e 的条数。文献[18, 19]中以图论的角度证明：网络容量是否被充分利用取决于路由选择的优劣，而所谓优质的路由选择应该有小的边转发指数。

基于上述思想，设计一种基于流量需求的负载均衡(TM-LB，traffic matrix based load balancing)算法。在控制平面的 RM 中执行该算法进行路径规划。RM 依照规划好的路径给位于数据传输层的各个转发组件设定转发规则。转发组件对后续到来的数据流（以路径族群 PID/终端 NID 标识的流）进行转发。该算法的设计原则为：优先给流量需求高的 OD 对分配路径，并保留其所需链路带宽；尽量减小网络中各链路的边转发指数。

基于流量需求的负载均衡（TM-LB）算法描述及步骤见算法 1。

算法 1　TM-LB

//初始化流量需求为本自治域测算出的流量矩阵；

1) $\boldsymbol{TM}=\{r_1,r_2,\cdots,r_M\,|\,r_{ij}\neq 0,\ i\in V, j\in V\}$

//初始化本自治域所有 L 条链路容量 u_l；链路容量变化标志 f_l=0;

2) $U=\{(u_1,f_1),\cdots,(u_L,f_l)\,|\,u_l>0\}$

//初始化所有 L 条链路权重为 1

3) $W=\{w_1,w_2,\cdots,w_L|w_l=1\}$

4) $G=G(V, E, W)$

//设置一个阈值 K，对 $\boldsymbol{TM}$ 集中，流量值不小于该阈值的 OD 对进行路径权重改进

5)for ($r_m\geqslant K$, $r_m\in\{\boldsymbol{TM}\}$)　{

//找出当前 $\boldsymbol{TM}$ 集中高于 K 的最大流量值

6) $r_{\max}=\max\{\boldsymbol{TM}\}$，($r_{\max}=r_{sd}$)

//去掉不满足最大流量需求的链路

7)if　$u_l<r_{\max}$, $w_l=\infty$,(l=1,⋯,L)

//在 $r_{\max}$ 源点 s，目的点 d 之间选取最短路径

8)　$G=G(V, E, W)$, $q=0$

　　$S=\{s\}$, D[s]=0, $Path\{\}=\varnothing$

　　for each $j\in V$-S　{

　　if (j connect to s)

　　D[j]=min w_l ,l=(s,j)

　　else　D[j]=∞;　}

　　for each $m\in V$-S　{

　　D[k]=min{D[m]}

　　if(D[k]=d)

　　return $S=S\cup\{k\}$

for each $n \in V\text{-}S$ {
if(D[n]>D[k]+D[k][n])
D[n]=D[k]+D[k][n] } }
$q=d$
while (q!=s) {
Path { $r_{\max}$ }= Path { } ∪ {q},
}}}

//为 $r_{\max}$ 保留所需带宽后，求剩余链路容量，链路容量只变化一次。

//并且在 $r_{\max}$ 途径的链路上增加边转发指数，即提高链路权重。

9)for all $l \in E$ ，
if $l \in$ Path {$r_{\max}$}
if (f_l =0){u_l= u_l−$r_{\max}$，f_l ++ }
w_l =w_l+1，

//流量需求 ***TM*** 集去掉元素 $r_{\max}$

10)***TM***={ ***TM*** }−{$r_{\max}$} }

//对 ***TM*** 集中，流量值小于阈值 K 的 OD 对，仅执行最短路径选择。

11)else 执行步骤 8）

5 TM-LB 算法开销

算法的开销包括两方面：1) 算法本身的时间复杂度；2) 在实际系统运行中为实现算法而额外产生的消息开销。

5.1 时间复杂度

TM-LB 算法是基于 Dijkstra 算法[20]改进而来，其时间复杂度取决于用来实现优先队列的数据结构以及用来表示输入图本身的数据结构。为了便于分析，假设输入网络有 N 个节点，L 条链路，那么{***TM***}集中 OD 对至多有 $N(N-1)$个，其中，流量需求大于等于阈值 K 的 OD 对有 M 个。如果将网络输入图用权重矩阵表示，优先队列用无序数组来实现，其时间复杂度主要受 2 部分影响。

第 1 部分为算法中步骤 5)~步骤 8)中内外嵌套的 2 个 for 循环语句。针对{***TM***}集中值大于等于阈值 K 的所有流量需求 OD 对，在源点和目的点之间寻找最短路径。在极端情况下，步骤 8)会遍历网络中的所有链路，其最坏时间复杂度不会大于 $O(L)$。外层 for 循环会执行 M 次，其时间复杂度为 $O(M)$。所以第 1 部分的时间复杂度为 $O(M)O(L)$。

第 2 部分是针对{***TM***}集中值小于阈值 K 的所有流量需求 OD 对，在源点和目的点之间寻找最短路径，也是内外嵌套的 for 循环语句。内层的 for 循环复杂度还是 $O(L)$。外层 for 循环会执行 $N(N-1)-M$ 次。第 2 部分的时间复杂度为 $O(N(N-1)-M)O(L)$。

$$O(M)O(L)+O(N(N-1)-M)O(L) = O(N(N-1)) = O(N^2) \quad (15)$$

所以 TM-LB 算法在极端最坏情况的总时间复杂度为 $O(N^2)$。

而现实环境中的网络一般是稀疏图[20]，所以如果将输入图用邻接链表表示，优先队列用最小堆来实现，时间复杂度可以改进为 $O(L\lg N)$。如果用斐波那契堆[20]来实现优先队列，其最坏时间复杂度可以提升到 $O(N\lg N+L)$。

总之 TM-LB 算法的时间复杂度可以提升到网络规模的线性对数阶多项式，完全能够应用于实时计算。

5.2 系统运行消息开销

作为一种控制与转发分离架构的新型网络，智慧协同网络中控制层 RM 对于转发层 FN 的指挥，也需要依靠类似 OpenFlow 的控制协议来实现，即智慧协同控制协议[21]。RM 通过智慧协同控制协议给 FN 配置流表，指挥组件层的数据转发。智慧协同控制协议支持 3 大类共 10 种消息：下行消息 4 种(属性请求消息、流控制消息、状态读取消息、分组转发消息)，上行消息 4 种(流移除消息、状态汇报消息、端口变化消息、出错消息)，以及同步消息 2 种(Hello 消息、Echo 消息)[21]。

与现有 SDN 架构网络一样，在 RM 中维护有全局网络拓扑信息，如果利用传统路径选择算法(OSPF)，则不需要额外的消息开销，直接在 RM 中利用全局拓扑计算最短路径即可。

而如果利用 TM-LB 算法计算路径，除了拓扑信息以外，还需要本域的流量矩阵信息。由第 2.3 节可知，流量矩阵的计算是由所有 FN 分布式并行进行的。各 FN 仅计算其自己到本域所有其他 FN 的局部流量矩阵，并维护至 FN 的 NBD 信息表中。然后 FN 通过文献[21]中介绍的 NBD 信息交互机制，利用“状态读取/汇报消息”，将局部流量矩阵传递至 RM，经 RM 整合组成本域的全局流量矩阵。

总之，智慧协同网络中运行 TM-LB 算法需要通过“状态读取消息”和“状态汇报消息”将局部流量矩阵从 FN 传递至 RM，也就是说，执行 TM-LB

算法产生了额外的消息开销。

假设在一个自治域中，1 个资源管理器 RM 控制 n 个转发组件 FN，RM 每隔 t s 向 FN 发送“状态读取消息”来问询各个 FN 的局部流量矩阵信息。定义 C 为传递流量矩阵信息产生的总消息数量，那么，在该自治域中执行 TM-LB 算法 T 时长所产生的消息数量为

$$C = 2\frac{T}{t}n \tag{16}$$

可见，执行 TM-LB 算法产生的额外消息数量与自治域中 FN 数量和问询频率成正比。

6　实验及分析

本节首先在 OMNET++仿真平台下，针对 TM-LB 算法的有效性开展仿真实验，将 TM-LB 算法和传统路径计算方法的负载均衡效果进行对比。然后，在运行智慧协同控制协议的 SINET 原型系统中，对执行 TM-LB 算法产生的额外消息开销进行分析。

6.1　TM-LB 算法有效性分析

本文在OMNET++环境下建立NSFNet（national science foundation network）拓扑，并运行智慧协同网络机制。智慧协同网络协议应用于 2 种功能实体：资源管理器 RM 和转发组件 FN。RM 主要功能是进行服务请求分组的查询、映射和转发，以及域间路由族群的封装。FN 主要功能为：监测网络状态和按照 RM 配置的路由表转发数据。

如图 4 所示，该仿真实验拓扑包含 14 个 FN，每个 FN 都连接一个客户端，该模块同时实现服务请求者及服务提供者功能，（图 4 中将客户端与 FN 合并为一个节点）；1 个 RM；42 条链路。各链路带宽为 100 Mbit/s，每一个 FN 会向其他节点发送随机流量，流量大小在 10~70 Mbit/s 随机分布，发送时延设置为从 10~30 ms 随机分布。在该仿真环境下，每个 FN 将该节点到本域其他节点的流量信息发送给 RM。RM 集合各节点计算出的流量形成本域全局流量矩阵。RM 分别运行传统最短路径优先（OSPF）算法和 TM-LB 算法，计算出路径并为本域内各 FN 配置流表。在执行 TM-LB 机制时，RM 中流量矩阵的更新时间间隔为 2 s。阈值 K 分别取 70 Mbit/s、60 Mbit/s 和 55 Mbit/s。

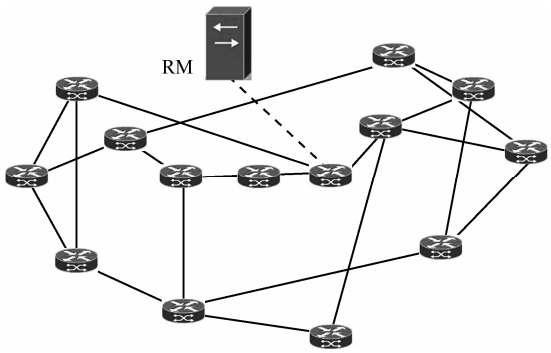

图 4　NSFnet 网络拓扑

图 5 为智慧协同网络架构下，运行传统最短路径选择机制及 K=70 Mbit/s 时 TM-LB 机制的网络链路利用率情况比较。运行传统最短路径选择机制时，链路最低带宽利用率为 9.5%，链路最高带宽利用率为 84.2%，带宽利用率变化方差为 22%。K=70 Mbit/s 时运行 TM-LB 机制时，链路最低带宽利用率为 22.7%，链路最高带宽利用率为 65.1%，带宽利用率变化方差为 9%。TM-LB 机制为需求大于 70 Mbit/s 的 OD 对优先规划路径，保留带宽并增加途经链路权重后，再为其他 OD 对分配传输路径，带宽利用率变化方差可以反映出网络中链路负载均衡情况，TM-LB 机制将带宽利用率方差从 22%降低到 9%，使网络链路利用率得到明显优化。

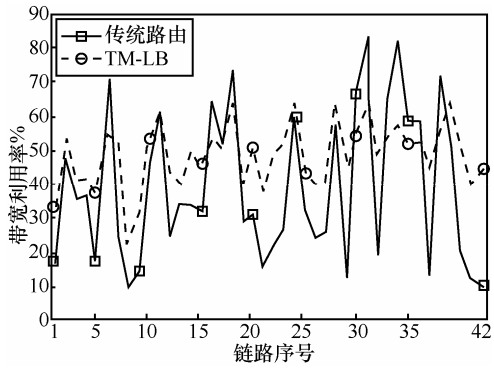

图 5　K=70 Mbit/s 时带宽利用率

图 6 为传统最短路径选择机制及 K=60 Mbit/s 时 TM-LB 机制比较。仿真设置由每一个节点向本域其他节点发送大小为 10~70 Mbit/s 随机流量，所以传统最短路径选择机制的各链路带宽利用率具体数值有所变化，但高负载及低负载链路基本与图 5 一致，其最低及最高链路利用率分别为 10.2%和 86.6%,带宽利用率变化方差为 23%。K=60 Mbit/s 时运行 TM-LB 机制时，链路最低带

宽利用率为 32.6%，链路最高带宽利用率为 57.8%，带宽利用率变化方差为 6%。TM-LB 机制阈值 *K* 从 70 Mbit/s 降低到 60 Mbit/s，将带宽利用率方差从 9%降低到 6%。

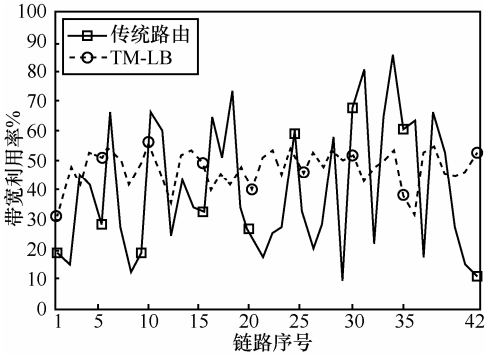

图 6 *K*=60 Mbit/s 时带宽利用率

图 7 为 TM-LB 机制在 *K*=70 Mbit/s、60 Mbit/s 和 55 Mbit/s 链路带宽利用率情况对比。*K*=55 Mbit/s 其最低及最高链路利用率分别为 35.7%和 56.4%, 带宽利用率变化方差为 5.6%。TM-LB 机制中，阈值 *K* 从 70 Mbit/s 降低到 60 Mbit/s、55 Mbit/s，RM 对更多的 OD 对进行流量路径规划，减小链路的转发指数，使本域带宽利用率方差从 9%降低到 6%、5.6%，链路负载均衡效果得到提升。但是当 *K* 值降低到一定程度，TM-LB 机制的负载均衡效果不会再提高，如图 7 中 *K*=70 Mbit/s 和 *K*=60 Mbit/s。

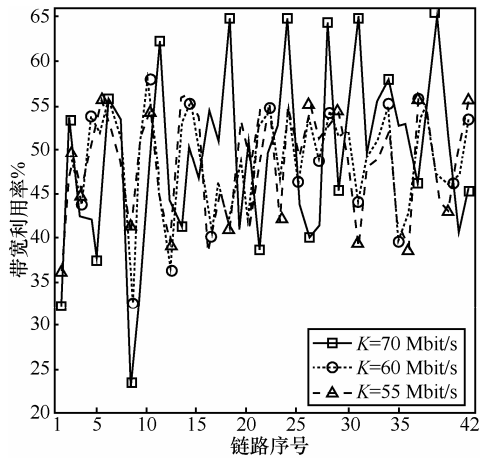

图 7 *K* 值变化时 TM-LB 机制带宽利用率

总之，TM-LB 能够根据实时流量矩阵合理规划数据传输路径，相比于传统路由算法，有效降低带宽利用率变化方差，实现负载均衡。通过阈值 *K* 取值变化时 TM-LB 实现效果对比可知，合理设定阈值 *K*，能在实现网络负载均衡的同时，尽量减小计算开销。

6.2 TM-LB 算法消息开销分析

智慧协同网络的 RM 中维护有全局网络拓扑信息，如果执行传统路径选择算法(OSPF)，则直接利用全局拓扑计算最短路径即可，不需要额外的消息开销。而执行 TM-LB 算法计算路径，除了拓扑信息以外，还需要本域的流量矩阵信息。

为了分析 TM-LB 算法产生的额外消息开销，本文搭建了与图 4 拓扑结构相同的 SINET 原型系统。SINET 原型系统主要功能实体为 1 个资源管理器 RM 和 14 个转发组件 FN，42 条链路，其操作系统均为 Linux2.6.29.5，运行类似 OpenFlow 的智慧协同控制协议实现 RM 与 FN 之间的通信控制，并且，在 RM 和 FN 之间利用“状态读取消息”和“状态汇报消息”来实现局部流量矩阵信息从 FN 到 RM 的传递。

在智慧协同网络中，“状态读取消息”和“状态汇报消息”采用相同格式的数据分组头，长度为 8 byte。“读取消息”内容部分为 16 byte 的 FN 组件标识 NID，用来指明被读取的 FN。“汇报消息”的内容部分包括 16 byte 的 NID 和 FN 提供的局部流量矩阵数据，在当前测试环境中流量矩阵数据部分统一设定为 336 byte（42×8 byte，NSFnet 中 OD 对数量为 42 个）。

通过第 5.2 节中的算法消息开销分析可知，执行 TM-LB 算法产生的额外消息数量与自治域中 FN 数量和问询频率成正比，而通常 FN 的数量是一定的，所以本文研究问询频率与系统中消息开销的关系。

在测试过程中，RM 以 *t* 为时间间隔，向每个 FN 发送“状态读取消息”，而 FN 则返回携带着局部流量矩阵的“状态汇报消息”。本文在 RM 上用 TcpDump 工具捕获“读取消息”和“汇报消息”数据分组来进行分析。图 8 为问询时间间隔 *t* 分别取 2 s、4 s 和 6 s 情况下，RM 中执行 TM-LB 算法所需要处理的额外消息开销情况。如图 8 所示，在问询间隔取 2 s 时，RM 处理“读取消息”和“汇报消息”数据的平均开销为 21 kbit/s，也就是说，为了利用 TM-LB 算法来计算传输路径，RM 需要处理 21 kbit/s 的数据。在问询间隔取 4 s 和 6 s 时，RM 执行 TM-LB 算法产生的消息开销分别为 10.5 kbit/s 和 7 kbit/s。可见，随着问询时间间隔的增加，执行 TM-LB 算法的消息处理开销

会降低。但问询频率过低，会导致 RM 中维护的全局流量矩阵的及时性得不到保证。

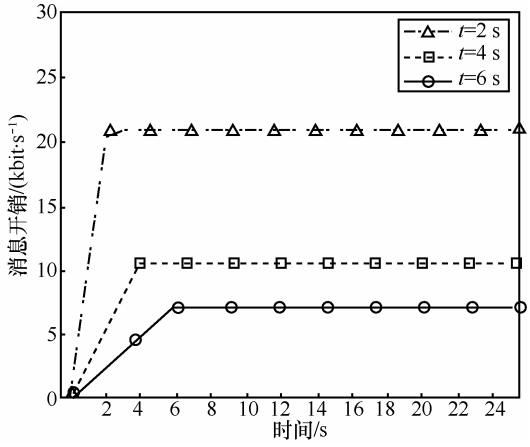

图 8　TM-LB 算法消息开销

另一个影响算法开销的因素是 FN 的数量，由于利用 TM-LB 算法进行路径分配，通常针对的是域间流量占多数的骨干传输网络。而本测试环境采用的 NFSnet 拓扑正是美国 Tier-1 骨干传输网络拓扑，所以基于该拓扑结构进行的开销分析是具有代表性的。如图 8 所示，即使问询时间间隔取值 2 s，执行 TM-LB 算法给 RM 带来的额外消息处理开销仅为 21 kbit/s，这个开销大小对于 RM 的处理能力来说是完全能够接受的。

此外，像 NFSnet 这种骨干传输网络拓扑通常不会采用全联通模式，其 14 个骨干路由器由 42 条链路相连，存在 42 个 OD 对，所以“状态汇报消息”提供的局部流量矩阵数据长度为 42 个 8 byte，共 336 byte。假设，将某骨干网络中路由器数量提升至 30 台，OD 对数量提高至 100，那么，“状态汇报消息”头部、NID 加上流量矩阵数据部分，长度共计 824 byte，如果问询频率仍然设为 2 s 一次的话，TM-LB 算法的额外消息处理开销仍然不超过 100 kbit/s。

总之，在智慧协同网络架构下的传输网络中利用 TM-LB 算法计算路径所产生的消息开销非常有限，是完全可以接受的。

7　结束语

智慧协同网络中的流量矩阵测算，采用分布式的并行计算模式，并且能够在数据转发的同时进行线速计算。本文以负载均衡路由最优化作为优化目标，以智慧协同网络测算出的实时流量矩阵作为流量守恒约束，对负载均衡路由优化问题进行数学建模。再通过拉格朗日对偶法将原问题转化为对偶问题。设计出一种基于流量矩阵的负载均衡路由（TM-LB）算法，将对偶问题的优化目标转化为网络中实际的路由机制，从而实现负载均衡路由问题的最优规划。在 OMNET++仿真平台上，将 TM-LB 算法和传统路经计算方法的负载均衡效果进行对比，结果表明 TM-LB 能够有效避免拥塞，实现链路负载均衡。最后，为了分析 TM-LB 算法的实际消息开销，本文在 SINET 原型系统中基于 NFSnet 拓扑结构进行测试，得出结论：在传输网络中利用 TM-LB 算法计算路径的话，给 RM 带来的额外消息处理开销不会超过 100 kbit/s，是完全能够接受的。

本文研究重点是域内路由优化，下一步工作将进行多域路由优化的研究。传统 BGP 域间流量工程的实现机制是在各域边界路由器上建立代理。各域代理负责处理本域的路由策略、通告、网络状态，并且与其他域的共同作为逻辑集中的控制层面对跨域的路由进行优化，是一种覆盖（overlay）机制。而智慧协同网络本身具有资源管理器 RM，能够直接实现集中控制。所以如何结合服务行为描述（SBD）信息来对跨域的全局路由进行优化，是下一步的研究重点。

参考文献：

[1] PAN J L, SUBHARTHI P, RAJ J. A survey of the research on future internet architectures[J]. IEEE Communications Magazine, 2011, 49(7):26-36.

[2] MEDINA A, TAFT N, SALAMATIAN K, et al. Traffic matrix estimation: existing techniques and new directions[J]. ACM SIGCOMM Computer Communication Review. 2002, 32(4): 161-174.

[3] 赵国锋, 王灵矫, 唐红, 等. 基于 IP / MPLS 网络的动态业务流量矩阵测量模型[J]. 通信学报, 2003, 24(10):145-152.
ZHAO G F, WANG L J, TANG H, et al. An architecture for traffic matrix measurement in IP/MPLS based network [J]. Journal on Communications, 2003, 24(10):145-152.

[4] 杨扬, 周静静, 杨家海,等. 流量矩阵估算算法研究[J]. 计算机科学, 2009, 36(7):42-45.
YANG Y, ZHOU J J, YANG J H, et al. Traffic matrix estimation algorithm based on square root filtering[J]. Computer Science, 2009, 36(7):42-45.

[5] MCKEOWN N, ANDERSON T, BALAKRISHNAN H, et al. OpenFlow: enabling innovation in campus networks[J]. ACM SIGCOMM Computer Communication Review, 2008, 38(2): 69-74.

[6] TOOTOONCHIAN A, GHOBADI M, GANJALI Y. OpenTM: traffic matrix estimator for OpenFlow networks[C]//Passive and active measurement. Springer Berlin Heidelberg, c2010: 201-210.

[7] LUO H B, CHEN Z, ZHANG H K, et al. An approach for efficient,

accurate, and timely estimation of traffic matrices[C]//The INFOCOM WKSHPS 2014. Toronto, Canada, 2014.

[8] LUO H B, CHEN Z, ZHANG H K, et al. CoLoR: an information-centric Internet architecture for innovation[J]. IEEE Network Magazine, 2014, 28(3):4-10.

[9] 张宏科, 罗洪斌. 智慧协同网络体系基础研究[J]. 电子学报, 2013,4(7): 1249-1254.
ZHANG H K, LUO H B. Fundamental research on theories of smart and cooperative networks[J]. Acta Electronic Sinica, 2013,41(7): 1249-1254

[10] ZHANG H K, QUAN W, CHAO H J, et al. Smart identifier network: a collaborative architecture for the future Internet[J]. IEEE Network, under Review, 2015.

[11] 苏伟, 陈佳, 张宏科. 智慧协同网络中的服务机理研究[J]. 电子学报, 2013,41(7): 1255-1260.
SU W, CHEN J, ZHANG H K, et al. Research on the service mechanisms in smart and cooperative networks[J]. Acta Electronic Sinica, 2013,41(7): 1255-1260.

[12] 郜帅, 王洪超, 王凯, 等. 智慧网络组件协同机制研究[J]. 电子学报, 2013,41(7): 1261-1267.
GAO S, WANG H C, WANG K, et al. Research on cooperation mechanisms of smart network components[J]. Acta Electronic Sinica, 2013,41(7): 1261-1267.

[13] JIA R, CHEN Z, LUO H B, et al. Status-aware resource adaptation in Information-centric and software-defined network[J]. China Communications, 2013, 10(12):66-76.

[14] CHEN Z, LUO H B, CUI J B, et al. Security analysis of a future internet architecture[C]//The 21th IEEE International Conference on Network Protocols. Gottingen, Germany, c2013:1-6.

[15] 陈宝林. 最优化理论与算法[M]. 北京：清华大学出版社, 2005.
CHEN B L. Theory and algorithms of optimization[M]. Beijing: Tsinghua University Press, 2005.

[16] 谢金星, 邢文训, 王振波, 等. 网络优化[M]. 北京：清华大学出版社, 2009.
XIE J X, XING W X, WANG Z B, et al. Network optimization [M]. Beijing: Tsinghua University Press, 2009.

[17] 杨洋, 杨家海, 王会, 等. IP 网络时延敏感型业务流自适应负载均衡算法[J]. 通信学报, 2015, 36(3): 2015082.
YANG Y, YANG J H, WANG H, et al. Towards load adaptive routing based on link critical degree for delay-sensitive traffic in IP networks[J]. Journal on Communications, 2015, 36(3): 2015082.

[18] 徐俊明. 图论及其应用[M]. 北京: 中国科学技术大学出版社, 2004.
XU J M. Graph theory with applications [M]. Beijing: Press of University of Science and Technology of China, 2004.

[19] 徐俊明. 组合网络理论[M]. 北京: 科学出版社, 2007.
XU J M. Combinational network [M]. Beijing: Science Press, 2007.

[20] THOMAS H C, CHARLES E L, RONALD L R, et al. 算法导论[M]. 潘金贵, 译. 北京：机械工业出版社, 2006.
THOMAS H C, CHARLES E L, RONALD L R, et al. Introduction to algorithms[M]. Beijing: China Machine Press, 2006.

[21] JIA R, SU W, LUO H B, et al. Status message transmission mechanism in SINET[J]. Journal of Internet Technology, 2015, 16(4): 727-734.

作者简介：

贾濡（1988-），女，吉林长春人，北京交通大学博士生，主要研究方向为未来互联网体系架构、软件定义网络、智慧路由等。

郜帅（1980-），男，河南济源人，博士，北京交通大学副教授、硕士生导师，主要研究方向为下一代互联网关键理论与技术。

罗洪斌（1977-），男，重庆人，博士，北京交通大学教授、博士生导师，主要研究方向为通信网络技术、未来互联网体系结构、网络生存性。

张宏科（1957-），男，山西大同人，博士，北京交通大学教授、博士生导师，北京交通大学下一代互联网互联设备国家工程实验室主任，主要研究方向为下一代信息网络关键理论与技术。

万明（1984-），男，内蒙古通辽人，博士，中国科学院沈阳自动化研究所副研究员，主要研究方向为未来网络与信息安全、工业控制系统信息安全。