doi:10.11959/j.issn.1000-436x.2015082

IP 网络时延敏感型业务流自适应负载均衡算法

杨洋[1,2,3]，杨家海[1,2]，王会[1,2]，李晨曦[1,2]，王于丁[1,2]

（1. 清华大学 网络科学与网络空间研究院，北京 100084;
2. 清华信息科学与技术国家实验室（筹），北京 100084；3. 西安通信学院 信息管理中心，陕西 西安 710106）

摘 要：互联网对时延敏感的业务数据流，要求具有较低的端到端时延，但是网络拥塞的发生，将会使服务质量无法保证。基于链路关键度提出了一种新的自适应负载均衡路由算法(LARA，load adaptive routing algorithm)，能最大限度地避开拥塞链路从而减少端到端延迟。该算法通过得到一个优化目标函数，并利用凸优化理论将优化目标函数分解为若干个子函数，最终得到一个简单的分布式协议。利用 NS2 仿真器在基于 CERNET2 真实的拓扑结构上进行仿真实验，同时与网络中能普遍部署的等开销多路径（ECMP，equal-cost multi-path）算法相比较，通过测试反馈时延、分组丢失率、流量负载，结果表明 LARA 具有更好的自适应性和健壮性，性能相比更优。

关键词：网络拥塞；关键链路；链路关键度；多路径路由；负载均衡

中图分类号：TP393.1 **文献标识码**：A

Towards load adaptive routing based on link critical degree for delay-sensitive traffic in IP networks

YANG Yang[1,2,3], YANG Jia-hai[1,2], WANG Hui[1,2], LI Chen-xi[1,2], WANG Yu-ding[1,2]

(1. Institute for the Network Sciences and Cyberspace, Tsinghua University, Beijing 100084, China;
2. Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing 100084, China;
3. Information Management Center, Xi'an Communication Institute, Xi'an 710106, China)

Abstract: Delay-sensitive traffic requires lower end-to-end delay in IP networks, such as online video, VoIP, video conference. Based on the criticality degree of link. A load adaptive routing algorithm (LARA) was presented which could avoid the link to be congested to reduce the end-to-end delay. Firstly, an optimization objective function has been put forward; and then decomposed into several sub-functions by using convex optimization theory; finally, the optimization objective function and sub-functions were transformed into a simple distributed protocol. LARA with ECMP (equal-cost multipath) routing strategy was compared which was widely deployed in the network by using NS2 simulation under CERNET2 topology. By evaluating the feedback delay, packet loss rate and traffic load, the results show that LARA can exhibit good performance and achieve excellent load balance, and meanwhile improve the robustness of the link when using multipath routing technology.

Key words: network congestion; critical link; criticality degree of link; multipath routing; load balance

1 引言

伴随着宽带互联网增值业务进入消费者市场，交互式应用程序变得越来越流行，如视频直播、VoIP、多媒体会议、在线游戏等。截止到 2013 年 12 月，中国网络视频用户已达 4.28 亿，网民使用网络视频业务的比例上升至 69.3%，在选择网站时，约四成的用户考虑了“播放流畅”因素[1]。在

收稿日期：2014-01-08；**修回日期**：2014-05-28

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（2012CB315806）；国家自然科学基金资助项目（61170211，61202356, 61161140454）；教育部博士学科专项基金资助项目（20110002110056, 20130002110058）

Foundation Items: The National Basic Research Program of China (973 Program) (2012CB315806); The National Natural Science Foundation of China (61170211, 61202356, 61161140454); Specialized Research Fund for the Doctoral Program of Higher Education (20110002110056, 20130002110058)

全球范围内，思科预计在2018年[2]，互联网视频流量将会占到所有互联网流量的80%到90%，相比2013年增长了66%。随着我国“三网融合”的加快，实时多媒体业务对承载网络的服务质量(QoS)能力提出了更高的要求。在QoS的指标当中，最小化网络时延是针对视频类网络服务的关键度量指标[3]。为了保障视频类应用的互动性以及能够做到实时回放，数据分组交付的低时延是必须要保证的，即使是一个短暂的突发延迟都可能会影响到用户体验。但是目前的互联网设备对这种时延敏感型的数据流传输并没有做到很好的支持，数据流在链路中传输容易遭受瞬时的链路拥塞，造成突发的时延或分组丢失，从而降低视频流的质量。

之前的研究工作，用QoS的方法能够对有延迟和带宽需求的应用程序提供保障[4]，满足某个应用程序在特定的带宽和延迟下数据分组的交付，例如RSVP协议。然而，这种方法需要网络中的资源协调工作，因为它需要沿数据传输路径上的每一个节点为应用程序的请求预约保留资源。此外，还需要使用准入控制，即在数据源端根据端到端的网络特性和所需要保证的QoS对进入网络的流进行分类，如果时延的增加超过某一阈值，那么这些数据源端将不允许发送数据。本文试图找到一个更简单更实用的解决方案，不需要为每流进行资源预订，同时允许用户在他认为需要的时候能够不受限制地传输数据。

目前，由于多路径路由[5~7]技术与传统单路径路由相比，能够提供链路的负载均衡、容错、聚合可用带宽以及提高端主机的吞吐量[8]等优点，使其越来越受到重视。文献[9,10]主要针对路由快速恢复和路由保护，当一条链路发生拥塞或者不可达，多路径路由技术能对数据流进行重路由，从而避开拥塞或者失效链路，提高数据交付的可靠性和健壮性。但是，如果是因为链路的暂时故障，例如路由器断电等因素，等到故障链路恢复后，路由又切换回首选主路由，这样路由的频繁切换，容易引起路由震荡[11]。无论是路由快速恢复或路由保护，都是为主路由选择备份路由，其实质还是单路径路由。并行的多路径路由，例如ECMP[12]算法采用简单的轮询方式将数据分组均匀地分布到多条等代价的路径上，达到负载均衡的目的。并行多路径路由还可以根据应用服务的需求，首先对数据分组进行分类，例如需要低时延的业务、高吞吐量的业务或者安全度高的业务，然后同时在多条路径上并行地传输数据。并行多路径传输机制相比于单路径能够保障低时延业务需求，例如VoIP的应用[13]，只是对数据分组进行分类将会增加额外的数据平面开销。而本文使用的关键技术也是并行多路径路由，不同的是，为减少数据平面的开销，并不对数据分组进行分类，而是直接对基于链路关键度的开销进行优化，最小化链路总时延。

文献[14~16]为最小化网络传播时延采用的方法，是为每个源节点指定大概有多少流量从该节点离开并进入与该节点连接的链路。虽然文献的研究工作是采用多路径路由并在理论上能达到时延的最优值，但是还存在如下几个问题。首先，文献[14]无法做到在流量发生波动的情况下提供一个合适的步长值使算法收敛，这个问题在文献[15]中得到解决，但是增加了算法的复杂度，如算法需要估计时延二阶导数的下限和上限值；其次，测量开销的问题[16]，算法需要测量链路的传播时延、剩余带宽等信息，无论主动测量还是被动测量都需要增加一定的开销值并对实际流量产生影响，测量造成的误差容易对链路的性能产生误判，影响算法的准确性。相比之下，本文的方法更实用，因为不需对当前的路由协议进行太多的修改，其次充分利用链路状态协议的扩展信息[17]，降低算法复杂度，减轻路由器计算负担。

文献[18]提出了最小干扰路由算法(MIRA, minimum interference routing algorithm)，它的关键思想是在源和目的节点对之间选择一条对未来业务产生最小干扰的路径。MIRA算法中关键链路的定义是当链路容量减少一个单位时，源目的节点对之间的最大流也减少。该算法的目标是选择包含尽可能少的关键链路的路径，一般可以避免瓶颈效应。在本文中，将重新定义关键链路，并基于链路关键度提出了一种新的算法能最大限度地避开拥塞链路，从而减少端到端延迟。本文优化的目标函数是以链路关键度作为链路流量分配的权值函数，通过将目标函数的优化分解[19]从而获得一个简单的分布式解决方案，同时，将它转换成一个实际的协议，即基于链路关键度的自适应负载均衡路由算法(LARA, load adaptive routing algorithm)，使它能够被网络中的路由器和数据源端很好地执行。算法的设计目标是：1)能最小化网络流量的时延；2)通

过避免拥塞链路，尽可能满足用户对流量传输速率的需求；3)提高网络的健壮性，能够适应网络瞬时的性能下降，降低分组丢失率。算法的设计将确保数据流所选择的路径相对较短且通过拥塞避免能保证链路的最小时延甚至是在链路负载发生变化的时候。最后通过在 NS2 中进行模拟评估，展示了在真实的网络拓扑和反馈延迟的环境下 LARA 的性能，证明了该协议更简单、更实用。

综上所述，本文设计的 LARA 算法协议具备如下特点。

1) 提出基于链路关键度的凸优化问题，将算法分为离线预计算和在线计算 2 部分，减少数据平面的开销，降低算法复杂度，减轻路由器计算负担。通过对链路权值进行优化，最小化链路时延，保证时延敏感业务端到端的服务质量。

2) 采用并行多路径路由技术，基于反馈路由节点的 OSPF 扩展域信息，通过凸优化理论求解数据源端最佳的速率分配值，做到自适应调节并行多路径上的速率分配问题，使得链路总时延最小。该协议并不需对当前的路由协议进行太多的修改。

2　算法关键技术分析

一个好的路由算法要能适应网络拓扑和流量条件的改变，并且只有在网络发生拥塞时才能充分体现出它的优点。本文提出了基于链路关键度算法，根据最大流−最小割定理，考虑节点对之间最大流流经每条链路的影响，每条链路上可能经过的路由的影响以及链路剩余带宽的影响，使尽可能多的路由均衡地通过网络，并且采用预先计算的方法减少动态路由时的计算复杂度，从而使该算法高效快捷。

2.1　关键链路描述

目前的互联网域内路由协议广泛采用的是 OSPF、IS-IS 等协议，均属于单下一跳路由机制。路由协议仅计算数据源和目的节点之间的最优路径，节点对间的分组信息传输时总是沿最优路径传输，这样会导致网络中不同源目节点对间的最优路径趋于重合。依据这种路由机制所选的最优路径会保持相对的稳定，但是却容易使某些链路被长时间过多占用，形成关键链路。这些关键链路上传输负载过大，导致出现局部网络拥塞。而在这些链路负载过大甚至出现拥塞的同时，其他可用链路却基本闲置。以欧洲的科研教育骨干网 Geant[20]为例，其平均链路利用率仅有 2%左右，但其网络中关键链路的带宽占用率却高达 90%。

在图 1 所示的简单网络拓扑中，假设节点间链路代价均为 1，源目节点对（*S*1，*D*1）、（*S*2，*D*2）、（*S*3，*D*3）根据最短路径优先算法都将选择 7-8 这段链路，那么 7-8 链路势必成为关键链路，也越容易发生拥塞。对于节点对（*S*3，*D*3），在链路 7-8 发生拥塞的时候，同样能到达目的的路径 5-9-10-6，负载较轻，那么提出的新算法将对这一现象加以改善，通过定义链路的关键度，即关键度高的链路越容易成为关键链路，那么对链路关键度相对较高、剩余容量相对较低的链路提高其惩罚因子，这样使流量能在有效链路之间做到均衡分配。

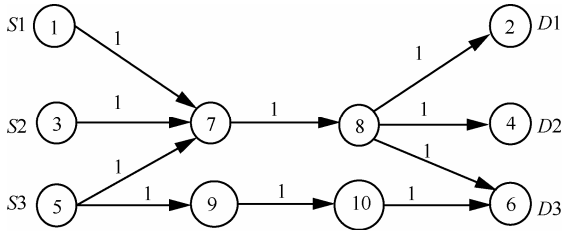


图 1　关键链路

2.2　链路关键度定义

设置一个无向连通图 $G=(V,E)$作为网络模型来定义链路关键度。其中 V 代表网络中所有节点的集合，E 代表所有节点之间的链路集合，P 表示节点对的集合，图中每个节点都有唯一的标识。s 和 d 表示图 G 中的任意两点，其中 $s,d\in V$，$(s,d)\in P$。

对于给定一个流网络，目标就是最大化网络效用，找出一个具有最大值的流，基于最大流−最小割定理，首先定义链路 l 的平均期望负载 $AVE(l)$这个概念。定义链路的平均期望负载，即链路的平均每流大小，就是这条链路期望的最大流值，这能反映出一条链路质量的好坏，对于期望值越高的链路，说明链路质量越好，源端都选择这条链路的概率值就大。本文提出的路由协议，由离线预计算和在线计算 2 个阶段来完成，这样做的好处是减轻路由器的计算开销。对链路关键度的定义是在离线预计算阶段完成。在离线预计算阶段，链路 l 的平均期望负载$AVE(l)$就是所有节点对(s,d)之间最大流中通过链路 l 的流量之和与通过链路 l 的最大流路径数目 m 的比值。用公式表示为

$$AVE(l)=\sum_{(s,d\in P)} f_l(s,d)/m \tag{1}$$

其中，$AVE(l)$表示链路 l 的平均期望负载，$f_l(s,d)$ 表示节点对(s,d)之间的最大流中通过链路 l 的流量，m 表示所有节点对中通过链路 l 的最大流路径数目。根据链路的平均期望负载，确定链路的关键度。链路的关键度定义为链路的平均期望负载与链路的容量比值。公式如下

$$\rho(l) = AVE(l)/C_l \tag{2}$$

其中，$\rho(l)$ 表示链路 l 的关键度，$C(l)$ 表示链路 l 的容量。可以看出，链路 l 的关键度的值越大，表示链路 l 越关键，相应地链路就容易造成堵塞。以图 1 为例，假设每条链路带宽均为 100 Mbit/s，链路代价均为 1，由于源端 $S3$ 存在等代价链路，将 $S3$ 发送的最大流流量平均分配，经计算，节点对（$S3$，$D3$）之间所有链路中，链路 7-8 的关键度值最高 $\rho(l_{7\text{-}8}) \approx 0.83$，其余链路关键度值为 0.5，那么链路 7-8 即为关键链路也是最容易产生拥塞的链路，符合上一节的推断。下面将在算法的在线计算阶段讨论基于链路关键度的链路成本开销计算。

2.3 数据源流量分割

流量分割是指网络中某一个路由节点，对到达相同目的节点的数据分组实现多下一跳链路并行的转发，使网络中各链路的资源利用率趋于均衡，最大程度地降低网络数据传输中的拥塞。采用多路径传输最主要的好处是能够降低端到端延迟并保证更好的传输可靠性。在本文中没有指定数据源是否为边缘路由器或端主机以及新的协议是否只能在域间或者域内进行部署，将分别针对这些不同的场景讨论使用多路径的可行性。

在域内采用多路径路由相对容易完成，因为所有流量的起始节点和目的节点都在一个自治域内，路由器能够计算出K条最短路径或者通过域内网络管理员在端节点之间设置多隧道。当然，域间的多路径路由也是可以做到的[5,21]，例如在重叠路由中，数据源可以将流量直接路由到某一指定路径上，如果数据源是多宿主的，那么将流量根据业务的不同需求路由到某一上游提供商提供的有效链路上变得更加容易实现。

当源端能够对瞬时低时延链路进行选择的时候，那么针对链路负载的动态流量分割将能提高网络传输性能。另外，这种方式在链路发生故障时，源端可以避免失效链路或由于链路的高关键度而造成拥塞的链路。所以，灵活的流量分割技术能够在多条不同质量的链路上根据业务需求进行流量配置，很好地做到链路负载均衡。如何在有效的多路径上进行流量分割，将在下一章中重点讨论。

3 核心优化问题定义

目前，将最优化理论应用于网络研究领域已取得了显著的成果。总体上，它在为网络效用最大化(NUM)[19]，尤其是对拥塞控制[22]和覆盖网络的优化[23]设计新的分布式协议时发挥了重要作用。本文中，将利用凸优化理论对时延敏感业务流量进行优化。上一节介绍了算法设计的关键技术，本节将确定优化的核心问题，并利用优化分解理论来解决核心优化问题。优化分解技术分为主问题分解和对偶问题分解这 2 种方法，前者是分解原始的主要问题，而后者利用拉格朗日对偶函数法求解问题，本文将采用后者方法。通过提出优化目标函数以及约束条件，同时能满足是一个凸优化问题，那么其最优解就是数据源端在多路径上的最佳流量分配值，并能够保证端到端路径时延最小化。

3.1 问题的相关定义

上节中已经定义了网络中某一特定的链路 l，以及链路的容量C_l。假设网络能提供多路径路由，并在源目节点对间存在多条有效路径。这里定义路由矩阵 $\boldsymbol{H}^i$，其中 i 代表源目节点对，H_{lj}^i 表示节点对 i 选择的路径 j 中某条链路 l。如果流量流经链路 l，则 $H_{lj}^i = 1$，否则 $H_{lj}^i = 0$。路由矩阵 $\boldsymbol{H}$ 没有必要代表物理拓扑中所有可能的路径，但是却可以代表由网络管理员挑选出来的路径。

由于每个源目节点对之间可以将流量在多条有效路径上进行分配，用符号 r_j^i 表示连接源目节点对 i 的路径 j 上分配速率，用符号 $(Hr)_l$ 表示链路 l 的总负载，用 R_i 表示每个源目节点对 i 之间的流量需求。观察到，对于视频流的带宽需求在 10~30 s 的间隔内是恒定不变的，所以在本文中假设 R_i 是恒定比特率，但同时必须认识到在这段时间间隔之间，由于视频压缩速率的变化，带宽的需求还是有变化的，在后面的实验中会反映出这一点。

3.2 核心优化问题

本文的优化目标是最小化端到端的网络延迟。换句话说，需要优化的目标是在网络中保证分配在有效多路径上的所有时延敏感型数据流其端到端的时延最小化，同时也需要数据源需求流量恒定来

确保满足优化的一定条件，以及链路的实际负载不能超过链路的带宽容量等约束条件。

在本文第 2 节中，为了减轻路由器的计算负担，针对路由算法的设计提出了离线预计算和在线计算 2 个阶段。在离线阶段主要完成链路关键度值的计算。目前的路由协议中，最常用的就是定义链路容量或可用带宽的倒数作为链路的边权值。例如，某一刻，一条链路剩余带宽较大，按照以前定义的倒数权值关系，那么权值相对较小，越容易吸引流量，但是如果这是一条关键度值很高的链路，根据第 2 节定义 AVE 的物理意义，在下一刻这条链路是容易产生拥塞的，给它分配过多的流量并不一定合适。如果将这条链路的关键度值作为剩余带宽倒数的修正因子，权值就不一定小了。所以，为了更好地刻画链路的实时状态，将离线阶段得到的链路关键度值与该链路可用带宽相比，其值能很好地反映出链路的动态特性，提高算法的准确性。在线计算阶段，可以获取链路的可用带宽 $C_l^{'}$。根据式(1)、式(2)以及链路的可用带宽，定义链路的成本开销为链路的关键度与链路的可用带宽的比值。公式如下

$$cost(l)=\frac{\rho(l)}{C_l'} \tag{3}$$

其中，$cost(l)$表示链路 l 的开销。式(3)将链路关键度与该链路可用带宽的比值定义为链路的成本，相当于惩罚因子，那些高关键度、低可用带宽的链路就是本文要进行拥塞避免的链路，可以将一部分流量引导至相对值较低的有效链路上，保证端到端的低时延。根据这个设计思路，最终提出的优化目标函数为 $f(r)=\sum_i\sum_j r_j^i\sum_l H_{lj}^i\frac{\rho_l}{C_l'}$，即 $\text{Mininize}\sum_i\sum_j r_j^i\sum_l H_{lj}^i\frac{\rho_l}{C_l'}$，其中，$r$ 为变量。优化目标函数的提出将伴随着一定的约束条件，首先链路的负载不能超过链路的承载能力，即 $(Hr)_l\leqslant C_l$；其次要确保分配在每一个源目节点对 i 流量的总和等于用户需求流量，即 $\sum_j r_j^i=R_i$，其中 R_i 应为常量；最后是链路流量分配的非负取值约束，即 $r\geqslant 0$ 。

对于约束条件 $\sum_j r_j^i=R_i$，这里假设 R_i 是常量。首先，在前文中提到大部分视频流的带宽需求在 10~30 s 的间隔内是恒定不变的，也符合本文的假设条件，如果时间间隔外需求流量发生改变，那么算法将重新计算并收敛得到最优解；其次，如果 R_i 是变量，那么优化问题必将引入流量矩阵，将使算法的复杂度增加，加重路由计算的负担。基于以上原因，本文中定义 R_i 为常量。

3.3　凸优化证明

为了使本文提出的优化目标函数有唯一解，那么需要证明提出的核心优化问题是凸函数。对于不等式约束条件

$$(Hr)_l\leqslant C_l,\forall l \tag{4}$$

因为是线性的，所以是凸函数。同样对于等式约束条件

$$\sum_j r_j^i=R_i,\forall i \tag{5}$$

是仿射函数，所以只需要证明优化目标函数

$$\sum_i\sum_j r_j^i\sum_l H_{lj}^i\frac{\rho_l}{C_l'} \tag{6}$$

是凸函数即可。由于目标函数是线性函数，根据凸函数性质可以证明式(6)是凸函数。

3.4　优化问题分解

首先将提出的核心优化问题利用拉格朗日对偶法重新定义为

$$L(r,\beta,\delta)=\sum_i\sum_j r_j^i\sum_l H_{lj}^i\frac{\rho_l}{C_l'}+\sum_l\beta_l((Hr)_l-C_l)+\sum_i\delta_i(R_i-\sum_j r_j^i) \tag{7}$$

式中引入 2 个新的对偶变量 β_l 和 δ_i，它们也叫拉格朗日乘子，分别关联着链路的不等式约束条件以及数据源端的等式约束条件。这 2 个对偶变量的引入可以认为是当流量分配背离约束条件时的惩罚代价。在分布式算法中，这种惩罚性的代价值可以在节点路由器上通过次梯度方法计算得到，并可以将计算得到的代价值反馈给数据流源端。

上一节中已经证明了本文提出的优化问题属于凸优化，且满足利用 KKT(karush-kuhn-tucker)条件来找到多路径上最优的流量分配速率，即 r_j^i 的最优解，KKT 条件是拉格朗日乘子法的推广。算法中，假设在 t 时刻，源端 i 获得来自不同链路的代价值 $\beta_l(t)$，根据不同链路的代价相应地分别计算源端各自发送速率的代价值 δ_i。即在满足 KKT 条件下，通过拉格朗日对偶法分解得到链路分配的最优速率 $r_j^{i*}(t)$ 应满足式(8)，即满足式(8)的解为最优解。

$$\frac{\partial}{\partial r_j^i}(L(r,\beta,\delta))=0 \tag{8}$$

将式（7）整理如下

$$L(r,\beta,\delta)=\sum_i\sum_j r_j^i\sum_l H_{lj}^i\frac{\rho_l}{C_l'}+$$
$$\sum_l\beta_l((Hr)_l-C_l)+\sum_i\delta_i(R_i-\sum_j r_j^i)$$
$$=\sum_i\sum_j r_j^i\{\sum_l H_{lj}^i(\frac{\rho_l}{C_l'}+\beta_l)-\delta_i\}-\sum_l\beta_l C_l+\sum_i\delta_i R_i$$

则由初始问题通过拉格朗日变化得到函数 $L(r,\beta,\delta)$，那么其基于变量 r 上的对偶目标函数 $D(\beta,\delta)$ 则可以定义为

$$D(\beta,\delta)=\inf_r L(r,\beta,\delta) \tag{9}$$

可以证明初始目标函数和对偶目标函数对于任意的可行解 r 和 (β,δ) 都满足 $f(r)\geqslant D(\beta,\delta)$，那么当原始优化问题在最优解处得到下限值 f^* 的时候，其对偶函数将获得最大值。其对偶形式为

$$\max_{\beta,\delta} D(\beta,\delta)$$
$$\text{s.t.}\quad \beta_l\geqslant 0 \tag{10}$$

可知即使初始目标函数不是凸函数，式(10)也是凸优化问题[19]。将式(9)通过变形得到

$$D(\beta,\delta)=\inf_r L(r,\beta,\delta)$$
$$=\sum_i\sum_j r_j^i\{\sum_l H_{lj}^i(\frac{\rho_l}{C_l'}+\beta_l)-\delta_i\}-\sum_l\beta_l C_l+\sum_i\delta_i R_i$$
$$=A(r)-\sum_l\beta_l C_l+\sum_i\delta_i R_i \tag{11}$$

其中，

$$A(r)=\min\sum_i\sum_j r_j^i\{\sum_l H_{lj}^i(\frac{\rho_l}{C_l'}+\beta_l)-\delta_i\}$$

通过初始问题的对偶形式可以看到，对偶问题的目标函数 $D(\beta,\delta)$ 被分解为独立的子问题 $A(r)$、$\sum_l\beta_l C_l$ 和 $\sum_i\delta_i R_i$。这就表明通过求解对偶问题，各源端只需要知道局部信息，而不需要知道全局信息，这样就可以通过分布式算法来设计新的协议。

由于式(10)属于凸优化问题，则对偶目标函数 $D(\beta,\delta)$ 属于凸函数，其导数存在。对代表链路代价的变量求偏导数为

$$\frac{\partial D}{\partial\beta}=(Hr)_l-C_l=\sum_i\sum_j H_{lj}^i r_j^i-C_l$$

利用次梯度算法可以得到反馈回来的链路代价更新算法

$$\beta_l(t+1)=[\beta_l(t)-\varepsilon_\beta(C_l-\sum_l H_{lj}^i r_j^i)]^+ \tag{12}$$

其中，ε_β 是迭代步长，每一条链路上 β_l 的更新是由链路负载和链路容量之间的差异而决定的。如果源端对链路 l 的带宽需求超过了链路的自身的带宽，例如链路发生拥塞，只有在这种情况下表达式 $[\ ]^+$ 代表正值，相当于提高链路代价；否则就降低链路代价。

同样通过求解变量 δ 的偏导数，再利用次梯度算法可以得到数据源端的发送代价

$$\delta_i(t+1)=[\delta_i(t)-\varepsilon_\delta(\sum_j r_j^i-R_i)]^+ \tag{13}$$

根据式(11)中的 $A(r)$ 可以获得源端 i 在路径 j 上的速率表达式，同样在对偶的目标函数中对变量 r 求偏导数为

$$\frac{\partial D}{\partial r}=\sum_l H_{lj}^i(\frac{\rho_l}{C_l'}+\beta_l)-\delta_i$$

那么为了选择一个目标函数值下降最快的方向以利于尽快达到极小值点，利用最速下降法得到源端速率的更新算法为

$$r_j^i(t+1)=r_j^i(t)+\varepsilon_r[\delta_i-\sum_l H_{lj}^i(\frac{\rho_l}{C_l'}+\beta_l)] \tag{14}$$

当 $\delta_i=\sum_l H_{lj}^i(\frac{\rho_l}{C_l'}+\beta_l)]$ 时，源端 i 在路径 j 上的速率则不需要更新。式(12)~式(14)中，ε_β、ε_δ、ε_r 分别是在迭代的 t 时刻定义的迭代步长。如果逐步减小步长值，例如当 $t\to\infty$ 时 $\varepsilon\to 0$，分布式算法的目标函数将收敛于全局目标函数。

4 分布式多路径协议设计

上节中，提出了优化问题的目标函数，并利用拉格朗日对偶函数法求解该优化问题，对偶问题的目标函数 $D(\beta,\delta)$ 分解为独立的子问题 $A(r)$、$\sum_l\beta_l C_l$ 和 $\sum_i\delta_i R_i$。这就表明通过求解对偶问题，各数据源端只需要知道局部信息，而不需要知道全局信息，这样就可以通过分布式算法来设计新的协议，最终利用最优化理论与数学方法得到式(12)~式(14)。本节将利用上一节得到的可调参数以及数据源端速率的更新算法，同时根据网络实际运行情况对链路负载进行合理地替换，最终转换成一个实际的算法协议，即基于链路关键度的自适应负载均衡路由算法(LARA，load adaptive routing algorithm)，使其能够收敛于目标函数的最

优解，并使 LARA 能够在路由器和数据源端部署运行。

路由器在网络中能做到监控与其连接的链路性能，计算链路代价以及将计算的代价反馈回数据源端。路由器更新链路代价是在粒度为 T 的时间间隔内迭代更新的，那么在这个时间间隔内到达链路 l 的比特数 B_T 作为该时刻链路的负载并与该链路的容量进行比较，即式(12)中链路负载 $\sum_l H_{lj}^i r_j^i$ 由 $\frac{B_T}{T}$ 替换。

数据源端是根据路由器反馈的链路代价信息进行链路分配速率的调整，值得注意的是源端收到的路由器反馈信息是有延迟的。例如，源端 i 收到来自路径 j 的代价反馈信息是伴随着往返时间延迟，即一个 RTT_j^i 周期。因此，让源端 i 更新所有的路径发送速率是以这些路径中 RTT 值最长的作为更新触发时刻，即 $T_i = \max(RTT_j^i), \forall j$ 。最终结合式(12)~式(14)得到分布式算法协议，如算法 1 所示。

算法 1　LARA

//采样时刻 t=1,2,3,…,每条链路

1) 接收网络中经过链路l的各源端发送速率 r_j^i // $i \in P, j \in E$

2) 更新反馈链路代价：

$\beta_l(t+T) = [\beta_l(t) - \varepsilon_\beta(C_l - \sum_l H_{lj}^i r_j^i)]^+$　　// ε_β 为反馈链路代价步长，$\sum_l H_{lj}^i r_j^i = B_T / T$

3) 将新的链路代价值传给数据源端

4) 计算数据源端 i 的需求代价：

$\delta_i(t+T_i) = [\delta_i(t) - \varepsilon_\delta(\sum_j r_j^i - R_i)]^+$　　// ε_δ 为源端需求代价步长

5) 数据源端 i 在路径 j 上分配速率更新计算：

$r_j^i(t+T_i) = r_j^i(t) + \varepsilon_r[\delta_i - \sum_l H_{lj}^i(\rho_l / C_l' + \beta_l)]$

// ε_r 为迭代步长，$T_i = \max(RTT_j^i)$

在算法 1 中，将使 ε_β、ε_δ、ε_r 步长参数之间相互独立，减少关联性。LARA 中要强调实现步长的递减是不切实际的，因为无论何时一条新的数据流进入网络或者随着一条数据流的离开，步长参数都会增加，所以将选择一个恒定的步长参数来简化协议，但需要寻找到合适的步长值。同时要注意到，如果步长值选取太大，其解决方案可能最终离最优值相差较多，而如果选取太小，则协议的收敛速度会变得非常缓慢。在下一节仿真实验中将对步长值的选取做进一步讨论。

5　仿真与评估

仿真实验的目标是能够展示出所设计协议的性能以及它的动态特性。本文利用 NS2 仿真模拟器在基于真实的网络拓扑上进行仿真。首先，介绍仿真中使用的网络拓扑以及具体 NS2 参数的设置；其次，将选择设置协议中的步长参数并讨论其设置的合理性；最后，通过与 ECMP 比较结果展示 LARA 设计的优点。在实验过程中，曾对美国 Internet2 骨干网 Abilene 进行拓扑仿真（共 20 个主节点，30 条链路），因为其拓扑结构较为简单，满足实验条件链路较好遴选，实验结果完全符合预期。同样，还从 CAIDA 网站获取真实的 AS 拓扑数据生成较为复杂的拓扑（随机选取 50 个 AS 节点，生成 100 条链路），依然取得较好的实验结果，说明本协议是能适应较大规模网络部署的。因篇幅有限，本文选择更接近实际生活的 CERNET2 主干网进行实验仿真。

5.1　NS2 仿真及拓扑

分布式路由协议的研究需要大规模网络拓扑环境的评估和验证，但是受资源、技术条件和场地等因素限制，往往很难在实际的网络系统中完成 LARA 的验证和测试工作。NS2 仿真器具有强大的数据处理功能，可扩展性强，执行效率高，且仿真结果的可靠性高，本文将用 NS2 对提出的 LARA 协议进行仿真评估。仿真的拓扑图是基于第二代中国教育和科研计算机网 CERNET2 骨干网，其覆盖中国几十个主要城市，其拓扑结构包括 25 个主节点(其中 20 个城市节点)和 28 条链路。如图 2 所示。

图中路由器代表 20 个城市节点，连接这 25 个主节点的链路为 28 条。其中，粗实线代表链路带宽为 10 Gbit/s，细实线为 2.5 Gbit/s。实际模拟中，将整个 CERNET2 视为一个 AS，遴选 PoP (point of presence)的原则是：1)每个 PoP 对至少包含一条主干链路，即图 2 中代表 10 Gbit/s 链路的粗实线；2)每个 PoP 对之间都有多于一条的备选路径，并具备相同路径开销。基于上述原则，选择 4 对源目节点对作为 PoP，其主节点分别是沈阳到南京、北京大学到广州、北京邮电大学到重庆、兰州到长沙，其局部拓扑如图 3 所示。

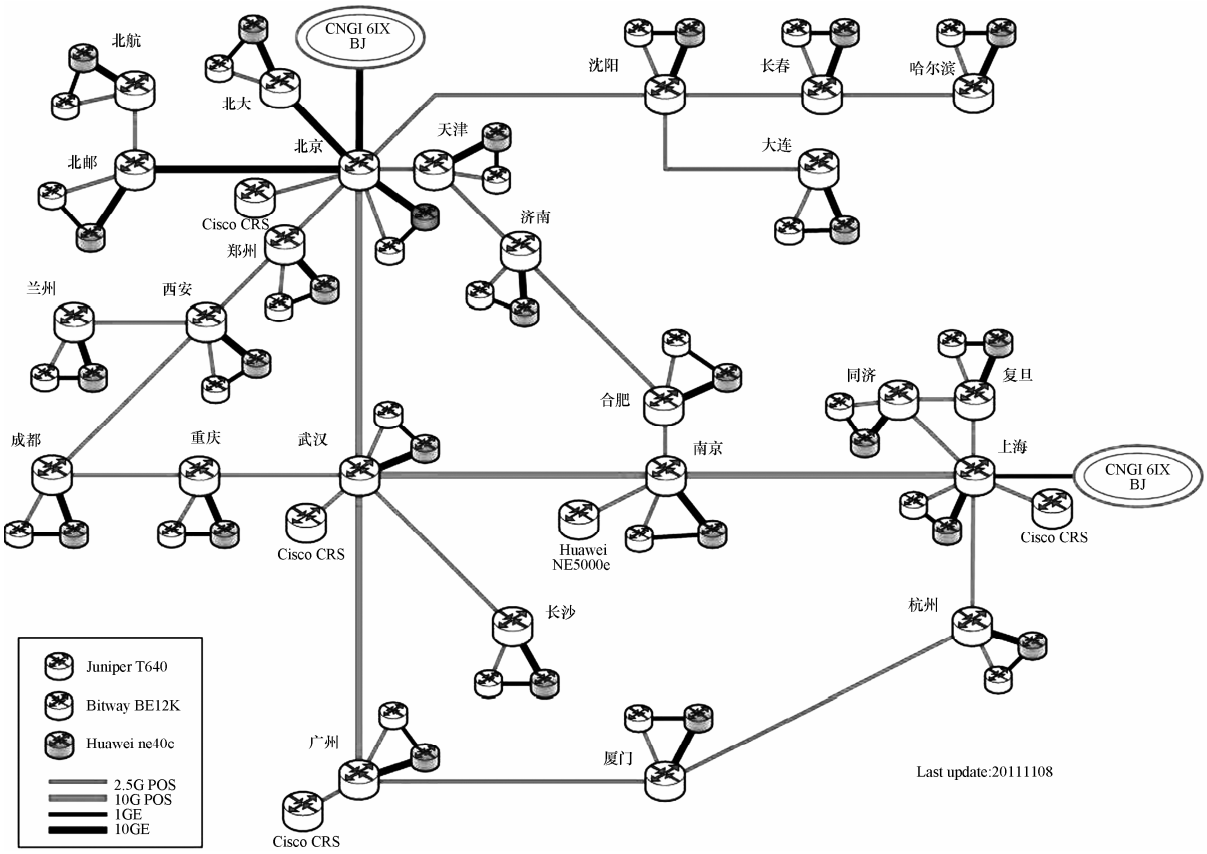


图 2 CERNET2 网络拓扑

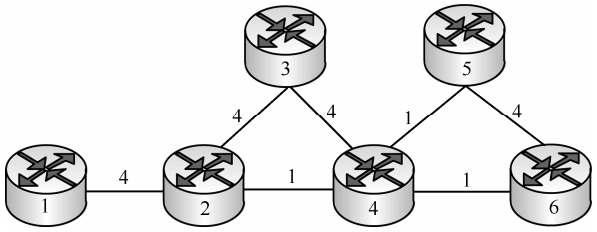


图 3 CERNET2 局部网络拓扑

根据实际链路的带宽、时延，以及时延与带宽的反比关系，图中链路上的值即为链路的代价值，例如节点 1 和 2 之间链路带宽代表原拓扑图中 2.5 Gbit/s 链路，节点 2 和 4 之间链路带宽代表原拓扑图中 10 Gbit/s 链路。实际仿真中，首先，设定 100 Mbit/s 链路容量代表实际 10 Gbit/s 链路，25 Mbit/s 链路容量代表实际 2.5 Gbit/s 链路，严格满足实际链路的带宽以及开销的比例关系；其次，将在源目节点对之间选择使用 TCP 与 UDP 协议混合的传输模式，因为 UDP 协议通信开销较小，所以目前大部分对时延要求较高，而对可靠性要求不高的应用，例如视频点播等，都采用 UDP 协议进行通信。

LARA 还涉及到链路剩余带宽的信息获取，知道在一个运行 OSPF 协议的自治域内，每个路由器都维护有路由信息数据库，路由表中的每一条记录都包含有链路的当前状态，OSPF 的所有 LSA(链路状态广播)都包含有 Option 域字节，那么链路剩余带宽信息可以通过对 Option 域中 ToS(type of service)域的扩展[17]得到，并经过周期性的链路通告广播出去。

5.2 步长参数选取

LARA 中涉及到 3 个步长参数的选择，即 ε_β、ε_δ、ε_r 分别代表反馈链路代价步长、源端需求代价步长以及源端在路径 j 上分配速率迭代步长。在前文中提到如果步长值选取太大，其最终解可能离最优值相差较多，而如果选取太小，则协议的收敛速度会变得非常缓慢。利用 Matlab 仿真工具对步长值进行选取，其好处是能快速的遍历所设置的参数空间，并进行结果的比对。以反馈链路代价步长 ε_β 为例，假定一条链路带宽分别为 10 Mbit/s、100 Mbit/s 和 1 000 Mbit/s，通过实验结果发现链路带宽越小，所需步长值相对与高带宽链路步长值越大，最终可以得到反馈链路代价步长 $\varepsilon_\beta \approx 1/C_l$，即步长值与该链路带宽成反比关系，用

同样的方法可以确定源端需求代价步长近似满足 $\varepsilon_\delta \approx 0.5/R_i$ 的关系。在获取更新后的反馈链路代价以及源端需求速率更新代价值之后，最终需要确定源端在路径 j 上分配速率迭代步长 ε_r。对 ε_r 步长值的选取，假设速率的随机初始值在[0,10] Mbit/s 之间均匀选取 10 个，选取的每一个迭代步长值的迭代次数值都是 10 次计算平均的结果。在仿真实验中，考虑到某些初值选取会影响算法的收敛速度，实际操作时如果算法的迭代次数超过 200 则直接让算法迭代过程终止退出，最后得到 $\varepsilon_r \approx 10^{-5}$ 比较合适。至此 LARA 所需步长值均得到确认。步长参数值的最终确认是经过对不同的拓扑仿真（Abilene、CAIDA），并经过多次实验得到，具有普适性。

5.3　性能比较

由于目前关于多路径路由协议的研究成果大部分都没有实际的部署，为了客观地评估性能，利用 NS2 仿真模拟器在基于 CERNET2 真实的网络拓扑上，将 LARA 与目前在网络中普遍部署的路由算法 ECMP 进行比较。在实际部署 ECMP 的网络中，数据源节点通过测量获取到所有能到达目的节点的最短路径，即这些路径的开销值都是相同的，那么数据源端可以将流量平均的分割到这些路径上进行传输，这即是 ECMP 的基本原理。

在实验中，将 4 个节点对流量需求以 5 Mbit/s 的迭代步长从 5 Mbit/s 增加到 70 Mbit/s 来观察目的节点平均产生的时延、分组丢失率以及吞吐量。同时，为了更真实模拟实际场景，通过配置 NS2 随机数产生器，使每个数据源端在 0~1 s 内由产生的随机数决定数据流开始传送的时刻。首先观察 LARA 平均传播时延的表现，实验结果如图 4 所示。发现当源端发送速率在 5 Mbit/s 至 11 Mbit/s 之间时，ECMP 与 LARA 时延表现相当，说明这一阶段网络链路状态良好，无拥塞链路出现。当源端发送速率超过 12 Mbit/s 时，ECMP 时延开始增大；当发送速率在 35 Mbit/s 到 60 Mbit/s 之间，时延产生较大抖动，网络性能开始下降，发送速率在 41 Mbit/s 时，延时出现峰值。实验现象符合预期，这是 ECMP 本身的设计所不能避免的结果，因为当链路负载增加到一定值，数据分组传输路径上开始出现拥塞链路，由于 ECMP 的算法并不能将数据流切换到链路负载较轻，且未发生拥塞的路径上，从而使分组丢失率开始增加，网络性能下降，这样对于视频服务来说已经开始影响到用户体验，ECMP 路由使数据分组交付时延增大，已不能满足用户对源端提供服务的需求。相比之下，LARA 则能将时延控制在有效范围内，总体平稳无抖动产生，能体现出 LARA 的优越性。

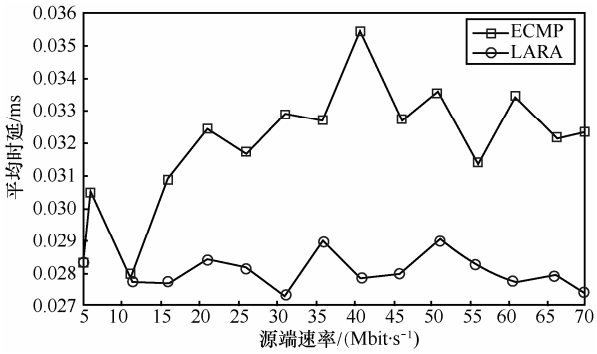


图 4　时延比较

LARA 分组丢失率的测试是在同一实验环境下进行测量，如图 5 所示。当源端发送速率低于 5 Mbit/s 时，ECMP 和 LARA 均无分组丢失产生；当发送速率在 5 Mbit/s 至 35 Mbit/s 之间时，ECMP 控制分组丢失率总体上优于 LARA，但是当发送速率大于 35 Mbit/s 时，LARA 部署下的分组丢失率开始明显下降，而 ECMP 部署下的分组丢失率开始增加，并在 51 Mbit/s 时刻，分组丢失率达到峰值。在对时延的测试中，当发送速率在 35 Mbit/s 到 60 Mbit/s 之间，ECMP 部署下的时延产生较大抖动，网络性能开始下降，其原因是随着负载增大，链路出现拥塞时，分组丢失率开始增加，当数据源端进入 TCP 拥塞避免时，分组丢失率开始下降，如此反复是造成 ECMP 实验结果波动的原因，这也是 ECMP 机制所无法避免的。这一现象与本次分组丢失率测试结果相一致，即在网络时延增大的情况下，分组丢失率也相应增大。由此可以得出：LARA 的网络部署在链路高负载的情况下能显示出对 ECMP 部署的优越性。

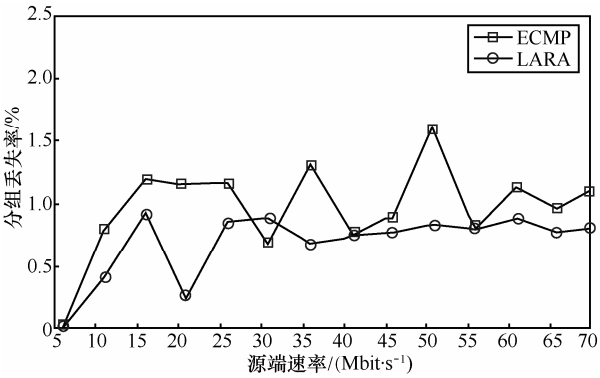


图 5　分组丢失率比较

同时，对网络吞吐量进行了比较测试，实验结果如图 6 所示。当源端发送速率在 5 Mbit/s 至 11 Mbit/s 之间时，ECMP 与 LARA 吞吐量表现相当，增长迅速，说明期间并无拥塞链路产生；当发送速率大于 12 Mbit/s 时，网络开始出现拥塞链路，ECMP 吞吐量回落迅速，LARA 则表现较平稳。实验结果表明，当节点路由器使用 LARA 进行部署时，与使用 ECMP 部署时相比，网络吞吐量明显增加，进一步验证了 LARA 的优越性。

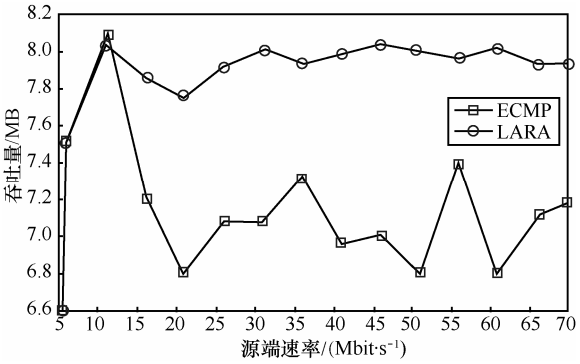


图 6 吞吐量比较

从以上的实验结果可以得出结论：当网络链路处于低负载、无拥塞产生时，LARA 与 ECMP 性能表现相当，传播时延较低；当链路负载增加并有拥塞产生时，LARA 能够使源端动态调整不同链路的发送速率，将部分流量引导至链路负载相对较低、链路质量较好且无拥塞产生的链路上去，保证端到端的时延控制在有效的范围之内，能够很好地进行链路负载均衡。

6 结束语

本文的主要目标是针对时延敏感型的业务流，例如在线视频等，为确保基于该业务的端到端时延控制在有效范围之内，而开发一个新的协议将更好满足这些对时延敏感的应用程序的需求。观察到这些应用程序在网络性能瞬时下降时将受益于链路较低的时延以及更好的健壮性，考虑到这些目标，利用优化论设计了一个新的路由算法协议 LARA。

LARA 利用凸优化理论，通过提出优化目标函数，将流量在多个有效路径上进行分割，确保关键链路不会成为产生拥塞的瓶颈路径。然后用优化分解的方法将优化问题转化为一个具有 3 个可调参数的分布式算法和实用协议。通过利用 NS2 仿真器在基于实际的 CERNET2 网络拓扑上模拟 LARA 的运行，验证了 LARA 在低时延和健壮性方面的优越表现。

LARA 在低时延和健壮性方面的优越表现是通过与目前在网络中普遍部署的 ECMP 路由策略进行比较并得出的结论，即在网络的关键路径产生拥塞之前，LARA 能够将部分流量引导至链路负载相对较低、链路质量较好且无拥塞产生的链路上，保证了端到端的时延，避免了关键链路拥塞的发生，同时在链路高负载的情况下能保证链路的低分组丢失率，提高网络的吞吐量。

在下一步的工作中针对本文提出的 LARA 还有 2 个可能的扩展工作可以做，首先在 LARA 中假设用户对服务提供商即源端的需求是常量，但是在实际情况下，需求量可能随时都会变化，而不是在一定时间间隔内是恒定的，因而可以进一步改进本文的协议；其次，LARA 可以与互联网经济学融合，由于协议中对链路反馈的代价消息十分敏感，对于低开销链路将会吸引更多的流量，尤其对于域间的跨网络部署，因为 ISP 之间存在市场竞争关系而通告虚假路径信息或者出于安全和隐私保护的目的不通告实际可达路由，那么如何让 ISP 提供诚实的路径信息以及如何通过激励机制使 ISP 提供更多的可达路由，都是下一步研究的重点。

参考文献：

[1] Chinese netizens network video application research report in 2013[EB/OL].http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/spbg/201406/t20140609_47180.htm.

[2] Cisco visual networking index: forecast and methodology[EB/OL]. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html.

[3] VOGEL A, KERHERVE B, *et al*. Distributed multimedia and QoS: a survey[J]. IEEE Multi-Media, 1995, 2(2): 10-19.

[4] XIAO X, NI L M. Internet QoS: a big picture[J]. IEEE Network, 1999,13(2):8-18.

[5] HE J, REXFORD J. Towards Internet-wide multipath routing[J]. IEEE Network Magazine, Special Issue on Internet Scalability, 2008, 22(2): 16-21

[6] KELLY F, VOICE T. Stability of end-to-end algorithms for joint routing and rate control[J]. ACM SIGCOMM Computer Communication Review, 2005,35(2):5-12.

[7] XU W, REXFORD J. MIRO: Multi-path interdomain routing[J]. ACM SIGCOMM Computer Communication Review, 2006,36(4):171-182.

[8] DAMON W, COSTIN R, ADAM G, *et al*. Design, implementation and evaluation of congestion control for multipath TCP[A]. Proc of the 8th USENIX Conference[C]. 2011. 99-112.

[9] SUCHARA M, XU D H, DOVERSPIKE R, *et al*. Network architecture for joint failure recovery and traffic engineering[J]. ACM SIGMETRICS Performance Evaluation Review, 2011,39(1):97-108.

[10] NGUYEN G T K, AGARWAL R, LIU J D, *et al*. Slick packets[J]. Performance Evaluation Review, 2011,39(1): 205-216.

[11] SUCHARA M, FABRIKANT A, REXFORD J. BGP safety with spurious updates[A]. IEEE INFOCOM[C]. 2011.2966-2974.

[12] HOPPS C. Analysis of an Equal-Cost Multi-Path Algorithm[S]. RFC 2992, 2002.

[13] ZLATOKRILOV H, LEVY H. Packet dispersion and the quality of voice over IP applications in IP networks[A]. IEEE INFOCOM[C]. 2004. 1170-1180.

[14] GALLAGER R. A minimum delay routing algorithm using distributed computation[J]. IEEE Transactions on Communications, 1977,25(1): 73-85.

[15] BERTSEKAS D, GAFNI E, GALLAGER R. Second derivative algorithms for minimum delay distributed routing in networks[J]. IEEE Transaction Communications, 1984,32(8):911-919.

[16] JAVED U, SUCHARA M, HE J Y, *et al*. Multipath protocol for delay-sensitive traffic[A]. Proc of the First International Conference of Communication Systems and Networks[C].2009.

[17] APOSTOLOPOULOS G，WILLIAMS D. QoS Routing Mechanism and OSPF Extensions[S]. RFC 2676, 1999.

[18] KODIALAM M, LAKSHMAN T V. Minimum interference routing with applications to MPLS traffic engineering[A]. IEEE INFOCOM[C]. 2000.884-893.

[19] PALOMAR D, CHIANG M. A tutorial on decomposition methods for network utility maximization[J]. IEEE Journal on Selected Areas in Communications, 2006, 24(8):1439-1451.

[20] UHLIG S, QUOITIN B, LEPROPRE J, *et al*. providing public intradomain traffic matrices to the research community[J]. ACM SIGCOMM Computer Communication Review, 2006, 36(1):83-86.

[21] HE J, SUCHARA M, BRESLER M. Rethinking Internet traffic management: from multiple decompositions to a practical protocol[A]. Proc of the ACM CoNEXT[C]. 2007.17.

[22] WEI X D, CHENG J, LOW H S, *et al*. FAST TCP: motivation, architecture, algorithms, performance[J]. Networking, IEEE/ACM Transactions on, 2006,14(6):1246-1259.

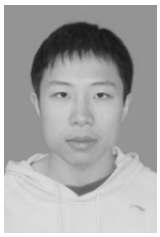
[23] KURIAN J, SARAC K. A survey on the design, applications, and enhancements of application-layer overlay networks[J]. ACM Computing Surveys, 2010, 43(1):5.

作者简介：

杨洋（1980-），男，江苏无锡人，清华大学博士生、主要研究方向为计算机网络、路由协议、流量工程等。

杨家海（1966-），男，浙江云和人，清华大学网络运行与管理技术研究室主任、教授、博士生导师，主要研究方向为计算机网络、网络管理与测量、网络安全、云计算与大数据等。

王会（1977-），女，河南南阳人，博士，清华大学副研究员，主要研究方向为互联网路由、流量工程等。



李晨曦（1991-），男，湖北武汉人，清华大学博士生，主要研究方向为网络安全、异常检测等。

王于丁（1984-），男，河北石家庄人，清华大学博士生，主要研究方向为计算机网络、云计算等。