

一种基于测量的启发式网络拓扑匹配优化算法

廖 怡^{1),2)} 盛益强¹⁾ 王劲林¹⁾

¹⁾(中国科学院声学研究所国家网络新媒体工程技术研究中心 北京 100190)

²⁾(中国科学院大学 北京 100049)

摘 要 在复杂多变的网络环境下,覆盖网络与物理网络之间普遍存在着拓扑不匹配问题. 拓扑不匹配问题会给网络造成不必要的压力,影响系统的效率和可扩展性等. 缓解拓扑不匹配问题有助于提高网络寻址效率、减少冗余流量、降低端到端时延. 随着计算机和通信技术的不断发展,互联网的规模不断增大,网络节点的地理位置分布范围扩大、移动性增强,极大地增加了网络的动态性,尤其是节点的频繁加入、退出和失效,严重地加剧了大规模网络中覆盖网络与物理网络的不匹配问题. 为了缓解该问题,该文提出了一种基于测量的启发式拓扑匹配优化算法(Measurement-based Heuristic Topology Matching Optimization Algorithm, MHTMOA),该算法包括了节点加入、退出和失效算法,用于维护一个或者多个树形覆盖网络. 该算法的主要优点在于:(1)通过网络测量技术获取底层物理网络中节点间的跳数信息,简单地利用跳数三角形的边长关系,就可有效地将相近节点逐渐地汇聚;(2)允许对跳数进行粗粒度的比较,并通过三角不等式违反(Triangle Inequality Violation, TIV)感知以及启发式规则选择邻居节点,每个节点最终可获得一个准确度较高的邻居节点集合;(3)在节点频繁加入、退出和失效的场景下,节点之间也能保持高一致性的近邻关系. 除了传统的时延伸缩比(Latency Stretch, LS)外,该文还定义了全局拓扑匹配比(Global Topological Matching Ratio, GTMR)和局部邻居节点准确率(Local Neighborhood Accuracy, LNA)这两个量化指标,以便更精确地衡量拓扑一致性. 评价结果表明,相较于现有算法,提议算法的 GTMR 和 LNA 提升显著,LS 降幅可达 53%,从而更好地缓解了拓扑不匹配问题.

关键词 拓扑不匹配;覆盖网络;物理网络;网络测量;三角不等式违反;时延伸缩比

中图法分类号 TP311 **DOI号** 10.11897/SP.J.1016.2018.02044

A Measurement-Based Heuristic Topology Matching Optimization Algorithm

LIAO Yi^{1),2)} SHENG Yi-Qiang¹⁾ WANG Jin-Lin¹⁾

¹⁾(National Network New Media Engineering Research Center, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190)

²⁾(University of Chinese Academy of Sciences, Beijing 100049)

Abstract Topology mismatch between the overlay network and the underlying network is a critical problem in the complex and unstable network environment. The topology mismatch problem means that the virtual link between two nodes in the overlay network is not consistent with the physical link in the underlying network. It will impose an extraordinary amount of redundant traffic and affects both the scalability and the efficiency for Internet service. Resolving the mismatch problem is the key to improving the efficiency of addressing, eliminating redundant traffic, and reducing the end-to-end delay. With the continuous development of computer and communication technologies, the Internet grows into a large scale by magnitude, the geographical locations of the network nodes are widely distributed and the mobility of nodes and the dynamics of networks are greatly increased. The mismatch problem is becoming increasingly serious in

收稿日期:2017-05-09;在线出版日期:2017-11-17. 本课题得到中国科学院先导专项基金(XDA06040602, XDA06040501)、国家科技重大专项“新一代宽带无线移动通信网”(2017ZX03001019)资助. 廖 怡,女,1991年生,博士研究生,主要研究方向为视频服务系统、网络感知算法和未来网络. E-mail: liaoy@dsp.ac.cn. 盛益强(通信作者),男,1978年生,博士,副研究员,硕士生导师,主要研究领域为分布式机器学习、未来网络和启发式算法. E-mail: shengyq@dsp.ac.cn. 王劲林,男,1964年生,硕士,研究员,博士生导师,主要研究领域为未来网络、流媒体网络应用技术和无线通信.

environments where the dynamic network nodes join, leave or fail frequently.

In order to solve the above problem, many researchers have investigated the overlay construction approaches. Some of them tried to get location information in the underlying network. When a new node joining in the overlay network, the neighbors would be selected according to the location information. However, most approaches are with fatal drawbacks. On the one hand, most of the existing approaches were lack of considering the hop as a major metric with its impact on construction, maintenance and evaluation to reflect the topological similarity and difference between global matching and local matching. On the other hand, few existing approaches took full advantages of the heuristic rules based on the Triangle Inequality Violation (TIV) by considering the topological consistency.

In this paper, we propose a Measurement-based Heuristic Topology Matching Optimization Algorithm (MHTMOA) to alleviate the topology mismatch problem more effectively. The proposed MHTMOA includes nodes joining, leaving and failure processing methods to maintain one or more tree-based overlays. The main merits of MHTMOA include: (1) By utilizing the network measurement technology to obtain the hop information of the nodes in the underlying networks, it becomes easier to converge the nodes located in proximity positions by simply using the edge triangular relations of the hop counts. (2) The neighboring set of a joining node has higher accuracy by means of TIV awareness and heuristic selection rules, even though the coarse-grained comparisons are used among hop counts. (3) The neighborhood relationships with relatively high consistency can be remained even under a highly dynamic network environment. Besides, the MHTMOA does not need to obtain the global information of the underlying network, and it only needs to measure a part of network nodes on demand. The MHTMOA can be used for centralized, distributed or hybrid deployment according to different application requirements. The logical trees generated by different deployment methods turn out the same, but the maintenance costs might be different. The long routing path in the overlay network would be reduced with the TIV awareness, thus the MHTMOA is available to wide area networks.

For evaluation, the Latency Stretch (*LS*) is used to test the performance on mitigating topology mismatch problem. Furthermore, we define two additional metrics, named Global Topology Matching Ratio (*GTMR*) and Local Neighborhood Accuracy (*LNA*), to evaluate the global and local topological consistency more accurately. The results show that MHTMOA outperforms the state-of-the-art algorithms for topology matching optimization with respect to *GTMR*, *LNA* and *LS*. The improvement of *LS* is up to 53% based on the evaluation.

Keywords topology mismatch; overlay network; underlying network; network measurement; triangle Inequality violation; latency stretch

1 引言

互联网的底层网络设备具有异构性、动态性、移动性等特点,网络连接关系复杂,导致网络管理的难度增大,因而普遍采用覆盖网络构造技术,即在底层物理网络的上层构建一个虚拟网络,通过建立节点之间的逻辑关系,可较好地规避底层物理设备之间

的异构性等问题,为用户提供统一的节点访问方法以及相对稳定的网络拓扑结构,并可利用节点间的合作与协同服务,实现有效的负载均衡,避免网络中部分节点资源消耗过快,进而提高系统的性能。

覆盖网络构造技术广泛地应用于点对点(Peer-to-Peer,P2P)网络、内容分发网络(Content Delivery Network,CDN)、应用层多播、移动对等网络、云计算数据中心网以及新兴的软件定义网络(Software

Defined Network, SDN) 架构中, 现有的覆盖网络根据网络结构以及节点所维护邻居节点的组织方式可分为非结构化和结构化两类. 对于非结构化的覆盖网络, 节点维护的邻居节点是随意的、无结构的, 信息资源在网络中存放的位置与网络本身的拓扑无关; 对于结构化的覆盖网络, 节点维护的邻居是有规律的、网络拓扑结构受到严格控制^[1]. 覆盖网络的结构直接关系到整个网络的寻址效率, 也将影响建立在覆盖网络上层的网络应用的效率和扩展性^[2]. 然而无论是非结构化还是结构化的覆盖网络, 都普遍存在着较严重的拓扑不匹配问题. 拓扑不匹配主要指两个在覆盖网络中相邻的节点映射到实际物理网络中却相距很远, 造成逻辑链路与底层物理链路不一致的现象^[3]. 对于非结构化的覆盖网络, 节点通过随机选择得到的邻居节点丢失了物理网络中的位置信息, 故邻居节点与物理网络匹配程度较差, 准确性不高; 对于结构化的覆盖网络, 节点加入时没有考虑节点位置信息, 或是对物理拓扑信息搜集不全或不准确, 都易导致拓扑不匹配问题^[4].

拓扑不匹配将带来一系列的影响, 例如使端到端的时延增加、引入不必要的网络流量、使消息数目增长等, 尤其是消息数目的增长, 易占用较多的带宽资源, 给物理网络造成压力^[5]. 研究表明, 75% 以上的 P2P 系统都遇到过因拓扑不匹配导致的寻址效率下降等问题^[6]. 现如今 P2P 流量已占据了整个互联网流量的最大比重, 而这其中很大一部分的冗余流量是由拓扑不匹配造成的^[7]. 近些年来, 移动互联网发展迅猛, 网络节点的移动性加剧了网络的扰动(Churn), 并且导致物理网络拓扑变化频繁, 一些对节点位置准确度要求较高或对时延和抖动敏感的业务, 如视频直播、点播业务等, 因网络动态性导致的拓扑不匹配问题更加突出.

随着云计算的普及和虚拟化技术的发展, SDN 架构开始兴起, 覆盖网络技术与 SDN 相互融合, 形成了 SDN 覆盖网络. SDN 将控制平面与数据平面分离, 控制平面通常利用一个逻辑上中心化的控制器搜集网络拓扑信息以构建覆盖网络, 并可针对不同的网络流量或不同源节点和目的节点, 选择性能较优的覆盖网络路径^[8-9]. SDN 覆盖网络可应用于云计算数据中心网, 然而云数据中心之间通常是广域分布的, 云之间的数据传输多依赖于公网传输, 若覆盖网络与物理网络匹配度低, 易导致覆盖网络中存在较多的长路由路径, 不利于降低时延, 且增加流量成本, 因而对于含有大量广域分布节点的覆盖网

络, 解决拓扑不匹配问题尤为迫切. 结构化的覆盖网络, 因其对节点的有效组织以及拓扑结构可控等优点, 在网络流量控制和寻址效率上普遍优于非结构化的覆盖网络而受到持续地关注^[10]. 如何构造一种结构化覆盖网络, 既能满足应用的需求, 同时还能保持覆盖网络与物理网络的一致性, 适应如今分布式、大规模、高动态性的网络环境, 是学术界和工业界一直在探讨和解决的问题.

为有效缓解拓扑不匹配问题, 实现网络中邻近节点的汇聚, 提高拓扑一致性, 本文提出了一种基于测量的启发式拓扑匹配优化算法(Measurement-based Heuristic Topology Matching Optimization Algorithm, MHTMOA), 用于维护一种树形覆盖网络. 该算法利用网络测量技术获取物理网络中待加入节点到部分其它节点的跳数, 并根据跳数三角形的边长关系构建和维护覆盖网络. 本文的主要贡献有以下几点:

(1) 提出了一种新的可应用于广域网络的拓扑匹配算法, 该算法有效地降低了覆盖网络与物理网络的时延伸缩比, 从而更好地缓解了拓扑不匹配问题.

(2) 在构建覆盖网络时, 提议算法不仅考虑了邻居节点的接近性, 同时利用三角不等式违反(Triangle Inequality Violation, TIV)感知进行路由路径优化, 可有效减少长路由路径, 有利于提高网络服务质量(Quality of Service, QoS).

(3) 定义了全局拓扑匹配比(Global Topology Matching Ratio, GTMR)和局部邻居节点准确率(Local Neighborhood Accuracy, LNA)这两个精确度更高的指标, 从拓扑结构上描述覆盖网络与物理网络的全局和局部匹配程度.

本文第 2 节简要地介绍覆盖网络研究的相关工作; 第 3 节详细地介绍本文提议的基于测量的启发式拓扑匹配优化算法, 简称为 MHTMOA; 第 4 节介绍仿真实验和算法性能评估; 第 5 节为讨论; 第 6 节为结论和展望.

2 相关工作

基于 DHT(Distributed Hash Table)技术的覆盖网络构造算法是目前应用最广泛的结构化覆盖网络构造算法之一, 如 Chord^[11]、CAN^[12]、Pastry^[13]等, 这些算法不仅应用于传统的 P2P 网络, 更在近些年兴起的移动对等网络、移动自组织网络研究中占据举足轻重的地位, 许多文献在 DHT 算法的

基础上提出了新的路由算法以满足移动覆盖网络的组网需求^[14-17]. DHT 算法路由效率高,扩展性好,但大部分基于 DHT 的方法只考虑了覆盖网络中节点之间的逻辑关系,缺乏节点的物理位置信息,导致覆盖网络存在较严重的拓扑不匹配问题,故研究人员尝试在构建覆盖网络时通过获取节点在底层物理网络的位置信息,选择位置相近的节点作为邻居节点,以缓解拓扑不匹配问题.

目前用于缓解拓扑不匹配问题的覆盖网络构造方法可大致划分为以下四类^[18]:

第一类为界标法.该方法中节点首先与全局已知的几个界标节点进行 RTT(Round Trip Time)测量,再将 RTT 值按升序或降序排列,得到节点界标向量.该方法认为具有相同界标顺序的节点物理位置是相近的,通过对界标向量进行划分,可以有效地得到地理位置接近的邻居节点^[19].如文献[20]提出的 GNP 系统中,新节点通过与少量界标节点测距来获取节点的全局坐标,得到节点的位置信息.文献[21]提出了基于位置感知的 Chord 覆盖网络构造方法,该方法在 Chord 协议的基础上利用界标节点获取节点的位置关系,实现基于位置的路由以解决移动自组织网络中的拓扑不匹配问题.界标法的实质是建立节点在网络中的相对坐标系,其优点是节点加入时只需探测与界标节点有关的信息,算法的可扩展性好,复杂度低,但该算法最大的不足在于界标向量的划分是粗粒度的,对落入同一个区间的节点不能进一步的区分,在最坏情况下,系统中所有的节点可能落入同一个区间中,此时邻居节点的准确度较低^[22].

第二类为基于位置信息聚类法.典型的方法为基于 IP 的分层聚类法.根据 IP 分配的特点,IP 相近的节点地理位置也接近,故可通过 IP 地址与地理位置的映射来获取节点的位置信息,并利用聚类方法筛选出距离相近的节点集合,以缓解拓扑不匹配问题.如文献[23]提出的 Coral 系统中,新节点通过随机地选择节点并进行路径解析以发现 IP 簇,再选择最开始的 5 个节点作为最近的 IP 簇加入其中.文献[24]中提出的覆盖网络构造方法首先利用界标法得到时延向量对节点进行粗略的地理定位,再根据 IP 进一步筛选出同一区域内位置相近的节点.基于 IP 的分层聚类方法的主要缺点是 IP 可能随时间变化,邻居节点的准确性难以保证,难以适用于高动态性的网络.

第三类为基于位置关系的节点层次化命名方

法.该类方法将节点位置信息嵌入节点的命名中,在构建覆盖网络时,可通过节点命名获取网络距离以选择邻近节点.如文献[25]提出的 Anysee2 在节点编址时加入节点所在网络的 ISP、CITY、POST-CODE、IP 等信息,通过构建一棵以源节点为根的节点编址字典树,可方便选择不同层次接近性的节点.该方法中邻居节点的准确性很大程度上依赖于编址方法的设计,若编址位数不足,易导致节点位置划分的精度不足,且影响系统的可扩展性;若编址位数过长,则影响寻址效率,易造成资源的浪费.文献[26]中首次提出了利用三维坐标表示节点空间位置关系的覆盖网络构造方法 3DO,该方法在节点加入时,不仅考虑了节点间的距离,还考虑了邻居节点间的相对位置,并将相对位置关系映射至三维空间(例如 X, Y, Z),用一个三维的逻辑标识符(Logic Identifier, LID)表示.新节点先利用扩展环搜索算法获得距其 1 跳或 2 跳的节点,再执行以新节点为根的最小生成树算法,选择最小生成树中与其直接相连的节点作为邻居节点,之后根据新节点与其邻居间的相对位置关系,利用邻居节点的 LID 计算得到自身的 LID.3DO 网络中的每个节点需维护基于 LID 的多维度转发信息以实现最近邻转发.该方法虽能有效解决拓扑不匹配问题,但其维护机制过于复杂,若邻居节点间的相对位置发生变化,需要更新 LID 和路由转发表.当节点规模增大或网络动态性增强时,3DO 的维护和更新代价过高,因而 3DO 适用于小规模、低动态性的网络.

第四类为基于测量的方法.该类方法通过测量跳数、时延、带宽甚至节点操作的历史数据等信息构造覆盖网络.如文献[18]提出的 T2MC 算法,该算法中新节点首先随机选择系统中一个距离较远的节点,并利用 traceroute 测量工具探测,用三元组 $\langle \text{IP}, \text{跳数}, \text{时延} \rangle$ 记录节点 traceroute 所经路径的信息,再利用 K-means 聚类方法获得相近的邻居节点.文献[27]在 PeerStreamer 的架构下对比了不同的覆盖网络构造方法中邻居节点选择策略对 P2P-TV 系统 QoS 的影响. PeerStreamer 主要用于流媒体数据块的传输,该系统通过测量节点的上行带宽和链路的 RTT 来构建覆盖网络,并根据节点间内容传输记录在节点本地维护一个有向图,通过动态地建立一张候选邻居节点列表,筛选出位置相近、上行带宽较大的邻居节点以解决拓扑不匹配问题.基于网络测量的方法,其主要缺点是系统实现的难度较大,且在有限的时间内难以获取足够的邻居节点信息,故

通常采用局部信息测量的方式以降低维护开销。

三角形不等式违反(TIV)现象广泛地存在于现实网络中,它指在网络距离测量结果中,存在三个节点间的距离关系违反三角不等式关系,例如图 1 中所示的 $D(B,C)>D(A,B)+D(A,C)$ 。

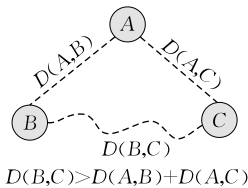


图 1 三角不等式违反示意图

若图 1 中节点间的距离用时延表示, $D(B,C)>D(A,B)+D(A,C)$ 则意味着从 B 途经 A 到达 C 的时延比直接从 B 到 C 的时延更短,故可在构建覆盖网络时充分利用 TIV 现象,用一条具有更低时延的路径替换直接路径以提高网络 QoS^[28]。文献[28-29]利用上述 TIV 的特性提出了基于时延 TIV 感知的覆盖网络路由性能优化方法。文献[30]讨论了在车联网路由中的 TIV 现象,并分析了不同路由算法中,TIV 发生的频率和密度对路由性能的影响。MHTMOA 在构建覆盖网络时也借鉴 TIV 的思想,但区别于上述文献,本提议方法的不同主要有以下两点:

(1) 文献[28-30]利用 TIV 感知进行路径替换,并没有单独利用三角形的其它边长关系,如 $D(A,B)$ 和 $D(A,C)$ 。MHTMOA 在构建覆盖网络时,不仅利用 TIV 进行路径优化,还利用了 $D(A,B)$ 和 $D(A,C)$ 并结合贪心策略以及启发式规则选择邻近的节点作为父节点。

(2) 文献[28-30]主要对时延 TIV 进行感知,而 MHTMOA 选用跳数度量节点间的距离,并进行跳数 TIV 感知。此外,MHTMOA 还设置了跳数三角形边长关系所满足的若干情况的优先级处理机制,以实现与物理网络的匹配。

3 一种基于测量的启发式拓扑匹配优化算法

本节主要介绍一种基于测量的启发式拓扑匹配优化算法(MHTMOA),包括节点加入、节点退出和节点失效算法,生成一种树形覆盖网络。本文采用路由跳数表示节点间的距离,在实际网络中可利用基于 traceroute 的网络测量技术得到节点间的路由跳数,或简单地利用 ping 指令获取 TTL 值计算得到

节点间的路由跳数。

3.1 节点加入

3.1.1 节点信息

对于每个加入树形覆盖网络的节点,都拥有一个全局唯一的节点标识。令 V_i 表示第 i 个节点, P_i 表示其父节点, Num_i^C 表示其子节点个数, S_i^C 表示其子节点集合, C_i^k 表示其第 k 个子节点,其中 $C_i^k \in S_i^C$ 且 $0 \leq k \leq Num_i^C - 1$ 。

3.1.2 节点加入算法

记加入节点为 J ,树结构的根节点为 R 。在节点加入过程中,设有一个动态根节点 DR ,用于表示新节点的候选父节点,候选父节点并不是固定的,而是动态变化的。用 C_{DR}^k 表示 DR 的第 k 子节点, $H(J, DR)$ 表示 J 到 DR 的路由跳数, $H(C_{DR}^k, DR)$ 表示 DR 到 C_{DR}^k 的路由跳数, $H(J, C_{DR}^k)$ 表示 J 到 C_{DR}^k 的路由跳数。表 1 所示为算法的符号表。

表 1 节点加入算法符号表

符号	意义	符号	意义
V_i	第 i 个节点	S_i^C	V_i 的子节点集合
J	新加入节点	Num_i^C	V_i 的子节点个数
R	根节点	C_i^k	V_i 的第 k 个子节点
DR	动态根节点	$H(i, j)$	V_i 到 V_j 的跳数
P_i	V_i 的父节点	—	—

在构建覆盖网络时,通常设置根节点服务器作为初始的根节点 R 。算法初始化时,先将 DR 设为 R 。当 J 加入时,先进行跳数测量,得到 J 到 DR 以及 DR 所有子节点 C_{DR}^k 间的路由跳数,并构建如图 2(a)所示的跳数三角形,再根据三角形的边长关系分为图 2(b)、图 2(c)、图 2(d)三种情况处理,节点加入算法伪代码如算法 1 所示。

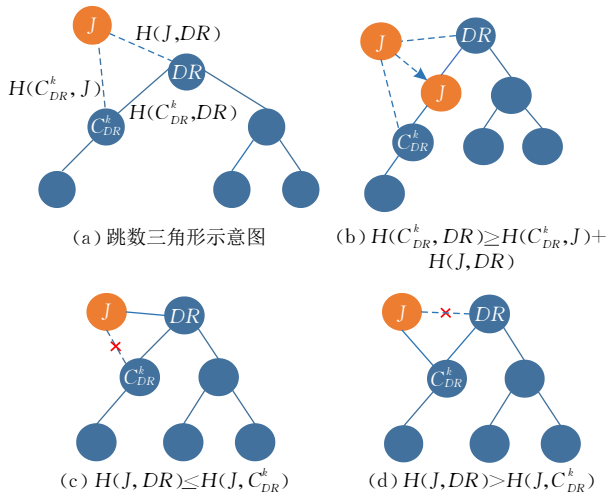


图 2 节点加入算法原理示意图

算法 1. 节点加入.

输入: J, R, DR, S_R^C, Num_R^C ,

$$H(J, DR), H(C_{DR}^k, DR), H(J, C_{DR}^k)$$

输出: P_J, Num_J^C, S_J^C

```

1. 初始化:  $DR=R, P_J=\text{null}, Num_J^C=0, S_J^C=\emptyset$ 
2. IF  $Num_{DR}^C=0$  THEN
    //J 添加为 DR 的叶子节点
3.  $P_J \leftarrow DR$ 
4.  $S_J^C \leftarrow S_J^C \cup \{J\}$ 
5.  $Num_{DR}^C \leftarrow Num_{DR}^C + 1$ 
6. RETURN  $P_J, Num_J^C, S_J^C$ 
7. END IF
8. ELSE
9. IF  $\exists C_{DR}^k \in S_{DR}^C, 0 \leq k \leq Num_{DR}^C - 1$  satisfy
     $H(C_{DR}^k, DR) \geq H(C_{DR}^k, J) + H(J, DR)$  THEN
    //将 J 插入 DR 和  $C_{DR}^k$  之间
10.  $P_J \leftarrow DR$ 
     $S_{DR}^C \leftarrow S_{DR}^C \cup \{J\} - \{C_{DR}^k\}$ 
11. FOREACH satisfied  $C_{DR}^k$  DO
12.  $P_{C_{DR}^k} \leftarrow J$ 
13.  $S_J^C \leftarrow S_J^C \cup \{C_{DR}^k\}$ 
14.  $Num_J^C \leftarrow Num_J^C + 1$ 
15. END FOR
16. RETURN  $P_J, Num_J^C, S_J^C$ 
17. END IF
18. ELSE IF  $\forall C_{DR}^k \in S_{DR}^C, 0 \leq k \leq Num_{DR}^C - 1$  satisfy
     $H(J, DR) \leq H(J, C_{DR}^k)$  THEN
    //J 添加为 DR 的叶子节点
19.  $P_J \leftarrow DR$ 
20.  $S_{DR}^C \leftarrow S_{DR}^C \cup \{J\}$ 
21.  $Num_{DR}^C \leftarrow Num_{DR}^C + 1$ 
22. RETURN  $P_J, Num_J^C, S_J^C$ 
23. END IF
24. ELSE //J 处在新 DR 的子树中
25. IF more than one  $C_{DR}^k \in S_{DR}^C, 0 \leq k \leq Num_{DR}^C - 1$ 
    satisfy  $H(J, DR) > H(J, C_{DR}^k)$  THEN
26.  $d \leftarrow \text{MIN}(H(J, C_{DR}^k))$ 
27.  $DR \leftarrow C_{DR}^d$ 
28. IF minimal  $H(J, C_{DR}^k)$  are equal THEN
29.  $d \leftarrow \text{MAXDEGREE}(C_{DR}^k)$ 
30.  $DR \leftarrow C_{DR}^d$ 
31. IF maximal degrees are equal THEN
32. randomly choose a  $C_{DR}^k$  satisfies
    above as DR
33.  $DR \leftarrow C_{DR}^k$ 
34. END IF
35. END IF
36. return TO 2

```

```

37. END IF
38. ELSE
39. choose the unique  $C_{DR}^k$  satisfies
     $H(J, DR) > H(J, C_{DR}^k)$  as DR
40.  $DR \leftarrow C_{DR}^k$ 
41. RETURN TO 2
42. END IF
43. END IF
44. END IF

```

(1) 若存在 $C_{DR}^k \in S_{DR}^C$ 满足 $H(C_{DR}^k, DR) \geq H(C_{DR}^k, J) + H(J, DR)$, 则将 J 插入到 DR 和 C_{DR}^k 之间; 若存在多个 $C_{DR}^k \in S_{DR}^C$ 满足上述跳数关系, 则将 C_{DR}^k 都变成 J 的子节点; 否则判断是否满足情况 2.

(2) 若对任意 $C_{DR}^k \in S_{DR}^C$ 都满足 $H(J, DR) \leq H(J, C_{DR}^k)$, 则将 J 作为 DR 的叶子节点加入; 否则判断是否满足情况 3.

(3) 若只存在一个节点 $C_{DR}^k \in S_{DR}^C$ 满足 $H(J, DR) > H(J, C_{DR}^k)$, 则选择该 C_{DR}^k 作为新的 DR 即 $DR = C_{DR}^k$; 若存在多个 $C_{DR}^k \in S_{DR}^C$ 满足, 则选择 $H(J, C_{DR}^k)$ 最小的节点作为新的 DR ; 若最小值也相等, 则选择节点度最大的节点作为新的 DR ; 若节点度也相等, 则随机选择满足上述所有条件的节点作为新的 DR .

DR 变化后, 再按情况 1~3 判断 J 与新的 DR 及其子节点间的关系, 直到节点找到合适的父节点, 加入覆盖网络中.

图 3 所示是以 R 为根节点, 节点 $A \sim G$ 以字母顺序加入覆盖网络的过程示意图. 在图 2 所示的三种情况中, 情况 2 和情况 3 利用贪心策略和启发式规则选择距离最近的候选父节点. 若新节点离动态根节点最近, 即满足情况 2, 则选择动态根节点为父节点, 如节点 C ; 若离某一个子节点最近, 即满足情况 3, 则选择子节点作为新的动态根节点, 再继续比较新节点与动态根节点及其子节点的跳数, 直到节点加入树结构合适的位置, 如节点 G . 情况 1 主要利用跳数 TIV 感知进行节点位置的调整 and 路径替换, 其优先级最高. 若新节点满足情况 1, 新节点较某个子节点更接近动态根节点, 并且从动态根节点途经新节点可用更短的跳数到达子节点, 则将新节点作为动态根节点的子节点加入, 并调整对应子节点的位置, 如节点 D . 情况 1 保证了节点即使后加入, 也可调整至距其更近的父节点处.

节点加入算法中的贪心策略和启发式规则保证了节点加入时总是往距其最近的动态根节点处移

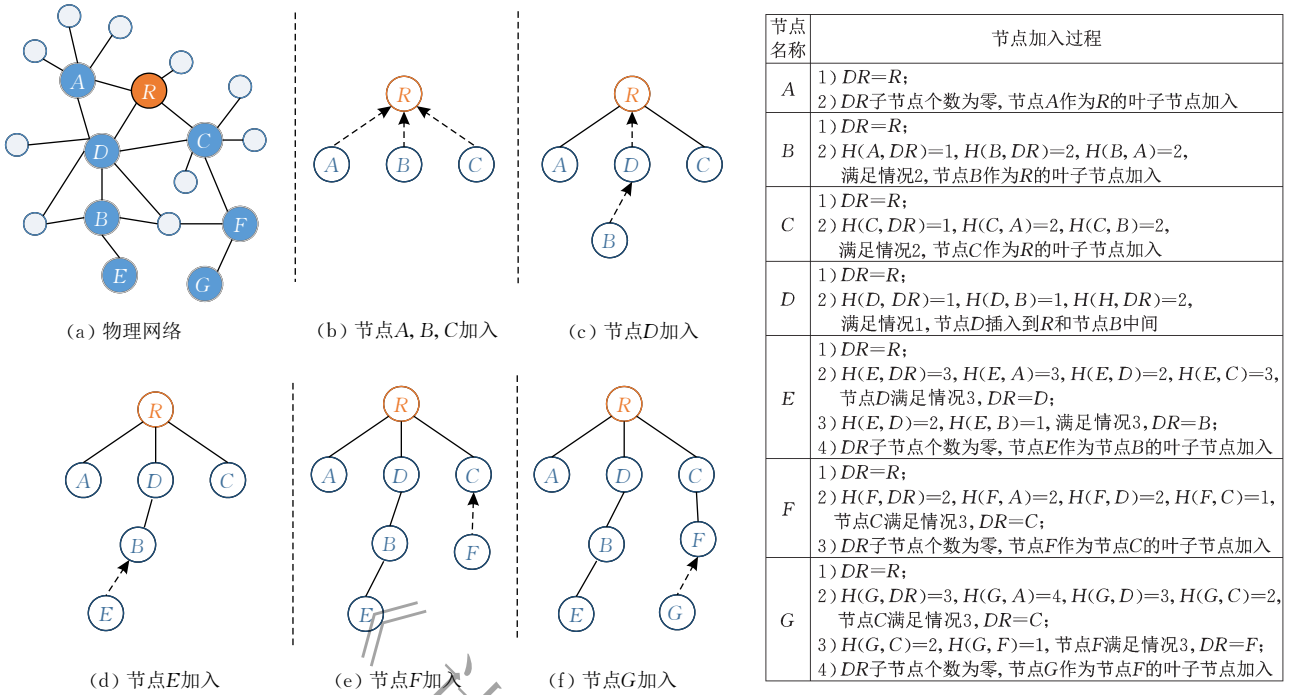


图 3 节点加入过程示意图

动,故位置相近的节点加入时所经过的动态根节点路径相似,最终位置相近的节点可汇聚在同一个父节点下. MHTMOA 节点加入算法其实质是根据跳数构建新节点到根节点的最短路径,实现网络中相近节点的动态聚类.从图 3 可看出,按节点加入方法构造的树形覆盖网络较好地保持了物理拓扑结构的连接关系,节点可获取到物理位置相近的邻居节点.即便是对于跳数间隔很大的节点,也可在节点刚加入覆盖网络时建立一种准确度较低的逻辑关系,并在之后的维护过程中,使物理位置接近的节点逐渐汇聚,邻居节点的准确度逐渐提高,所以 MHTMOA 可应用于节点广域分布的网络.

MHTMOA 节点加入算法的优点可概括为以下 2 点:(1)无需测量物理网络的全局信息,只需获取部分节点信息即可完成节点加入;(2)利用 TIV 可有效减少长路由路径,减少冗余流量,降低时延,同时可有效降低节点加入顺序对树结构的影响.

MHTMOA 既可采用集中式部署,也可采用纯分布式部署.若 MHTMOA 采用集中式部署,树结构的存储和维护依赖于一个中央服务器,系统内所有节点均可访问该服务器.为降低通信代价和计算开销,除了存储树结构外,可在中央服务器中缓存已测量的跳数信息以及跳数比较结果.若树结构发生变化或因物理网络改变导致跳数变化,可在中央服务器中统一进行更新,避免重复测量以降低维护开销.集中式部署可应用于例如基于 SDN 的覆盖网络

中,且为满足不同流量类型或应用服务需求,可在中央服务器中构建多重的覆盖网络并实现多重覆盖网络间的协同^[31],但该方法的缺陷是中心管理服务器易成为系统性能的瓶颈.

若采用纯分布式部署,树结构的维护依赖于单个节点,每个节点需维护其邻居节点和父节点信息,并缓存测距以及跳数比较结果.当新节点加入时,需先与树的根节点建立连接,获取其邻居节点列表,再执行节点加入算法.若节点加入过程中动态根节点变化,待加入节点再与新的动态根节点建立连接,重复上述过程直到节点完成加入为止.纯分布式部署方式与集中式部署相比,二者生成的树结构一致,但前者中新节点加入时可能需与多个动态根节点建立连接以获取邻居节点信息,故其通信代价和维护代价比集中式高,因而纯分布式部署方法更适用于静态或者低移动性的网络,如 CDN 覆盖网络,但纯分布式部署缺乏节点测距的全局信息.

综上,MHTMOA 也可采用集中式和分布式相结合的混合式部署,在中央服务器处缓存所有测量结果,有利于拓扑结构的调整、优化以及全局资源调度;同时,在每个节点处缓存邻居节点、父节点信息以及部分测量结果,可降低中央服务器的访问压力,提高系统的健壮性.

3.2 节点退出

当节点退出时,为保证树形覆盖网络的连通性、树结构的层次关系以及节点的邻居关系,其父节

点和子节点都需做出相应调整. 本文中, 节点退出时按退出节点是否为叶子节点分别处理: (1) 若退出节点为叶子节点, 则直接退出, 并通知其父节点; (2) 若退出节点为非叶子节点, 首先通知其父节点, 再给子节点重新寻找恰当的替换父节点. 为降低拓扑维护代价, 兼顾拓扑一致性, 子节点以退出节点的父节点为动态根节点, 再按节点加入算法重新加入覆盖网络中.

3.3 节点失效

当节点失效时, 无法主动向其父节点和子节点发送消息, 故本文中, 当有节点失效时, 由第一个检测到其失效的节点向覆盖网络中其它节点广播消息, 以寻找失效节点的父节点和子节点. 若失效节点为叶子节点, 则由发现节点通知失效节点的父节点; 若失效节点为非叶子节点, 则其所有子节点按节点退出时的方法处理, 重新加入覆盖网络中.

4 仿真实验与算法性能评估

4.1 性能评价指标

现有的用于描述覆盖网络与物理网络匹配程度的指标主要为伸缩比, 用覆盖网络中任意两点间的平均距离与物理网络中任意两点间平均距离的比值表示, 常用的指标有路径伸缩比 (Path Stretch, PS) 和时延伸缩比 (Latency Stretch, LS). 本文采用 LS 表示树形覆盖网络与物理网络匹配程度, 其表达式如式(1)所示:

$$LS = \frac{\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} d_t(i, j)}{\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} d_p(i, j)} \quad (1)$$

其中, n 表示网络中节点的个数, i 和 j 表示节点序号, $0 \leq i, j < n$ 且 $i \neq j$, $d_p(i, j)$ 表示在物理网络中 V_i 和 V_j 的最短路径时延, $d_t(i, j)$ 表示在树形覆盖网络中 V_i 和 V_j 之间的时延, 可利用物理网络中两点间的最短路径时延计算得到. LS 理想值为 1, 其值越接近 1, 表明覆盖网络与物理网络的匹配性越好.

LS 主要从网络延迟方面反映覆盖网络与物理网络的全局匹配程度, 它只能粗略地反映拓扑匹配度, 无法获知拓扑匹配的局部特征, 故精确度不高. 然而, 拓扑结构的匹配特征对于网络结构的调整和优化至关重要, 若能获取网络拓扑结构匹配的特征, 尤其是局部特征, 不仅有利于提高 LS , 还可提高整

个网络的 QoS, 故本文还将从网络拓扑结构上描述覆盖网络与物理网络的匹配程度. 为此, 我们定义了全局拓扑匹配比 (Global Topological Matching Ratio, $GTMR$) 以及局部邻居节点准确率 (Local Neighborhood Accuracy, LNA) 这两个指标, 以便更精确地获取拓扑匹配的细节信息.

定义 1. $GTMR$ 用于表示覆盖网络与物理网络的链路匹配程度, 它的定义为覆盖网络中与物理链路相匹配的链路数与覆盖网络中链路总数的比值. $GTMR$ 表达式如式(2)所示:

$$GTMR = \frac{\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} L_o^p(i, j)}{\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} L_o(i, j)} \times 100\% \quad (2)$$

式(2)中, U 为树形覆盖网络的节点全集, n 表示节点的总数, 其中, $n = |U|$, $V_i, V_j \in U$ 且 $0 \leq i, j < n$, $i \neq j$. $L_o(i, j)$ 表示覆盖网络中的节点的连接关系, $L_o^p(i, j)$ 表示覆盖网络中与物理网络匹配的节点连接关系. 若 V_i, V_j 相连, $L(i, j) = 1$, 则 $\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} L_o(i, j)$

表示树形覆盖网络的链路总数, $\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} L_o^p(i, j)$ 表示在树形覆盖网络中与物理网络匹配的链路总数. $GTMR$ 的理想值为 1, 其值越大, 表示覆盖网络与物理网络的链路匹配度越高.

定义 2. LNA 表示树形覆盖网络中局部邻居节点的准确率, 其定义为覆盖网络中单个节点邻居节点与其物理网络中邻居节点之间的匹配比例. LNA 表达式如式(3)所示:

$$LNA_i = \frac{|S_i^{NM}|}{|S_i^{NO}|} \times 100\% \quad (3)$$

式(3)中, U 表示树形覆盖网络中节点全集, $V_i \in U$ 且 $0 \leq i < |U|$. S_i^{NO} 表示 V_i 在覆盖网络中 h ($h \leq 3$) 跳邻居节点集合, S_i^{NP} 表示 V_i 在物理网络中 h ($h \leq 3$) 跳邻居节点集合, 两个邻居节点集合的关系如图 4 所示.

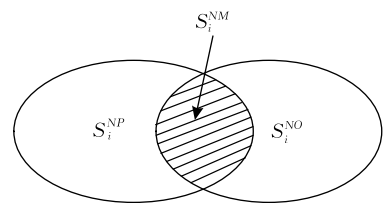


图 4 LNA 示意图

由图 4 可知, S_i^{NM} 表示集合 S_i^{NO} 与 S_i^{NP} 的交集, 其物理意义是 V_i 在覆盖网络中 h 跳邻居节点集合

中与其物理网络的 h 跳邻居节点集合中邻居关系保持一致的节点的集合. LNA_i 的理想值为 1, LNA_i 值越高, 表示 V_i 的邻居节点的位置越准确.

在本文中, 我们设置 $h=1$, 即统计节点的 1 跳邻居节点集合的准确率. 我们统计整个树形覆盖网络中 LNA_i 等于 1 的节点数与覆盖网络节点总数的百分比, 其表达式如式(4)所示:

$$LNA = \frac{\sum_{i=0}^{n-1} LNA_i}{n} \times 100\% \quad (4)$$

式(4)中, n 表示网络中的节点总数, $0 \leq i < n$. LNA 主要用于评估整个树形覆盖网络局部邻居节点位置的准确性, 表征树形覆盖网络局部邻居节点位置准确性的整体水平, 其理想值为 1, LNA 值越大, 则表示整个覆盖网络邻居节点的准确度越高.

4.2 仿真实验设计

4.2.1 物理拓扑

我们利用 BRITE 拓扑生成器^[32]提供的 Waxman 随机图模型^[33]和 BA 无标度模型^[34]各生成了六个不同节点规模的物理网络拓扑: {100, 200, 400, 800, 1600, 3200}, 设置节点最小度为 2, 节点位置满足随机分布. BRITE 是波士顿大学开发的一款通用型拓扑生成器, 其提供的 Waxman 和 BA 拓扑模型的幂率指数以及特征路径长度等方面的特性与 Internet 实际拓扑接近^[35].

4.2.2 仿真方案

在仿真时, 我们利用节点在物理拓扑中的最短路径跳数代替测量所得跳数, 作为节点加入的依据, 两节点间时延为物理拓扑中的最短路径时延. 值得说明的是, 本文提出节点加入算法并不要求实际测量所得的跳数为最短路径跳数, 只需利用跳数反映节点间的距离关系, 得到节点在物理网络中的位置信息.

对 MHTMOA 算法仿真时, 首先指定网络中的一个节点为根节点, 再随机抽取其它节点加入到树结构覆盖网络中, 直到所有节点都完成加入, 最后统计树形覆盖网络的 $GTMR$ 、 LNA 和 LS 、树的深度以及节点维护开销等指标, 并统计 MHTMOA 在高动态环境下 LS 的变化情况, 得到多次仿真实验结果. 其中, 节点维护开销用消息数目表示.

本文采用 OLT 算法^[36]、3DO 算法^[26]以及 Pan^[37]提出的结构化拓扑匹配算法(后文简称 Pan 方法)作为对照算法. OLT 利用 IP 地址到物理位置映射技术解决拓扑不匹配问题, 在该算法中新节点

选择满足带宽需求且物理位置最近的节点为父节点, 最终生成一种具有低时延特性的树结构覆盖网络. 3DO 利用扩展环搜索算法获取指定 TTL 范围内的节点并选择距离近的节点作为邻居节点, 利用新节点到邻居节点之间的跳数作为节点 LID 相关权值的计算依据. Pan 方法则是在 DHT 覆盖网络构造方法的基础上, 利用界标法与 IP 相结合的方法解决拓扑不匹配问题, 并利于基于事件驱动法维护和优化拓扑结构. 对照算法仿真时, 均采用与 MHTMOA 相同的网络拓扑, 节点带宽参数设置参照文献^[36], 3DO 算法中设置 TTL 最大值为 4, Pan 方法中界标节点数为 12.

本次仿真实验的硬件环境为奔腾双核 E5500 2.8GHz 处理器, 2 GB 内存, Windows7 操作系统, 在 VS2012 平台上进行算法仿真.

4.3 算法性能评估

(1) $GTMR$ 和 LNA 的平均值

图 5 分别表示不同网络模型、不同节点规模下 MHTMOA 和对照算法的 $GTMR$ 、 LNA 的平均值.

从图 5(a)、图 5(b)中可看出, 在 Waxman 和 BA 模型下, MHTMOA 的 $GTMR$ 平均值均保持在 72% 以上; 相较于 OLT 算法提升至少 30%, 提高比例可达 70%; 相较于 3DO 算法, 提升至少 5%, 提升比值可达 13%; 相较于 Pan 方法, 提升比可高达 90%. 从图 5(c)、图 5(d)中可以看出在两种模型中 MHTMOA 的 LNA 平均值均保持在 70% 以上; 相较于 OLT 提升至少 22%, 提升比例可高达 70%; 相较于 3DO 提升至少 5%, 提升比值可达 20%; 相较于 Pan 方法, 提升至少 60%. 由图 5 可知在 Waxman 和 BA 两种网络模型中, MHTMOA 的 $GTMR$ 和 LNA 变化趋势保持一致, 随着网络规模的增大, $GTMR$ 和 LNA 虽有下降, 但均高于 70%, 明显优于 OLT、3DO 和 Pan 方法, 体现了 MHTMOA 良好的健壮性, 其拓扑匹配特性不易受网络拓扑模型和节点规模的影响. 综上可得, MHTMOA 所构建的树节点之间的虚拟链路 with 物理网络链路可保持更高的匹配比例, 邻居节点集合具有更高的准确率, 因而 MHTMOA 能与物理网络保持良好地匹配关系.

MHTMOA 之所以具有良好的拓扑一致保持特性, 主要原因有以下 2 点:

① 同一父节点的子节点具有相对较近的位置关系, 故用动态根节点的子节点集合作为新加入节点的候选父节点集合, 其准确度较高.

② 在新节点加入时, 对物理网络拓扑的测量为

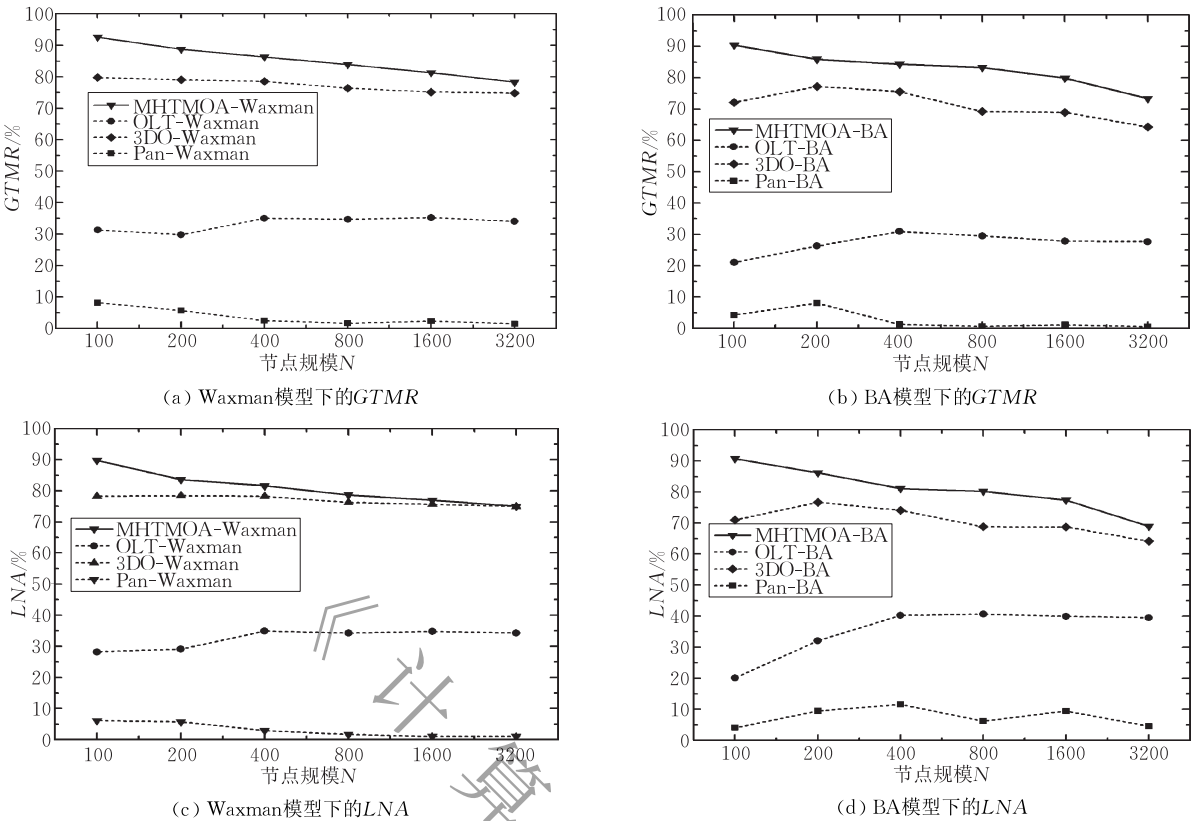


图 5 两种网络模型中不同算法的 GTMR 和 LNA 平均值

按需测量,即在跳数测量时以节点在覆盖网络中的邻居节点为依据,保证了测量信息的方向性和准确性.随着测量信息的增多,对底层物理拓扑特性描述得更全面,使得节点的位置准确度逐渐提升,最终整个树形覆盖网络取得更好的匹配效果.

从图 5 可看出,OLT 和 Pan 方法的 GTMR 和 LNA 平均值都较低,尤其是 Pan 方法的 GTMR 只有 3%左右,说明二者的链路匹配比和邻居节点准确度较低.OLT 的 GTMR 和 LNA 值较低是因为新节点选择邻居节点时,需在满足带宽条件的候选节点中选择,再利用 IP 信息进一步筛选邻居节点,这无法保证候选节点集合中包含位置最相近的节点,故其邻居节点准确度不高,与物理网络匹配度较差;同理,Pan 方法需先利用界标进行区域定位,界标的选择是随机的,倘若区域定位不准确,即使利用 IP 筛选邻近节点和基于事件驱动方法进行拓扑优化,也无法保证邻居节点的准确性.此外,Pan 方法的 GTMR 和 LNA 值还受到邻居节点个数的影响,DHT 所需维护的邻居节点个数较多,而其中准确的邻居节点个数少,故 GTMR 和 LNA 值偏低.在仿真时,我们尝试着增加 Pan 方法中界标节点的个数,但其 GTMR 和 LNA 未有明显提升.

在图 5 所示的不同的网络模型中,随着网络规模的增大,MHTMOA 的 GTMR 和 LNA 均呈下降趋势,这主要是由 MHTMOA 的节点加入机制易使节点位置准确度的误差随着规模的增大而累积造成的.由于树结构本身是层次结构,在节点加入时,需按树结构中节点位置顺序探测待加入节点到树中节点的路由跳数,故节点加入时对底层物理网络的跳数信息搜集是顺序相关的.若在树结构中,处在上层的节点位置与在物理网络中的位置有较大的偏差,则会影响下层节点进行路由探测的节点范围,进而影响新节点对父节点的选择,降低邻居节点的准确度,最终导致 GTMR 和 LNA 值下降.为降低节点加入顺序以及节点规模对邻居节点准确度的影响,可通过定期交换邻居节点信息进行局部地调整以提高邻居节点准确度.

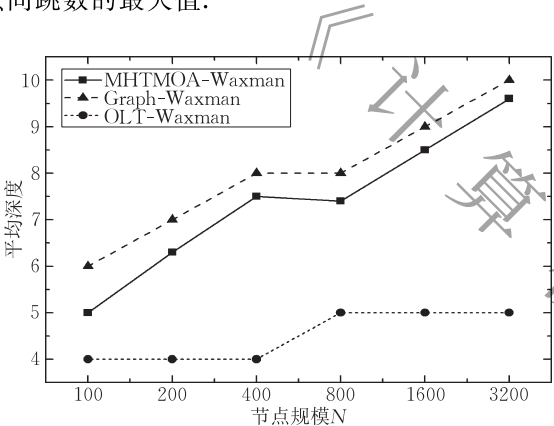
(2)GTMR 和 LNA 最大值

表 2、表 3 分别统计的是两种拓扑模型中,不同网络规模下 MHTMOA 和对照算法的 GTMR 和 LNA 最大值.从表 2、表 3 可看出,在不同网络模型、不同节点规模下 MHTMOA 生成的树形覆盖网络 GTMR 的最大值均保持在 93%以上,在最好的情况下甚至高达 99%,相较于 OLT 和 Pan 分别提升至

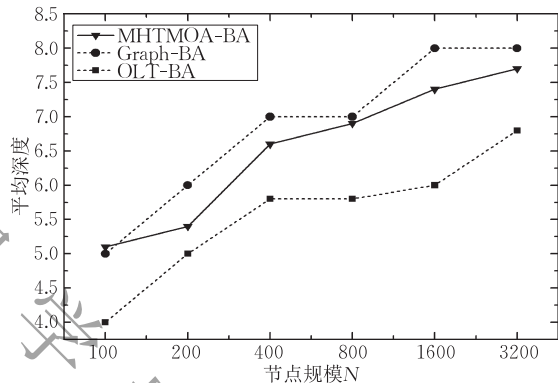
少 60%和 89%，最高提升比值可达 77%和 93%；不同规模的树形覆盖网络的 LNA 最大值均保持在 90%以上，相较于 OLT 和 Pan 提升至少提升 50%和 87%，最高可提升 75%和 92%。MHTMOA 与 3DO 相比，当节点规模小于 800 时，二者的 $GTMR$ 和 LNA 保持在同一水平；当节点规模大于 800 时，MHTMOA 的 $GTMR$ 和 LNA 明显高于 3DO。由此可得，在大规模网络中，MHTMOA 的拓扑匹配性能明显优于 3DO，因而 MHTMOA 更适用于大规模覆盖网络的构建。

(3) 树的深度

图 6(a)、图 6(b)分别统计的是 Waxman 和 BA 模型中 MHTMOA 和 OLT 算法生成树深度的平均值。其中,Graph 表示物理网络中根节点到其它任意节点间跳数的最大值。



(a) Waxman模型生成树深度



(b) BA模型生成树深度

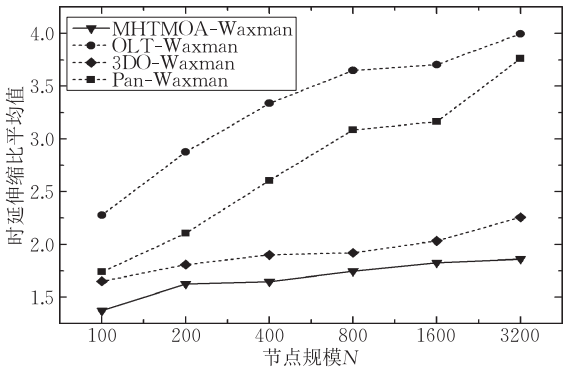
图 6 两种网络模型中不同算法的生成树深度

(4) LS 的平均值

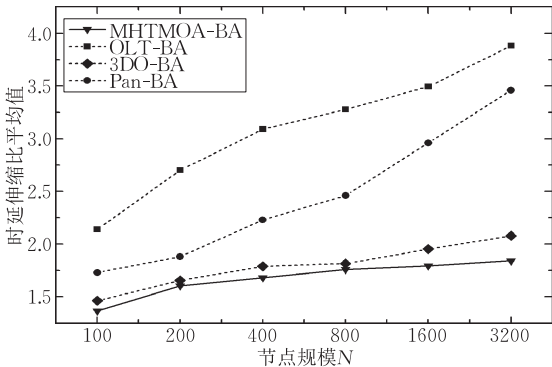
如图 7(a)、图 7(b)所示为 Waxman 和 BA 模型中不同节点规模下 MHTMOA 和对照算法 LS 的平均值。从图 7 可看出，在 Waxman 和 BA 网络模型中，MHTMOA 与对照算法的 LS 均随着节点规模

的增大而增大，但在相同的节点规模下，MHTMOA 的 LS 明显小于 OLT、3DO 和 Pan 方法。

例如，在图 7(a) Waxman 模型中，当 $N=1600$ 时，MHTMOA 的 LS 平均值约为 1.8，而 OLT 的 LS 约为 3.7，3DO 的 LS 约为 2.0，Pan 的 LS 约为



(a) Waxman模型的 LS 平均值



(b) BA模型的 LS 平均值

图 7 两种网络模型中不同算法的 LS 平均值

3. 0. 从图 7 还可看出,随着节点规模的增大,MHTMOA 的 LS 增幅较小. 例如,在图 7(b)BA 模型中,当节点个数从 100 增至 3200 时,MHTMOA 的 LS 从 1.4 增至 1.8,而 OLT 的 LS 从 2.2 增至 4.0,3DO 从 1.5 增至 2.1,Pan 的 LS 从 1.7 增至 3.5. 此外,相较于对照算法,在两种模型中,MHTMOA 的 LS 降幅皆可高达 53%. 综合上述可得,相较于对此照算法,MHTMOA 生成的树形覆盖网络的 LS 明显降低,更好地缓解了拓扑不匹配问题.

LS 可看做是 PS 和覆盖网络与物理网络平均响应时延比的乘积,它不仅可反映覆盖网络与物理网络的全局匹配程度,更能反映节点间响应时延的平均水平. 从 4.3 节(1)中对 $GTMR$ 和 LNA 的分析可知,MHTMOA 的邻居节点准确度高,逻辑链路与物理链路直接匹配程度高,故其逻辑路径与底层物理路径重合比例高,使得消息转发路径与底层物理网络传输路径趋于一致,故端到端的时延明显降低,最终 LS 减小. 3DO 采用基于 LID 的最近邻

路由方法,从图 5 中可看出其 $GTMR$ 和 LNA 较 MHTMOA 低,故 3DO 的覆盖网络与物理网络的匹配程度更差,导致 LS 比 MHTMOA 更高. 从 4.3 节的图 6 可看出,OLT 生成树的平均深度较 MHTMOA 低. 树的深度小意味着其平均跳数小,但由于 OLT 的 $GTMR$ 和 LNA 值低,其拓扑匹配程度差,导致覆盖网络路径时延较长,最终导致 LS 值增大. 同理可得,Pan 方法基于 DHT 技术构造覆盖网络,DHT 的平均路由跳数虽少,但由于拓扑不匹配,导致响应时延较长,故 LS 值较大.

(5) 节点加入效率

如图 8(a)、图 8(b)所示为 Waxman 和 BA 模型中 MHTMOA 采用集中式部署时与对照算法的维护开销比较. 从图 8 可看出,在两种模型中,MHTMOA 的维护开销均比 OLT 和 3DO 小,但比 Pan 的开销大. 需要指出的是,在 Waxman 模型下,MHTMOA、Pan 方法二者的维护开销相近,MHTMOA 的维护开销略小于 3DO.

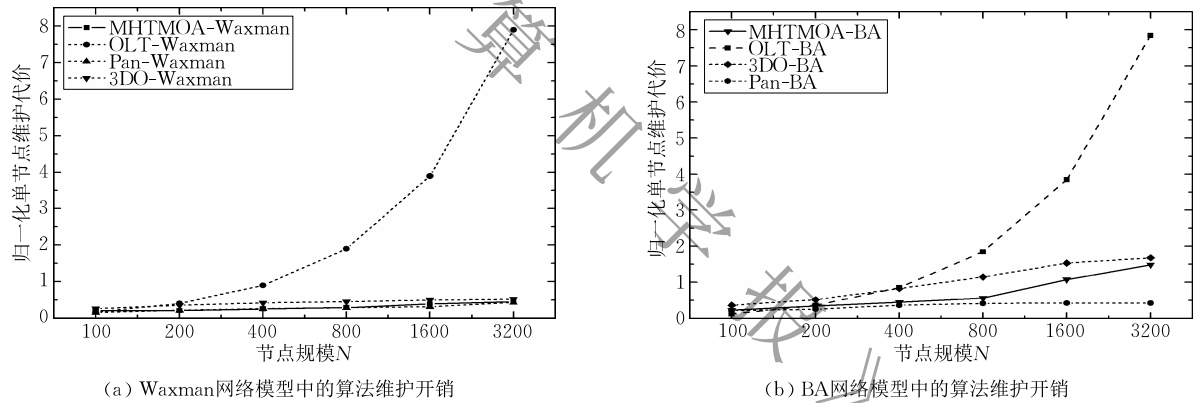


图 8 两种网络模型中不同算法的维护开销

在 MHTMOA 和对照算法中,MHTMOA、3DO、Pan 只需获取网络拓扑的局部信息,而 OLT 需获取网络的全局信息. 在 OLT 节点加入机制中,新节点加入时需先向所有已加入节点广播消息,并与候选父节点进行比较以选择位置最接近的节点,故其通信代价和计算开销随着规模的增大而急剧增加. MHTMOA 无需广播消息,只需与指定的节点进行测距和跳数比较即可,故其单节点维护代价较 OLT 低. 3DO 采用扩展环搜索发现邻居,即在局部范围内广播消息,它的维护代价主要受局部范围内节点个数的影响,其计算开销主要为节点 LID 的计算,同样受邻居节点个数的影响. Pan 方法中每个节点加入时先与界标节点进行时延测量,再在候选邻居节点集合中选择距离最近的节点作为邻居,其维护开销主要由界标节点个数决定,不易受节点规

模影响,故其通信开销和计算开销较小,维护代价较 MHTMOA 低.

在最坏的情况下,MHTMOA 在节点加入时,需从根节点开始沿着动态根节点的某个分支一直探测至树的最底层. 因而,随着节点规模的增大和树层次的加深,节点加入的通信开销和计算代价呈线性增长,易导致节点加入时间过长. 但通过仿真发现,这种情况出现的概率较低,大部分节点只需测量至树的中间层即可完成加入.

在本次仿真中 MHTMOA 采用集中式部署方法,若采用分布式部署,在节点加入过程中若动态根节点发生变化,节点需与新的动态根节点建立连接. 在仿真中发现,大部分节点只需测量至中间层即可完成加入,故动态根节点的平均变化次数较小,因而所需建立的连接数较少,故分布式的维护代价与集

中式相差不大。

(6) 动态环境下 LS 的变化

在仿真时,我们还比较了动态环境下 LS 的变化情况。由于 OLT、3DO 只提出了节点加入方法,未包含节点退出方法,故我们只比较 MHTMOA 与 Pan 方法在 Waxman 和 BA 模型中 LS 的变化情况。我们设置节点搅动率为每分钟加入 50 个,退出 50 个,得到 LS 随时间的变化曲线,如图 9 所示。

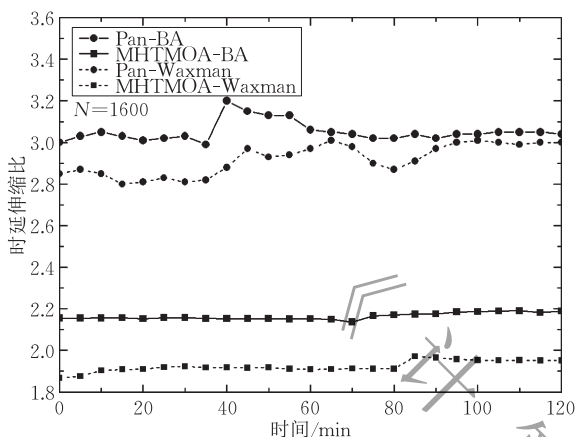


图 9 两种网络模型中 LS 随时间变化的示意图

从图 9 可看出,与 Pan 方法相比,在 Waxman 模型中,MHTMOA 的 LS 从 2.9 左右降至 1.9 左右;在 BA 模型中,MHTMOA 的 LS 从 3.0 左右降至 2.2 左右;在两种模型中 LS 均保持稳定,随着时间的变化波动较小。上述表明,在高动态环境中,MHTMOA 生成的树形覆盖网络的节点之间可更好地保持近邻关系,覆盖网络与物理网络仍保持高匹配特性。在 MHTMOA 动态拓扑维护过程中,若有节点退出,退出节点的子节点只需按新节点加入方法重新加入覆盖网络,节点维护代价同单节点维护开销,无需额外的维护开销。

综上,仿真结果表明 MHTMOA 生成的树形覆盖网络可保持较高的全局拓扑匹配比($GTMR$)和局部邻居节点准确率(LNA);同时, $GTMR$ 和 LNA 的提升也直接带来了 LS 的降低,从而更有效地缓解了覆盖网络与物理网络的不匹配问题。

5 讨 论

本文主要利用跳数三角形不等关系实现树形覆盖网络的构造和维护,除了跳数以外,MHTMOA 还可利用时延、带宽等作为度量指标。无论是跳数、时延还是带宽,都可表征节点的距离关系或是性能差异,MHTMOA 算法最终取得的效果都是使得网

络中位置相近或者性能相似的节点汇聚,以便更好地提供服务。

在 MHTMOA 的节点加入算法中时,情况 1 可能引起生成树结构中子树的合并,继而导致 LS 的抖动,影响网络 QoS。但树结构的优点即是节点变化时只需要修改其父节点,若子树合并,只需修改子树根节点的父节点信息即可,而不影响子树中节点间的位置关系,故子树合并的复杂性和维护代价较小。从本文高动态环境下 LS 的变化仿真实验结果来看,由节点动态变化引起的 LS 抖动较小。若子树合并引起拓扑结构变化较大,可通过局部节点位置调整来缓解 LS 的抖动。

在本文中,我们定义了 $GTMR$ 和 LNA 来反映覆盖网络链路和底层物理网络链路的匹配度以及局部邻居节点的准确度, $GTMR$ 、 LNA 、伸缩比都可较好地反映拓扑匹配性能,但三者又有所区别。伸缩比侧重反映全局匹配度,结合相关的 QoS 指标可直观地反映覆盖网络性能; $GTMR$ 和 LNA 侧重反映拓扑结构匹配的局部信息,单纯从结构上描述拓扑匹配特性。 $GTMR$ 和 LNA 可作为伸缩比的补充,三者结合可动态地反映拓扑特征变化和 network 性能变化。在实际应用中,可结合 $GTMR$ 和 LNA 针对不同的时延、带宽、负载均衡以及网络流量控制等需求对拓扑结构进行优化和调整。例如,可在保证一定程度的拓扑一致性的基础上,通过增加备份链路或是实现覆盖网络多路径路由等方法来缓解网络拥塞、降低端到端时延。

时延也是评估覆盖网络性能的一个重要的指标,解决拓扑不匹配问题可以有效降低寻址路径长度、降低端到端时延,但覆盖网络与物理网络的高度匹配,有可能使某些共享链路中的流量过于集中导致网络拥塞,继而导致传输时延增大。因而对于例如 CDN、应用层多播、P2P-TV、数据中心网等数据传输量大的网络,在构建覆盖网络时,需要折衷考虑寻址效率、时延、网络流量均衡等多方面的性能因素。如文献[38]提出了一种具有流量感知能力的启发式覆盖网络构造算法,在构造覆盖网络的同时考虑流量因素的影响。

本文重点关注 MHTMOA 的拓扑一致性保持特性,在不增加节点度约束的情况下,MHTMOA 易出现某个父节点所需维护的子节点个数过多,导致负载不均衡的问题。因而,后续我们还需制定相应的机制实现覆盖网络中节点的负载均衡。例如,可利用节点的可用带宽制定节点的度约束,或是同时利

用时延、带宽作为节点选择依据,实现动态服务邻居节点推荐实现负载均衡等。此外,我们还可构建基于 Mesh-Tree 的混合结构覆盖网络,通过增加节点间虚拟链路提高系统的鲁棒性,以降低父节点的压力。

6 结论和展望

本文提出了一种基于测量的启发式拓扑匹配优化算法(MHTMOA),旨在缓解覆盖网络与物理网络的拓扑不匹配问题。该算法包括节点加入、节点退出以及节点失效等处理方法,用于构造和维护一种树形结构覆盖网络。该算法通过测量底层物理网络中节点间的跳数,并根据跳数三角边长关系所满足情况的优先级别,结合 TIV 感知以及启发式规则构建覆盖网络。MHTMOA 在节点加入时无需获取网络的全局信息,只需测量网络中部分节点信息即可完成加入。MHTMOA 还可根据不同的应用需求采用不同的部署方式,不同部署方式生成的树形覆盖网络结构保持一致,但在维护代价上略有差别。

仿真的结果表明,在不同的网络拓扑模型、不同的节点规模下,相较现有算法,MHTMOA 的时延伸缩比(LS)、 $GTMR$ 、 LNA 值均有明显提升,尤其在节点规模增大时性能提升显著, LS 降幅可达 53%,从而更好地解决了拓扑不匹配问题。在高动态网络中,提议算法仍能保持良好的拓扑一致性。

MHTMOA 现已应用于“海云协同”服务创新平台^[39]中大规模海端节点的覆盖网络构造,并有望在未来应用于 5G 与信息中心网络融合的场景中,提供具有超低时延的服务以及基于位置的服务。下一步我们的研究工作主要是对 MHTMOA 进行优化,利用 $GTMR$ 以及 LNA 提升 MHTMOA 对拓扑动态变化的自适应性以及提高算法的可扩展性。同时,增加节点的度约束,并结合多项指标制定动态服务邻居节点推荐机制,实现负载均衡的目标。

致 谢 感谢中国科学院先导专项(XDA06040602, XDA06040501)、国家科技重大专项“新一代宽带无线移动通信网”5G 与信息中心网络融合技术研发专项(2017ZX03001019)对本论文工作的支持,感谢审稿专家对本文提出了宝贵的修改意见和建议!

参 考 文 献

[1] Zhang Chun-Hong, Qiu Xiao-Feng, Mi Wei. Comprehensive Analysis of P2P Technology. Beijing: The People's Posts and Telecommunications Press, 2010: 36-40(in Chinese)

(张春红, 裘晓峰, 弭伟. P2P 技术全面解析. 北京: 人民邮电出版社, 2010: 36-40)

[2] Moustakasa V, Akcanb H, Roussopoulosc M, et al. Alleviating the topology mismatch problem in distributed overlay networks: A survey. *Journal of Systems and Software*, 2016, 113: 216-245

[3] Malatras A. State-of-the-art survey on P2P overlay networks in pervasive computing environments. *Journal of Network and Computer Applications*, 2015, 55(55): 1-23

[4] Qiu Tong-Qing, Chen Gui-Hai, Mao Ye, et al. Towards location-aware topology in both unstructured and structured P2P systems//*Proceedings of the 2007 International Conference on Parallel Processing*. Xi'an, China, 2007: 30-41

[5] Lalitha B, Ch D, Rao S. Study of the topology mismatch problem in peer-to-peer networks. *Journal of Computer Engineering and Intelligent Systems*, 2012, 3(7): 145-157

[6] Shen Guo-Bin, Wang Ye, Xiong Yong-Qiang, et al. HPTP: Relieving the tension between ISPs and P2P//*Proceedings of the International Workshop on Peer-to-Peer Systems 2007*. Washington, USA, 2007: 1-6

[7] Sen S, Wang J. Analyzing peer-to-peer traffic across large networks. *IEEE/ACM Transactions on Networking*, 2004, 12(2): 219-232

[8] Tkachova O, Yahya A R, Muhi-Aldeen H M. A network load balancing algorithm for overlay-based SDN solutions//*Proceedings of the Problems of Infocommunications Science and Technology*. Kharkiv, Ukraine, 2016: 139-141

[9] Francois E, Gelenbe E. Optimizing secure sdn-enabled inter-data centre overlay networks through cognitive routing//*Proceedings of the 2016 IEEE 24th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*. London, UK, 2016: 283-288

[10] Mojamed M Al, Kolberg M. Structured peer-to-peer overlay deployment on MANET: A survey. *Computer Networks*, 2016, 96: 29-47

[11] Stoica I, Morris R, Karger D, et al. Chord: A scalable peer-to-peer lookup service for internet applications//*Proceedings of the the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*. San Diego, USA, 2001: 149-160

[12] Ratnasamy S, Francis P, Handley M, et al. A scalable content-addressable network//*Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*. San Diego, USA, 2001: 161-172

[13] Rowstron A, Druschel P. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems//*Proceedings of the Middleware 2001: IFIP/ACM International Conference on Distributed Systems Platforms*. Heidelberg, Germany, 2001: 329-350

- [14] Folino G, Pisani F, Trunfio P. Efficient discovery of data mining services over DHT-based overlays//Proceedings of the 2014 International Conference on High Performance Computing & Simulation. Bologna, Italy, 2014: 484-490
- [15] Isaac W, Fan-H T, Lin Y-H, et al. MR-chord: Improved chord lookup performance in structured mobile P2P networks. *IEEE Systems Journal*, 2015, 9(3): 743-751
- [16] Singh A, Dwivedi A, Singh Y N. Algorithms for faster overlay creation under high growth rate in query network based overlaid multicasting//Proceedings of the 2016 22nd National Conference on Communication. Guwahati, India, 2016: 1-6
- [17] Shin S, Lee U, Dressler F, et al. Motion-MiX DHT for wireless mobile networks. *IEEE Transactions on Mobile Computing*, 2016, 15(12): 3100-3113
- [18] Shi Guang-Yu, Long You-Shui, Chen Jian, et al. T2MC: A peer-to-peer mismatch reduction technique by traceroute and 2-means classification algorithm//Proceedings of the International Conference on Research in Networking. Singapore, 2008: 366-374
- [19] Ratnasamy S, Handley M, Karp R. Topologically-aware overlay construction and server selection//Proceedings of the IEEE International Conference on Computer Communication 2002 & 21st Annual Joint Conference of the IEEE Computer and Communications Societies. New York, USA, 2002: 1190-1199
- [20] Ng T S E, Zhang H. Predicting Internet network distance with coordinates-based approaches//Proceedings of the IEEE International Conference on Computer Communication 2002 & 21st Annual Joint Conference of the IEEE Computer and Communications Societies. New York, USA, 2002: 170-179
- [21] Fantar S G, Youssef H. Locality-aware chord over mobile ad hoc networks//Proceedings of the 2009 Global Information Infrastructure Symposium. Hammamet, Tunisia, 2009(009): 1-6
- [22] Li Zhen-Yu, Xie Gao-Gang, Min Ying-Hua. Topology matching method for structured P2P systems. *Journal of System Simulation*, 2006, 18(5): 1181-1185(in Chinese)
(李振宇, 谢高岗, 闵应骅. 一种结构化P2P系统的拓扑匹配算法. *系统仿真学报*, 2006, 18(5): 1181-1185)
- [23] Freedman M J, Mazieres D. Sloppy hashing and self-organizing clusters//Proceedings of the International Workshop on Peer-to-Peer Systems. California, USA, 2003: 45-55
- [24] Zhu Dan, Xu Xiang-Yang, Liang Yan, et al. Research on overlay network topology-matching. *Microcomputer Applications*, 2008, 29(6): 17-22(in Chinese)
(朱丹, 徐向阳, 梁燕等. 结构化P2P中覆盖网络拓扑匹配的研究. *微计算机应用*, 2008, 29(6): 17-22)
- [25] Huang Qi, Jin Hai, Liao Xiao-Fei. P2P live streaming with tree-mesh based hybrid overlay//Proceedings of the 2007 International Conference on Parallel Processing Workshops. Xi'an, China, 2007: 55
- [26] Abid S Akhtar, Othman M, Shah N. 3D P2P overlay over MANETs. *Computer Networks*, 2014, 64(4): 89-111
- [27] Traverso S, Abeni L, Birke R, et al. Neighbor-hood filtering strategies for overlay construction in P2P-TV systems: Design and experimental comparison. *IEEE/ACM Transactions on Networking*, 2015, 23(3): 741-754
- [28] Kawahara R, Lua E K, Uchida M, et al. On the quality of triangle inequality violation aware routing overlay architecture//Proceedings of the IEEE 28th International Conference on Computer Communications. Rio de Janeiro, Brazil, 2009: 2761-2765
- [29] Lumezanu C, Baden R, Spring N, et al. Triangle inequality and routing policy violations in the Internet//Proceedings of the 10th International Conference on Passive and Active Network Measurement. Seoul, Korea, 2009: 45-54
- [30] Fleming C L, Griffis S E, Bell J E. The effects of triangle inequality on the vehicle routing problem. *European Journal of Operational Research*, 2013, 224(1): 1-7
- [31] Imada N, Ueda K. Peer-to-peer network system and application design on multiple virtual networks//Proceedings of the 19th International Conference on Network-Based Information Systems. Ostrava, Czech, 2016: 298-302
- [32] Medina A, Lakhina A, Matta I. BRITE: An approach to universal topology generation//Proceedings of the 9th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems. Cincinnati, USA, 2001: 346-353
- [33] Waxman B M. Routing of multipoint connections. *IEEE journal on selected areas in communications*, 1988, 6(9): 1617-1622
- [34] Albert-László B, Albert R, Jeong H. Mean-field theory for scale-free random networks. *Physical A: Statistical Mechanics and its Applications*, 1999, 272(1): 173-187
- [35] Zhang Yu, Zhang Hong-Li, Fang Bin-Xing. A survey on Internet Topology Modeling. *Journal of Software*, 2004, 15(8): 1220-1226(in Chinese)
(张宇, 张宏莉, 方滨兴. Internet拓扑建模综述. *软件学报*, 2004, 15(8): 1220-1226)
- [36] Bagheri M, Movaghar A. Tree optimization in overlay multicast based on location and bandwidth//Proceedings of the 2009 International Conference on Computer Technology and Development. Kota Kinabalu, Malaysia, 2009: 46-49
- [37] Pan Li-Li. A topology matching algorithm of structured P2P network//Proceedings of the 2015 5th International Conference on IEEE Electronics Information and Emergency Communication. Beijing, China, 2015: 167-170
- [38] Adami D, Callegari C, Giordano S. Design and performance evaluation of service overlay networks topologies//Proceedings of the 2009 International Symposium on Performance Evaluation of Computer and Telecommunication Systems. Istanbul, Turkey, 2009: 296-303

[39] Tian Jing. Sea-Cloud coordinative and look ahead (preface). Journal of Network New Media, 2014, 3(1): 3-10(in Chinese)

(田静. 海云协同, 走向未来(序言). 网络新媒体技术, 2014, 3(1): 3-10)



LIAO Yi, born in 1991, Ph.D. candidate. Her current research interests are in video service system, network-aware algorithm and future networks.

SHENG Yi-Qiang, born in 1978, Ph.D., associate professor. His current research interests are in distributed machine learning, future networks, and heuristic algorithms.

WANG Jin-Lin, born in 1964, M.S., professor. His research interests are in future networks, technologies of media streaming-based applications in networks, and wireless communications.

Background

This work was mainly supported by the Special Fund for Strategic Pilot Technology of Chinese Academy of Sciences under Grant XDA06040602, Grant XDA06040501 and National Science and Technology Major Project under Grant 2017ZX03001019. The research aims at building a service platform, which provides an on-Site, Elastic, Autonomous (SEA) multimedia streaming service at the edge of the Internet. The SEA service platform would be used to collaborate with the existing Cloud computing at data center, named SEA-Cloud collaborative technology.

Nowadays, multimedia streaming traffic accounts for the highest proportion of the Internet traffic. According to Cisco's forecast, the portion of video traffic in the total amount of Internet traffic keeps increasing to reach around 80% or even more in 2019. It brings a big challenge for future Internet architecture to maintain high quality of

Internet service for both academic research and industrial applications.

This paper proposes a Measurement-based Heuristic Topology Matching Optimization Algorithm (MHTMOA) to build an overlay to manage a large scale of edge devices efficiently. The MHTMOA is the key to realizing collaboration, integrated control and intelligent interaction between the network elements of the SEA and the Cloud. Based on the MHTMOA, a public SEA-Cloud TV service would be built to support the construction of smart cloud television industry chain. In the near future, the MHTMOA would be used to the real applications in more scenarios such as the fifth generation mobile communications system (5G) and the information-centric networking (ICN) to implement ultra-low latency and location-based services.