DOI:10.13718/j.cnki.xsxb.2021.09.015

基于 SDN 的 Hadoop 数据中心网 DECMP 方法①

姚嘉鑫[1], 温佐承[2]

1. 四川旅游学院 信息中心, 成都 610100 ; 2. 四川旅游学院 信息与工程学院, 成都 610100

摘要: 提出一种基于软件定义网络(software defined network, SDN)的 Hadoop 数据中心网动态等价成本多路径路由(dynamic equal-cost multipath routing, DECMP)方法. 所提方法由 3 个模块组成: 链路监控模块、Hadoop 监视器引擎模块、基于软件定义网络 DECMP 的调度和路由模块. 在 DECMP 调度和路由模块中提出多路径 Dijkstra 路径查找算法以提供具有相同最小权重的多个路径. 然后, DECMP 调度和路由模块根据网络资源需求以及数量和大小, 来获得数据中心网络中每个流的有效带宽利用率, 为 Hadoop Map Reduce shuffle 阶段动态分配有效路径, 提高数据中心网络路由性能. 实验结果表示, 与其他现有方法比较, 本文 DECMP 方法在链路利用率和吞吐量方面都有所提升.

关 键 词: 数据中心网络; 软件定义网络; 动态等成本多路径路由; Hadoop

中图分类号: TP391 **文献标志码**: A **文章编号**: 1000-5471(2021)09-0115-06

DECMP Method of Hadoop Data Center Network Based on SDN

YAO Jiaxin[1], WEN Zuocheng[2]

1. Information Center of Sichuan Tourism University, Chengdu 610100, China;

2. Sichuan Tourism University School of Information and Engineering, Chengdu 610100, China

Abstract: A dynamic equal-cost multipath routing (DECMP) method based on software defined network for Hadoop data center network has been proposed. The proposed method consists of three modules: link monitoring module, Hadoop monitor engine module and scheduling and software-defined network DECMP scheduling and routing module. A multipath Dijkstra path finding algorithm has also been proposed in the DECMP scheduling and routing module to provide multiple paths with the same minimum weight. Then, DECMP obtains the effective bandwidth utilization of each stream in the data center network according to the network resource demand, quantity and size, dynamically allocates the effective path for the Hadoop MapReduce shuffle phase, improves the network routing performance in the data. The experimental results show that, compared with other existing methods, the proposed data center network DECMP method utilizes the leaf ridge topology designed in this paper, which has improved link utilization and throughput.

Key words: data center network; software-defined network; dynamic cost-cost multi-path routing; Hadoop

由于云服务的快速发展和物联网(internet of things, IoT)应用的广泛使用, 越来越多的存储设备、服务器和网络设备不断被添加到数据中心以存储、管理和分析数据[1-2]. 目前, 数据中心网络(data center net-

① 收稿日期: 2019-11-26
基金项目: 2017 年四川省教育厅重点项目"旅游过程实时智慧监管平台研究"(17ZA0290).
作者简介: 姚嘉鑫, 硕士, 副教授, 主要从事计算机应用研究.

work，DCN)内部带宽不能满足需求[3-4]．支持数据中心所需性能的一种方法是应用胖树拓扑[5]，在胖树拓扑中，有一个冗余链路用作故障转移机制，但单独应用拓扑不足以满足数据中心的带宽要求．多路径的路由技术为数据中心网络提供了更好的性能[6-13]．

数据中心海量数据集需要高效的存储工具和并行分布式计算，现有SDN的数据中心网络路由方法虽然提升了带宽利用率，但是没有考虑到Hadoop并行处理能力，且现有等价多路径路由协议(equal-cost multipath routing protocol，ECMP)方法不灵活．本文在此基础上提出一种基于SDN的Hadoop数据中心网动态等价成本多路径方法，该方法由链路监控模块、Hadoop监视器引擎模块和基于软件定义网络DECMP的调度和路由模块组成，通过动态调度和多路径路由方法，获得DCN中每个shuffle流的有效带宽利用率，并提供动态调度和路由，另外还可以加快shuffle阶段的执行时间，从而提高Hadoop作业的性能．

1 多路径Dijkstra算法

ECMP方法利用Dijkstra算法搜索最短路径，并在选择传递路径时使用模N哈希运算，但是在Dijkstra算法中，选择作为最佳路径的路径仅是一条路径，应用ECMP机制肯定是不够的，因为ECMP机制需要多条路径．本文对Dijkstra算法进行改进，以提供具有相同最小权重的多个路径．改进Dijkstra算法会选择多个路径，算法伪代码如下所示．

算法1 多路径Dijkstra算法

```
function MultipathDijkstra(Graph, source)
  for 有向图 Graph 中的每一个顶点 v
    dist(v)←INFINITY
    visited(v)←FALSE
    previous(v)←UNDEFINED
  end for
  dist(source)←0
  insert source into Q
  while Q 不空 do
    u←Q 中的顶点与 dist()中的最小顶点尚未访问
    从 Q 中将 u 移除，visited(u)←true
    for each neigbour v of u
      alt←dist(u)+ dist_between(u, v)
      if alt< dist(v)
        dist(v)←alt
        重置 previous(v)
        将 v 加入 previous(v)
      end if
      else if alt== dist(v)
        将 v 加入 previous(v)
      if ! visited(v)
        将 v 插入到 Q
    end for
  end while
  return dist
end function
```

获得最短路径以后进行哈希过程，哈希hash函数首先创建空数组，然后接收传入的数据包，数据包具有IPv4地址．如果数据包使用TCP或UDP协议，则源端口和目的端口信息将插入到阵列中．信息完成后，将调用CRC32函数进行哈希过程．哈希函数的输出结果是一个32位无符号整数，在哈希取模N过程中进行该过程，其中N是可用的路径数，模数的结果将指示使用哪条路径来发送数据包．随后控制器创建

路由规则并转发到与路由选择相关的所有交换机，然后将数据包转发回交换机，交换机将数据包转发到目的主机.

2　基于 SDN 的数据中心网 DECMP 方法

2.1　数据中心网络叶脊拓扑结构

目前数据中心网络中最常用的拓扑结构为三层胖树拓扑结构，缺点是限制了终端主机的位置，且由于使用多路径技术(如 ECMP)时连接终端主机的冗余路径会在网络中创建环路，最终导致网络利用率较差，并且会造成拥塞. 因此本文设计一种高效的数据中心架构：叶脊拓扑结构，具有可扩展性、可靠性和有效性. 该拓扑结构由两层组成，底层(叶层)中连接网络的终端主机交换机连接到顶层(脊层)，如图 1 所示.

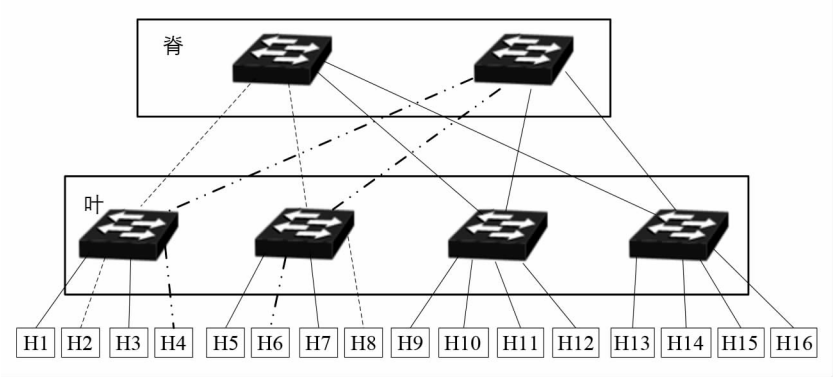

图 1　叶脊拓扑结构

从图 1 中可以看到，由于 ECMP 的分配技术，主机 3 可能会竞争叶交换机中相同的重载链路. 这可能会为两个大的数据流选择相同的重载链路，从而导致拥塞和冲突. 原因是 ECMP 缺乏整个网络的全局视图. 此外，使用 ECMP 算法，数据流根据其哈希值进行路由，可能会导致使用相同的路径，并在某些链路中造成拥塞. 另外，所有情况下的所有可能数据流路径都可能竞争相同的交换机，这会导致某些链路交换机上的过载. 在属于叶、脊交换机中用于数据流的分配路径的某些链路上可能会发生崩溃或故障. 因此，提出了一种有效的基于 SDN 的动态路由算法来执行路由过程，该算法考虑了每个流的网络资源需求以及大小和数量，能够在网络中任何链路发生故障或崩溃的情况下将流重新路由到另一可用路径.

2.2　基于 SDN 的 Hadoop 数据中心网 DECMP 方法

所提基于 SDN 的 Hadoop 数据中心网 DECMP 的方法包括 3 个模块，分别是：链路监控模块、Hadoop 监视器引擎模块、基于软件定义网络 DECMP 的调度和路由模块.

2.2.1　链路监控模块

该模块监控网络链路状态，定期从所有连接的 Open Flow 交换机以特定的时间间隔获取数据中心网络中加载的所有链路的统计信息，收集每个表、每个流和每个端口等统计信息并将其存储为快照.

由于 SDN 控制器缺少交换机之间链路所需的信息，因此没有链路层发现协议(Link Layer Discovery Protocol，LLDP) 用于识别网络拓扑中所有链路和交换机层所需的信息. 路由模块使用有关链路加载的统计信息来计算路径，数据中心网络中每个链路的当前负载计算方法如下所示：

$$L_{LK}=\frac{t\text{ 时间内传输的总字节数}}{B} \tag{1}$$

其中：B 是链路的带宽，t 是最近的间隔. 假设所有链路具有相同的带宽，并且每个链路都具有固定的权重($W=1$)，检查每个链路的当前负载(L_{LK})是否达到链路峰值：如果 $L_{LK}<1$，则表示尚未达到链路峰值；如果 $L_{LK}=1$，则可能由于一些更重的流而导致链路过载，从而导致不正确的路径分配. 因此，应该基于流的数量和每个链路的吞吐量来估计每个链路的权重，流的自然需求由 Hadoop 引擎模块估计. 如果当前负载超过已设置链路容量的 90% 的阈值 γ，则当前负载达到链路容量. 此外，通过使用每个链路的最大负载来计算叶、脊交换机中所有流路径的路径负载：

$$L_p=\max_{l\in p} f_l \tag{2}$$

其中：p 表示用于将流从源路由到目的地的路径，l 表示属于路径 p 的每个链路，并且 f_l 表示从源节点到目的地节点的遍历叶、脊交换机路径的每个链路的负载. 一旦计算了每个链路的路径负载，所有信息都被传递到调度和路由模块，以选择路径负载最小的方便路径 L_p，然后将流条目安装到所选路径的一组交换机中.

2.2.2 Hadoop 监视器引擎模块

此模块负责记录来自所有连接 Hadoop 服务器流的所有必需信息. 在 Hadoop 集群环境中，当 mapper 节点中的 map 任务将其输出数据写入 reducer 节点时，在 Hadoop 作业的 shuffle 阶段会生成 shuffle 流量，此流量需要足够的网络带宽来加速 Hadoop 作业的处理时间. 但是，主 Hadoop 框架不包含有关所需网络资源的足够信息. 因此，建议使用此模块在 Hadoop 作业的 shuffle 阶段识别从 mapper 节点传输到 Hadoop 集群中的 reducer 节点的数据.

之后，Hadoop 引擎将为每个 shuffle 流获取所需的网络带宽. 此外，所有收集的 shuffle 信息包括源 IP 地址、目的 IP 地址和每个 shuffle 流的大小. Hadoop 监控引擎还确定了 shuffle 数据的总量以及通过每个链路传输 shuffle 流的数量. 有关流的所有信息由 Hadoop 监控器提供给调度和路由模块，以根据每个流所需的带宽和链路利用率的当前负载来分配适当的路径.

2.2.3 基于软件定义网络 DECMP 的调度和路由模块

图 2 给出了基于 SDN 的 DECMP 调度和路由流程.

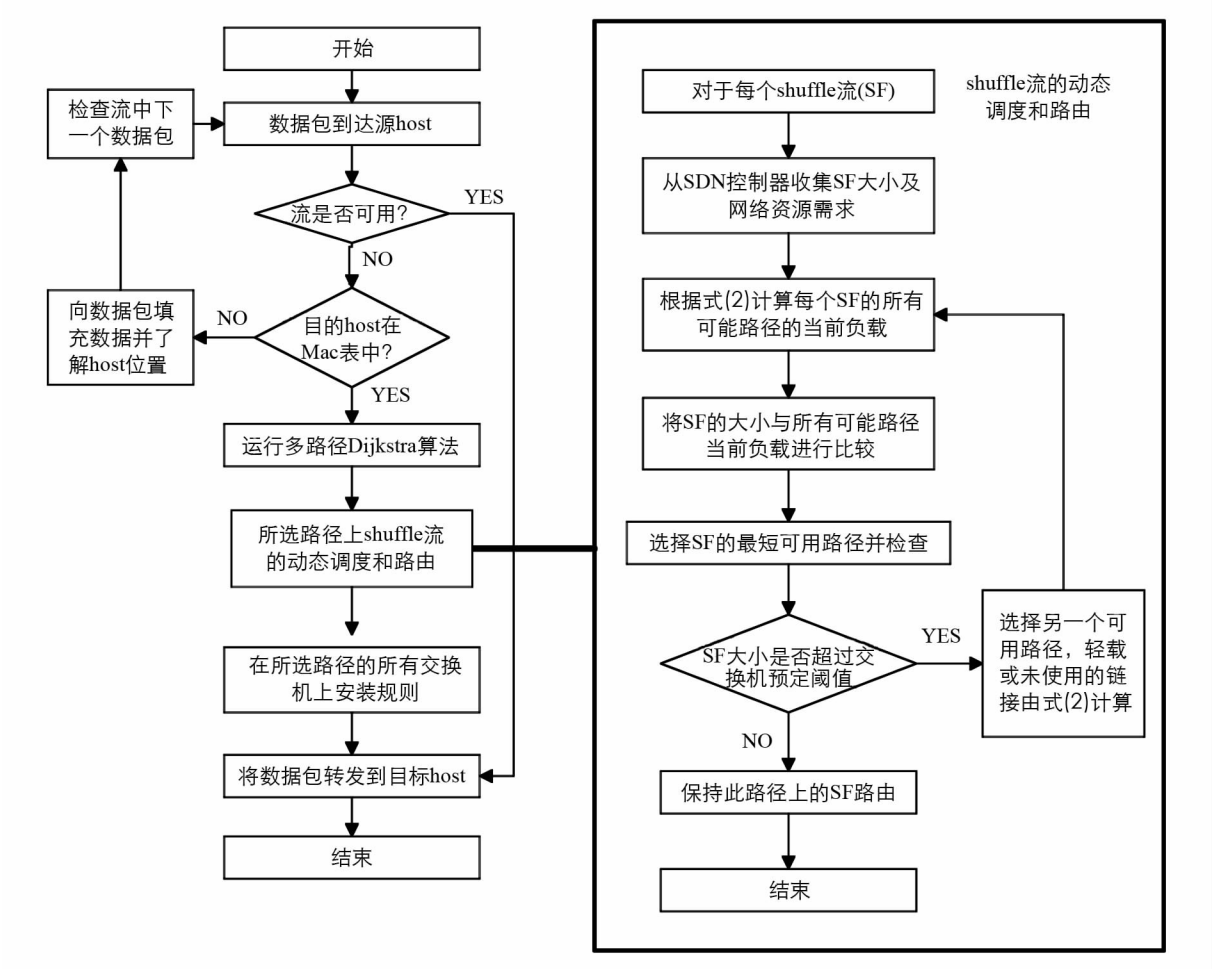

图 2 基于 SDN 的 DECMP 调度和路由流程

该过程从数据包到达发送 host 的第一台交换机开始. 然后交换机检查其流表，当传入数据包的流表存在时，立即将数据包转发到目的交换机. 但如果不存在，则数据包将被封装，然后发送到控制器. 已封装的数据包称为 packet-in. Packet-in 是控制器将接受的处理内容. 然后，控制器将检查传入的数据包. 检查目的 MAC 地址是否托管在 MAC 表中，如果不存在，则控制器将向其所有子交换机发送 packet-in，以查明所述主机的存在. 如果目的主机 MAC 地址已经在 MAC 表中，则可以执行路径搜索过程. 然后执行多路径

Dijkstra 算法，进行所选路径上 shuffle 流的动态调度和路由，结果将指示选择哪个路径发送数据包.

下一步是控制器创建一个路由规则并将该路由规则转发给与路由和转发数据包相关联的所有交换机，然后交换机将数据包转发给目标主机.

shuffle 流的动态调度和路由过程中，Open Flow floodlight 控制器转发模块中生成 packet-in 消息，通知控制器新流已到达 Open Flow 交换机. 交换机检查数据包，如果与其流条目不匹配，则将数据包转发给控制器. 另一方面，当流到期时，也生成流删除消息. DECMP 调度和路由模块为数据中心网络中不同主机之间的可交换 shuffle 流分配有效路径. 该模块执行所选路径上 shuffle 流的调度和路由，具有两个任务：计算可能路径任务和分配最佳路径任务. 计算可能路径任务是根据链路监控模块的统计信息计算可能的路径. 分配最佳路径任务是根据每个流所需的带宽，为所有 shuffle 流有效地分配最佳路径，防止拥塞.

3　实验结果与分析

使用 EstiNet 仿真器评估了所提方法，该软件模拟了一个叶脊拓扑，由 6 个连接的交换机组成，使用 1Gbps 的链路. 叶脊拓扑包括两层，叶层负责提供与端点的连接，脊层提供叶交换机之间的连接. 叶层中每个交换机都连接到脊层中的交换机. 脊层中交换机彼此不互连，叶层中交换机也不相互连接. EstiNet 软件还模拟了胖树拓扑，其中包括使用 1Gbps 链路的三层交换机，胖树拓扑包含了 20 个 4 端口的交换机.

ApacheHadoop 安装在 16 台主机上，所有主机都通过 Estinet 软件连接到模拟的叶脊拓扑和胖树拓扑上. 所有交换机都根据从 Floodlight 控制器接收的流规则转发数据包，这些交换机使用 TCP 连接到 Floodlight 控制器. 在 Floodlight 控制器和 Hadoop 主机的引擎之间建立另一个 TCP 连接，用来收集所有需要的 shuffle 流信息.

将本文所提方方法与现有数据中心网络路由方法进行比较，比较方法有：文献[11]中基于 SDN 的大象流负载均衡(elephant flow load balancing，EFLB)方法，文献[13]中基于 SDN 的胖树数据中心网络链路实时状态和流量特征的多路径路由算法(multipath routing algorithm based on Link real time status and traffic Feature，MLF)方法和文献[8]中的 ECMP 方法. 性能指标选择平均链路利用率、网络吞吐量和 shuffle 阶段执行时间作为性能评价指标.

图 3 给出了所提方法在叶脊拓扑和胖树拓扑下，shuffle 流的执行时间.

可以看出，叶脊拓扑的 shuffle 流执行时间比胖树拓扑的执行时间要少，这是因为胖树拓扑主要用于处理核心交换机到边缘交换机的流量，且胖树拓扑中两个主机之间的通信需要遍历从边缘层到核心层的分层路径，从而导致延迟和流量瓶颈. 而叶脊拓扑只有两层结构，减少冗余路径，减少时间延迟，最终减少 shuffle 流执行时间. 图 4 给出了不同方法在不同负载情况下的链路利用率对比结果.

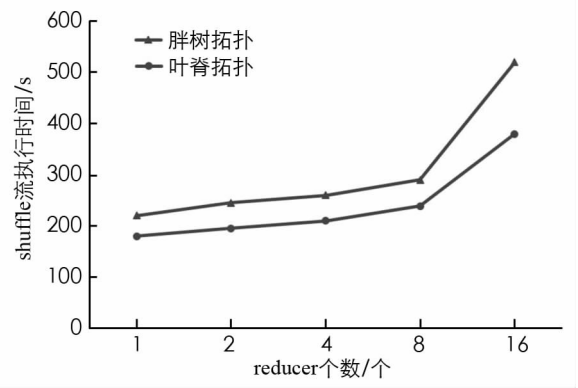

图 3　不同拓扑结构的 shuffle 流执行时间对比

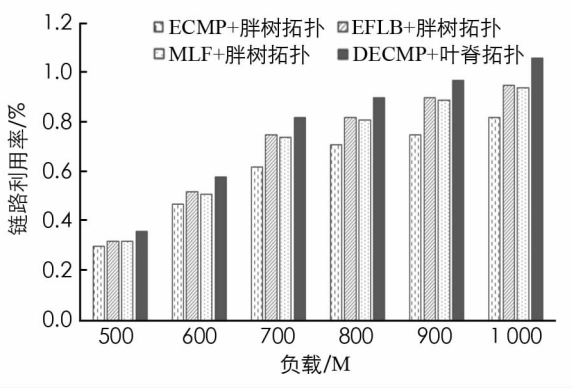

图 4　不同方法的链路利用率对比

从图 4 中可以看出，不同方法的链路利用率都随着负载的增加而增加. 本文叶脊拓扑下的基于 SDN 的 DECMP 方法链路利用率优于其他 3 种方法，这是因为：首先，其他三种方法中使用的是胖树拓扑，没有本文方法叶脊拓扑性能好；其次，ECMP 为非动态路由方法，且没有考虑流大小和链路实时状态，导致利用率最低. EFLB 算法对大象流进行负载均衡，MLF 算法将大小流分开，考虑了流大小，使得两种算法的链路利用率较接近，本文方法中使用了高效的叶脊拓扑，在 Hadoop 框架下，基于 SDN 的 DECMP 调度和路

由方法在SDN控制器中收集shuffle流大小和网络资源需求，动态选择SF最短路径，避免了链路拥塞，进一步提高网络链路利用率.

从图5可知，本文所提的基于SDN的DECMP路由方法在叶脊拓扑下的吞吐量优于其他三种方法，这是由于所提方法计算每个SF所有可能路径的当前负载，并通过SF大小与路径对比，动态选择可用路径，大大减少网络拥塞，提高了网络吞吐量.

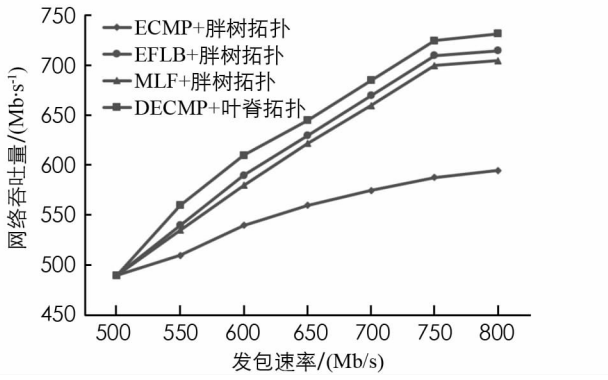

图5 不同方法的吞吐量对比结果

4 结 论

为解决数据中心网络中网络链路利用率和吞吐量问题，设计了叶脊拓扑结构，提出一种Hadoop下的基于SDN的DECMP路由方法，该方法由三个模块组成，分别是：链路监控模块、Hadoop监视器引擎模块和基于软件定义网络DECMP的调度和路由模块. 在两层叶脊拓扑中，提高了胖树拓扑性能，通过计算SF的所有路径当前负载，并进行比较，动态调度和多路径路由，获得数据中心网络中每个shuffle流的有效带宽利用率，并提供动态调度和路由，提高链路利用率和网络吞吐量. 另外还可以加快shuffle阶段的执行时间，提高Hadoop作业的性能.

参考文献：

[1] 母泽平. 一种虚拟化网络功能启发式动态编排算法[J]. 西南师范大学学报(自然科学版)，2019，44(6)：92-102.

[2] 杨 杰. 云计算在企业部署模式的研究[J]. 西南师范大学学报(自然科学版)，2017，42(6)：32-39.

[3] ZHANG Z，DENG Y H，MIN G Y，et al. HSDC：a Highly Scalable Data Center Network Architecture for Greater Incremental Scalability[J]. IEEE Transactions on Parallel and Distributed Systems，2019，30(5)：1105-1119.

[4] ZHENG K Y，WANG X R，LIU J. DISCO：Distributed Traffic Flow Consolidation for Power Efficient Data Center Network[C]//2017 IFIP Networking Conference (IFIP Networking) and Workshops. New York：IEEE Press，2017：1-9.

[5] QIAN Z M，FAN F J，HU B，et al. Global round Robin：Efficient Routing with Cut-through Switching in Fat-Tree Data Center Networks[J]. IEEE/ACM Transactions on Networking，2018，26(5)：2230-2241.

[6] 付应辉，刘必果，束永安. 基于SDN的胖树数据中心网络多路径负载均衡算法研究[J]. 计算机应用与软件，2017，34(9)：147-152.

[7] FARRUGIA N，BUTTIGIEG V，BRIFFA J A. A Globally Optimised Multipath Routing Algorithm Using SDN[C]//2018 21st Conference on Innovation in Clouds，Internet and Networks and Workshops (ICIN). New York：IEEE Press，2018：1-8.

[8] RHAMDANI F，SUWASTIKA N A，NUGROHO M A. Equal-Cost Multipath Routing in Data Center Network Based on Software Defined Network[C]//2018 6th International Conference on Information and Communication Technology (ICoICT). New York：IEEE Press，2018：222-226.

[9] DEWANTO R，MUNADI R，NEGARA R M. Improved Load Balancing on Software Defined Network-Based Equal Cost Multipath Routing in Data Center Network[J]. Jurnal Infotel，2018，10(3)：157-162.

[10] PALIWAL M，SHRIMANKAR D. Effective Resource Management in SDN Enabled Data Center Network Based on Traffic Demand[J]. IEEE Access，2019，7：69698-69706.

[11] 刘 毅，李凯心，李国燕，等. 基于SDN的动态负载均衡策略[J]. 计算机应用研究，2020，37(10)：3147-3152.

[12] 王红运，束永安. 数据中心网络中基于蚁群算法的动态多路径负载均衡[J]. 计算机应用研究，2020，37(7)：2148-2150，2166.

[13] 彭大芹，赖香武，刘艳林. 基于SDN的胖树数据中心网络多路径路由算法[J]. 计算机工程，2018，44(4)：41-45，65.

责任编辑 张 枸