PNFVNbM:一种基于 Metis 划分大规模 NFV 网络的方法

邓 理[1)] 戴宁赟[2)] 许 博[2)] 邢长友[2)] 陈 鸣[1),2)]

[1)](南京航空航天大学计算机科学与技术学院 南京 211106)

[2)](陆军工程大学指挥控制工程学院 南京 210007)

摘 要 随着虚拟网络的规模变大和虚拟网络的功能增多,单台服务器的资源可能无法支持一个较大规模 NFV 网络的运行,需要将其划分为多块并分别部署到服务器集群上,而且要使该 NFV 网络呈现出良好的性能.本文通过网络测量手段研究了 NFV 网络的资源需求和宿主服务器的资源供给以及两者的映射关系,探讨了利用多权值约束的图划分算法描述 NFV 网络中的关键计算资源的问题;提出了一种基于 Metis 划分 NFV 网络算法(Partitioning NFV Networks based on Metis,PNFVNbM),该算法在划分前对多权值进行融合,在划分后评估划分效果以改进融合参数,以提高各个部分的平衡度.最后建立了原型系统,对多个大规模 NFV 网络进行了实验验证.实验的结果表明了 PNFVNbM 算法能够均衡地将大规模 NFV 网络划分在多台服务器上,使得服务器集群能够支撑大规模 NFV 网络及其应用.

关键词 NFV 网络;Metis 算法;资源描述;多权值约束;服务器集群

中图法分类号 TP393 **DOI 号** 10.11897/SP.J.1016.2020.01958

PNFVNbM: A Method Partitioning a Larger Scale of NFV Network Based on Metis

DENG Li[1)] DAI Ning-Yun[2)] XU Bo[2)] XING Chang-You[2)] CHEN Ming[1),2)]

[1)](*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing* 211106)

[2)](*College of Command Control Engineering, Army Engineering University, Nanjing* 210007)

Abstract The widespread application of NFV technology in cloud computing and data center networks has made the scale of virtual networks larger and the functions of virtual networks increased. According to results in a production environment, the resources of a single server may not be able to support the operation of a large-scale NFV network. A key technical challenge is to divide a large scale NFV network into several parts and deploy them to a server cluster that each server runs a subnet of the NFV network. The multi-level partition algorithm Metis proposed by Karypis and Kumar is considered to be a very effective weighted graph partition algorithm, which can achieve the balance of the number of edge cuts and each partition. However, there are still many specific technical problems that need to be solved in the scheme for applying the Metis algorithm to the division of NFV networks, such as reasonably representing resources as weights in the diagram, and how to reflect the characteristics of NFV networks. There have been some researches on the algorithm of Metis to divide the network topology at home and abroad, but there is still a lack of research on the use of graph division algorithms to deploy large-scale NFV networks. Based on this idea, this paper studies the approximate resource

收稿日期:2019-01-05;在线发布日期:2019-05-06.本课题得到国家自然科学基金项目(61772271,61379149)资助.**邓 理**,硕士研究生,主要研究方向为网络功能虚拟化、软件定义网络.E-mail:dengli.bg@qq.com.**戴宁赟**,硕士,工程师,主要研究方向为网络体系结构、分布式计算.**许 博**,博士,副教授,主要研究方向为网络测量与网络应用.**邢长友**,博士,副教授,主要研究方向为网络测量与未来网络.**陈 鸣**(通信作者),博士,教授,博士生导师,IEEE Fellow,主要研究领域为网络测量、网络性能分析与建模、未来网络.E-mail:mingchennj@163.com.

requirements of servers and NFV networks through network measurement methods and discusses the problem of describing key computing resources when using the Metis graph partitioning algorithm. We quantified the CPU, memory, and disk usage of typical applications of NFV networks built using LXC and analyzed the CPU usage of the forwarding chain consisting of multi-hop virtual routers. According to the measurement results, we believe that the CPU of the host server is a critical shortage of resources in the system. Based on some modeling analysis, an NFV network partitioning algorithm based on a multi-weight constraint called Partitioning NFV Networks based on Metis (PNFVNbM for short) is proposed. This algorithm fuses multiple weights before partitioning and evaluates them after partitioning. Divide the effect to improve the fusion parameters to improve the balance of each part. Finally, we used a popular virtualization technology, LXC, to build a fairly large-scale NFV prototype system, and deployed multiple NFV networks on the system to test our partitioning algorithm. We compared the performance of NFV networks with regular and irregular shapes and different numbers of subnets. In addition, we also measured connectivity and communication bandwidth between parts deployed on different servers. The experimental results show that the PNFVNbM algorithm can evenly partition a large scale of NFV network. Large-scale networks divided by this method can ensure the performance of the network and make the work load-balanced on all available servers, which mean that this method makes the server cluster support the large-scale NFV network and its applications effectively. In the future, we will use this method to study the technology of deploying large-scale NFV networks on server clusters automatically and parallelly, and do some research on processing technologies when the workload changes dynamically.

Keywords　Network Function Virtualization network; Metis algorithm; resource description; multi-weight constraint; server cluster

1　引　言

网络功能虚拟化(Network Function Virtualization,NFV)技术是指用软件安装、控制、操作运行在通用硬件上的网络功能,以代替传统的专用网络设备[1],用软件实现传统上要用专用硬件与软件实现的网络设备如路由器、防火墙、入侵检测设备等.随着 NFV 在越来越多的运营商网络、数据中心、移动网络和企业网等场景中得到应用[2].NFV 已经为网络应用、网络安全、网络管理、网络科研等领域提供了虚拟化的用户终端设备、接入设备、网络转发设备以及防火墙、入侵检测系统等功能,同时这些虚拟网络设备还能够与实际网络设备互联在一起,极大地改善了网络技术的有效性、经济性和灵活性.

我们将由虚拟网络设备或由虚拟网络设备与实际网络设备结合所形成的网络称为 NFV 网络.目前基于 NFV 技术搭建网络平台的问题已经有了不少研究,大多采用虚拟机或容器技术进行部署.在 NFV 网络中,尽管所有虚拟主机、路由器、交换机、链路和控制器等都是完全由软件实现的实体,而非像传统网络设备一样是由软硬件结合体实现的,但是它们与传统网络设备的功能完全一样,各种协议的交互过程、分组传输过程等都是实际存在的.唯一不同的是,当 NFV 网络中某些虚拟设备消耗资源过多,而宿主服务器无法有效提供时,该 NFV 网络会存在一定的性能瓶颈[3-5].特别是近年来迅速发展的 OPNFV(Open Platform for NFV)、OpenStack[6]等技术,促进了 NFV 技术在云计算、数据中心网络中的应用,为 NFV 的发展提供了更广阔的应用场景,也使得 NFV 网络的规模也越来越大.由于大规模 NFV 网络将包括数量越来越多的虚拟网络设备,以及运行在虚拟网络设备上类型越来越多的网络应用,运行这些虚拟设备和应用都需要消耗宿主服务器上不同量的各种实体资源.一般而言,当单台服务器无法提供如此大量的资源,就需要将该大规模 NFV 网络部署到由多台服务器构成的服务器集群上,每台服务器上运行由该 NFV 网络划分出来

的一个子网.一方面该子网的功能能够正常运行,另一方面这些子网的功能之和能够等效于该 NFV 网络运行在一台资源充足的大型计算机之上.因此,这种技术方案中存在的问题包括:如何将该大规模 NFV 网络科学合理地分割多个子网,使得其各子网能够分别有效地运行在一台服务器上,并且这些子网集合能够等效于该大规模 NFV 网络运行在一台资源充足的大型计算机上所具有的功能和运行性能.划分 NFV 网络使其能够运行在服务器集群之上是当前 NFV 技术发展面临的一个关键技术挑战.

我们考虑是否能够将划分大规模 NFV 网络的问题转换为图划分问题,因为图划分问题广泛存在于日常生活和科研领域中,并且图划分的理论问题已经得到了较为充分的研究.图划分的思路是将设备抽象成图中的点,设备之间的链路抽象成图中的边,同时将每种设备占用的关键资源量抽象成图中点的权值.然而,图划分的目标依赖于具体问题的性质以及具体应用的目标,在各类实际问题中也并不相同,例如在生产专用集成电路(ASIC)上实现多核心的片上网络时,需用将 ASIC 划分到多片 FPGA 上,需要用 I/O 接口数目、触发器数目和查找表数目来描述.划分时,一方面划分的网络图显然应当保证子网之间的通信量尽可能小,另一方面也应当满足单个划分所需资源小于单个 FPGA 所能提供的资源量,以满足整个 ASIC 的性能需求.如果采用图划分方法用于划分 NFV 网络,我们面临许多具体的棘手的技术问题需要研究并解决,例如:

(1)采用何种图划分算法?解决图划分问题已经有许多方法,每种方法都要面对复杂性、精度、有效性等因素进行折中.划分 NFV 网络的问题采用哪种算法为宜呢?

(2)如何满足资源表示要求?图划分的本质是资源分配问题,处理好服务器的资源供给与 NFV 网络的资源需求是关键.NFV 网络关键资源包括哪些?这些资源如何表示才能映射成为图?

(3)如何满足图划分要求?首先,每台服务器供给的资源要能够承载运行子网所需,以满足 NFV 网络的性能要求.其次,服务器之间通过设备的物理接口相连,应尽可能减少子网与子网之间的通信量,使得服务器之间的流量小于服务器之间设备连接的最大传输带宽.第三,划分的各子网应保持负载均衡,使得每台服务器的负载量大致相当.上述三种要求如何体现在算法中?

(4)如果每个 NFV 网络设备需要多种资源,这些资源的关系如何处理?

本文针对在服务器集群上运行大规模 NFV 网络的实际需求,研究所需主要网络资源及其数量,提出解决上述问题的整体解决方案.本文的主要贡献包括:提出了基于网络测量手段定量分析服务器与 NFV 网络的资源需求的方法;提出了一种基于 Metis 多权值约束的 NFV 网络划分算法 PNFVNbM;设计实现了原型系统,并通过对多个不同规模的 NFV 网络的实验验证了本文方法.

本文第 2 节综述相关工作;第 3 节探讨 NFV 网络的资源描述与向服务器的资源映射;第 4 节研究基于 Metis 算法划分大规模 NFV 网络的模型及其算法;第 5 节通过原型系统对相关模型与算法进行实验验证;最后第 6 节总结全文.

2 相关工作

图划分问题及其算法的研究已有较长的历史,图划分问题及其算法的研究属于优化及启发式算法领域.Kernighan 认为图划分问题可以转化为整数条件下的线性规划问题[7].Stockmeyer 证明了图划分问题是 NP-complete 问题[8],因此寻找时间复杂度较低的启发式算法成为此类研究的重点.使用启发式算法将一个图划分为 k 部分(称为 k 路),通常使用分而治之的递归二分法.尽管递归二分法存在局限性,但仍是目前主要的图划分手段之一.为解决用于大规模集成电路的大规模图的划分以及优化问题,产生了多级算法[9].所谓多级算法是指先将规模较大的原图划分成若干较小的图,小图划分结束后,算法会设法将小划分还原成原始图的划分,并在还原的过程中不断进行优化和改进.Karypis 和 Kumar[10] 提出 Metis 算法就是多层多路算法中的一种精度较高、用时较短的有效算法,该算法是本节研究的基础.

由于大规模集成电路设计制造等领域实际需求的刺激,多年来图划分的研究成了国内学术界研究的热点问题.这些研究大致可以分为两类:一类将各种其他问题转化为图划分问题,然后应用图划分算法[11-13];另一类则是研究图划分算法本身,寻求提升算法的效率和降低时间复杂度[14].本节研究属于前一类问题.

国内外对于基于 Metis 划分网络拓扑的算法已有一些研究.Yocum 和 Eade 等人采用如随机、集群和 Metis 这三种图划分算法,对不同拓扑结构进行

了对比，发现 Metis 的划分效率最高、流量通信的效果最好[15]. 文献[16]通过分析影响并行网络仿真性能的因素，对 Metis 进行改进，实现了并行网络模拟拓扑的优化划分方法. 文献[17]针对片上网络的应用，提出了基于 Metis 的多权值约束的拓扑划分算法. 文献[18]将该算法运用到超图划分当中，设计了一个超图划分公式.

到目前为止，就我们所知尚未发现将图划分算法用于划分大规模 NFV 网络的研究. 但根据上述分析，用 Metis 算法划分大规模 NFV 网络具有可行性.

3　NFV 网络的资源描述与映射

3.1　NFV 网络的资源

Metis 算法是一种图划分算法，用该算法划分图有两个主要原则[10]：第一是最小切边原则，即划分块之间通信量最小；第二是负载均衡，即每个划分块规模相当. Metis 主要采用多级 K 路的划分方法，主要包含了三个步骤：粗化、初始化分和细化. 当图规模较大时，如包括数千个节点，若直接采用初始化分过程中的多级 K 路划分方法往往容易出现局部最优的情况，不能获得较好的划分结果. 而对规模较小的图，该方法有较好的效果. 因此，Metis 算法首先进行粗化步骤，将图的规模压缩至 100 个节点左右，以形成粗化图；再通过初始化分步骤，分割粗化图；最后，采用细化步骤，将粗化图的划分结果映射回原图，并通过微调优化结果.

在划分之前，我们需要将 NFV 网络与图关联起来. 对应于图的节点是 NFV 网络的网络设备，对应于图的边是 NFV 网络中网络设备间的链路. 其中节点的权重与该节点所占资源的多少正相关，边的权重与节点之间的通信流量大小正相关.

在对 NFV 网络进行划分时，我们需要针对 NFV 网络的关键设备、关键应用的资源需求进行定量分析. 然而，大规模 NFV 网络的拓扑复杂，应用多样，如何建立各种应用与多种资源的定量关系是个难题. 为此，我们采用网络测量方法，通过在典型的商用服务器上进行大量测量实验，以理解基于主流虚拟化技术建立的 NFV 网络中的关键设备、关键应用对各种资源的典型需求及其量化关系.

我们假设服务器主要提供三种关键资源：CPU、内存和磁盘储存空间. 表 1 列表给出了两种典型服务器的资源供应量. 以当前流行的虚拟化技术 Linux 容器（LXC）作为 NFV 网络的基础设施支撑，NFV 网络的各种虚拟网络功能（VNF）承载在 LXC 之上. 网络测量的结果表明，LXC 在不承载其他 VNF 的情况下，所消耗的内存约为 471 MB，存储空间约为 159.43 MB，而 CPU 利用率小到可忽略不计.

表 1　典型服务器的资源

服务器型号	CPU	内存/GB	磁盘/TB
Lenovo ThinkServer RD550	Intel(R) Xeon(R0) CPU E5-2620 v3@ 2.40 GHz (24 cores)	32	4
System X3650 M5 8871AC1	Intel(R) Xeon(R) CPU E5-2630 v4@2.20 GHz v3@2.40 GHz(40 cores)	100	4

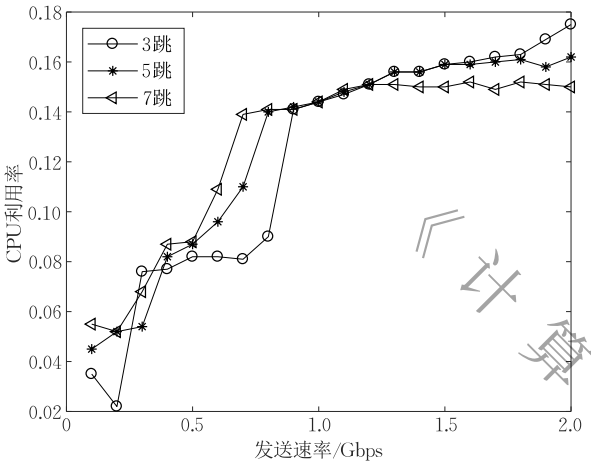
3.2　资源的量化评估

我们也对由 LXC 承载的构建 NFV 网络的主要构件如虚拟路由器、虚拟主机进行了测量. 统计测量结果表明，它们在 NFV 网络中具有数量庞大但每个个体消耗的资源很小的特点. 进一步研究发现，随着这些虚拟设备处理的流量增大，CPU 资源的消耗会急剧增大. 对比表 1，在三种资源中，宿主服务器的内存及存储器资源充足，但 CPU 却可能成为紧缺资源. 尽管各种服务器的 CPU 资源有差异，但更强大的 CPU 总是能够支持运行更大的 NFV 网络.

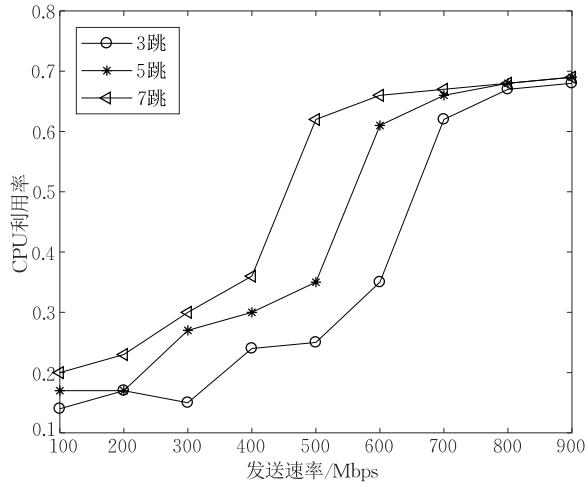
下面我们利用网络测量技术重点对于 NFV 网络中的典型设备和应用所消耗的各种资源，特别是 CPU 资源进行了分析，利用统计方法对它们进行定量描述. 典型的虚拟网络设备例如路由器、防火墙和 IDS/IPS 等. 注意到对现实的 NFV 应用而言，NFV 网络是其基础，PNFVNbM 算法首先研究的是大规模 NFV 网络构建问题. 仅当该基础性问题解决后，才能研究与其他 NFV 应用相关的问题. 研究中我们发现，在虚拟设备及其路径上转发大量分组时，会消耗宿主服务器的大量 CPU 资源，而内存及存储器资源消耗较少.

为此，我们定义了一个新概念转发链，用以衡量宿主服务器的资源消耗量. 该转发链定义为：首尾相连的多跳虚拟路由器所形成的分组转发路径，该路径两端分别具有一个模拟流量发送端和一个接收端，这些虚拟路由器通过运行网络路由协议（如 OSPF 或 BGP 等）形成分组转发路由. 当发送流量增大时，该转发链对 CPU 资源的消耗急剧增加. 在测量中我们将转发链看作一种虚拟网络设备，来考察它的 CPU 资源使用情况，转发链能力恰好能够与 NFV 网络的带宽资源正相关映射.

图 1 显示了在一款常用服务器(型号为 Lenovo ThinkServer RD550)对转发链的测量结果. 图 1(a)显示了在单条转发服务链且转发链的长度分别为 3、5、7 跳的情况下,当转发速率不断增加时,使得宿主服务器的 CPU 利用率迅速增加. 当发送速率大于 0.7Gbps 时,CPU 利用率可达 15%. 此后随着速率进一步增加,CPU 利用率基本呈线性增长. 图 1(b)显示了在 5 条并行转发服务链且转发链的长度分别为 3、5、7 跳的情况下,CPU 利用率增长趋势. 此时,宿主服务器的 CPU 利用率可以到达 0.7,可见转发服务链已经对宿主服务器性能产生了重大影响.



(a) 单条转发服务链的情况



(b) 5条并行转发服务链的情况

图 1 转发服务链对某服务器 CPU 利用率的影响

表 2 给出了在该服务器上运行不同虚拟设备时对于各种资源消耗的测量结果. 其中,为了便于对 CPU 资源的比较统计,我们将基于 LXC 的虚拟路由器定为 1 个 CPU 资源单位,而该服务器能够供给的 CPU 资源总量计为 2500,即计算 CPU 资源时采用相对值计算. 而计算内存和磁盘资源时则采用实际使用量的绝对值. 显然,对于不同的 VNF 如转发链、防火墙、IDS 等,所消耗的 CPU 资源量存在差异,而且都比虚拟路由器/主机消耗的 CPU 高 100 多倍,但它们彼此之间的差异并不太大. 对于内存和存储资源,不同的 VNF 也比路由器/主机消耗高 50 倍,但它们彼此之间的资源开销基本相同.

表 2 典型应用在测试服务器上的资源使用量

典型虚拟设备/应用	CPU(相对值)	内存/MB(绝对值)	存储/MB(绝对值)
7 跳转发链	120	200	200
防火墙	112	200	200
IDS/IPS	100	200	200
路由器/主机	1	4	4

观察表 2 的资源占用情况,能够发现宿主服务器的 CPU 是系统的关键性紧缺资源. 这表明单台服务器可以承载的 NFV 子网规模主要取决于其 CPU 资源能否满足需求. 由于 NFV 网络中的设备很多,我们可以主要考虑典型消耗 CPU 资源多的设备/应用;普通虚拟路由器/主机在设计时可以忽略不计,而是通过增加一个系数来进行调整.

4 基于 Metis 划分大规模 NFV 网络的工作过程

基于第 3 节的工作,我们可以估算出运行一个大规模 NFV 网络所需要服务器的数量. 接下来,需要解决将 NFV 网络转化为带权的无向拓扑图的问题,才能够实际使用 Metis 划分算法大规模网络划分问题. 图 2 给出了基于 Metis 划分 NFV 网络(Partitioning NFV Networks based on Metis,

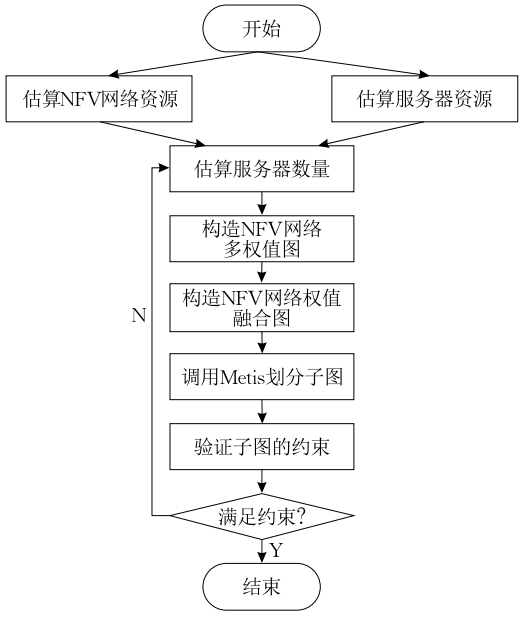


图 2 PNFVNbM 的工作过程

PNFVNbM)的工作流程.

4.1　工作过程描述

4.1.1　估算所需的服务器资源

假设某服务器能够提供 m 种资源，分别记为 $R_1,R_2,\cdots,R_m$，NFV 网络中共有 k 种应用(每种应用对应一条服务链)，每种应用 $A_i(i=i\sim k)$ 占用资源 R_j 的资源量为 S_{ij}. 若每台服务器中每种资源 R_j 的资源量为 C_j. 那么，对于每种资源来说，服务器集合所提供的资源总量必定要大于 NFV 网络需要消耗的资源量. 即若共有 N 台服务器，则应满足：

$$\sum_{i=1}^{k} S_{ij} < N \cdot C_j,\ j=1,2,\cdots,m \tag{1}$$

考虑到实际部署 NFV 网络时，不能将每台服务器的资源全部用完. 由常识可知，当服务器的 CPU 利用率长期达到 70%时，该服务器就可能无法维持其操作系统和应用程序的正常工作. 因此在估算服务器数目时，需要除以系数 $\varepsilon(0<\varepsilon<1)$，即每台服务器中的负载量最多为总资源量的 $1/\varepsilon$ 倍. 因此理论上来说，若要满足服务器承载能力需求，对于每种资源来说，至少需要的服务器数目为 $\left\lceil \dfrac{\sum_{i=1}^{k} S_{ij}}{\varepsilon \cdot C_j} \right\rceil$. 在所有资源中，取该值的最大值作为服务器的数量，即初始化数量为

$$N=\max\left(\left\lceil \frac{\sum_{i=1}^{k} S_{ij}}{\varepsilon \cdot C_j} \right\rceil\right),\ j=1,2,\cdots,m \tag{2}$$

我们考虑 CPU 利用率达到 70%时会影响 NFV 网络中应用的性能，故令 $\varepsilon=0.6$ 作为匹配值.

4.1.2　构建 NFV 网络的多权值图

由于 Metis 只能处理单节点权重的图，因此需要通过点权值的融合，将 CPU、内存、存储这 3 种资源需求融合为一个点权值，然后才能进行后继处理.

如果将每个节点的不同资源分别用点权值表示的话，我们能够得到表示三种关键资源的多权值图，每个节点有三个点权值. 这样，将 NFV 网络图转换为无向加权图需遵循以下规则：

(1) NFV 网络中的每个网络设备或应用均映射为图中的一个节点，只保留节点之间的连接关系而不包含其他信息.

(2) 由于 NFV 网络中大部分节点均为虚拟 Router/Host 节点，因此所有节点的初始权值均为(1,1,1)；然后再根据具体设备与应用情况调整其三种资源的权值. 特别是对于高带宽转发链，由于资源占用涉及所有节点，因此对于该应用所有节点的资源占用量均为(20,100,100).

(3) 节点之间的链路对应为节点之间连接的边，边权对应于带宽值，初始值可设为 100；对于高带宽的转发链应用，其边权值可根据实际情况设置，如设为 1000.

举例来说，对于图 3(a)所示的一个简单的 NFV 网络，Router1 为边缘路由器，其上部署防火墙. 图 3(b)给出了相应的多权值无向拓扑图，图的存储以 CSR 格式为标准，单个节点按照以下格式输入：

$$(w_1 w_2 \cdots w_{ncon}\ v_1\ e_1\ v_2\ e_2 \cdots\ v_k\ e_k),$$

其中，$w_1 w_2 \cdots w_{ncon}$ 是每个节点的多个权值，v_i、e_i 为与其相连接的邻接节点以及边权值.

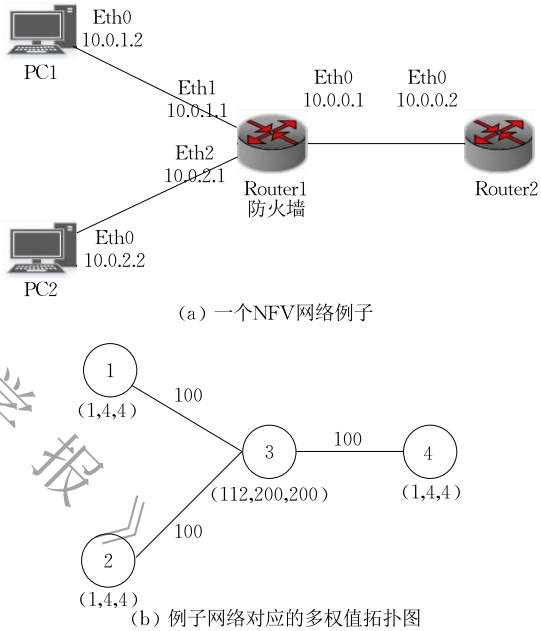


图 3　NFV 网络与其多权值拓扑图举例

据此，图 3 所对应的图文件为：

```
4  3
1  4  4  3  100
1  4  4  3  100
112  200  200  1  100  2  100  4  100
1  4  4  3  100
```

其中第一行记录了图的节点总数与边数. 下面第 n 行则依次记录了节点号为 n 的节点的三个点权值，以及相邻接点和边权值. CSR 格式广泛用于图的存储，同时也是 Metis 可以处理的标准格式. 但对于具有多个点权值的拓扑图来说，各权值的重要程度存在差异. 因此，需要通过权值融合，依据多个权值的重要程度合成为单一节点权重，才能调用 Metis 算法进行进一步处理.

4.1.3 多权值融合

本节将每个节点的3个权值融合成1个，获得一个新的拓扑文件，再交给Metis进行处理.

考虑节点 v 具有三个权重(r,s,t)，C_1、C_2、C_3是三个权重在每个划分中的上限.定义 α,β,γ 为每个权重的因数.它们遵循以下规则[7]：

$$\alpha+\beta+\gamma=1 \tag{3}$$

$$\alpha:\beta:\gamma=\frac{n\cdot C_3}{\sum_{i}^{1}t}:\frac{n\cdot C_2}{\sum_{i}^{1}s}:\frac{n\cdot C_1}{\sum_{i}^{1}r} \tag{4}$$

式(3)保证当三个权重重要性相同，均为1时，原权重不变；式(4)中$\frac{n\cdot C_j}{\sum_{i}^{1}w}$表示每个权重的上限与所有点的权重和之比，该值越大，表明这种资源越充足，限制越容易满足，与之相对应的权重重要性越低，因此三个权重重要性因数应为该值的反比.通过以上方法计算出权重的重要性因数后，可通过式(5)得到每个节点的新权值：

$$Z=\alpha\cdot r+\beta\cdot s+\gamma\cdot t \tag{5}$$

通过这一步，更新了原先获得的图文件，得到了具有单个点权值的图文件.

4.1.4 验证约束

根据先前估算的划分数量，采用Metis算法对4.1.2节中所得的图文件进行处理，得到一个划分结果.Metis的输出为一列数字，第 k 行的数字表示节点号为 k 的节点所在的划分号.这样我们可以清楚地得知每个划分子块中的所有节点号.

接下来，我们需要验证划分结果是否满足两种约束：一是统计每个划分子块中三种资源的总使用量，并与每台服务器的资源供应量进行对比.二是计算每个划分子网的不平衡系数.不平衡系数为划分子块中节点权重和最大值与最小值的比值，该参数反应了划分子块之间的差异.若每个子网的三种资源约束条件均得到满足，且不平衡系数小于1.05，则划分成功；反之，划分失败.若划分失败，需要将划分数量增加1，然后再用上述步骤对拓扑进行重新划分.值得注意的是，不平衡系数可以根据实际服务器集群的情况进行调整，本文中由于各服务器性能相同，因此设置为1.05.如果服务器的承载能力大小不同，可以适当提高不平衡系数，改进划分结果.

4.2 算法描述

本节以基于Metis划分NFV网络(Partitioning NFV Network based on Metis，PNFVNbM)算法的形式对4.1节的工作过程进行描述，假设NFV网络有 N 个节点、m 种资源.

算法1. PNFVNbM.

输入：大规模NFV网络节点数 N，模型 model

输出：划分结果

```
1.  topo.brite=Brite(N,model)
                    //输入网络节点数量N以及网络模式，通
                      过拓扑生成器Brite生成网络拓扑文件
2.  topo=transform(topo.brite)
                    //拓扑转换，转换为CSR格式
3.  add typical network applications APP_i(i=1~k) to
    topo                    //为网络添加典型应用
4.  n=max(⌈Σ_{i=1}^{N} S_ij / (ε×C_j)⌉), j=1,2,…,m
     //根据m种资源需求量和提供情况，估算划分数目n
5.  merge the node weights, update topo
                    //权值融合，更新拓扑文件topo中的权重
6.  execute gpmetis topo n
7.  FOR i=1 TO N
8.    node[i].part=temp
9.    FOR j=1 TO m
10.     part[temp].s[j]=part[temp].s[j]+
                        node[i].s[j]
11.   END FOR
12. END FOR
13. FOR i=0 TO n
14.   FOR j=1 TO m
15.     IF (part[i].s[j]<0.6×Cj&&
            balance<1.05) THEN
16.       output partition result
17.     ELSE
18.       n=n+1
19.       GOTO execute gpmetis
20.     END IF
21.   END FOR
22. END FOR
```

在PNFVNbM算法中，第1行通过网络拓扑生成器Brite，输入网络节点数量、网络模型(如幂率分布、层次结构或随机等模型)等参数，生成具有.brite格式的拓扑文件.如果用户已经设计好某个NFV网络，也可以直接给出相应的拓扑文件.第2行通过格式转换，将拓扑文件转换为CSR格式.第3行根据网络应用的具体资源消耗，修改网络拓扑文件的相关节点赋予点权值以及边权，点权值的数量等于消耗资源的类型数量.第4行通过计算NFV网络

的资源整体需求以及单台服务器资源总供应量，估算运行该 NFV 网络需要的服务器数量. 其中为使 NFV 网络性能保持在较好水平，要求所需的资源总量仅能达到服务器集群供给量的 60%，我们将单台服务器资源使用量的上限设为其资源供给量也设为 60%. 第 5～6 行是根据各资源的重要因子，将节点的多个点权值融合为一个，更新拓扑文件，然后进行 Metis 划分，获得划分结果. 第 7～12 行，对所有节点进行扫描，统计各划分块中的 m 种资源的使用情况，第 13～18 行表示约束验证的过程，检查每个划分中的各种资源使用情况，如果每种资源的使用量都在约束范围之内，同时平衡因子小于 1.05，那么划分成功；否则将划分数量加 1，对拓扑进行重新划分. 迭代进行上述过程直至划分成功，输出划分结果.

5　实验及其分析

本节设计实现了基于 NFV 的原型系统，为验证 PNFVNbM 算法的有效性提供了条件. 搭建本实验平台共使用了 3 台服务器，服务器的型号为 Lenovo ThinkServer RD550，性能参数见表 1.

5.1　划分一个对称的 NFV 网络及其讨论

为了便于判断，我们给出了如图 4 所示的一个规模为 60 个节点的 NFV 网络，该网络具有对称的结构，部署了 6 条高带宽（如 1000 Mbps 速率）转发链，其他链路则均为 100 Mbps 速率.

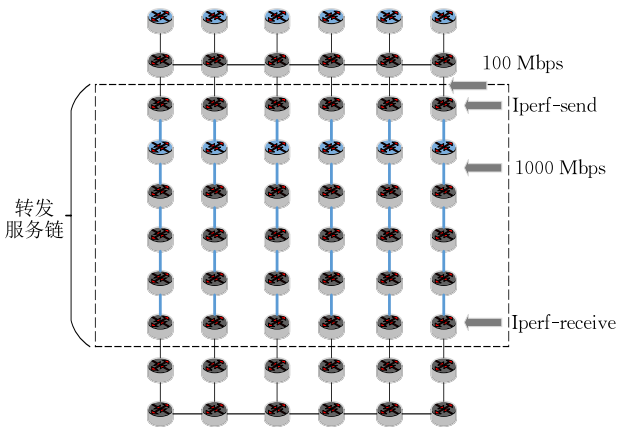


图 4　一个对称的实验 NFV 网络

先将该 NFV 网络转化为具有相应点权和边权的网络拓扑图，再采用 PNFVNbM 算法对该图进行分割. 根据 Metis 算法的最小切边原则，划分过程应当会尽量避开权值较重的边进行网络切割，而根据 Metis 算法另一个负载均衡的划分目标，每台服务器上的资源占用情况应当均衡. 显然，如果将该网络部署在 2 台服务器上，第 1、2 块都应当部署 3 条高速率转发链. 如果将该网络部署在 3 台服务器上，第 1、2、3 块都应当部署 2 条高速率转发链.

使用 PNFVNbM 算法分割该图的结果表明，该算法的确能够正确分割该图. 根据 PNFVNbM 算法将该图分割为 2 块或 3 块的结果，我们基于 LXC 虚拟化技术，对 NFV 网络在三台服务器上进行了实际部署，得到原型系统. 由于这部分内容超出了本文研究范围，这里不再讨论.

设 Part1、Part2、Part3 分别表示该 NFV 网络分别部署在 1 台、2 台、3 台服务器上的情况，子网之间多条 100 Mbps 的链路通过服务器之间的两条以太网千兆双绞线进行连接.

为了比较子网分割前后对服务器负载的影响，我们让各条高速率转发链承载不同强度的流量. 图 5 显示了在不同的流量速率下，测量得到转发链承载不同流量的 UDP 带宽值.

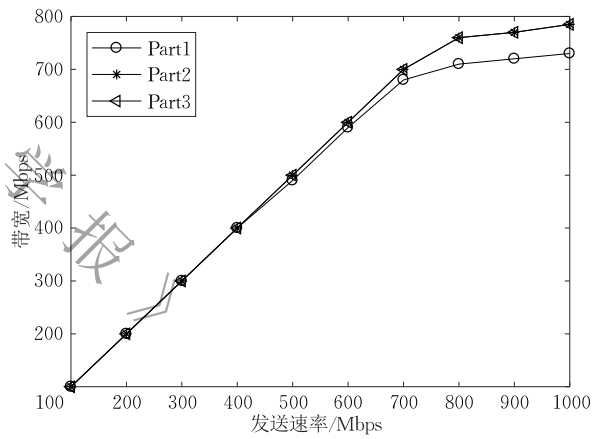


图 5　不同划分时高速率转发链的带宽测量

如图 5 所示，在不同划分的情况下，随着发送速率的增加，带宽值也随之线性增加. 但当发送速率达到一定值（如约 730 Mbps）时，出现随发送速率增加带宽增长放缓的趋势. 随着划分数目的增加，最大带宽值也随之增加. 对于 NFV 网络分别部署在 1 台、2 台、3 台服务器上的三种情况，上述情况大致相同. 只是对于将 NFV 网络部署在 1 台服务器（对应于 Part1 的曲线）的情况，当流量较大时，测量到的带宽值更小一些. 这是因为这台服务器所承受的负载比在其他两种情况下更大所致，而当该网络部署到多台服务器上时，各服务器分担的负载较小，远远达不到服务器的能力，服务器没有进入输入/输出的非线性区域.

此外，我们还测量了部署在不同服务器之上的各块之间的连通性和通信带宽. 无论 NFV 网络如何部署，测量显示网络节点之间都能够保持连通，并且通信带宽能够达到预定的 100 Mbps.

表 3 不同划分时各服务器 CPU 资源利用率

(a) Part1

发送速率/Mbps	CPU0	CPU1	CPU2	CPU 总
100	9.0	/	/	9.0
200	14.6	/	/	14.6
300	15.9	/	/	15.9
400	20.4	/	/	20.4
500	25.6	/	/	25.6
600	26.0	/	/	26.0
700	28.8	/	/	28.8
800	31.0	/	/	31.0
900	31.1	/	/	31.1
1000	31.1	/	/	31.1

(b) Part2

发送速率/Mbps	CPU0	CPU1	CPU2	CPU 总
100	4.5	3.7	/	8.2
200	7.5	6.5	/	14.0
300	8.0	6.5	/	14.5
400	9.3	8.5	/	17.8
500	11.2	9.5	/	20.7
600	12.5	11.0	/	23.5
700	14.6	12.0	/	26.6
800	15.3	13.7	/	29.0
900	15.5	13.5	/	29.0
1000	15.5	13.5	/	29.0

(c) Part3

发送速率/Mbps	CPU0	CPU1	CPU2	CPU 总
100	3.6	3.4	3.1	10.1
200	5.2	4.5	4.4	14.1
300	6.5	5.4	5.2	17.1
400	7.0	6.9	6.2	20.1
500	8.7	7.4	6.5	22.6
600	9.6	8.5	7.6	25.7
700	10.2	9.7	8.0	27.9
800	9.8	10.2	8.5	28.5
900	9.9	10.4	8.8	29.1
1000	10.5	10.0	8.6	29.1

由于服务器的内存及存储资源通常较为充足，CPU 是最为紧缺的资源，为此我们着重研究了多服务器部署对 CPU 利用率的影响. 表 3 列出了将 NFV 网络部署在 1、2、3 台服务器情况下，3 台服务器的 CPU 使用情况，其中 CPU 显示的是当前 NFV 网络所耗 CPU 值占总 CPU 的百分比. 其中 CPU0、CPU1、CPU2 分别表示第 1、2、3 台服务器的 CPU 资源.

从表 3 的实验测量结果，我们可得如下结论：

(1) 随着发送速率的增加，每种划分中的各服务器的 CPU 资源利用率随之增加.

(2) 当发送速率较小时，各种划分下的各服务器的 CPU 利用率之和大致相同，即多台服务器分担原先由一台服务器承载的计算量.

(3) 当发送速率较高时，随着划分数量的增加，各服务器的 CPU 利用率之和比起只适用一台服务器的情况有所降低. 这说明将较大规模网络部署在多台服务器上，可以减轻每台服务器的 CPU 负载，从而可以承载子网上更多的网络应用.

5.2 任意拓扑网络的划分及其讨论

为了对任意拓扑的 NFV 网络进行划分，我们利用 Brite 拓扑生成器基于 Waxman 随机模型，生成一个具有 100 个节点的网络拓扑文件. 然后在所生成的拓扑文件上通过标注节点和边属性，添加了 15 条千兆带宽、6 跳转发链，这些转发链中每个节点的点权值初始时为(6,33,33).

首先对划分数量进行预估. 为此，根据式(2)计算每种资源所需服务器数量. 结合表 2 中的典型应用资源使用量统计，对 CPU、内存、存储空间三种资源分别计算如下：

$$N=\max\left\lceil\frac{120\times15+1\times62}{2500\times0.6},\frac{200\times15+4\times62}{32\times10^3},\frac{200\times15+4\times62}{4\times10^6}\right\rceil.$$

由于 $N=\max\{2,1,1\}=2$，考虑增加系统可用性，我们先取划分数为 2.

其次对三种资源的权重进行融合. 对所有节点的权值进行融合，将三个权值融合成一个. 根据式(4)有：

$$\alpha:\beta:\gamma=\frac{4\times10^6}{120\times15+1\times62}:\frac{32\times10^3}{200\times15+4\times62}:\frac{2500\times0.6}{200\times15+4\times62}\approx221:1:0.048.$$

从上述计算可知，在三种资源中的 CPU 资源所占的比重远大于其他两种资源，可以考虑直接将 CPU 对应的权值作为节点的权值. 由此，我们获得了一个 Metis 算法可处理的拓扑文件 topo，再执行命令“gpmetis topo 2”，由 Metis 算法将该拓扑划分成 2 个拓扑文件. 接下来，我们根据这 2 个子拓扑进行资源的计算以及约束的验证. 因此，当 NFV 网络所要求 CPU 资源所占比重很大时，可以将 CPU 作为节点的唯一权值，以简化 PNFVNbM 算法. 接着，我们将上述拓扑文件划分为 2 块，并且块中所有节点进行统计，查看每块中的资源使用情况是否超出服务器的承载能力，得到如表 4 所示的数据.

表 4　划分块资源统计表

划分块	CPU	内存	存储
Part0	929	9720	9720
Part1	933	8528	8528

表 4 数据显示，将该拓扑文件划分为两块. 对于划分块 Part0 和 Part1 来说，其资源使用情况均在服务器的承载能力之内，小于资源提供量的 60%. 例如对于 Part0 来说，满足：

$$929<0.6\times2500,$$

$$9720<0.6\times32000,$$

$$9720<0.6\times4000000.$$

上述三个不等式分别表示在划分块 0 中，CPU、内存、存储三种资源的使用量均小于服务器所能提供的对应资源量的 60%. 对于 Part1 来说，也有类似情况. 因而，可以保证各划分块中应用的性能，满足约束条件，且不平衡系数 $=\frac{933}{(929+933)/2}=1.002<1.05$，因此划分成功.

我们又将上述拓扑文件划分为 3 块，并且与划分为 2 块的情况进行了对比，将该数据与划分为 3 个部分时的情况进行了分析，得到如表 5. 其中的不平衡系数是指划分块中最大资源占用量与平均资源占用量的比值，边割边数是指跨划分块的边权值之和，通信量是指跨划分块的边的端节点权值之和.

表 5　不同划分数目下的划分结果统计

划分的块数	不平衡系数	边割边数	通信量
2	1.002	13900	118
3	1.030	15300	183

表 5 说明，将上述拓扑文件划分为 3 块时的不平衡系数、边割边数以及通信量均高于划分为 2 块时的情况，这说明划分块数并非越多越好. 在本示例场景下，将该拓扑文件划分为 2 块，却是能够满足服务器资源约束场景的更为合理的方案.

6　结束语

基于虚拟化技术的 NFV 网络具有可软件定制、效果真实、经济有效等优点，在网络工程、网络管理、网络安全和网络研究等领域有着广阔的研究前景. 当单台服务器的资源无法支持一个较大规模 NFV 网络的运行，就需要将其划分并部署到服务器集群上. 本文利用网络测量研究了服务器中搭建 NFV 网络的主要资源需求量计算方法，探讨了利用 Metis 图划分算法解决问题时关键的计算资源描述问题；提出了一种基于 Metis 多权值约束的 NFV 网络划分算法 PNFVNbM，该算法在划分前对多权值进行融合，在划分后评估划分效果以改进融合参数，以提高各个部分的平衡度. 本文还建立了原型系统，对多个 NFV 网络进行了实验验证. 实验的结果表明了：PNFVNbM 算法能够均衡地将大规模 NFV 网络划分在多台服务器上；当 NFV 网络所要求 CPU 资源所占比重很大时，可以将 CPU 作为节点的唯一权值；需要利用 PNFVNbM 算法得到几种类似的方案，比较分析后得到较佳方案. 在下一步的工作中，我们将基于 PNFVNbM 方法，研究在服务器集群上自动分布式部署大规模 NFV 网络的技术以及在工作负荷动态变化时的处理技术.

参　考　文　献

[1] Mijumbi R, Serrat J, Gorricho J L, et al. Network function virtualization: State-of-the-art and research challenges. IEEE Communications Surveys & Tutorials, 2017, 18(1): 236-262

[2] Sekar V, Egi N, Ratnasamy S, Reiter M K, et al. Design and implementation of a consolidated middlebox architecture//Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation. San Jose, USA, 2012: 323-336

[3] Ahrenholz J, Danilov C, Henderson T R, et al. CORE: A real-time network emulator//Proceedings of the Military Communications Conference. San Diego, USA, 2008: 1-7

[4] Salopek D, Vasic V, Zec M, et al. A network testbed for commercial telecommunications product testing//Proceedings of the 2014 22nd International Conference on Software, Telecommunications and Computer Networks. Split, Croatia, 2014: 372-377

[5] Yan L, McKeown N. Learning networking by reproducing research results. ACM SIGCOMM Computer Communication Review, 2017, 47(2): 19-26

[6] Sefraoui O, Aissaoui M, Eleuldj M. OpenStack: Toward an open-source solution for cloud computing. International Journal of Computer Applications, 2012, 55(3): 38-42

[7] Kernighan B W. Some Graph Partitioning Problems Related to Program Segmentation [Ph. D. dissertation]. Princeton University, Princeton, New Jersey, 1969

[8] Garey M R, Johnson D S, Stockmeyer L. Some simplified NP-complete graph problems. Theoretical Computer Science, 2010, 1(3): 237-267

[9] Hendrickson B, Leland R. A multilevel algorithm for partitioning graphs//Proceedings of the 1995 ACM/IEEE Conference on Supercomputing. San Diego, USA, 1995: 28

[10] Karypis G, Kumar V. A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM Journal on Scientific Computing, 1998, 20(1): 359-392

[11] Leng Ming, Sun Ling-Yu, Yu Song-Nian. Research and implementation of VLSI partitioner. Computer Engineering and Applications, 2010, 46(3): 62-66(in Chinese)
(冷明, 孙凌宇, 郁松年. 一种 VLSI 剖分系统的研究与实现. 计算机工程与应用, 2010, 46(3): 62-66)

[12] Yang H, Mao X, Yan Z, et al. A reduced-order modeling of multi-port RC networks by means of graph partitioning. Chinese Journal of Semiconductors, 2002, 23(10): 1037-1040

[13] Tao Wen-Bing, Jin Hai, Tian Jin-Wen, et al. Real-time segmentation technology of man-made infrared object based on normalized cuts. Journal of Infrared and Millimeter Waves, 2005, 24(2): 114-118(in Chinese)
(陶文兵, 金海, 田金文等. 一种基于图划分的人造红外目标实时分割算法. 红外与毫米波学报, 2005, 24(2): 114-118)

[14] Liu Xu, Mo Ze-Yao. Multilevel MLA algorithm and its application in graph partitioning problem. Journal on Numerical Methods and Computer Applications, 2008, 29(3): 226-240 (in Chinese)
(刘旭, 莫则尧. 多层次图排序算法及其在图剖分中的应用. 数值计算与计算机应用, 2008, 29(3): 226-240)

[15] Yocum K, Eade E, Degesys J, et al. Toward scaling network emulation using topology partitioning//Proceedings of 11th IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer Telecommunications Systems. Orlando, USA, 2003: 242-245

[16] Wang Xiao-Feng, Fang Bin-Xing, Yun Xiao-Chun, et al. An approach for topology partitioning in parallel network simulation. Journal on Communications, 2006, 27(2): 16-21 (in Chinese)
(王晓锋, 方滨兴, 云晓春等. 并行网络模拟中的一种拓扑划分方法. 通信学报, 2006, 27(2): 16-21)

[17] Zhao Yi-Ming. A Density Balance-Aware Partition Methodology Based on Graph Partitioning Program Metis [M. S. dissertation]. Xidian University, Xi'an, 2014(in Chinese)
(赵一明. 基于 Metis 图划分算法的图平衡划分方法[硕士学位论文]. 西安电子科技大学, 西安, 2014)

[18] Kayaaslan E, Pinar A, Catalyurek U, et al. Partitioning hypergraphs in scientific computing applications through vertex separators on graphs. SIAM Journal on Scientific Computing, 2012, 34(2): 970-992

DENG Li, M. S. candidate. His research interests include network function virtualization, software-defined network.

DAI Ning-Yun, M. S., engineer. Her research interests include Network Function Virtualization, Distributed computing.

XU Bo, Ph. D., associate professor. His research interests include network measurement and network applications.

XING Chang-You, Ph. D., associate professor. His research interests include network measurement and future networks.

CHEN Ming, Ph. D., professor, Ph. D. supervisor. His research interests include network measurement, network performance analysis and modeling, future networks.

Background

Before new network equipment and protocols being entered a practical stage, the most critical process should be to test and validate them in the real network environment thoroughly. However, it may be very difficult to test in such way for the important infrastructure such as the space-air-ground network. To settle the problem, this paper proposed the concept of test fidelity, the architecture of network functions virtualization based network testing platform (NFVNTP) and its component model, and a method of designing NFVNTP was given.

This work is supported by the National Natural Science Foundation of China (Nos. 61772271 and 61379149) and the National Key Basic Research and Development Program of China (No. 2012CB315806), which research measurement technology of future networks. During the period, we found there exist many shortcomings for current simulation methods; especially they are hard to guarantee the fidelity of simulating behaviors and performance of large scale of or complex networks.