基于虚拟化的网络创新实验环境研究

周 烨,李 勇,苏 厉,金德鹏,曾烈光
(清华大学电子工程系,北京 100084)

摘 要: 研究者针对未来网络创新研究而提出的各种创新性解决方案,需要在大规模真实网络实验环境中测试、验证.现有互联网无法支持基于后 IP 新型体系结构的创新实验,因此需要构建全面支持未来网络创新研究的实验环境.清华大学 TUNIE 平台,基于虚拟化技术,是拥有自主知识产权的未来网络创新实验环境.本文介绍网络创新实验环境的相关研究、关键技术,以及 TUNIE 设计目标、关键技术实现方案、平台部署情况等.目前,TUNIE 平台已实现初步部署,应用于教学和科研中,并且承载一些网络创新实验.

关键词: 网络虚拟化; 网络创新实验环境; 下一代网络; 清华大学网络创新环境
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2012)11-2152-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.11.002

Research of Network Innovation Experimental Environment Based on Network Virtualization

ZHOU Ye, LI Yong, SU Li, JIN De-peng, ZENG Lie-guang
(*Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

Abstract: In order to solve the problems related to future network innovation, researchers propose numerous solutions. However, these solutions need to be tested in an large-scale network experimental environment, which cannot be totally supported by the current Internet. Thus, constructing a network innovation experimental environment is meaningful for the further develop of the Internet. We propose Tsinghua University Network Innovation Environment (TUNIE), a network innovation experimental environment in China, aiming to support research in future network based on network virtualization. In this paper, we survey the past and the state-of-the-art of network innovation experimental environment, as well as certain key topics. Then we introduce design goals and key technologies of TUNIE. TUNIE is already deployed in several clusters, applied in teaching and research, and supports some network innovation experiments.

Key words: network virtualization; network innovation experimental environment; next generation network; tsinghua university network innovation environment(TUNIE)

1 引言

互联网在快速发展的同时,也暴露出许多问题,这意味着未来网络创新研究势在必行.研究者提出的各种创新性解决方案,需要大规模真实网络实验环境来进行测试.现有互联网仅能对基于 IP 网的创新实验进行验证,却无法支持基于后 IP 新型体系结构的创新实验.因此,构建全面支持未来网络创新研究的实验环境值得深入研究.

网络虚拟化正逐步得到学术界的高度重视:从长期来看,网络虚拟化可能演进成为支持未来网络发展的基石;从短期来看,网络虚拟化能够用于构建网络创新实验环境,来测试、验证各种创新性解决方案[1].基于网络虚拟化的网络创新实验环境研究,在美国、欧洲、日本的国家级科研项目中占据重要地位,比如 GENI[2]、FIRE[3]、AKARI[4].

尽管 GENI 等项目已取得一定进展,但是基于虚拟化的网络创新实验环境研究尚有很多关键问题亟待解决.另外,依托国外实验环境进行创新研究,将难以保证创新技术的独立性与自主性.因此,我国自主建设支持未来网络研究的网络创新实验环境,是国家战略发展的必然选择.清华大学在国内率先展开相关研究,搭建清

收稿日期:2011-12-13;修回日期:2012-02-17
基金项目:宽带移动通信重大专项(No.2010ZX03004-002-02);国家自然科学基金(No.61171065)

华大学网络创新环境(Tsinghua University Network Innovation Environment,简称 TUNIE);目前,TUNIE[5]已实现初步部署,并开始运行一些网络创新实验.

2　相关研究

2.1　PlanetLab

PlanetLab[6]基于层叠网技术,是一个分布式、大规模的测试平台,目前涵盖 1074 个网络节点.PlanetLab 将网络节点上的物理主机虚拟化成多个虚拟主机,然后将独立的虚拟机提供给各个并行实验,而且并行实验的虚拟机之间互相隔离.PlanetLab 中只实现主机节点虚拟化,链路和核心网络设备都没有虚拟化,无法形成完整的网络虚拟化环境.

2.2　VINI

VINI[7]允许研究者在真实的环境中测试、实验各种新技术和新服务.与 PlanetLab 相比,VINI 在路由层次提供更多自由度,将虚拟化的概念延伸到网络核心设备.VINI 通过虚拟路由器,将各个虚拟主机相连接,形成虚拟网络,为实验用户提供独立实验环境,并允许各虚拟网络自定义拓扑、路由协议和转发策略等.

2.3　Emulab

作为网络实验仿真环境,Emulab[8]被研究者广泛用于网络领域和分布式系统的实验研究.Emulab 向实验者提供由可定义操作系统镜像的主机组成的网络,实验者可以按需对操作系统进行修改.在建设初期,Emulab 没有采用虚拟化技术,近年来开始逐步研究物理资源如何高效虚拟化、虚拟资源如何高效分配等问题.

2.4　GENI、FIRE、AKARI

GENI 旨在基于网络虚拟化技术,建立一个大规模、真实的全球范围网络实验环境,以虚拟网络方式支持多个面向未来网络创新研究的并行实验.除了主机、链路的虚拟化,GENI 还关注核心网络设备的虚拟化:OpenFlow 关注让交换机支持网络虚拟化,ShadowNet 则关注虚拟路由器的研发.目前,GENI 已完成全美范围基于 OpenFlow 的骨干网建设.但是,GENI 在网络设备可编程可配置、实验环境的有效测量管理控制等方面尚未有成熟方案.

与 GENI 类似,欧盟 FIRE 和日本 AKARI 均致力于利用网络虚拟化技术搭建面向未来网络研究的实验环境.这两个项目与 GENI 都有深度合作,在部署进度、实施效果等方面尚不及 GENI 成熟.

2.5　相关研究总结

PlanetLab 只提供虚拟主机,VINI 能提供虚拟网络.然而,两者采用层叠网技术,只能支持应用层创新,无法支持链路层、网络层创新.Emulab 作为一个网络实验仿真环境,建立之初未考虑虚拟化技术,导致整个实验环境的资源利用效率低下,近年来开始利用虚拟化技术提升资源利用效率.

GENI 等项目设计思路基本一致,关注现有的各种类似实验环境的互联、融合,比如 PlanetLab、Emulab 等,基于这些异构实验环境,并利用虚拟化技术,以支持未来网络实验创新.目前,GENI 等项目的重点仍然是如何让异构实验环境有效互联、融合;同时,以 OpenFlow 技术为核心进行虚拟化可编程设备的研发,正在以 OpenFlow 骨干网为基础进行全美范围部署.然而,GENI 等无法对基于后 IP 网络的创新实验进行支持,因为 OpenFlow 技术不支持 IP 包格式的改变.

综上所述,国际上较成熟的网络创新实验环境尚未出现,很多问题都亟待解决;网络虚拟化技术已是其中的重要概念,实验环境的可管理性、资源利用效率等问题都是重要研究内容,如何更好支持基于后 IP 网络的创新实验更是需要考虑的新问题.TUNIE 平台即是在总结国外先进经验、分析现存问题的基础上,旨在建立基于虚拟化技术的网络创新实验环境,支持广泛范围的创新实验,尤其是面向后 IP 网络的创新实验.

3　关键技术

3.1　物理资源虚拟化

物理资源只有在虚拟化之后,才能提供虚拟资源,因此物理资源虚拟化是基础研究内容.具体地,分为物理节点虚拟化和物理链路虚拟化部分.

在虚拟节点研究中,虚拟路由器研究尤其重要,其主要研究内容包括:可编程性,深度可编程虚拟路由设备包括软件虚拟路由和硬件虚拟路由,比如 Click、VERA 等;性能表现,这对网络虚拟化环境的整体性能表现有重要影响;新的路由、存储、转发机制,比如通过新型存储结构来高效存储多个虚拟路由器中的路由表.

虚拟链路的研究,主要有如下内容:物理链路支持虚拟链路的机制,比如一条虚拟链路对应一条物理链路或者多条物理链路;隔离机制,实现不同实验网数据流量的互相隔离;性能指标,这直接影响整个实验环境的性能表现.

3.2　实验网资源映射

建立实验网,需要将包括虚拟节点、虚拟链路在内的实验网络拓扑在物理资源中顺利映射,这就需要研究实验网资源映射问题.具体研究内容如下:

(1)特定约束条件的映射算法:比如拓扑结构固定、静态虚拟网络需求、无限资源条件等,Ricci 等提出基于节点资源无限的虚拟链路映射算法[9].

(2)无特定约束条件的映射算法:不对映射问题做任何特定约束,进而提出启发式算法,程祥等建立了映射问题的整数线性规划模型,提出基于粒子群优化的

映射算法[10].

(3)容错映射算法:物理网络可能因为恶意攻击、突发故障等因素失效,如何让实验环境具备一定的容灾容错能力值得研究,Guo 等提出基于预留未分配资源和已分配资源共享的容错映射算法[11].

3.3 实验网资源调度

在运营过程中,实验需求将动态变化,比如需要新增实验节点;在多个实验网之间,根据动态需求来高效分配资源,即是资源调度问题.具体地,可以从如下角度入手:资源需求方和提供方的交互关系、资源评价策略等,陶军等运用博弈理论对资源分配进行研究,提出基于竞价的资源调度机制[12].

3.4 实验环境的测量管理控制

如何实现实验环境的有效测量、管理、控制,值得研究.在实验环境中,需要对实验中的各种数据和参数进行测量,并反馈给研究者,因此需要研究高速测量算法和反馈机制.在控制管理体系中,采用面向管理、控制网络并存的思想,构建多个独立的并行虚拟网络,分别实现控制、管理功能;研究相关控制管理算法,构建整个控制管理体系,是具体研究内容.

4 平台设计与实现

4.1 设计目标与结构

TUNIE 平台的主要设计目标如下:

(1)支持多个互相隔离的并行实验:同时支持多个实验,各实验间实现物理资源、实验数据、测量控制管理功能的良好隔离.

(2)支持实验可编程可配置:需要物理设备可编程可配置,尤其是核心网络设备.基于可编程可配置的物理设备,才能方便地运行网络架构、协议各异的实验,尤其支持基于后 IP 新型体系结构的网络创新实验.

(3)实现资源高效利用:物理资源的稀缺性,以及各实验虚拟资源需求的动态变化,使得资源高效利用成为难题,需要研究符合实验环境特点的虚拟资源映射、分配等算法.

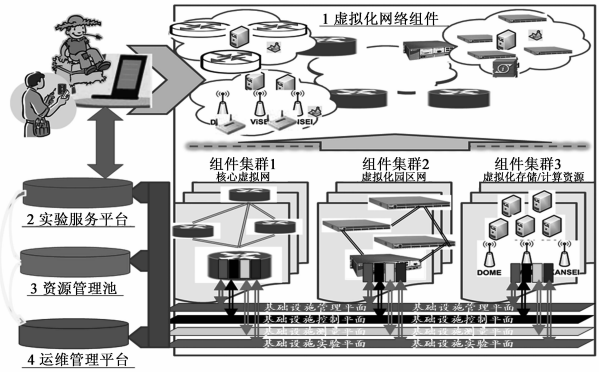

图1 TUNIE平台结构

(4)实现平台有效管控:以有效的开放机制,通过标准统一的形式,对虚拟化可编程设备进行有效的监测、管理和控制,进而提供给实验用户一个真实的网络环境.

图 1 是 TUNIE 平台结构示意图:物理资源包括硬件虚拟化、软件虚拟化的可编程设备,支持 OpenFlow 的硬件设备;实验用户通过统一接口,从实验环境中获取物理资源、构建虚拟实验网、进行创新实验;多个独立逻辑平面,分别进行测量、控制与管理功能;实验环境中有高效的虚拟资源映射、分配算法,以高效利用资源.

4.2 关键技术的实现方案

4.2.1 节点虚拟化

TUNIE 采取软硬件协同的节点虚拟化方案,综合软件虚拟路由器和硬件虚拟路由器的相关特点,提出基于可编程硬件的高性能虚拟路由器结构(如图 2 所示).数据平面由 FPGA 负责数据包转发,能够兼顾线速转发速率和高度可编程性.在 FPGA 中使用 VLAN 技术来隔离不同数据平面,在使用 VLAN 的同时保留了对一般以太网帧的处理.控制平面采用操作系统虚拟化技术,控制平面软件主要包括主机控制程序、虚拟机,主机软件负责对数据平面和虚拟机的配置和管理,虚拟机通过路由协议管理软件和数据包处理软件来实现路由信息和数据包的处理.

4.2.2 链路虚拟化

TUNIE 通过设置 VLAN 标签的方式,来进行链路隔离.在 TUNIE 中,一条虚拟链路只能被映射到一条物理链路中;一条物理链路可以支撑多条虚拟链路.为了实

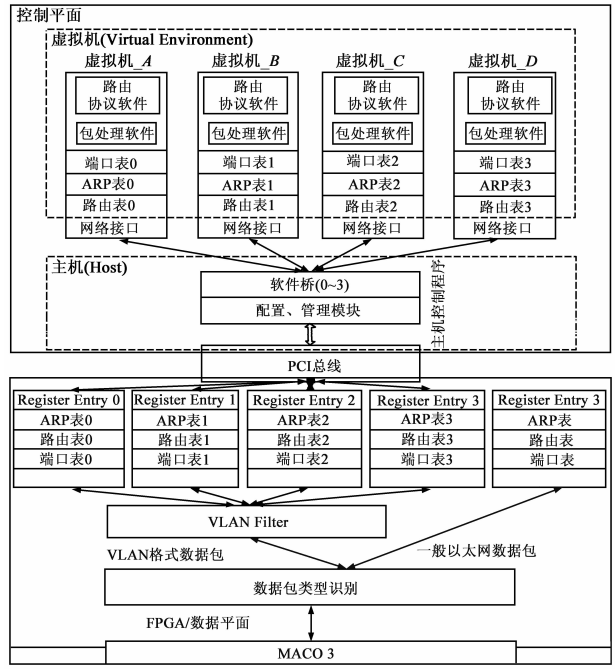

图2 基于可编程硬件的高性能虚拟路由器设计

现不同虚拟链路之间流量的隔离,为每个虚拟实验网中任两虚拟节点之间的链路连接,都分配 VLAN 标签.这样,同一条物理链路的多个虚拟链路,由于 VLAN 标签不同,各个虚拟链路的数据流量能够互相隔离.

4.2.3　虚拟资源映射

TUNIE 提出了一种基于混合蛙跳的映射算法,具有计算速度快、全局寻优能力强、易于实现等特点.在进行具体映射时,选择剩余资源相对较多的节点和链路,这样将有效提高物理资源使用效率,同时均衡实验环境中的网络负荷.该算法基于 JAVA 开发并集成到 TUNIE 平台中,主要流程如图 3 所示:

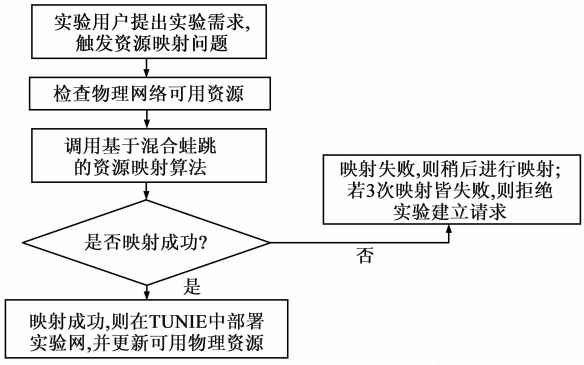

图3　基于混合蛙跳的资源映射算法

4.2.4　虚拟资源分配

TUNIE 提出基于虚拟资源动态变化的资源分配算法(如图 4 所示),基于已有虚拟资源映射结果,根据更新的虚拟资源需求,从单个虚拟实验网的角度出发,以最小化资源分配过程中的资源消耗代价为优化目标.该算法只针对原有映射结果的部分虚拟节点、虚拟链路进行再映射,可以有效降低计算复杂度、时间复杂度.

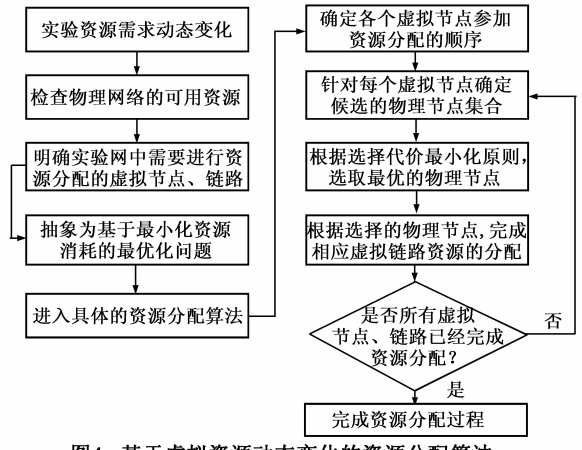

图4　基于虚拟资源动态变化的资源分配算法

4.2.5　测控管系统

管理测量系统是 TUNIE 平台的管理核心,具体分为管理系统和测量系统两部分.

管理系统主要用于实现虚拟实验网络的初始化配置、对虚拟机的控制和再配置等.实验网初始化配置是核心功能,包括网络拓扑设计、资源参数配置、路由协议配置等.管理系统为用户提供丰富的 API 与底层虚拟实验网络完成交互,这些 API 通过 Web 页面在线提交表单的方式提供给用户;管理系统会将用户在 Web 页面提交的请求发送至对应节点,对应节点接收后执行相应操作.

测量系统基于开源软件 Cacti[13] 实现,搜集实验平台的运行状态和实时指标.测量系统不直接对用户开放,通过管理系统的解析将监测状态、监测指标在 Web 页面上呈现给用户.

5　平台部署与展示

5.1　平台部署

TUNIE 平台已经在清华、联通初步部署,已有 4 个资源簇(如图 5 所示),提供 96 个线速虚拟路由器、超过 1000 个普通虚拟节点,能够支持 40 个一定规模的并行实验.TUNIE 通过 3 个 OpenFlow 高性能交换机和 1 个自主研发的高性能可重构路由器,形成骨干网;与 GENI 完全由 OpenFlow 交换机作为骨干网的方案相比,TUNIE 方案有如下特点:数据平面有可重构特性,可以支持非 IP 实验;具有良好的隔离特性,各个并行虚拟资源之间互不干扰.除了骨干网设备,TUNIE 物理资源还包括可编程集群,WiFi 虚拟化节点,传感器节点,OpenFlow 设备等.

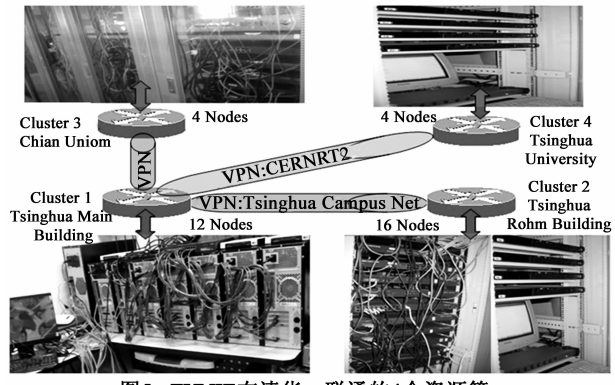

图5　TUNIE在清华、联通的4个资源簇

5.2　平台展示

5.2.1　设备研发

TUNIE 正在自主研发虚拟路由设备原型,以满足实验环境的具体需求,如表 1 所示.另外,TUNIE 正在研发 160G 的高性能可重构路由器,已完成 8G 板卡研发,能够支持 20 块板卡的背板结构.

TUNIE 对自主研发的高性能可重构路由器的虚拟化性能进行测试,比较单一数据平面和 2 个并行数据平面时的数据包转发速率测试,测试结果显示,当存在 2

个并行数据转发平面时,每个平面的峰值转发速率都与单一数据转发平面的速率相当,能够实现线速(1Gbps).如图 6 所示,在包长度不同时,利用 2 个并行数据平面转发,转发能力总是单一数据平面时的 2 倍.

表 1 TUNIE 自主研发的虚拟路由设备

设备名称	功能定位	端口数/端口速率	支持虚拟节点数
EdgeDev901A	接入边缘设备	16×1G	16 个线速
EdgeDev901B	接入边缘设备	40×1G	40 个线速
CoreDev901C	核心网路由器	4×10G	40 个线速 64 个普通

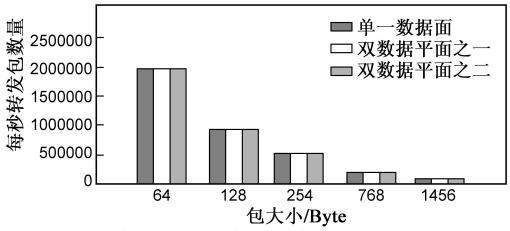

图6 高性能可重构虚拟路由器的测试曲线

5.2.2 教学和科研应用

TUNIE 已经在清华大学教学中具体应用,在课程《计算机网络实践》中开设基于可编程硬件的路由器设计环节.TUNIE 正在和国外展开深度合作:与美国西北大学、克莱姆森、韦恩州立等高校进行资源互联、实验共享,并开始考虑和 GENI 计划的进一步合作事宜.在科研中,TUNIE 已经能够用于网络新协议、算法、架构的验证.目前,已经部署网络层析算法、SDN 网络架构、基于虚拟化的三网融合架构等创新实验.下面,具体介绍三网融合实验,其架构如图 7 所示.

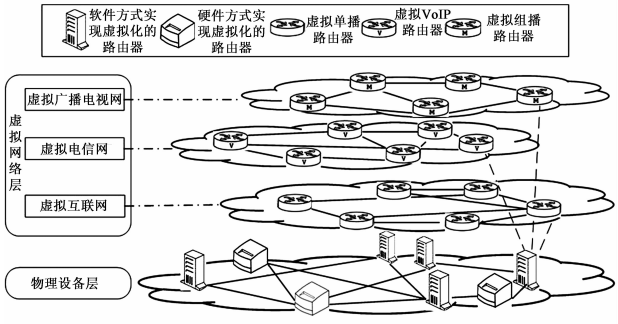

图7 基于虚拟化的三网融合架构示意图

由于电信网、广播电视网、互联网各自业务有不同的业务特性及业务质量需求,在现有互联网中很难做到有效融合.在 TUNIE 中,分别搭建虚拟互联网、虚拟电信网、虚拟广播电视网,各自运行相应路由协议并实现相应路由功能:在虚拟互联网中运行 RIP 路由协议,实现视频单播服务;在虚拟电信网中运行 OSPF 路由协议,实现 VoIP 服务;在虚拟广播电视网中运行 PIM-SM 组播协议,实现视频广播服务.

5.2.3 平台创新小结

鉴于 GENI 项目是目前国际上影响力最大、最成熟的网络创新实验环境,在此将 TUNIE 与 GENI 进行对比:

首先,GENI 将多个已有的异构实验环境进行互联、融合,这些异构实验环境在虚拟化程度、可编程深度、性能表现等方面各不相同,将制约 GENI 的整体性能.与 GENI 不同,TUNIE 设计虚拟化可编程整体方案,全网支持虚拟化、硬件设备深度可编程,可以通过物理资源高性能,来保证整个实验环境的高性能.

其次,GENI 目前正在推广、部署以 OpenFlow 骨干网为基础的大规模实验环境,虽然 OpenFlow 技术有很多优点,但是它无法支持 IP 包格式的改变.这就意味着 GENI 无法支持基于后 IP 网络架构的各种创新实验.随着未来网络研究的逐步深入,基于后 IP 网络架构的创新实验将越来越多,这将使得 GENI 的局限性日益凸显.TUNIE 通过可重构路由器的研发,支持比特微度可编程,可以有效支持基于后 IP 网络架构的广泛创新,这是相对于 GENI 的第二个创新之处.

6 结论与展望

利用网络虚拟化技术来搭建网络创新实验环境,已经被认为是有效途径;GENI 等计划正在逐渐开展研究,但仍有很多关键技术亟待解决,成型的实验环境也尚未出现.TUNIE 平台,旨在提供基于虚拟化的网络创新实验环境,为下一代网络研究尽绵薄之力.目前,TUNIE 平台已有小规模部署,并在重要关键技术领域提出若干解决方案.在未来工作中,TUNIE 一方面要继续研究关键技术,一方面要加强国内外合作,扩大影响力,集思广益共建平台.

参考文献

[1] Anderson T, Peterson L, Shenker S, et al. Overcoming the Internet impasse through virtualization [J]. IEEE Computer, 2005, 38(4): 34 – 41.

[2] GENI: Global environment for network innovations [DB/OL]. http://www.geni.net/, 2011-10-15.

[3] FIRE: Future internet research and experimentation [DB/OL]. http://cordis.europa.eu/fp7/ict/fire/, 2011-10-15.

[4] AKARI architecture design project [DB/OL]. http://akari-project.nict.go.jp/eng/index2.htm, 2011-10-15.

[5] TUNIE: Enabling Network Innovation [DB/OL]. http://fi.ee.tsinghua.edu.cn:8080/myweb/, 2011-10-15.

[6] PlanetLab: An open platform for developing, deploying, and accessing planetary-scale services [DB/OL]. http://www.planet-lab.org/, 2011-10-15.

[7] Bavier A, Feamster N, Huang M, et al. In VINI veritas: Realis-

tic and controlled network experimentation [A]. Proc of SIGCOMM'06[C]. New York: ACM, 2006. 3 – 14.

[8] Emulab-network emulation testbed home [DB/OL]. http://www.emulab.net/, 2011-10-15.

[9] Ricci R, Alfeld C, Lepreau J. A solver for the network testbed mapping problem [J]. ACM SIGCOMM Computer Communication Review, 2003, 33(2): 65 – 81.

[10] 程祥,张忠宝,苏森,等.基于粒子群优化的虚拟网络映射算法[J].电子学报,2011,39(10):2240 – 2244.
Cheng Xiang, Zhang Zhong-bao, Su Sen, et al. Virtual network embedding based on particle swarm optimization [J]. Acta Electronica Sinica, 2011, 39(10): 2240 – 2244. (in Chinese)

[11] Guo Tao, Wang Ning, Moessner K, et al. Shared Backup Network Provision for Virtual Network Embedding [A]. Proc of IEEE ICC'11 [C]. Kyoto: IEEE, 2011. 1 – 5.

[12] 陶军,吴清亮,吴强.基于非合作竞价博弈的网络资源分配算法的应用研究[J].电子学报,2006,34(2):241 – 246.
Tao Jun, Wu Qing-liang, Wu Qiang. Application research of network resource allocation algorithm based on non-cooperative bidding game [J]. Acta Electronica Sinica, 2006, 34(2): 241 – 246. (in Chinese)

[13] Cacti: The complete rrdtool-based graphing solution. [DB/OL]. http://www.cacti.net/screenshots.php, 2011-10-15.

作者简介

周　烨　男,1986 年出生于江西南昌,博士生.2008 年于清华大学获得工学学士学位,2008 年至今在清华大学电子工程系攻读博士学位.主要研究领域为未来网络、下一代 IP 网络体系结构、网络虚拟化等.
E-mail: zhouye04@mails.thu.edu.cn

李　勇　男,1985 年出生于湖南长沙,博士生.2007 年于华中科技大学获得工学学士学位,2007 年至今在清华大学电子工程系攻读博士学位.主要研究领域为未来网络、下一代 IP 网络体系结构、移动管理、移动容迟网络、网络虚拟化等.