基于 LXC 的大规模 IP 网络仿真和配置方法

李大伟

(中国电子科技集团公司第二十八研究所信息系统工程重点实验室,江苏 南京 210007)

[**摘要**] 大规模 IP 网络仿真技术在对大型分布式信息系统的研发和测试方面具有重要意义. 在分析基于 LXC 轻量级虚拟化技术优势的基础上,讨论了基于 LXC 的大规模 IP 网络仿真方法,提出了仿真网络的配置和优化框架及方法,并对关键网络指标进行了试验验证. 结果表明,使用 LXC 技术构建并合理配置后的大规模 IP 仿真网络可满足大型网络化信息系统的研发和测试需求.

[**关键词**] LXC 虚拟化技术,大规模 IP 网络仿真,无分类域间路由选择,网络性能
[**中图分类号**]TP393　[**文献标志码**]A　[**文章编号**]1672-1292(2014)04-0071-06

LXC-Based Large Scale IP Network Emulation and Configuration

Li Dawei

(Laboratory of Science and Technology on Information Systems Engineering, The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing 210007, China)

Abstract: Large scale IP network emulation technique has important significance to the development and testing of large scale distributed information systems. This paper analyzes the advantages of a lightweight virtualization architecture named LXC, discusses the methods of LXC based large scale IP network emulation, proposes the strategy of configuration and optimization of emulated large scale IP networks, and examines the performance index of the network by simulation experiments. The results show that the large scale IP network generated by proposed methods can satisfy the requirements of R&D and testing of network based information systems.

Key words: LXC virtualization technique, large-scale IP network emulation, CIDR, network performance

随着计算机网络和云计算技术的发展,信息系统,特别是分布式信息系统,对大规模 IP 网络的要求和依赖程度日益增高,为系统研发和测试带来了一系列挑战[1]. 一方面,在用系统节点分布广,业务关联度高,系统上线后无法开展新的研发和测试工作;另一方面,仿真环境下的系统测试又面临逼真度和可信度偏低的问题. 因此,在实验室环境下快速构建大规模 IP 网络仿真和试验平台,为网络化信息系统研发、测试提供基础网络运行环境,具有十分重要的意义[2-4].

大规模 IP 仿真网络为网络化信息系统提供基础网络化运行环境,如何高逼真度构建大规模 IP 仿真网络已成为国内外研究机构研究的热点问题. 目前该领域比较成熟的研究成果主要分为节点模拟、覆盖网络和虚拟化构建 3 种类型,分别以 Emulab[5]、Planetlab[6]和 CORE 为代表.

基于节点的模拟技术通过动态配置物理上相互联通的物理节点实现网络环境的构建. 美国 Utah 大学的 Emulab 项目、台湾成功大学的 Testbed@ TWISC、美国国防部的信息防御技术试验研究网络(DETER)都属于基于节点模拟技术构建的大规模网络试验环境. 基于覆盖网络的大规模 IP 网络仿真以 Planetlab 为典型代表,该技术通过在底层物理网络之上构建逻辑层实现物理上分布异构、逻辑上相邻的网络环境. 基于虚拟化技术的构建方法可为用户提供独立、隔离的虚拟计算环境,为管理人员提供硬件资源、软件资源的集中管理功能,逐渐成为大规模 IP 网络仿真的主流技术.

大规模 IP 网络由大量节点和链路组成,基于 LXC 的轻量级虚拟化技术可在主流服务器上实现大量虚拟网络节点,通过多台服务器组建集群可实现具有数千个节点的大规模 IP 网络,满足了高密度、大规模

收稿日期:2014-06-20.
基金项目:总装预研基金(9140A040413DZ380001).
通讯联系人:李大伟,博士,工程师,研究方向:云计算和虚拟化技术. E-mail:lidw1981@163.com

部署的需求. 例如,CORE 模拟器可在主流服务器上模拟数百个节点,每秒发送和接收数据包的总和超过 300 000 个[7]. 另一方面,虚拟化技术实现的节点可方便地接入真实流量、安装协议,具有理想的逼真度.

网络规模的扩大使每个虚拟节点占有的资源减少,难以应对大规模网络寻址产生的开销,致使模拟网络连通性、稳定性达不到可用要求. 而大规模 IP 网络中运行动态路由协议时,会遇到路由表庞大、路由器之间路由信息交换频繁、路由节点负担过重的问题,造成路由收敛过程漫长、地址冲突的概率增加、部分网络节点不可达的情况,严重影响网络运行的效果. 此外,相对于小规模网络,大规模 IP 网络接入的用户数和传输的数据量也相应增加,更增加了服务器 I/O 的负担,造成系统稳定性下降. 因此必须对仿真网络进行合理的配置和优化,达到最优运行状态.

本文针对大规模 IP 网络仿真的需求和面临的问题,探讨了基于 LXC 的网络仿真的基本原理和方法. 在此基础上,基于大规模组网特征,研究了大规模 IP 网络仿真中网络优化配置方法,并通过仿真试验验证了网络仿真的可行性、可用性及网络链路的逼真度.

1　基于 LXC 的网络仿真

Linux 容器(LXC)技术属于操作系统虚拟化,是一种轻量级的内核虚拟化技术,可在服务器上实现高密度虚拟化部署. 其通过将操作系统资源进行隔离和划分,在单一控制主机上同时提供多个貌似独立实则共享的虚拟运行环境(VE)或容器,每个虚拟环境拥有自己的进程和独立的名称空间,可有效地将由单个操作系统管理的资源划分到隔离的组中,以更好地在隔离的组之间平衡有冲突的资源占用需求. 从用户的角度上看,容器的运行等同于一台 Linux 服务器.

由于大规模 IP 网络节点可看作具有多端口的专用服务器,其基本功能是对网络数据包的处理和存储转发. 目前大多数网络设备都基于 Linux 内核构建,因此可通过在 LXC 容器中配置多网络接口和软路由的方式实现大规模 IP 网络仿真.

LXC 技术的实现包括名称空间和资源控制两方面. 名称空间(namespace)为容器进程提供独立的进程号、用户标识和组标识,建立独立的协议栈、根目录挂载点等,并保存到单独的配置文件中. 资源控制系统(CGroup)基于 Linux 内核为容器分配共享的资源,保证资源的隔离和对资源的控制. LXC 为大规模 IP 网络仿真提供一个用于使用和管理 LXC 容器的用户空间工具集,网络仿真用户可以使用这些工具按需构建仿真网络节点.

基于 LXC 的网络仿真原理是将网络节点运行所需的各类资源,如用户、文件系统、网络等,分配给不同名称空间,并基于共享内核的原理统一部署 Quagga、XORP 等软路由系统,通过桥接的方式实现不同 VE 之间的网络通信,从而实现网络节点的构建.

对于网络节点,主要采用文件挂载模块(MNT)、进程管理模块(PID)、网络连接模块(NET)分别实现节点文件系统、进程和网络的虚拟化构建. 其中,MNT 模块首先在宿主机文件系统中创建主目录,然后在主目录中为每个节点分别建立相应根目录,而对于共享目录,则通过映射的方式挂载到节点的根文件系统中;PID 采用层次结构对系统进程进行组织,共享的进程在不同名称空间内有不同的数据表示,需要进程号和名称空间信息进行标识;NET 模块将系统路由表和网络设备关联到相应名称空间中,采用桥接的方式与其他节点通信. 资源控制系统(CGroup)提供对进程所使用的系统资源,如 CPU、内存、I/O 进行限制、记录和隔离. 基于 LXC 和软路由系统,可将每个名称空间虚拟为独立的虚拟网络节点,实现大规模节点的构建.

对于网络链路,LXC 网络配置管理工具创建和维护了一系列虚拟网络设备,包括 veth、macvlan、vlan、phys 和 empty,其功能如表 1 所示. 通过这些设备结合 Linux 桥接和绑定机制,可实现虚拟节点之间、虚拟节点与物理接口之间的数据包转发.

基于 LXC 的大规模 IP 网络仿真系统结构如图 1 所示. 由于单台主机所支持的网络节点数量有限,大规模网络仿真需要采用多服务器分布式仿真实现. 每个服务器上部署 LXC 网络仿真系统,通过高速互联的真实网络设备按需接入到真实网络中. 真实网络提供了大量仿真网络和真实网络或设备接入的端口,可与真实系统互联或接入物理服务器实现管理控制、状态监控、数据采集及其他服务功能. 网络仿真服务器中运行的仿真网络可看作大规模 IP 仿真网络的子网或自治域进行配置和管理.

表 1　LXC 中虚拟网络设备

Table 1　Virtual network devices of LXC

名称	功能描述
veth	节点的虚拟网卡设备,可调用主机网卡资源为桥接提供端口
macvlan	通过 MAC 地址区分的 vlan,可实现多网卡绑定,macvlan 根据 MAC 地址判断转发的虚拟网卡端口
vlan	通过 IP 地址区分的 vlan
phys	将宿主机物理接口映射到虚拟节点中
empty	loopback 接口

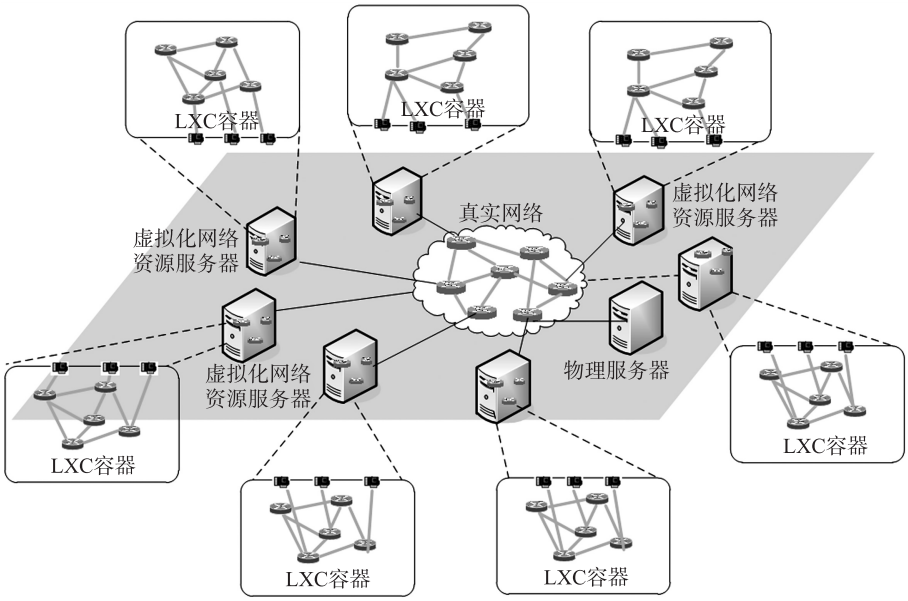

图 1　基于 LXC 的大规模 IP 网络仿真系统结构

Fig. 1　System structure of large-scale IP network emulation based on LXC

2　网络配置

随着仿真网络规模的扩大,需要对网络进行配置和优化以提升网络的整体性能. 目前已有的研究成果往往忽略了仿真网络的配置,虽然在服务器上构建了足够多的节点和链路,但网络性能远远低于实际网络.

大规模 IP 仿真网络中网络节点数量巨大,运行路由协议时单个节点所维护的路由表中表项数量随网络节点的增加而增加,节点存储的路由信息和信息交换量急剧增加,节点在路由表检索、节点间路由信息交换时消耗的资源大幅度增加,对资源相对较少的虚拟节点的性能带来严重影响,进而使仿真网络的性能大幅度降低. 另一方面,相比于一般网络仿真,大规模的网络仿真中服务器为每个节点分配的资源相应减少,使节点的网络性能进一步恶化. 因此合理的网络配置和路由寻址优化对提高大规模 IP 网络仿真的效率和改善其性能至关重要.

基于 LXC 技术实现的模拟网络将所有可配置的参数存放到一个配置文件中,网络初始化时根据配置文件的内容确立网络拓扑、链路、带宽,以及网络标识、路由协议等. 大规模 IP 网络配置框架如图 2 所示. 当模拟网络构建完成后,使用构建平台提供的配置工具对网络进行配置和优化. 配置过程包括硬件规划、IP 地址规划、路由协议配置、网络特征和链路特征设置等步骤. 配置完成后,首先通过物理设备测试工具验证物理设备之间的连通性和可用性,保证模拟网络的底层硬件环境的正常和高效运行;其次使用冲突检测工具对 IP 地址、MAC 地址、ID 标识等需要满足唯一性的参数进行验证,避免因地址冲突导致的大面积网络不通的现象发生;第三,对配置的路由协议进行测试和验证,保证协议的正常运行和节点路由信息的完整性,例如,通过读取路由表并与配置进行比对,检查路由信息是否覆盖所有的节点;最后,通过网络测试工具对网络特征,如链路带宽、时延、抖动等进行逼真度测试,确定各项参数符合用户需求. 验证完成后的模拟网络可以提供给网络用户使用.

硬件规划根据服务器负载和网络流量分布情况,通过硬件映射和节点迁移的方式对服务器资源进行

分配和优化,保证硬件资源的高效利用. 例如,对网络节点按照负载情况进行加权分类,合理映射到不同性能的服务器上,最大程度提高服务器计算效率,进而从总体上提高模拟网络的性能.

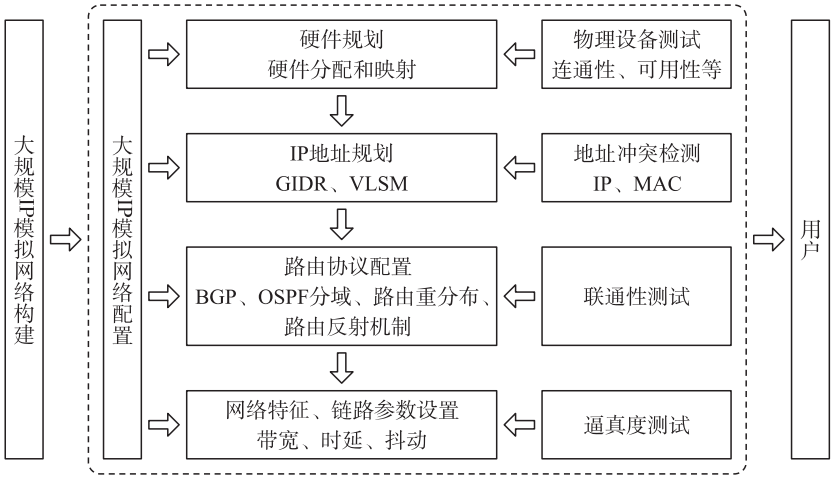

图 2　大规模 IP 网络配置框架

Fig. 2　Configuration frame of large-scale IP networks

IP 地址规划要满足全网地址的唯一性和规律性. 唯一性保证全网地址不冲突;规律性根据统一的规则进行 IP 地址规划,方便路由聚合、地址分配和故障定位.

在 IP 地址规划时采用变长子网掩码(VLSM)和无分类域间路由选择(CIDR)技术,将一个连续的地址形成地址块,每块的地址压缩为一个路由表项,可提高 IP 地址利用率并大幅度减少每个节点路由表的长度. CIDR 将 IP 地址前缀表示网络,后面部分表示主机,记为:IP 地址::={<网络前缀>,<主机号>},把前缀相同的连续 IP 地址组成一个地址块,每个地址块的大小是 2 的幂. 已知 CIDR 地址块中的任何一个地址,就可以知道这个地址块的起始地址和最大地址,以及地址块中的地址数,可以根据大规模 IP 网络的管理需要进行地址分配和定位. CIDR 技术的细节可参考 RFC1517、RFC1518、RFC1519、RFC1520 等标准[8-11].

在大规模 IP 网络仿真时,根据每个节点所在的拓扑位置和 area 按照 CIDR 规则进行编址,并将结果存储到目标网络配置文件中,网络仿真器启动时根据配置文件内容初始化网络,为每个节点的接口分配相应的 IP 地址和 area. 同时,配置生成工具对生成的 IP 地址和 Mac 地址、节点 ID 的唯一性进行检测.

大规模 IP 模拟仿真的路由协议根据目标网络的实际规模确定. 由于大规模 IP 网络节点数量多、拓扑关系复杂,一般采用划分自治域(AS)的方式进行路由管理. 将目标网络按照需求划分为不同的 AS,AS 之间配置 BGP 协议,使任何有效地 IP 都能在路由表中找到匹配的目的网络;AS 内部配置内部网关协议(RIP 或 OSPF),确保域内节点路由可达. 例如,在 OSPF 协议中配置路由聚合使路由表数量减少的配置项为:area X range A. B. C. D/前缀位数. 配置好的目标网络使用网络连通性测试软件进行测试,根据测试结果对配置进行调整,达到全网节点路由的可达状态.

网络特性的配置包括指定链路的带宽、试验、丢包率等参数的设置. 在大规模 IP 网络仿真系统实现时预留相应接口,通过在网络初始化时读取配置文件中的信息进行配置. 配置完成后通过商用网络性能测试平台进行网络参数和网络性能的逼真度测试.

3　仿真分析

为验证大规模 IP 网络仿真系统的功能和性能,在由 3 台配置 Intel Xeon E5-2603 CPU(主频 1 066 MHz,4 核心)、16G DDR3 内存、CentOS6. 2 操作系统的联想 D30 工作站组成的计算集群上构建 1 200 个节点规模的 IP 模拟网络,其拓扑结构如图 3 所示. 网络拓扑由核心层、汇聚层和接入层 3 层组成,分为 3 个自治域,域间部署 BGP 协议,域内部署 OSPF 路由协议.

3.1　目标网络构建的性能测试

大规模 IP 网络仿真环境在投入使用之前需要进行节点生成、配置生效、路由建立,称之为网络初始化过程. 初始化所消耗的时间由于系统硬件、操作系统版本、网络应用环境的差异而不同. 模拟网络构建的性

能测试通过测试给定软硬件条件下初始化时间测试模拟网络的构建性能. 基于前述测试环境,3 个自治域的边界路由器配置 BGP 路由协议,域内配置 OSPF 路由协议,IP 地址基于 CIDR 规划,所得的测试结果如图 4 所示,其中路由收敛时间统计的是从网络初始化开始到路由收敛所消耗的总时间.

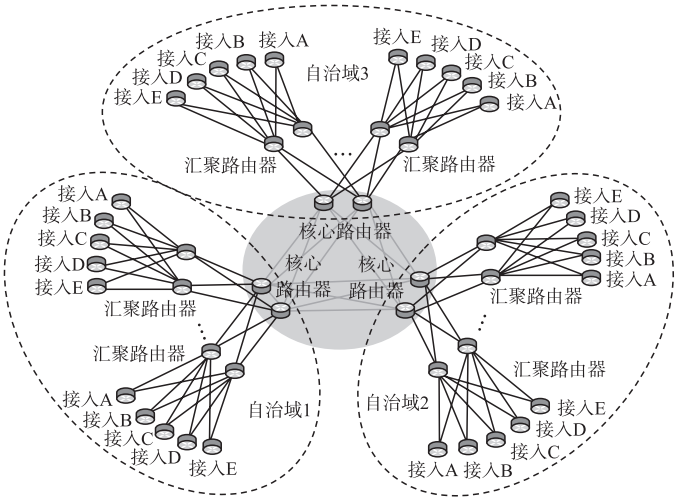

图 3　网络拓扑示意图

Fig. 3　Topological graph of emulated network

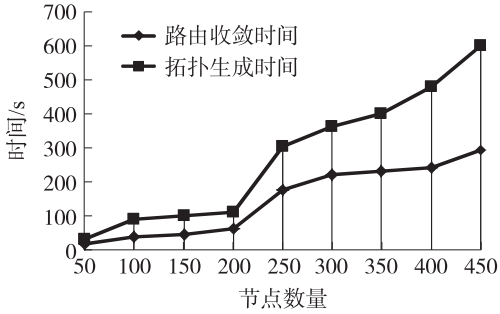

图 4　目标网络构建性能曲线

Fig. 4　Performance curve of target network

仿真结果看出,拓扑生成所消耗的时间随网络规模的扩大而增加,其中路由收敛时间增加的速度高于拓扑生成时间增长的速度. 这是因为随着网络规模的扩大,拓扑生成所消耗的资源呈线性增长,而网络规模扩大所带来的链路复杂度急剧增加,路由条目呈非线性增长趋势. 随着硬件技术的发展,硬件资源对于构建大规模 IP 模拟网络已不是瓶颈,而合理规划 IP 地址、配置路由协议、在配置过程中对网络进行优化,是大幅度提高模拟网络性能的关键.

使用 iperf 软件测试单跳链路传输不同速率的 UDP 数据包时的丢包率,试验结果如图 5 所示. 由于系统采用千兆物理交换机进行组网,传输带宽的最大值为 1 Gbps. 在 UDP 传输速率低于 80 Mbps 时,丢包率为 0,链路可视为理想链路;当 UDP 传输速率达到 200 Mbps 时,丢包率大于 50%,网络性能较低;当速率大于 800 Mbp 时,有 90% 的数据包丢失,此时已达到链路传输能力的极限,继续加大 UDP 发送速率,丢包率数值不再上升. 可见,模拟链路具有物理链路等同的丢包特征,且在常用带宽范围内的丢包率满足使用要求.

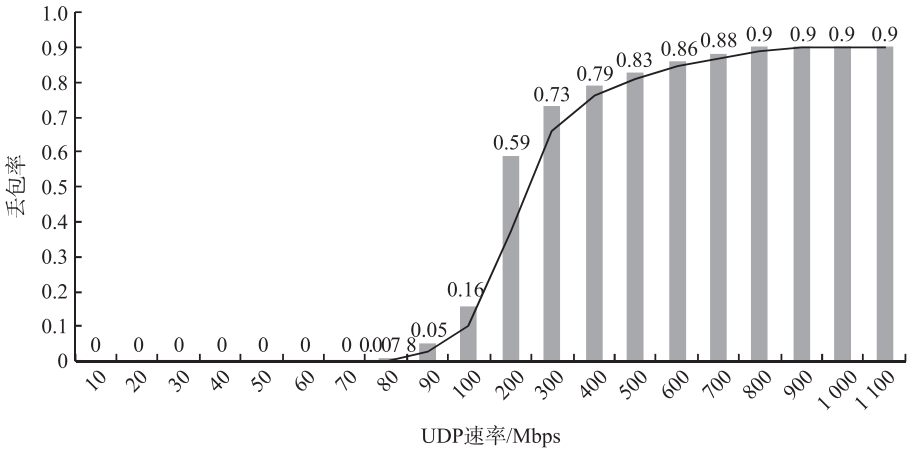

图 5　不同传输条件下的链路丢包率

Fig. 5　Packet lost rate under different UDP traffics

3.2　逼真度测试

逼真度测试检测目标 IP 网络同真实 IP 网络环境在带宽、时延等网络特征上的相似程度.

首先使用系统提供的配置工具对目标网络特性进行配置,使用网络测试工具对目标网络进行测试,测试 10 次取均值,结果与相同配置的真实环境中测得的数据相比对,如表 2 所示. 可以看出,链路带宽设置具有相当高的逼真度,与真实网络的相似度在 96% 以上.

表 2　目标网络带宽对照表

Table 2　Comparison of bandwidth between measurement values and setting values

设计带宽/bps	实测带宽/bps	参考带宽/bps	带宽符合度/%	设计带宽/bps	实测带宽/bps	参考带宽/bps	带宽符合度/%
限 64 k	60.8 k	62.2 k	97.7	限 10 M	9.2 M	9.28 M	99.1
限 256 k	240.8 k	245 k	98.3	限 100 M	40 M	41 M	97.6
限 512 k	440 k	438 k	99.3	限 1 000 M	576 M	600 M	96.0

使用相同的试验方法测试链路时延,得到的数据如图 6 所示. 数据显示,实测时延比设定值有所增加,这是因为物理链路时延由发送时延和传输时延构成,而仿真网络环境中的时延值由 CPU 计算得出,包含了 CPU 处理时间,因此实测值比设定值增大. 在实际应用中可根据需求适当减小设定时延,使实测时延达到试验要求.

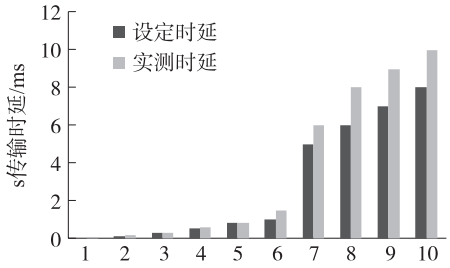

图 6　目标链路时延对照

Fig. 6　Comparison of link delay between measurement values and setting values

4　结语

大规模 IP 网络仿真对网络仿真、系统测试、新技术研发都具有非常重要的意义. LXC 虚拟化技术在提供资源隔离的同时还通过共享内核资源节省开销,因此基于 LXC 的 IP 网络仿真可实现大规模部署,结合 Linux 系统对网络功能的开放性,可实现规模大、逼真度高的 IP 网络仿真运行环境. 同时,对于大规模的 IP 仿真网络,在不增加硬件资源的前提下,只有进行合理的配置和优化才能达到所需的传输性能. 而仿真网络的自动化配置以及提高抗路由震荡等攻击能力是进一步研究的方向.

[参考文献](References)

[1] 邓克波,毛少杰. C4ISR 系统试验设计与分析[J]. 指挥信息系统与技术,2012,3(6):1-6.
Deng Kebo, Mao Shaojie. Design and analysis of C4ISR system experiment[J]. Command and Information System and Technology,2012,3(6):1-6. (in Chinese)

[2] Li Dawei, Mao Shaojie, Zhu Lixin. VNSP: a virtual network based simulation platform[C]//Asia Simulation Conference 2012 (AsiaSim2012). Berlin: Springer,2012:182-189.

[3] 李大伟. 大规模 IP 网络仿真试验环境构建方法[J]. 指挥信息系统与技术,2013,4(6):70-74.
Li Dawei. System building method for large-scale IP network simulation and test environment[J]. Command and Information System and Technology,2013,4(6):70-74. (in Chinese)

[4] 陈文龙,徐明伟,杨扬,等. 高性能虚拟网络 VegaNet[J]. 计算机学报,2010,33(1):63-73.
Chen Wenlong, Xu Mingwei, Yang Yang, et al. Virtual network with high performance: VegaNet[J]. Chinese Journal of Computers,2010,33(1):63-73. (in Chinese)

[5] Flux Group, School of Computing at the University of Utah. Emulab project[EB/OL]. [2000-02-10] http://www.emulab.net.

[6] Princeton University. PlanetLab project[EB/OL]. [2002-10-01] http://www.planet-lab.org.

[7] Ahrenholz J, Danilov C, Henderson T R, et al. Core: a real-time network emulator[C]//Military Communications Conference 2008(MILCOM 2008). San Diego: IEEE Press,2008:1-7.

[8] Hinden R. RFC1517 Applicability statement for the implementation of Classless Inter-Domain Routing(CIDR)[S]. Sun Microsystems,1993.

[9] Rekhter Y, Li T. RFC1518 An architecture for IP address allocation with CIDR[S]. T J Watson Research Center, IBM Corp, CISCO Systems,1993.

[10] Fuller V, Li T, Yu J, et al. RFC 1519 Classless Inter-Domain Routing(CIDR): an address assignment and aggregation strategy [S]. BARRNet, CISCO, Merit, and OARnet,1993.

[11] Yakov R. RFC1520 Exchanging routing information across provider boundaries in the CIDR environment[S]. T J Watson Research Center, IBM Corporation,1993.

[责任编辑:严海琳]