第38卷 第24期
Vol.38 No.24

计 算 机 工 程
Computer Engineering

2012年12月
December 2012

·云计算专题·

文章编号：1000—3428(2012)24—0005—04 文献标识码：A 中图分类号：TP393.02

基于云计算的网络创新实验平台

龚 宇，李 帅，李 勇，苏 厉，金德鹏，曾烈光

(清华大学电子工程系，北京 100084)

摘 要：传统网络实验平台通过直接物理设施或层叠网构建，但这类构建方案无法同时保障较高的物理资源利用率和实验网络链路质量。为此，提出一种基于云计算平台和虚拟化技术的网络创新实验平台设计方案，并给出其原型系统 TUNIE 的实现。应用结果表明，TUNIE 在设计实现上能兼顾功能支持的灵活性与使用的便捷性，而且提供高性能的网络链路和具有良好网络隔离的多用户并发运行环境。

关键词：云计算；虚拟化；网络隔离；网络创新实验平台；网络体系结构；虚拟路由器

Network Innovation Experiment Platform Based on Cloud Computing

GONG Yu, LI Shuai, 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 problem that the conventional approaches for network experiment platforms, which construct an experiment platform either as a physical network directly or in the form of overlay networks, cannot guarantee a high resource utilization and a well network link quality at the same time, this paper proposes a scheme for network experiment platform based on cloud computing and virtualization technology, and provides its preliminary implementation called TUNIE. Result shows that the platform is both flexible in function support and convenience for use, as well as providing a high bandwidth connection and a well-isolated network environment for multi-user support.

【Key words】 cloud computing; virtualization; network isolation; network innovation experiment platform; network architecture; virtual router

DOI: 10.3969/j.issn.1000-3428.2012.24.002

1 概述

计算机网络架构和相关协议研究成果在过去 50 年不断涌现。随着互联网不断发展与规模扩大，互联网地址不足、移动性支持不够等根本性的问题不断显现与加剧，无法在原有体系架构内通过演进性方案有效解决。因此，研究人员将精力转而投向革命性的解决方案，进行新一代网络体系结构研究。新体系结构投入实际运行前必须确保在相对较大规模的网络中可靠运行，但现有互联网无法支持新网络体系结构验证，因而导致研究界对于能够支持网络体系结构验证的实验平台的迫切需求。在国外已有此类项目启动。VINI[1]是美国早期的网络实验平台项目，通过在层叠网络上构建虚拟网络支持网络实验。GENI[2]是美国最新的有关网络实验平台建设的项目，通过在物理设施上进行创新为未来网络研究创建覆盖全球范围的网络实验平台。

搭建网络实验平台的思路一般可以归为 2 类：(1)利用物理设施直接搭建实验网络；(2)利用层叠网思路构建。直接基于物理设施搭建实验平台需要大量资金投入，且实验网络与物理硬件紧耦合，既不利于实验需求的多样化也不利于实验资源共享，资源利用率较低。层叠网的方式通过互联网连接了分布在不同地理区域的资源，实验网络利用正常的互联网数据包封装实验链路数据帧，并在应用层进行处理。因此，任何一条实验链路都通过将有限的物理链路带宽资源划分并顺次拼接多个物理链路虚拟而成，其链路性能严重受制于性能最差的物理链路。层叠网方案无法提供具有质量保障的底层链路。

构建网络创新实验平台需解决如下问题：足够数目网络节点提供，节点间网络连接及拓扑形成，资源在不同实验用户之间管理分配。近些年兴起的云计算[3]为此提供了新的解决思路。利用虚拟化技术，云计算平台划分有限物理计算节点获得大量虚拟节点，并支持虚拟节点之间的网络连接。云计算平台可以根据需求对虚拟节点进行灵活的控制、配置与管理以适应资源在不同用户之间的共享。亚马逊的 EC2 平台[4]是公共云计算平台的典型例子，但对于网络实验而言，需要云计算平台提供更多额外支持，尤其在网络隔离方面。

基金项目：国家自然科学基金资助项目(61171065)；宽带移动通信重大专项基金资助项目(2010ZX03004-002-02)

作者简介：龚 宇(1987－)，男，硕士研究生，主研方向：下一代网络体系结构，无线网络路由算法；李 帅，硕士研究生；李 勇，博士研究生；苏 厉，讲师、博士；金德鹏、曾烈光，教授、博士生导师

收稿日期：2012-03-30 **修回日期**：2012-04-27 **E-mail**：gongy06@mails.tsinghua.edu.cn

本文提出了一种基于云计算模型的网络实验平台设计方案，并给出初步系统实现——清华大学网络创新平台(Tsinghua University Network Innovation Environment, TUNIE)。该设计方案采用集中式控制，系统管理平台内从资源分配到实验网络配置的各个方面利用虚拟链路和虚拟路由器支持灵活的网络实验，通过单独网络测量系统进行平台运行监控。

2 设计目标与系统结构

2.1 设计目标

网络主要由 2 个部分组成：网络节点与网络链路。网络节点一般指计算机和路由器，网络链路指能够提供节点间连接和通信的线路。对于网络实验平台而言，还应该包括管理系统以协调不同的网络实验。

研究人员需要灵活配置网络节点。在硬件方面，研究人员希望灵活配置网络节点硬件资源，如中央处理器、内存、硬盘、网络接口和网络带宽等资源的配置。直接基于物理设施的网络实验，网络节点的硬件资源配置在设备集入系统时已经确定，后续实验不容易进行调整。在软件方面，研究人员希望能够对操作系统进行不受限制的开发及配置，包括内核模块编译及更新。对于传统基于操作系统共享的实验平台，用户权限局限于应用程序内，很难对操作系统进行功能设置。

对于网络链路，研究人员需要强大的控制支持。在一般网络中，链路连接决定网络拓扑和可能的最大带宽，无法提供链路质量保障。对于实验，灵活的连接关系调整可以进行可控动态拓扑调整，而动态带宽控制则可支持流量相关研究及实验。对于平台管理，网络链接控制提供了控制和监测网络节点运行的通道，为隔离不同实验网络以减少干扰提供支持。

管理系统一方面帮助平台管理员更好地管理用户和分配资源，另一方面又能有效地支持用户在平台的实验操作，包括实验管理、节点配置和运行状态监控等，对于实验平台正常有序运行至关重要。

基于以上实验需求分析，本文提出在网络创新实验平台建设方面需满足的 3 个基本目标：隔离性，灵活性和便捷性。

(1)隔离性：在链路层，阻止不希望的连接以形成特定的网络拓扑，并对上层提供动态链路调整支持。在网络层消除并行网络之间影响，保障平台运行的稳定性，避免单个实验网崩溃影响平台其余部分正常运行。

(2)灵活性：对于硬件资源的灵活配置可以方便用户按照需求增强网络节点性能或调整网络规模。独立的操作系统环境和不受限制的权限开放允许用户实现自定义设置。灵活的链路调整功能允许用户设计可控的网络链路环境，进行动态拓扑网络的相关研究。

(3)便捷性：网络实验通常涉及数十个节点的相似配置，便捷的辅助操作配置可以减少大量重复操作，极大地提升效率，同时有效地降低人工操作失误。

2.2 系统结构

TUNIE 是基于云计算模型的网络实验平台系统设计的初步实现，以交换机为中心的 TUNIE 将所有物理机构成全连通的物理资源簇。物理机与交换机之间通过千兆以太网链路连接，物理机通过多网卡配置增大网络吞吐容量。全连通的物理网络为所有的实验网数据流提供底层物理通道支持。另外一个独立的物理网络连接所有物理机、虚拟机及中心控制平台构成单独的控制网。TUNIE 规模易扩展，向资源簇添加物理主机，或通过虚拟专用网络(Virtual Private Network, VPN)通道连接多个资源簇都可实现规模扩大。TUNIE 的硬件组成结构如图 1 所示。

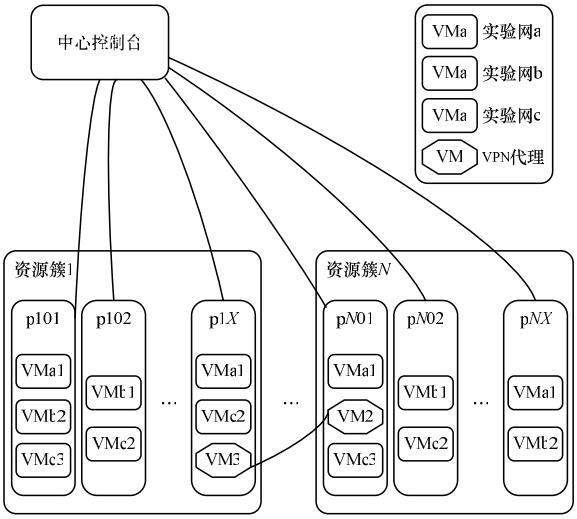

图 1 TUNIE 的硬件组成结构

图 1 所示的系统包含了虚拟机之间 3 种不同的链路连接。运行在同一物理节点的虚拟机之间链路为 PCIe (Peripheral Component Interconnect express)总线连接，相比以太网链路具有更大的带宽和更小的延时特性。运行在同一资源簇内不同物理节点的虚拟机之间链路为千兆以太网链路，可根据实验设计对带宽进行约束。运行于不同资源簇的虚拟机之间通过以太网链路连接，由于穿越公网，带宽较低延时较大，且有限带宽被所有位于 2 个资源簇的虚拟机之间的流量共享，造成网络带宽瓶颈。对此，TUNIE 采用 2 个策略平衡流量：(1)TUNIE 优先分配同一实验网络到同一资源簇，以避免资源簇之间的通信。(2)TUNIE 优先分配同一个实验网虚拟节点到不同的物理机，以减弱虚拟机之间 PCIe 总线链路连接对于网络环境的真实性影响。

除云计算平台通用设计之外，TUNIE 采纳一系列设计以实现本文 2.1 节中提出的网络实验平台 3 个设计目标。

(1)TUNIE 在物理节点以太网层实现了链路隔离。在物理节点上向以太网包添加/移除标签实现网络隔离，这一方案在实现隔离的同时对于实验网透明。隔离依托以太网帧而非网络层，使得平台能够更好地支持网络体系结构创新。

(2)TUNIE 支持虚拟路由器的用户自定义。TUNIE 将路由器设计分解为控制平面和数据平面 2 个部分。借鉴 VINI 的设计方案，TUNIE 利用 XORP[5]进行控制平面的用户自定义，利用 Click 模块[6]支持数据平面的自定义。除了在用户自定义方面提供辅助与支持外，独立的虚拟节点操作系统允许用户直接加载自定制网络模块到内核中编译运行。

(3)TUNIE 集成独立模块以简化实验网络设置和平台管理。模块功能包括基于浏览器的图形化界面进行网络拓扑定制和网络节点硬件资源配置、多样的路由器定制方式支持、独立的控制网实现网络测量和网络节点操作等。

3 系统实现

依据平台控制流程，TUNIE 自顶向下划分为 3 个层：中心控制台，物理节点层与虚拟节点层。图 2 是 TUNIE 软件系统结构的抽象，各部分详细的讨论在本节剩余部分给出。

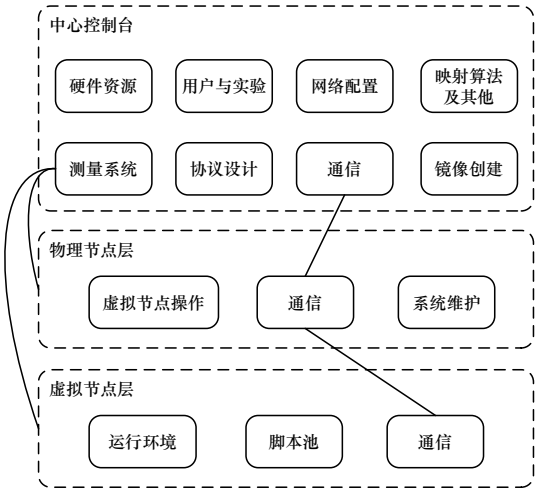

图 2 TUNIE 的软件系统结构

3.1 中心控制台

中心控制台通过各种功能模块进行平台的管理与控制。依据功能和运行环境，这些功能模块可以划分为三大类：平台管理，实验控制与辅助工具集。

3.1.1 平台管理

网络节点硬件资源、实验用户账户和实验网络是管理的主要对象。通信模块负责平台各部分间的通信。

硬件资源主要指网络节点的中央处理器、内存、存储、网卡设备及网络带宽等资源。在添加新的物理资源时，TUNIE 将相应信息存入自建数据库，并持续跟踪记录资源的使用情况，快速回收再利用被其他实验网络释放的物理资源。

TUNIE 通过集成注册管理系统对用户账户进行管理。注册阶段要求用户提供基本信息与简单的实验说明。网络实验归入用户账户进行管理，不同网络实验也可归入项目工程进行管理和用户权限设置。

中心控制台与物理节点之间的通信基于服务器-客户端模型，主要负责中心控制台向物理节点的命令发送及物理节点的执行结果回馈。

3.1.2 实验控制

主要实现实验设计和配置，包括镜像创建、协议设计和网络配置 3 个方面。

镜像创建允许用户部署自己的操作系统以进行自定义模块加载或复杂配置。镜像创建后，用户指定 TUNIE 使用自定义镜像部署实验网络。开放镜像创建功能，极大地提升了用户在网络节点配置上的灵活性。

对于常见的路由协议和转发模式，TUNIE 将设计简化为控制平面与数据平面的方案组合——用户从候选集中选择需要的路由和转发策略，组合成自己的路由器方案并由 TUNIE 在虚拟路由器中自动部署，由此平台实现了对于网络实验部署的快速支持。TUNIE 通过镜像创建方式支持复杂的路由协议和转发设计，以获得最大的灵活性。

网络配置模块实现实验网络拓扑定制和虚拟节点设置。TUNIE 提供图形化界面与文本文件 2 种设置接口。图形化界面操作简单直观，文本文件接口便于大规模网络配置和配置策略的程序优化。虚拟节点的设置同时包括操作系统镜像和网络协议方案的设置。

3.1.3 辅助工具集

辅助工具集包括除以上 2 类之外所有涉及平台管理和实验控制的模块，目前主要有网络测量系统和资源映射算法。

TUNIE 基于开源项目 Cacti[7]集成了独立的测量监控系统。系统通过不同权限设置同时向管理员和实验用户开放。平台管理员具有系统所有读写权限。实验用户可以查看实验网络节点的运行数据统计。测量系统的集成极大地便利了用户对于实验网络的监控和性能分析。

资源映射算法以独立模块形式集成，通过解耦合的设计一方面保证平台的运行，另一方面又便于进行算法的研究和实验算法在平台的测试运行。其他功能模块，如虚拟机在线迁移、备份系统等，都可以通过类似的集入方式同时用于平台实践和相关研究。

3.2 物理节点层

物理节点层执行中心控制台发送的命令。目前此部分包括 2 个功能模块：虚拟节点操作和系统维护。

虚拟节点操作模块在平台中为用户提供针对虚拟机的控制操作接口，如创建虚拟机、开关虚拟机电源等，以及设备操作，如存储或网络接口的增加删除、虚拟机的整机迁移等。虚拟节点操作为用户提供了对于虚拟节点灵活的控制和动态调整。在整个操作流程中，物理节点既作为命令的执行单元执行特定操作，同时也作为命令的转述单元将中心控制台的部分操作命令及参数转发到虚拟节点中执行。TUNIE 利用终端套接字方式在物理节点和虚拟节点之间进行通信。

系统维护模块处理本地资源分配及物理机灾难恢复。本地资源分配主要目标是避免资源使用冲突，提高资源利

用率，其具体实现依赖于对本地资源动态使用记录表的维护。TUNIE 记录所有针对虚拟机建立和配置的历史操作，并由独立的进程周期性地查看虚拟机的运行状态。结合两部分信息，机器意外重启后，可以在没有中心控制台参与的情况下由物理节点独立恢复之前的运行状态。

3.3 虚拟节点层

虚拟节点通过 2 个模块优化支持系统控制和网络实验：脚本池和预建立的运行环境。

脚本池有 3 个功能：网络设置，软件路由器配置文件生成和网络服务初始化。虚拟机建立后，物理机通过终端套接字调用特定脚本启动网络服务，建立与中心控制台通信连接，建立用户账户，创建配置文件并启动虚拟路由器。

TUNIE 目前主要为路由和转发创建运行环境。默认镜像中预置了 XORP 与 Click。通过集成这 2 个软件，用户可以灵活地进行控制平面和数据平面的设计及配置。其他运行环境可以按照需求加载到默认镜像中。

4 实验流程

用户在使用 TUNIE 之前需注册并通过平台管理员审核。TUNIE 按照项目进行资源分配和不同实验网络的管理。用户在建立项目的同时提交资源使用申请，在可用资源范围内可以建立多个实验网络。

项目申请通过后，用户可以开始建立实验网络：创建镜像，设计路由转发协议，定制网络拓扑，配置虚拟节点。完成之后，即可启动网络运行实验。整个实验流程概括如图 3 所示。用户可以自行决定是否创建镜像。使用默认镜像，完成配置后，完整实验网的启动可以在 1 min 内实现。

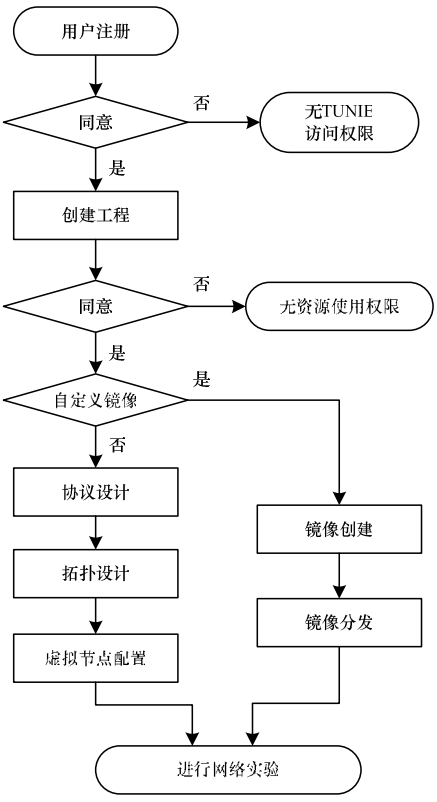

图 3 使用 TUNIE 的实验流程

5 TUNIE 性能实测

目前 TUNIE 原型系统已经部署到分布于 3 个资源簇的 30 台商用计算机中。实际测试结果显示，位于同一资源簇不同物理节点的虚拟节点之间的链路带宽最高可以达到 837 Mb/s。虚拟路由器转发速率在丢包率低于 0.001%的条件下可以达到 140 Mb/s，直接虚拟链路延时低于 0.3 ms。

为验证 TUNIE 对于网络隔离性的支持，利用平台设计模拟三网融合实验——同时在部署于相同物理节点的 3 个不同拓扑虚拟实验网中运行不同路由协议。实验拓扑设计如图 4 所示，不同形状的标签用以区分虚拟节点所依附的不同物理节点。

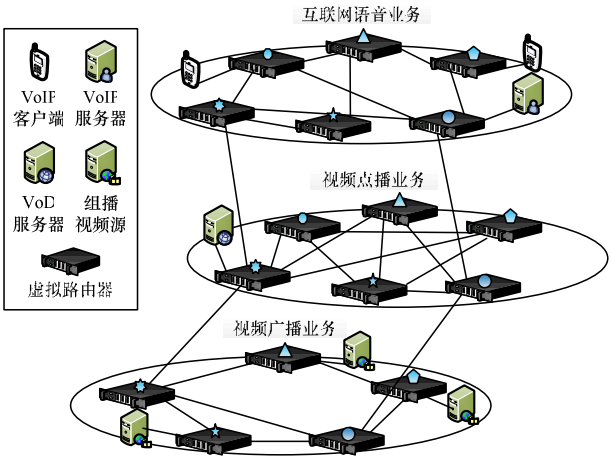

图 4 三网融合拓扑结构

在第 1 个虚拟网中，在 RIP(Routing Information Protocol)路由协议支持下运行 VoIP(Voice over Internet Protocol)业务。VoIP 服务器运行在单独的虚拟节点中，采用开源软件 Asterisk[8]实现，客户端节点运行 Twinkle[9]并连接至该虚拟路由器。第 2 个虚拟网在 OSPF(Open Shortest Path First)路由协议的基础上运行视频点播业务，服务器与客户端均采用 VLC[10]实现。第 3 个虚拟网构建了具有 3 个独立视频源和 6 个独立接收客户端的组播网，三路组播信息同时传输。组播协议采用 IGMP(Internet Group Management Protocol)与 PIM-SM(Protocol Independent Multicast-Sparse Mode)的组合方案，VLC 用 来搭建组播服务器和客户端。实验中每一路视频源均具有至少 640×360 像素分辨率和 30 帧每秒帧速率。在 3 个虚拟网中，使用 XORP 构建所有控制平面，利用 Click 做单播网络软件的转发数据平面，Linux 内核做组播网络的数据平面转发。如设计预期，3 个网络同时正常运行，没有任何彼此之间的干扰。在模拟意外情形的实验中，单个虚拟节点强制停止运行，另外 2 个虚拟网正常运行不受影响；中断其中一个物理节点，导致 3 个虚拟网各自有一个节点无法正常运行，但除与停止运行的路由器直接相连的客户端或数据源外，其余节点均能正常运行。

(下转第 13 页)