高性能虚拟网络 VegaNet

陈文龙[1),2)] 徐明伟[2)] 杨 扬[1)] 李 琦[2)] 马东超[3)]

[1)](北京科技大学信息工程学院 北京 100083)

[2)](清华大学计算机科学与技术系 北京 100084)

[3)](北京邮电大学光通信与光电子学研究院 北京 100876)

摘 要 现有网络研究的实验一般在模拟工具或原型系统上完成.然而,这些实验环境与真实网络环境总有较大的差别,主要包括没有真实用户流量、缺乏真实网络中的丰富网络事件、没有运行真正的协议栈软硬件平台等.另一方面,对于运营网络的研究,研究者经常希望分析网络事件发生前后的网络状况,例如更改拓扑或配置,注入网络故障等,显然这些措施难以在真实运营网络中实施.针对上述问题设计了一种为网络研究提供真实实验环境以及对核心网络进行模拟分析的高性能虚拟网络 VegaNet(Virtual Gigabit Network).文中详细介绍了 VegaNet 结构模型及设计实现.目前,已完成 VegaNet 在 CERNET2 清华校园网的初步部署.实验表明,VegaNet 能提供一个接近于真实网络状况的网络实验环境,并能灵活支持对核心网络的模拟分析.

关键词 VegaNet;虚拟路由器;网络实验;虚拟链路状态协议

中图法分类号 TP393 **DOI号**: 10.3724/SP.J.1016.2010.00063

Virtual Network with High Performance: VegaNet

CHEN Wen-Long[1),2)] XU Ming-Wei[2)] YANG Yang[1)] LI Qi[2)] MA Dong-Chao[3)]

[1)](*School of Information Engineering, University of Science and Technology Beijing, Beijing* 100083)

[2)](*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

[3)](*Institute of Optical Communications and Optoelectronics, Beijing University of Posts and Telecommunications, Beijing* 100876)

Abstract Network experiments are usually implemented through imitation or prototype. However, great difference exists between reality and experimental environment, such as no real user traffic, no rich network events, and no real routing platform. On the other hand, researchers prefer to analyze current network status after injecting failures and changes about configuration or network fabric. To solve the above issue, VegaNet(Virtual Gigabit Network) is designed, which is a virtual network with high performance. It supports the following aspects: real user traffic load, node and link failure injection, state synchronization within substrate network, high-performance forwarding, and so on. VegaNet can be regarded as a real experimental environment, and through which net researchers can simulate and analyze current network. This paper presents model and design of VegaNet. Experiments show that VegaNet with high performance can possess the above functions.

Keywords VegaNet; virtual router; network experiment; VLSP

收稿日期:2009-07-15;最终修改稿收到日期:2009-10-12.本课题得到国家"九七三"重点基础研究发展规划项目基金(2009CB320502)、国家"八六三"高技术研究发展计划项目基金(2007AA01Z2a2,2009AA01Z251,2007AA01Z234)和国家自然科学基金(60873192)资助.

陈文龙,男,1976年生,博士研究生,主要研究方向为网络体系结构、网络协议.E-mail: chenwenlong2008@gmail.com; wenlongchen@sina.com. **徐明伟**,男,1971年生,博士,教授,主要研究领域为网络体系结构、高速路由器体系结构和协议测试. **杨 扬**,男,1955年生,博士,教授,主要研究领域为网络通信、图像处理. **李 琦**,男,1979年生,博士研究生,主要研究领域为网络体系结构. **马东超**,男,1980年生,博士研究生,主要研究方向为计算机网络.

1 引 言

互联网高速发展，关于网络体系结构、路由协议及网络服务的新型研究层出不穷. 对于研究成果的分析评价，主要通过两种方式完成：(1) 在模拟工具中实现研究模型并人为设计网络拓扑进行实验；(2) 在实验室中完成原型系统并搭建简单环境完成实验. 然而，实验环境与现实状况有着较大的差距，主要包括没有真实用户流量，缺乏真实网络中的丰富网络事件，没有运行真正的协议栈软硬件平台等. 另一方面，对于运营网络的研究，面对 24h 不间断的长期稳定的网络服务，无法随意地更改网络拓扑或配置，更不能人为构造网络故障，使得大量优化方案无法验证其效能. 总之，网络研究者的新思想或新模型难有理想的实验环境，也难以获取有价值的评价依据.

研究者希望有一种虚拟网络架构，它既能提供几尽真实的网络环境又不影响任何实际网络正常运营，继而还能透明地对运营网络进行模拟分析. Overlay 虚拟网络架构是一个极佳的选择. Overlay 网络是一种部署于现存网络上独立的虚拟网络，就像在原有网络上叠加了一层新的网络. 典型的 Overlay 网络包括进行快速路由检测和恢复的 RON[1]、在现有 Internet 上提供端到端服务质量保证的 SON[2] 以及在 SON 基础上为了要减少性能限制提出的 SOI 架构[3].

继而，VINI (A Virtual Network Infrastructure)[4] 被提出，它为网络研究者提供了一个可控的、能够面对真实环境的底层网络架构. VINI 允许研究者将新型思想在具有真实路由软件、真实流量负载及真实网络事件的环境下进行实施，并得以评估. VINI 提供的网络实验支撑架构的主要目标包括运行真实路由软件，面对真实网络环境，网络事件可控，实验网络承载真实用户流量. VINI 在 PlanetLab[5-6] 节点上构造基本原型系统. 原型系统利用 XORP(eXtensible Open Router Platform) 开源路由协议软件[7] 实现虚拟系统路由控制层面的处理，利用 CLICK[8] 软件完成虚拟数据层面的报文处理. 节点间的虚拟链路通道是通过 UDP SOCKET 通信完成的. VINI 在 PlanetLab 测试平台上搭建完成了一个实例：PL-VINI，基于它的一系列实验证明了 PL-VINI 的正确性、有效性、可用性，并且能做到对真实网络状况真实反映. 然而，VINI 不是基于商业路由软硬件平台，没有实现虚拟网络到真实网络的拓扑映射以及链路故障的状态同步，也未实现虚拟网络与真实路由实体的信息交互.

我们设计的 VegaNet(Virtual Gigabit Network) 和 VINI 一样，也是一种 Overlay 结构的虚拟网络，它为网络研究提供真实实验环境. 此外，它还能透明地对当前运营的核心网络进行模拟分析. VegaNet 以实际运营的某个网络体系作为底层架构，在网络边缘接入若干虚拟路由器，虚拟路由器之间通过虚拟链路连接，构成一个虚拟网络平台. 该虚拟网络平台既可提供网络实验环境，又可对底层核心网络进行模拟分析. VegaNet 主要特性包括引入真实的用户流量，支持节点及链路故障的注入，同步底层网络故障，虚拟路由器基于真正的商业路由平台实现，支持高带宽的虚拟网络流量，虚拟网络协议族独立于底层网络，虚拟网络相对底层网络透明等. 其中，在报文处理性能、虚拟网络与底层网络故障同步、虚拟链路状态管理等方面，相对 VINI 有了显著的提高. 目前，VegaNet 已完成初步实施及实验. 实验结果表明，VegaNet 能提供一个如同真实物理网络一样的网络实验环境，也能对底层网络实现模拟分析和性能评价.

本文第 2 节给出了 VegaNet 的设计思想及结构模型；第 3 节给出了以 CERNET2 为底层支撑网络的 VegaNet 实例设计；第 4 节介绍了 VegaNet 核心设备虚拟路由器的软、硬件体系结构设计以及虚链路状态协议的设计；第 5 节描述了 VegaNet 在 CERNET2 清华校园网的初步部署情况以及基于它的实验及实验分析；第 6 节对全文进行了总结，并介绍了项目下一步的工作方向.

2 VegaNet 模型

VegaNet 总体目标包括两个方面：(1) 为新型网络协议或服务提供尽量真实的实验环境，保证实验能不断改进、反复进行；(2) 通过 VegaNet 对特定网络体系进行模拟，反映实际网络运行状况，并对该网络体系结构进行分析. VegaNet 模型设计有以下基本原则：(1) 它总是基于某个底层网络进行构建，并且对底层网络透明，即底层网络不关心 VegaNet 上的网络行为；(2) 能实现 VegaNet 中虚拟链路与其对应的真实物理链路的故障同步；(3) VegaNet 有真实用户并可承载真实网络一样的网络行为；(4) VegaNet 网络拓扑可重构并具有较高的可控

性，管理员根据实验需要只需简单配置便可改变虚拟网络拓扑，还可方便地获取 VegaNet 网络事件信息.

本文关于 VegaNet 的描述有表 1 所述的符号定义，图 1 是 VegaNet 的网络体系结构模型. VegaNet 是在 SUN 网络核心节点下面部署若干台虚拟路由器 VR，VR 像普通端系统一样接入 SUN，VR 之间通过虚链路 Vlink 连接. 图 1 中 VegaNet 是所有 VR 的 Full-mesh 连接，实际部署时可根据不同需求构建不同拓扑. VegaNet 就是由若干 VR 和 Vlink 构成的虚拟网络，每台 VR 通过 DCI 接口下连局域网. VegaNet 对于底层网络来说是透明运行的，并且 VegaNet 的网络协议族与 SUN 网络的协议族是相互独立的. 假设 SUN 网络协议族为 F_S，报文头格式为 $Header(F_S)$；VegaNet 网络协议族为 F_V，报文头格式为 $Header(F_V)$，那么，VegaNet 的网络报文在 SUN 网络传输时采取如图 2 所示直接封装的格式. VegaNet 中无论数据转发报文或是协议控制报文，只要通过 VR 中虚链路收发就采用图 2 的报文封装格式. 当然，对于边缘局域网内部的报文交互，仍然采用 VegaNet 协议族原始报文格式.

表 1　VegaNet 符号定义

符号	描述
SUN	Substrate Network，VegaNet 的底层支撑网络
VR	Virtual Router，虚拟路由器，作为端系统接入 SUN 网络，向下连 VegaNet 局域网提供数据转发服务
UCI	Up-Connecting Interface，VR 上连 SUN 网络的接口，一台 VR 只能配置一个 UCI
DCI	Down-Connecting Interface，VR 下连 VegaNet 局域网的接口，一台 VR 可能会有多个 DCI
Vlink	Virtual Link，虚链路，利用封装解封装技术、穿越 SUN 网络进行 VegaNet 协议报文透明传输的虚拟通道
VI	Virtual Interface，VR 上收发封装报文的虚拟接口，总与某条虚链路对应

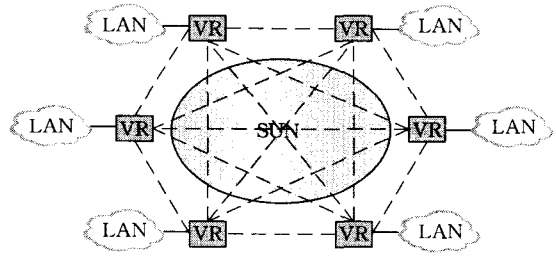


图 1　VegaNet 体系结构模型

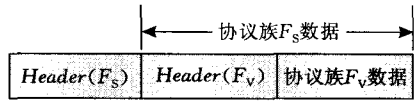


图 2　VegaNet 报文封装格式

对于 VegaNet 中穿越 SUN 网络的一次端到端通信，包含 3 种 VR 角色：

$$\begin{cases} VR_{send}, & num(VR_{send})=1 \\ VR_{rcv}, & num(VR_{rcv})=1 \\ VR_{fwd}, & num(VR_{fwd})\geq 0 \end{cases}.$$

源端系统所属的虚拟路由器称为 VR_{send}，它从 DCI 接收协议族 F_V 报文进行封装，并按照协议族 F_S 的寻径方式从 UCI 发送报文. 目的端系统所属的虚拟路由器称为 VR_{rcv}，它从 UCI 接收到达本地的协议族 F_S 报文进行解封装，并按照协议族 F_V 的寻径方式从 DCI 发送报文. 从 VR_{send} 到 VR_{rcv} 可能经过的若干个转发点称为 VR_{fwd}，它对协议族 F_S 的封装报文进行重新封装并向 VR_{rcv} 方向发送. 图 3 是 VR 的报文处理模型，它集成了上述 3 种 VR 角色的报文处理功能.

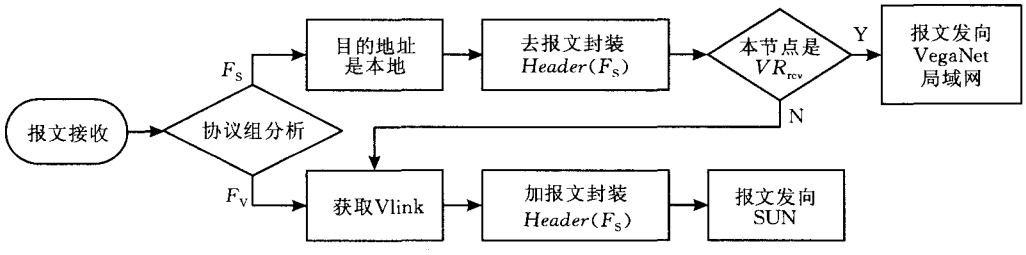


图 3　VR 报文处理模型

VegaNet 的一个重要目标是对 SUN 核心网络进行模拟，通过 VegaNet 运行情况反映 SUN 网络的实际运行状况，并对 VegaNet 网络结构及网络协议进行改进实验，给 SUN 网络的发展提供建设性的建议. VegaNet 部署时，SUN 网络每台核心路由器下连一台 VR，该 VR 用来模拟其上连的核心路由器. 对照 SUN 网络真实链路，VegaNet 配置相应的虚链路，一条 Vlink 对应 SUN 核心网中的一条真实链路.

我们希望 VegaNet 能做到虚拟网络与底层网络的链路故障同步，而其实施的关键是如何分析底层网络真实链路是否发生故障. 因为对于一条虚链路两端的 VR，SUN 网络中可能有多条可达路径，而实验者只希望虚链路与其模拟的那条真实链路保

持状态一致. 令集合 $S(R,L)$ 表示 SUN 网络核心设备连接拓扑，其中，$R=\{r_1,r_2,r_3,\cdots,r_n\}$ 是路由器集合，$L=\{l_{ij}\mid r_i\in R,r_j\in R,i\neq j,0<i,j\leq n\}$ 是物理链路集合，且对于 R 中任意两个不同的节点最多只有一条物理链路. VegaNet 网络对 SUN 进行模拟时，SUN 核心网中每一台路由器和每一条物理链路都在 VegaNet 中有一台 VR 和一条 Vlink 与之对应. 集合 $V(R',L')$ 表示 VegaNet 连接拓扑，它是集合 S 的等价类，其中 $R'=\{r'_1,r'_2,r'_3,\cdots,r'_n\}$ 是 VR 集合，$L'=\{l'_{ij}\mid r'_i\in R',r'_j\in R',i\neq j,0<i,j\leq n\}$ 是 Vlink 集合，同样对于 R' 中任两个不同的虚节点最多有一条虚拟链路. 另外，VegaNet 部署于 SUN 网络时，每台 VR 通过一条物理链路连接其对应的核心路由器，称其为 VR 接入链路，该类链路的数量与 VR 数量相等. 我们将 VR 接入链路集合定义为 $K=\{k_i\mid 0<i\leq n\}$，链路 k_i 的两个端节点为 r_i 和 r'_i，它们分别来自集合 R 和 R'. 在 VR 接入链路集合 K 无故障且 SUN 网络路由总是优先选择最短路径的情况下，有定理 1. 本文不考虑链路权值等评价参数，跳数最少即为最短路径.

定理 1. 物理 l_{ij} 链路正常当且仅当虚拟路径 $Path(l'_{ij})$ 的跳数为 3.

证明. VegaNet 模拟环境中，虚链路总是针对某条真实物理链路而建立的. 所以在 l_{ij} 链路正常时，其对应的虚链路 l'_{ij}（即虚拟路由器 r'_i 和 r'_j 之间的虚链路）所对应的最短实际路径为 $Path(l'_{ij})=\{k_i,l_{ij},k_j\}$，路径跳数为 3，其经过的节点序列为 $\{r'_i,r_i,r_j,r'_j\}$. 当 l_{ij} 链路发生故障时，假设虚链路 l'_{ij} 仍然可达，则说明核心路由器 r_i 和 r_j 间存在通信路径. 由于 r_i 和 r_j 间只有一条直连链路 l_{ij} 且已发生故障，所以 $Path(l_{ij})$ 的路径跳数必定大于等于 2；加上链路 k_i 和 k_j，$Path(l'_{ij})$ 的路径跳数必定大于等于 4. 定理 1 成立. 证毕.

基于定理 1，在 VegaNet 对特定 SUN 网络进行模拟时，只要 SUN 网络采取最短路径路由算法，我们可以通过分析虚链路的两端通信的实际转发跳数是否为 3 来判断其模拟的真实物理链路是否正常，从而决定虚链路的状态，最终实现 Vlink 与其模拟的真实链路的状态一致性.

3 VegaNet 设计

VegaNet 设计的主要思想就是部署若干 VR 通过 SUN 网络核心节点接入，VR 之间通过虚链路 Vlink 进行连接. VegaNet 数据通过虚链路传输时，需要用 SUN 网络协议族的 IP 头部进行封装发送. 本文实现了一个 VegaNet 实例设计，实例中 VegaNet 运行 IPv4 协议族，CERNET2 为其 SUN 网络. VegaNet 虚链路的 IPv4 数据传输通过 IPv6 封装技术实现，该传输对 CERNET2 网络完全透明. VegaNet 中地址及路由配置细节如下：

(1) IPv6 Address. VR 如同普通终端有且只有一个全局 IPv6 地址，配置在 UCI 上；

(2) IPv4 Address. VR 除了 UCI 以外的其它接口都可配置 VegaNet 中唯一的 IPv4 地址；

(3) Vlink. 两个 VR 之间要建立一条虚链路时，需要双方配置 Vlink；Vlink 配置主要就是指定对端 VR 的 IPv6 全局地址；

(4) IPv4 Routing. VR 通过 DCI 把所属的 IPv4 网络连入 VegaNet，VR 可在 DCI 连接的真实链路及 Vlink 上运行 IPv4 路由协议，实现整个 VegaNet 的路由信息交互；

(5) IPv6 Routing. 由于 VR 单点接入 CERNET2，VR 只需一条 IPv6 默认路由实现 IPv6 报文发送.

另外，VegaNet 支持人为对其网络节点及链路注入故障及故障恢复事件. 在 VR 或 Vlink 运行正常的情况下，有时因为实验需要希望 VR 或 Vlink 发生故障. 此时，管理员可以远程通过私有协议通告 VR 进行某种行为约束，达到网络故障发生的效果. 当然，管理员还可通告取消人为故障使网络恢复正常. 网络中各种故障都有其对应的表现形式，VegaNet 正是以此为据实现对网络故障的控制，包括以下处理：

(1) VI Down. VI 的接口状态为 Down，其后续行为和真实接口 Down 所导致的行为一样，不再发送或接收任何报文；

(2) VR Down. VR 上所有虚接口状态都为 Down，都不进行报文收发；

(3) Vlink Down. 虚链路两端对应 VI 的接口状态都为 Down；

相对于其它虚拟网络模型，本文的 VegaNet 设计主要有以下特点：

(1) 实现了真实网络故障在虚拟网络中的实时体现. VegaNet 希望网络研究及实验基于尽量真实的网络环境完成，真实环境的考验是对网络协议及网络体系结构各方面特性评估最有力的论据. VegaNet 通过 VLSP 协议，分析虚链路跳数或传输路径，实现了虚拟网络与所模拟的真实网络的链路故障同步.

(2) 监控体系可以实现深度分析并注入故障. 除了让虚拟网络同步真实网络的链路故障及恢复，VegaNet 还支持人为控制虚拟网络中的故障产生和故障取消事件. 监控平台通过私有协议(一种基于 IPv6 TCP 的私有协议，本文不做详细描述)对虚拟路由器发送控制消息，从而操作故障状态. 另外，监控平台还利用私有协议从虚拟路由器中采集各种网络事件，并在后台研究分析.

(3) VR 直接基于商业路由系统平台进行开发. VegaNet 中的虚拟路由器是一台真正的路由设备，它基于清华大学自行研制的 BWOS 路由平台改进而成(详见本文"虚拟路由器设计"部分). 该架构中，虚拟路由器可以面对和真实路由器一样的系统平台问题，如高负载任务对处理器的长期占用、路由设备中数据消息与控制消息的冲突、分布式体系结构中消息同步等问题. 所以，基于 VegaNet 的网络实验及研究与实际状况更贴近.

(4) 支持 Gibit 用户流量. 网络实验中，是否有真实的高带宽用户流量是非常关键的影响因素. 商业路由设备中，报文处理主要包括高速数据转发和协议报文收发，这两者共享外部网络接口并产生竞争关系. 所以，支持高速用户转发流量将使网络协议实验分析更具参考价值.

(5) 虚链路状态协议(VLSP)为模拟真实网络及链路控制提供灵活支持. VegaNet 中，虚拟链路像物理链路一样具有一些链路属性参数，如 MTU，我们希望虚链路两端链路参数协商一致. 另外，虚链路通常跨越多跳物理链路，它的可达性难以检测. VLSP 协议就是为解决上述问题而设计的，它是虚拟网络平台中重要的设计之一. VLSP 的功能包括维护虚拟邻居状态、实现虚拟链路两端的参数协商、虚拟链路双向可达性检测、虚拟链路所用真实路径监控等.

图 4 是 VegaNet 对 CERNET2[①] 的模拟的拓扑子图，我们假设所有 VR 接入链路总是正常，因为任一接入链路发生故障，整个网络模拟就无效. 对于虚拟路由器 JN′到 HF′的虚链路 *Vlink*(JN′，HF′)，有两条可达路径：$Path_1$ 经过节点序列{JN′，JN，HF，HF′}，$Path_2$ 经过节点序列{JN′，JN，TJ，BJ，WH，NJ，HF，HF′}. CERNET2 核心网为一个自治域，域内运行 OSPF 协议选择最短路径. 基于定理 1，若 *Vlink*(JN′，HF′)的物理跳数为 3 表明 JN 到 HF 的物理链路正常，否则发生链路故障. VegaNet 中由 VLSP 协议通过监测 Vlink 的实际物理跳数是否为 3，来判断其模拟的物理链路是否发生故障，从而决定 Vlink 的链路状态.

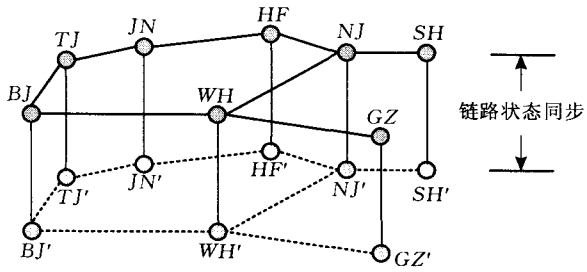


图 4　VegaNet 对 CERNET2 模拟拓扑子图

3.1　虚拟网络数据转发

图 5 是由 5 台 VR 通过虚链路组成的一个虚拟网络，我们以此拓扑详细描述 VegaNet 中路由学习及数据转发细节. 图 5 中每台 VR 有两个邻居，VR 之间的 Vlink 全部运行 OSPFv2 协议，各虚链路权值全设置为 1，VR 将所属的 IPv4 局域网路由引入 OSPF. 这样，VegaNet 只是通过常规的路由配置，就实现了虚拟网全网路由的学习. 下面以主机 A 与主机 C 通信为例来分析 VR 下连端系统通过虚拟网络的通信过程. 主机 A 到主机 C 必经 VR1、VR3，而 VR1 到 VR3 有两条可达路径 VR1—VR2—VR3 以及 VR1—VR5—VR4—VR3，VR1 进行 SPF 计算得出路径开销分别为 2 和 3，将选取第 1 条路作为去往 VR3 的转发路径. 同样，VR3 也会选取该路径的反向路径作为去往 VR1 的数据转发路径. 从主机 A 通过虚拟网络发送 IPv4 报文给主机 C，共有如表 2 所列步骤.

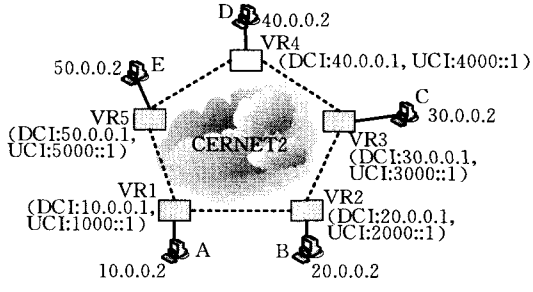


图 5　VegaNet 路由及转发示例拓扑

上述报文变换过程如图 6 所示，需要重点说明的是 VR2 接收和转发的 IPv6 封装报文是更换了封装头部的，也就是说 VR2 进行了重新封装. 由主机 C 发往 A 的转发过程与上述步骤类似. 整个转发过程中的关键点是 VR 对报文的封装及解封装. 报文从源 IPv4 网络进入 IPv6 网络时，由源端所在 VR 进行 IPv6 封装；报文离开 IPv6 网络进入目的 IPv4

① http://www.cernet2.edu.cn/, http://nms-v6.cernet2.edu.cn/fault/

网络时，由目的端所在 VR 进行解封装；而中间经过的 VR 要对封装报文进行解封装后重新封装发往另一个 VR. 虚拟网络中的 IPv4 通信对于承载它的 IPv6 网络来说是完全透明的.

表 2 图 5 环境下主机 A 发送 IPv4 报文至主机 C 所经步骤

步骤	角色	行为描述	报文处理细节
1	主机 A	发出原始 IPv4 报文	*Dst*(30.0.0.2)、*Src*(10.0.0.2)、*Nexthop*(10.0.0.1)
2	VR1	接收并转发 IPv4 报文	查找 IPv4 转发表，出接口为连接 VR2 的虚接口
3	VR1	VI 发送 IPv4 报文	IPv6 封装后发出，封装头中 Dst(2000::1)、Src(1000::1)
4	VR2	VI 接收封装报文	接收封装报文进行解封装操作，得到原始 IPv4 报文
5	VR2	VI 发送 IPv4 报文	查 IPv4 转发表出接口为 VI，IPv6 封装后发出，封装头中 *Dst*(3000::1)、*Src*(2000::1)
6	VR3	VI 接收封装报文	接收封装报文进行解封装操作，得到原始 IPv4 报文
7	VR3	DIC 发送 IPv4 报文	查 IPv4 转发表出接口为 DIC，发送原始 IPv4 报文
8	主机 C	接收原始 IPv4 报文	接收到主机 A 发出的原始 IPv4 报文

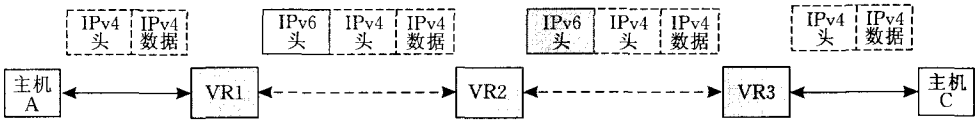


图 6 VegaNet 报文变化过程

4 虚拟路由器设计

4.1 虚拟路由器体系结构

虚拟路由器系统基于 BWOS 平台[9]改造完成. 可扩展路由器操作系统 BWOS 支持 IPv6 和 IPv4 双协议栈，实现了多种网络接口下的高速分组转发功能，并通过硬件冗余和软件状态备份技术实现了路由器系统的高可用性.

虚拟路由器系统相对 BWOS 平台，主要是在控制层面增加了虚接口管理，并屏蔽虚接口对其它模块的影响. 用户配置虚接口后，系统在接口管理模块增加虚接口处理. 接口管理向其它协议模块通告接口状态信息时与其它物理接口完全一致，协议模块不关心虚接口报文收发具体细节. 数据层面，虚拟路由器将常规报文转发与封装/解封装结合起来，保证 VegaNet 报文的快速处理. 虚拟路由器主要功能模块结构如图 7 所示.

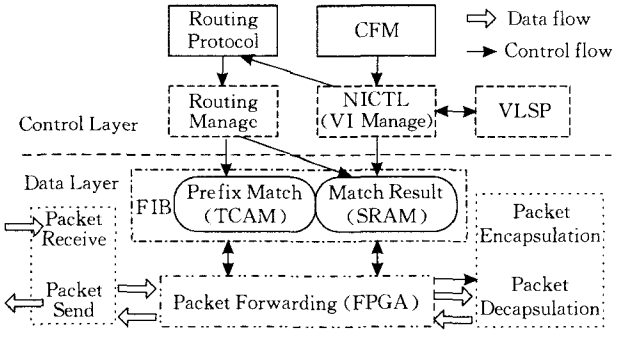


图 7 虚拟路由器主要功能模块结构图

各模块描述如下：

(1) 网络接口控制(NICTL). 网络接口管理模块中增加虚接口管理. 虚接口和其它物理接口一样具有各种接口状态信息：管理状态、链路协议状态、MTU 等等，NICTL 模块会将虚接口状态信息通告系统中关心接口状态的模块. 另外，该模块还与数据层面有消息交互，主要是将虚接口的对端 IPv6 地址信息下发给数据层的 SRAM 模块，用作封装头中的目的地址.

(2) 配置管理(CFM). 用户通过命令行配置创建或取消虚接口，针对每个生成的虚接口用户还需配置虚链路对端的 IPv6 地址. 由于 VegaNet 中 VR 只有一个连入 CERNET2 的 IPv6 接口：UCI，所以虚链路关心的对端地址就是 UCI 上的 IPv6 全局地址.

(3) VLSP. 在虚拟链路上收发 VLSP 协议报文，按 4.2 节所述实现虚接口的可达性检测、跳数测量、参数协商等功能.

(4) 路由协议模块(Routing Protocol). 路由协议模块不关心虚接口行为，按传统方式运行. 各协议模块可以从虚接口收发报文，操作方式和物理接口完全一样. 虚接口报文收发涉及的封装和解封装都由数据层完成，对上层模块是透明的.

(5) 路由管理(Routing Manage). 路由协议可以通过虚接口收发协议报文，会生成以虚接口为出接口的路由项. 路由管理模块将核心路由表项通告给数据层，前缀信息通知给 TCAM 模块，匹配处理信息通知给 SRAM 存储.

(6) 转发表存储及数据转发(FIB & Packet Forwarding). 三元内容可寻址存储器(TCAM) 是一种特定类型的完全相联存储器，允许完全并行的查找转发表或分类器数据库. 静态随机存储器(SRAM) 是一种广泛应用的高性能存储器，在路由器查找系

统中，可以配合 TCAM 用于存储匹配表项信息. 简而言之，TCAM 模块实现快速的路由前缀最长匹配，SRAM 中存储了 TCAM 查询结果对应的处理信息. 为了提高硬件处理速度，VR 通过一次查表获取封装及转发相关所有信息. 而且因为 VR 只通过 UCI 连入 CERNET2，封装后的 IPv6 报文转发无需路由查询，固定 IPv6 下一跳发送. 所以，数据层面只需要 IPv4 转发表. 对于需要封装处理的报文，它所匹配的路由前缀在对应 SRAM 的存储中会增加用于封装的地址信息. 概括起来，虚拟路由器中数据层面主要有 3 种报文处理行为：

① 接收 IPv4 报文，路由匹配结果是进行常规 IPv4 转发.

② 接收 IPv4 报文，路由匹配结果是 IPv6 封装发送，则从 SRAM 存储中获取封装地址，封装成 IPv6 报文后无需查表，固定下一跳发送.

③ 接收 IPv6 封装报文，先解封装，然后进行 IPv4 转发处理.

(7) 封装/解封装(Encapsulation & Decapsulation). 将 IPv4 报文加上 IPv6 头封装成 IPv6 报文，或者将 IPv6 封装报文解封装成 IPv4 报文.

4.2 虚链路管理

基于 VegaNet 的网络实验可能需要获取与虚链路相关的数据，如虚链路两端往返时间等. 另外，对于虚链路实际转发路径或转发跳数的测量有助于实现 VegaNet 与 CERNET2 的网络故障同步. VegaNet 通过在 VR 上部署 VLSP 协议来实现以上目标，具体包括以下子功能：

(1) 双向通信可达性检测. 对于虚链路，传统设计中通过路由可达性确定链路通信的有效性，而且多为单向检测. VegaNet 中，VLSP 协议通过发送请求报文并接收应答报文，达到双向通信可达性检测的目的.

(2) 往返时间测量. 仍然利用 VLSP 协议的请求、响应报文完成，每一个响应报文总是和一个请求报文对应，它们的 Seq Num 相等. 发送请求报文时 VR 记下发送时间，所以收到应答报文的一端通过计算请求报文发送时间与应答报文的到达时间的差值就能得到虚链路的往返时间.

(3) 虚链路实际转发跳数测量. 强制 VLSP 协议请求报文的初始跳数总为一固定值(如 64)，虚链路两端可根据接收的协议请求报文的到达跳数，获得虚链路到达方向的实际转发跳数.

(4) 精确路径跟踪. 在 VLSP 功能模块中集成 IPv6 Traceroute 机制.

(5) 参数协商. 同一条虚链路两端的虚接口可以协商接口 MTU、Bandwidth 以及进行地址冲突及同一网段的检测. 另外，虚链路加密传输时还可协商密钥和加密算法.

VLSP 协议报文交互机制非常简单：VR 定时向虚链路对端发送 VLSP 请求报文；若 VR 从某个虚接口收到 VLSP 请求报文，则按上述机制回送应答报文，每一个应答报文唯一对应一个请求报文. VLSP 报文格式如图 8 所示. VLSP 协议作为 VegaNet 中虚拟链路的链路层协议，服务的对象是 Ipv4 虚拟网络，但其协议数据直接紧跟 Ipv6 包头. VLSP 协议现有设计已能完成上述前 4 项功能. 当然，最后一项功能利用保留字段并对协议进行简单扩展后也能完成.

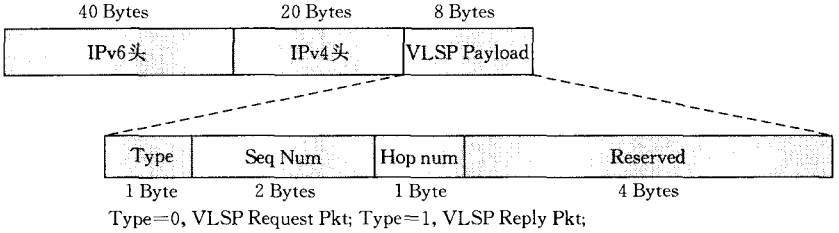


图 8　VLSP 协议报文格式

5 部署及实验

虚拟路由器系统现已完成 V1.0 版本的研发. 实现的功能包括虚接口的配置管理、VLSP 协议、基于虚链路的 OSPF 协议及 BGP 协议、硬件实现 VegaNet 报文的快速转发、远程对 VegaNet 网络的故障注入及故障消除. 而且，我们在 CERNET2 清华大学校园网部署了 3 台虚拟路由器进行实验. CERNET2 是一个纯 IPv6 网络，2004 年开通，现已在清华大学校园实现了全面覆盖，并且日常有大量实验或应用流量出入. 图 9 左边部分是虚拟路由器部署的物理连接拓扑，VR1 直连清华大学第 2 边界路由器，VR2 通过 3 跳连接清华大学第 1 边界路由器，VR3 通过 3 跳连接清华大学核心路由器，VR 接入 CERNET2 全是通过 100Mb/s 以太网链路进行

IPv6 连接. 在每个 VR 下面通过 100Mb/s 以太网链路连接一台实验 PC 机，为 IPv4 连接. 图 9 右边部分就是以左边网络连接为基础生成的虚拟网络拓扑，它是一个品字形的 IPv4 连接拓扑. 可知，Vlink1、Vlink2 和 Vlink3 分别对应着 5 跳、5 跳、7 跳物理链路. 由于条件有限，初步实验中 VR 只能百兆接入 CERNET2，3 台 PC 的硬件配置均为 Pentium 1.5GHz 处理器、512MB 内存，安装 Windows XP 系统. 为了对照，我们在实验室也搭建了对应的简化 VegaNet 网络拓扑进行完全一样的实验，如图 10 所示. 实验室网络环境中，PC 机到虚拟路由器是百兆连接，其它都是千兆连接.

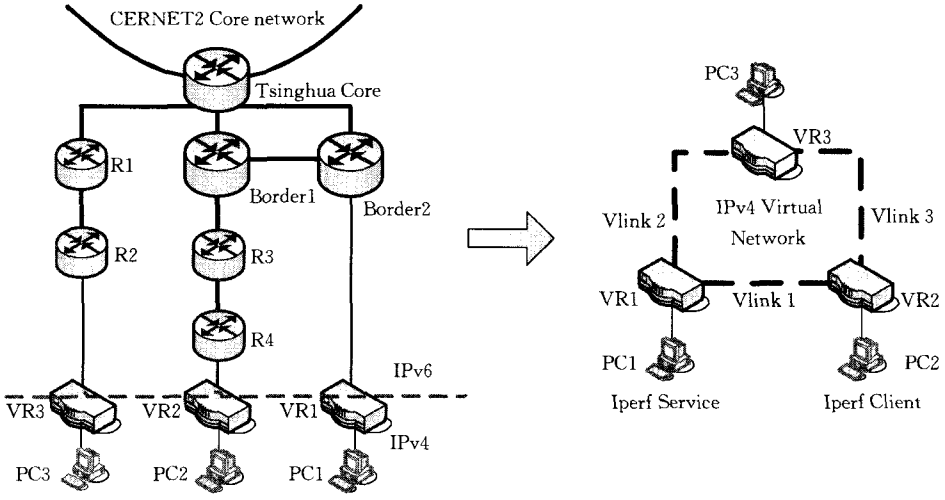


图 9 VegaNet 清华校园网网络拓扑

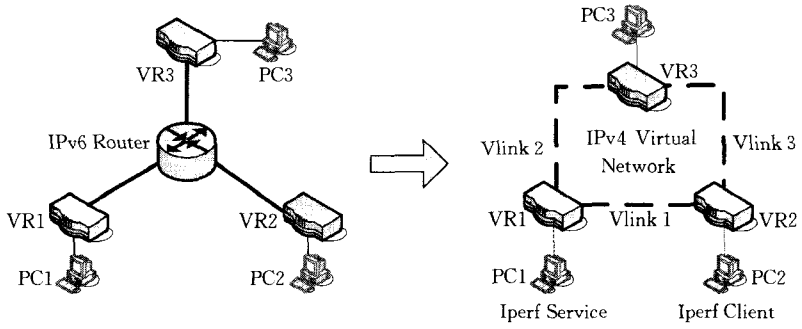


图 10 VegaNet 实验室网络拓扑

实验 1. 我们通过 Iperf 工具进行虚拟网络上的端系统之间的 UDP 带宽和 Jitter 值测试. PC1、PC2 分别充当 Iperf 服务器端和 Iperf 客服端. 首先，只在 VR1 与 VR2 之间运行 OSPF 协议，测量单跳虚链路通信相关数据，对应的 Iperf 实验路径为

PC1<—>VR1<—>VR2 <—>PC2.

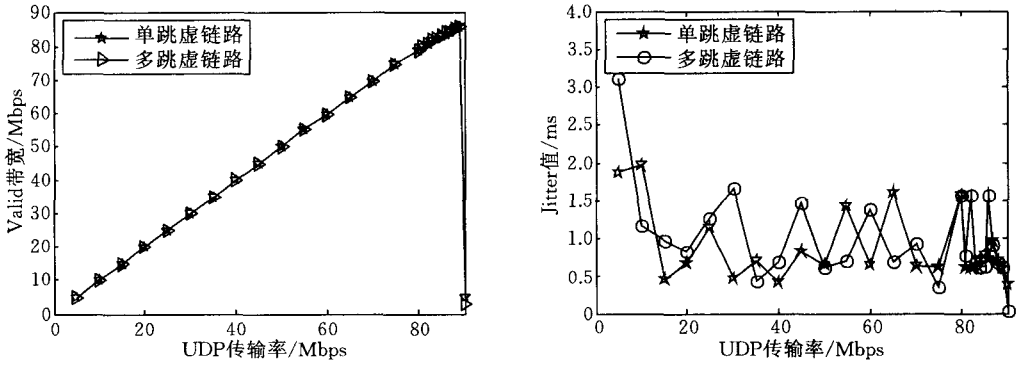
接着，只在 VR1 与 VR3 以及 VR2 与 VR3 之间运行 OSPF 协议，测量多跳虚链路通信的测量值，对应的 Iperf 实验路径为

PC1 <—>VR1 <—>VR3 <—>VR2 <—>PC2.

测试中 Iperf 配置带宽依次增加实验流量. 图 11 是 CERNET2 环境下的测试结果. 当流量配置超过 90Mbps，发现 iperf 显示有效测试流量急剧下降. 所以，90Mbps 流量是实验中端系统所能发出的流量上限. 而且，虚拟网络可以为百兆链路下连的端系统提供线速 UDP 带宽，且无论单跳虚链路或多跳虚链路，UDP 带宽测试结果基本一致. CERNET2 环境下端到端 UDP 通信中平均 Jitter 值测量结果都在 3.2ms 以内变化. 图 12 是实验室环境下的测试结果，UDP 带宽测试结果与 CERNET2 环境类似，但 Jitter 值变化范围缩小在 0～1.7ms，说明 VegaNet 实际网络部署时所面对的网络环境流量更为复杂.

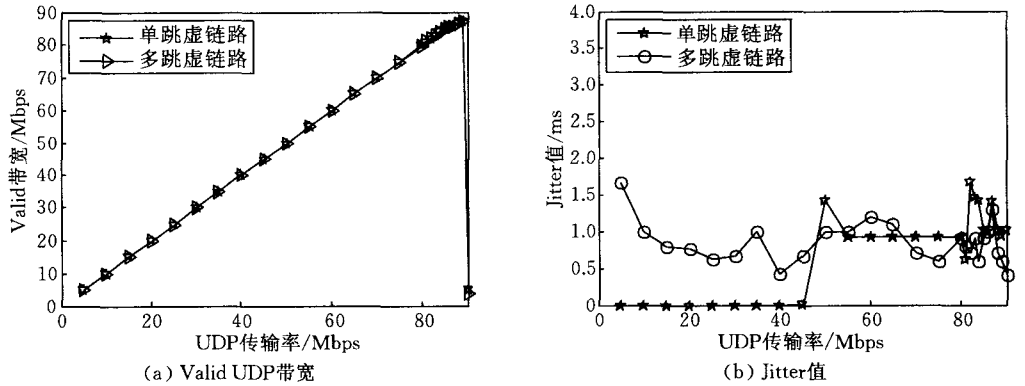
实验 2. 我们分析 OSPF 协议在虚拟网络上的运行状况以及运行状况变化对端到端流量的影响. 首先，在各虚拟路由器的所有接口上都使能 OSPF 协议，VR 之间相互学习路由. 实验中，利用 iperf 测试 PC1 到 PC2 之间的 TCP 吞吐量；同时，利用 hrPING 工具测试 PC1 到 PC2 的 RTT 值，hrPING 是一个高精度的 PING 工具. OSPF 稳定后开始测试数据，在第 50s 向虚拟网络注入 Vlink1 链路故障

事件，并在第 100s 注入 Vlink1 链路恢复正常事件，实验 2 总测试持续时间为 200s. 图 13 显示了实验结果，对于 iperf 端到端 RTT 值，可以看出前 50 多秒 RTT 都保持在 0.7ms 左右，Vlink1 断连后 RTT 在 1ms 左右，当 Vlink1 恢复正常后 RTT 又回落到 0.7ms. 同样，TCP 吞吐量刚开始能达到 90Mbps，随着虚链路故障发生而略有降低，故障恢复正常后 TCP 吞吐量也回到 90Mbps 以上. 初始情况中，OSPF 协议为 VR1 与 VR2 选择的最短路径为 1 跳虚链路路径：Vlink1；当 Vlink1 故障发生，OSPF 收敛后为 VR1 与 VR2 选择的最短路径为 2 跳虚链路路径：Vlink2—Vlink3；当故障恢复正常，该通信又回到 1 跳虚链路路径. 此过程中，虽然虚拟链路跳数只有增、减 1 跳的变化，但结合图 9 可知其对应的物理链路跳数变化为 7 跳. 所以，无论是端到端 RTT 值，或是 TCP 吞吐量，都随着故障变化而有着较为明显的变化. 另外，两个子实验数据都显示：当在 50s 时刻注入链路故障后，端到端通信在几秒之内就立刻受到影响，而在 100s 时刻恢复正常后，大约在几十秒后端到端通信才恢复到故障前的状态. 这是因为，链路故障发生后 OSPF 可以立刻选择备用路径通知数据层面，而链路正常后OSPF要经过邻



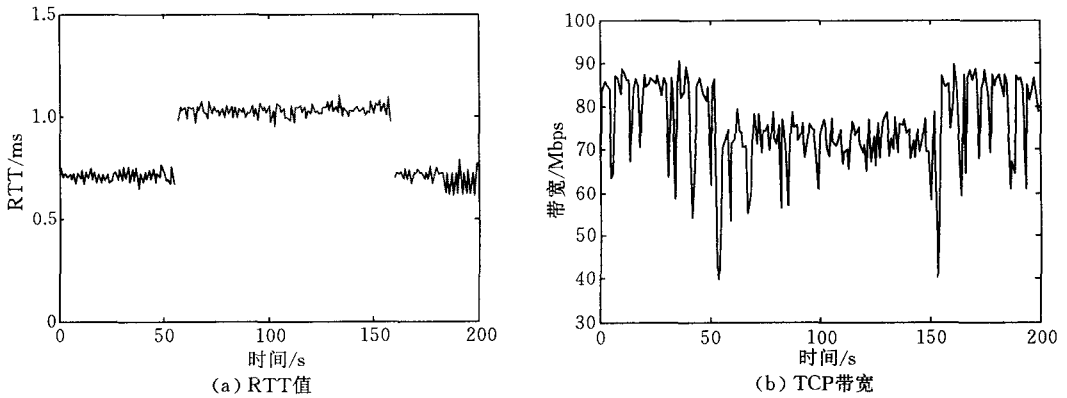
(a) Valid UDP 带宽　　(b) Jitter值

图 11　基于 CERNET2 的 VegaNet 中带宽和 Jitter 值测量



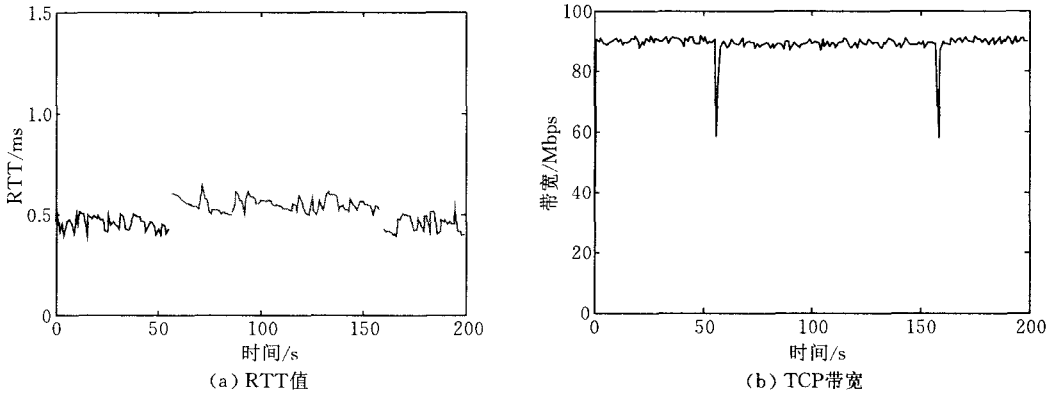
(a) Valid UDP带宽　　(b) Jitter值

图 12　实验室环境中 VegaNet 带宽和 Jitter 值测量



(a) RTT值　　(b) TCP带宽

图 13　基于 CERNET2 的 VegaNet 中 RTT 和 TCP 带宽测量



(a) RTT值 (b) TCP带宽

图 14 实验室环境 VegaNet 中 RTT 和 TCP 带宽测量

居建连并互发 UPDATE 报文才能学到新的路由，该过程耗时较长. 这与真实网络中 OSPF 运行情况完全一致. 同样的实验也在图 10 所示的实验室环境下完成，图 14 是测试结果，RTT 测量结果明显低于 CERNET2 环境，而且 RTT 值和 TCP 带宽在 OSPF 链路故障收敛过程的变化不明显.

另外，我们在图 10 环境中测试了虚拟路由器的极限转发能力. IXIA 测试仪的两个千兆接口分别连接 *VR*1 和 *VR*2 的千兆接口，形成一个全千兆连接的测试环境. 我们测试环境下 *VR*1 到 *VR*2 的 IPv4 数据转发能力. 测试结果如表 3 所示，平均能达到近 800Mbps 的 IPv4 报文转发性能. 由于 VegaNet 中报文转发过程，需要 IPv6 头部封装，不可避免会有相应带宽的损耗. 所以，VegaNet 基本达到线速转发能力.

表 3 VR 转发性能测试

包长/Bytes	吞吐率/Mbps
64	516.13
160	650.41
260	764.71
360	786.03
460	815.60
512	863.60
660	863.80
1000	866.55
1200	867.35
1300	867.42
1478	867.80
64～1478(随机)	798.14

最后，我们还利用图 9 环境中的端系统，进行了视频播放、FTP 等常规应用实验，都能正常完成. 综合所有实验，我们可以得出以下结论：

(1) VegaNet 网络能够为用户提供真实、高速的端到端流量转发.

(2) VegaNet 网络中路由协议运行情况及其对转发流量的影响与实际网络基本相同.

(3) 部署于实际网络的 VegaNet，能够让网络实验面对象实际网络一样的复杂环境.

(4) VegaNet 支持人为注入链路故障事件及故障恢复事件.

6 总结及下一步工作

本文所提出 VegaNet 的设计目标是为新型网络协议实验提供一个架构平台以及对运营网络进行模拟仿真并展开网络性能研究和分析. 本文首先提出了通用的 VegaNet 模型，并且分析了 VageNet 虚链路的属性. VageNet 通过分析虚链路的实际转发跳数来决定的链路状态，VegaNet 与 CERNET2 的故障同步就是以此为据设计的. 此外，介绍了一种以 CERNET2 为底层物理网络的 VegaNet 设计实例及部署方案，包括具体的地址配置、路由交互、报文封装及数据转发过程. VegaNet 的核心设备：虚拟路由器，是以 BWOS 平台为基础而开发的，本文给出了优化后的系统体系结构及重要模块的设计细节. 本文还设计了 VLSP 协议，实现虚拟链路两端的参数协商、可达性检测、路径跟踪等功能.

目前，虚拟路由器已完成 V1.0 版本的研发. 实现功能包括虚接口的配置管理、VLSP 协议、基于虚链路的 OSPF 协议及 BGP 协议、硬件实现 VegaNet 报文的快速转发、远程对 VegaNet 网络的故障注入及故障恢复. 此外，我们已在 CERNET2 清华大学校园网内部部署了 3 台虚拟路由器进行实验. 它运行在真正的商用路由平台上，承载着真实、高速的用户流量，面对着如真实网络一样的复杂情况；而且还支持人为向 VegaNet 注入链路故障事件及故障恢复事件. 实验结果说明 VegaNet 网络可以为用户提供几尽真实的网络实验及模拟分析环境.

VegaNet 下一步工作将在以下几个方面进行：(1) 大范围部署，设计激励机制引入更多真实流量；(2) 新型路由协议的实施，课题组提出的自愈路由协议将以 VegaNet 为实验网络环境验证其改进特性；(3) 对 CERNET2 进行深度监控及分析，设计实验方

案、收集实验数据，并对 CERNET2 网络体系结构、协议运行、运营流量进行研究分析，并提出改进建议；(4)改善虚拟网络及虚拟设备设计. 改进虚拟路由器报文转发能力，从而提高整个虚拟网络的带宽；VegaNet 及 VR 将支持更丰富的管理员注入事件.

致　谢　本论文工作在清华大学完成，感谢清华大学计算机网络研究所提供的研究项目支持！

参　考　文　献

[1] Andersen D, Balakrishnan H, Kaashoek F, Morris R. Resilient overlay network//Proceedings of the 18th ACM Symposium on Operating Systems Principles (SOSP). Banff, Canada, 2001: 131-145

[2] Duan Z H, Zhang Z L, Hou Y T. Service overlay networks: SLA, QoS, and bandwidth provisioning. IEEE/ACM Transactions on Networking, 2003, 11(6): 870-883

[3] Chandrashekar J, Zhang Z L, Hou Y T, Duan Z H. Service oriented Internet, towards a service-oriented Internet. IEICE Transactions on Communications, Special Section on Networking Technologies for Overlay Networks, 2006, E89-B(9): 2292-2299

[4] Bavier A, Feamster N, Huang M, Peterson L, Rexford J. In VINI veritas: Realistic and controlled network experimentation//Proceedings of the ACM SIGCOMM 2006 Conference. Pisa, Italy, 2006: 3-14

[5] Peterson L, Anderson T, Culler D, Roscoe T. A blueprint for introducing disruptive technology into the Internet//Proceedings of the 1st ACM Workshop on Hot Topics in Networking (HotNets-I). Princeton, New Jersey, USA, 2002

[6] Bavier A, Bowman M, Culler D, Chun B, Karlin S, Muir S, Peterson L, Roscoe T, Spalink T, Wawrzoniak M. Operating system support for planetary-scale network services//Proceedings of the 1st Symposium on Networked Systems Design and Implementation. San, Francisco, CA, USA, 2004

[7] Handley M, Hodchild O, Kohler E. XORP: An open platform for network research//Proceedings of the 1st Workshop on Hot Topics in Networks (HotNets-I). Princton, New Jersey, USA, 2002

[8] Morris R, Kohler E, Jannotti J, Kaashoek M F. The click modular router//Proceedings of the 17th Symposium on Operating Systems Principles (SOSP'99). Kiawah Island, SC, USA, 1999: 217－231.

[9] Wu Jian-Ping, Cui Yong. Development of BE12000 IPv6 core router. Telecommunications Science, 2005, 21(1): 18-23 (in Chinese)
(吴建平，崔勇. BE12000 系列 IPv6 核心路由器的研制. 电信科学，2005，21(1)：18-23)

CHEN Wen-Long, born in 1976, Ph. D. candidate. His research interests include network protocol and network architecture.

XU Ming-Wei, born in 1971, Ph. D., professor. His research interests include network architecture, high-performance router architecture and protocol test.

YANG Yang, born in 1955, Ph. D., professor. His research interests include network protocol and network architecture.

LI Qi, born in 1979, Ph. D. candidate. His research interests include network protocol and network architecture.

MA Dong-Chao, born in 1980, Ph. D. candidate. His research interests focus on computer network.

Background

This research is supported by the National Basic Research Program (973 Program) of China under grant No. 2009CB320502; the National High Technology Research and Development Program (863 Program) of China under grant Nos. 2007AA01Z2a2, 2007AA01Z234; 2009AA01Z251, and the National Natural Science Foundation of China under grant No. 60873192.

Virtual network is an important experimental infrastructure for network research. Most existing virtual networks[1-3] focus on study of failure detection and recovery, QoS, etc. They can not be used for evaluations of routing protocols. VINI[4] is proposed to address this issue and has been deployed on planetlab. However, VINI still have some shortcoming, such as network failure exposure and topology changes, failure control and low network performance. In this paper, a high performance virtual Network, VegaNet, is proposed. VegaNet realizes a network experimental infrastructure through which network researchers can evaluate their network protocols and services within a realistic environment. In addition, VegaNet can simulate current core networks and arbitrarily configure virtual networks. Thus, we can effectively evaluate improvements to current networks.

Virtual routers play an important role in virtual networks. Lots of study on virtual routers had come forth. There were different definitions and implementation of virtual router in the community. The main character of the authors' virtual router is that all virtual routers are transparent for substrate networks. Virtual network composed of these virtual routers can achieve the following goals: running real routing software, exposing realistic network conditions, controlling network events, carrying real user traffic.

As key part of the projects supported 973 and 863 programs, VegaNet is regarded as a basic network infrastructure. The authors will implement the proposed new/improved routing protocols in virtual routers and evaluate their performance in VegaNet.