嵌入虚拟机监视器的高性能虚拟网络

唐 源[1,2], 夏 磊[2], 崔 峥[3], John Lange[2], Peter Dinda[2], Patrick Bridges[3], 李建平[1]

(1 电子科技大学计算机科学与工程学院 成都 611731 中国;
2 西北大学电子工程与计算机学院 埃文斯顿 60201 美国;
3 新墨西哥大学计算机学院 阿布奎基 87131 美国)

摘 要: 由虚拟机和虚拟覆盖网相结合而构成的虚拟计算环境,在云计算和绿色计算中发挥非常重要的作用,然而,现有的虚拟计算环境在性能上难以满足高性能分布式计算的要求。设计和实现了一种高性能虚拟网络:VNET/P。基于网络第2层建立虚拟机互连模型,对一组虚拟机进行抽象,使其位于同一局域网中。与早期的用户层虚拟网络系统不同,VNET/P嵌入于可扩展、高性能的Palacios虚拟机监视器中。试验结果表明,VNET/P具有高带宽的特点,其性能接近于实际硬件的性能。

关键词: 虚拟化;虚拟机监视器;覆盖网络;高性能计算
中图分类号: TP338 **文献标识码:** A **国家标准学科分类代码:** 520.30

High performance virtual network embedding virtual machine monitor

Tang Yuan[1,2], Xia Lei[2], Cui Zheng[3], John Lange[2], Peter Dinda[2], Patrick Bridges[3], Li Jingping[1]

(*1 School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China;*
2 Department of EECS, Northwestern University, Evanston, IL 60201, USA;
3 Department of CS, University of New Mexico, Albuquerque, NM 87131, USA)

Abstract: The virtual computing environment integrated with virtual machines (VMs) and virtual overlay networks possesses many important advantages in cloud computing and green computing. However, the un-negligible overhead of existing virtual networks prevents their vast use in high-performance distributed computing. This paper has designed and implemented a high-speed virtual network system VNET/P. VNET/P is a layer 2 network abstract for virtual machines, which enables a distributed collection of virtual machines to be maintained in a location independent way across wide-area network. Different from the existing virtual network in user layer, VNET/P is embedded into an open-sourced Palacios virtual machine monitor (VMM), has much higher bandwidth and performance scalability, and achieves near-native performance and negligible overhead on 1Gbps Ethernet.

Key words: virtualization; virtual machine monitor; overlay network; high performance computing

1 引 言

近年来,随着互联网的快速发展,网络中的各种计算资源、存储资源以及服务资源等日益丰富,导致管理成本和能量消耗上升,而可靠性和资源利用率降低。然而实际应用中迫切需要将这些孤立的、异构的资源进行聚合,实现在广域、动态环境下的资源共享与协同工作,因此虚拟化(virtualization)、云计算(cloud computing)和绿色计算(green computing)这类满足超大规模、低成本和可扩展性的技术应运而生。目前,如何在虚拟环境下实现具有可扩展性(scalability)、高可靠性(reliability)、热迁移性(live migration)和低功耗的高性能分布式计算,已成为当前国际学术界和业界的热点研究问题。

虚拟机监视器(virtual machine monitor, VMM)提供的性能越来越接近实际硬件的性能,虚拟技术逐步引入

收稿日期:2011-09 Received Date:2011-09

到高性能分布式计算领域[1]。各项研究开始将虚拟机(virtual machine, VM)和广域环境下的覆盖网络(overlay)结合起来,实现以虚拟机和虚拟覆盖网为基础的广域虚拟计算环境,解决主机分散和网络异构的问题。典型的项目有:VIOLIN[2]、IPOP[3]以及本课题组早期研究的 Virtuoso[4]等。虚拟覆盖网络 VNET[5-7],是 Virtuoso 项目的重要组成部分,它实现于操作系统的用户层,网络带宽为 21.5 MB/s[7],虽然可以满足广域网的应用,但远不能达到高性能计算的要求。因此,本文设计和实现了一种新的嵌入于 Palacios VMM[8-10]中的虚拟网络系统,称之为 VNET/P。VNET/P 的网络开销(overhead)足够小,其性能接近于实际硬件的性能,因此在当前的云计算和绿色计算等工程应用中,可以将 VNET/P 引入紧耦合的集群,或存储分布的并行计算环境。同时,仪器仪表行业也逐步走向智能化、网络化和虚拟化[11-12]。本文设计的虚拟网络为计算机技术和仪表技术搭建了一座桥梁,可对当前的仪器仪表进行网络化和虚拟化扩展,用于远程测控、汽车电子[13]、核电站设备监测[14]等领域。特别是在云计算技术出现的今天,虚拟网络为仪器仪表利用云资源提供了实现基础。

2 相关研究

目前在高性能分布式计算和云计算中,越来越多的使用到虚拟技术,尤其在解决资源分散、管理复杂、性能瓶颈等问题上,将虚拟机和虚拟覆盖网络整合起来构成的虚拟计算环境具有非常突出的优势。

2.1 广域分布式平台的虚拟计算环境

结合虚拟机和虚拟覆盖网络的研究多数集中在操作系统用户层,其中典型的工作有 VIOLIN、IPOP、WOW、ViNe 和 Virtuoso。Purdue 大学的虚拟网格项目 VIOLIN[2],其目标是建立一个基于虚拟服务和虚拟网络的,并按需提供服务的网格架构。另一种虚拟网络是 Florida 大学的 IPOP[3],在 P2P 网络之上建立一个虚拟 IP 网络层,支持在 IP 管道里透明地进行 UDP 和 TCP 传输。此外,该项目组还进一步研究了 WOW[15],针对高吞吐量的计算建立一种基于虚拟工作站的可扩展的广域网络。类似的工作还有 ViNe[16],在同一基础设施上支持私有网络和多个独立的虚拟网络。Virtuoso 是本课题组早期研究的项目,其目标是建立虚拟机网格计算环境。Virtuoso 的本质任务是管理和控制分散于远程的虚拟机,并对自适应算法进行优化,文献[4]对 Virtuoso 的工作原理进行了详尽描述。

2.2 PALACIOS 虚拟机监视器

VNET/P 以模块方式嵌入 Palacios VMM。Palacios 由美国 Northwestern 大学和 New Mexico 大学合作研究,其目的是针对现代计算机结构而研制的应用于高性能计算的 VMM。目前,Palacios 支持具有硬件虚拟化特性(AMD SVM, Intel VT)的 x86 和 x86_64 等 CPU 结构,并且植入 Sandia 美国国家试验室开发的 Kitten 操作系统中。Palacios 设计的动机是应用于高性能计算环境,具有全系统虚拟、连续内存预分配、Passthrough 资源和干扰(noise)小等设计特点。本课题组在 Red Strom 超级计算机平台上,对 Palacios 和 Kitten 构成的系统性能进行了测试。其结果显示,与实际硬件性能相比,Palacios 带来的性能损失小于 5%[8],非常接近实际硬件的性能。

3 VNET/P 设计和实现

下面对 VNET/P 的设计和实现细节进行阐述。

3.1 VNET/P 结构

嵌入于 Palacios VMM 的 VNET/P 基本结构和数据交换过程如图 1 所示,它支持在同一物理机上运行多个虚拟机,其中 Dom0 是一个特殊的虚拟机,运行在该虚拟机上的 guest 操作系统,有访问物理主机全部设备的权限。VNET/P 使用 Dom0 作为后端(backend)服务,通过 Dom0 中的 Passthrough 网卡与 VMM 外界交流数据包,这样 Dom0 就成为了 VNET/P 和 host 之间的桥梁。非 Dom0 的其他虚拟机,称之为 DomU 虚拟机。Palacios 为每个 DomU 虚拟机创建一个 Virtio[17]虚拟网络设备,作为 DomU 的以太网卡。所有通过 Virtio 设备发送的网络包,首先传入 Palacios,然后 VNET/P 在 Palacios 里为数据包选择路由,决定将其传递到同一台物理机的其他 DomU 虚拟机的 Virtio 设备,或是通过 Dom0 虚拟机的桥传递到外部网络。

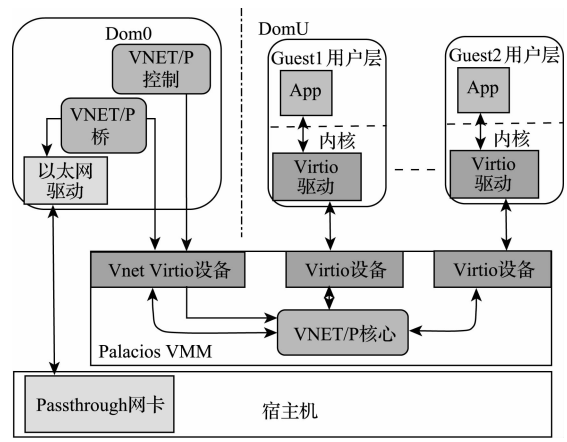


图 1 VNET/P 结构

Fig. 1 VNET/P structure

VNET/P 包含 3 个主要的组成部分:1) 嵌入 Palacios

VMM 的核心部分(VNET/P Core),它负责包的封装和路由;2) 桥(VNET/P Bridge)和控制(VNET/P Control)部分,位于 Dom0 的 guest 操作系统中;3) Virtio 虚拟设备。下面介绍 VNET/P 的关键技术。

3.2 VNET/P 核心

VNET/P 核心的基本功能是一个封包、路由和转发的过程。转发依据是基于网络第 2 层的 MAC 地址,转发目的地需要根据用户设置的路由规则进行路由。VNET/P 核心处理数据包的流程如图 2 所示,详细过程如下:

1) 接收数据包。VNET/P 将转发其接收到的所有包,包括从外部网络通过 Dom0 虚拟机桥(Vnet Virtio Device)进入 VNET/P 的包;从 DomU 虚拟机的虚拟网络设备(Virtual NIC)产生的包;或者从 VMM 的 Virtio 虚拟设备(Virtio device)获得的包。

2) VNET/P 核心将收到的数据包放到内部队列里(packets queue),采用先进先出的顺序访问数据包。

3) 包分配器(Packet Dispatcher)在路由表中查找表项,决定转发路径。路由表是用户通过 VNET/P Control 进行初始化的,可在运行时动态改变。

4) 根据收到包的来源和目的不同,将数据包进行拆封与封装。

5) 最后根据转发路径,将数据包发送到 Vnet Virtio Device、Virtual NIC 或 Virtio Device 等设备。

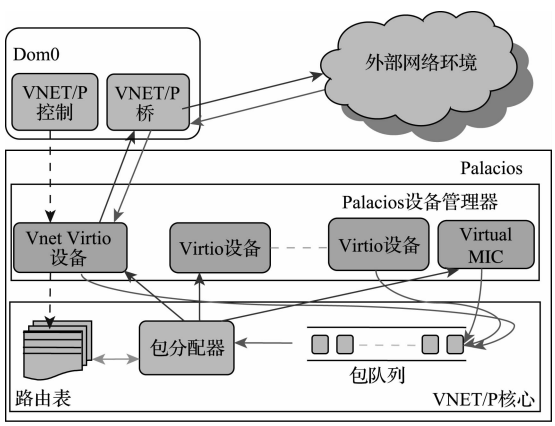


图 2　VNET/P 核心

Fig. 2 VNET/P core

3.2.1 路由(Routing)

VNET/P 的路由基于网络第 2 层,需要源虚拟机和目的虚拟机的 MAC 地址,以及虚拟设备接口(interface)或连接(link)的某些选项(如源和目的地址类型、目的地址优先级)等参数。VNET/P 核心路由时使用的转发表存储在 Palacios 里,其简略的路由过程如下:

1) 首先从虚拟接口捕获包,或是从某 link 上接收包,然后分析包的结构,得到路由必须的参数。

2) 按照一定的规则(在 VNET/P 控制部分设置转发规则),在路由表中匹配对应的表项。

3) 根据路由表中匹配而得的 interface/link 号,再从 interface 表或 link 表获得对应的目标 IP 或本机虚拟设备接口,即是要转发的目的。

4) 根据包的不同来源和目的,重新对包进行拆封与封装,最后将处理完成的包转发到目的 interface/link。

3.2.2 VNET/P Virtio 设备

VNET/P Virtio 有 2 种结构,一种结构前端为 Dom0 中的 VNET/P Bridge 和 Control,后端为 Palacios 中的 Vnet Virtio Device;另一种结构前端为 DomU 中的 Virtio Driver,后端为 Palacios 中的 Virtio Device。Vnet Virtio Device 一方面负责从 VNET/P Bridge 和 Control 获得的数据包和控制信息,并传递到 VNET/P 核心;另一方面负责从 VNET/P 核心获得的数据包,再传递到 VNET/P Bridge。Virtio Driver 和 Virtio Device 分别是针对 DomU 的 Virtio 虚拟网络设备的前端和后端,负责 DomU 和 Palacios 之间的数据交换。

3.3 Dom0 虚拟机

在 Dom0 虚拟机中主要实现了以 Passthrough 以太网卡来接收和发送数据的设备,以及以内核模块方式实现的 VNET/P 桥和控制。Vnet Virtio 设备接收来自 2 个方向的数据包:一方面是 VNET/P 桥传入的 IP 包;另一方面是 VNET/P 核心发送的以太包,其目的地址是本物理主机之外的主机。接收的以太包经过 VNET/P 核心封装,含有目的 link 信息,该包再通过 Vnet Virtio 设备传送到 VNET/P 桥。VNET/P 桥首先根据以太包中的 link 信息,查找对应的目的 IP 地址等信息,然后再通过 Passthrough 以太设备发送到目的主机的 VNET/P 端。VNET/P 桥从外网接收到数据,以上述相反的过程处理,其数据包传输流程如图 2 所示。VNET/P 桥根据目的主机在本地网或远程网的不同分 2 种:

1) 直接桥(Direct Bridge):发包过程中,假如包的目标主机和源主机在相同的局域网中,包不再需要进行封装,而直接发送到以太网中。该方式所有 Dom0 虚拟机的 Passthrough 以太网卡工作在混杂模式(promiscuous)。

2) UDP 封装(Encapsulation):假如包的目的主机和源主机不在同一局域网中, VNET/P 桥对以太包进行 UDP 封装,再以 UDP socket 的方式送到外部网络。

4 性能测试

VNET/P 系统的主要目的是使其构成的 overlay 具有最小化的开销,满足基于紧耦合集群或局域网的高性能计算的需要。由于 overlay 的性能在 Internet 上受到网络低带宽高延迟的限制,不能充分反映 VNET/P 的真实性能,而且 VNET/P 侧重于集群和云计算中心等高性能计

算,所以考虑在局域网中测试其性能。

4.1 测试环境

测试环境如图 3 所示,两个物理机(host)以 1 Gb/s 带宽的网卡通过以太网互连,host 主机具有 1.86GHz 双核 Intel Xeon 3400 处理器,1G 内存。Dom0 虚拟机使用 512M 内存,DomU 虚拟机使用 384M 内存。Dom0 和 DomU 分别运行于处理器的一个核中。实验中分别采用了 4 种测量方式:本地硬件(Native)、Passthrough 穿透、未进行 UDP 封装的直接桥方式和 UDP 封装方式。

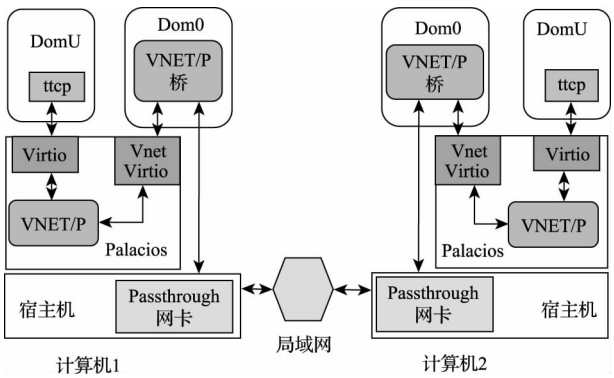


图 3 VNET/P 测试环境

Fig. 3 VNET/P test setup

1) Native 方式,两个物理机没有运行 Palacios/Kitten 和虚拟机,而是在硬件上直接运行 Linux,通过在 Linux 中执行 ttcp 工具测量两个物理机之间的网络性能。

2) Passthrough 方式,ttcp 分别运行于 Dom0 虚拟机中。发包和收包过程没有经过的处理,直接由 Passthrough 连接 Dom0 和以太网。

3) 直接桥和 UDP 封装方式,ttcp 分别运行于 DomU 虚拟机中,由 VNET/P 来处理包的发送和接收过程。

使用 ICMP 请求/响应的往返时延(round-trip delay),来计算平均延迟(使用 ping 命令);使用 ttcp 工具来测量平均通量,在 TCP 测试中,设置发送和接收窗口为 64 KB,而测试 UDP 时,设置发送的 UDP 包大小为 1 400 B。

4.2 结果分析

在带宽为 1 Gb/s 的局域网环境中,分别对上述 4 种方式的网络延迟和通量进行测试,每种情况测量若干组数据,取其平均值。

4.2.1 延迟

图 4 中,Native 或 Passthrough 的延迟分别是 0.14 ms 和 0.18 ms,Passthrough 的延迟非常接近 Native,可见 Palacios 及其 Passthrough 的性能很高,几乎不再额外消耗时间。直接桥和 UDP 封装分别是 0.303 ms 和 0.321 ms,其差距也相当小,由此可知 VNET/P 对包作 UDP 封装的速度很快。但直接桥或 UDP 封装方式的延迟与 Native 或 Passthrough 的延迟相比差距却较大,这对了解数据包收发过程时间消耗的细节很重要。

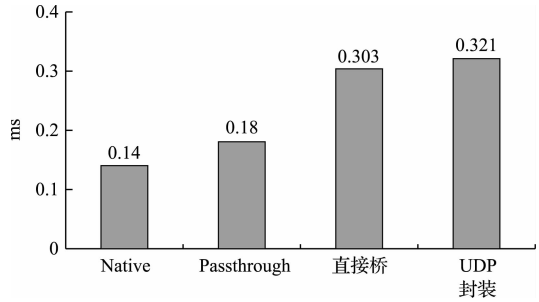


图 4 VNET/P 端到端延迟

Fig. 4 End to end latency of VNET/P

参考图 3,Machine-1 的 DomU 虚拟机的 guest 操作系统发送一个 ICMP 请求到 Machine-2 的 DomU 虚拟机的 guest 操作系统,其经过如下:

1) ICMP 请求从 Machine-1 的 DomU guest 通过 Virtio 发送到 VNET/P 核心。

2) 包在 VNET/P 核心进行路由、拆封和封装。

3) 包从 VNET/P 通过 Virtio 传递到 Dom0 的 VNET/P 桥。

4) 最后再由 Passthrough 发送到局域网。

5) 数据在局域网上传播。

6) Machine-2 的 Passthrough 网卡收到包,将其传递到 Dom0 的 VNET/P 桥。

7) VNET/P 桥将包通过 Virtio 转发到 VNET/P 核心。

8) 与第 2 步一样,VNET/P 核心对包进行路由、拆封和封装。

9) 最后 ICMP 请求通过 Virtio 发送到了 Machine-2 的 DomU guest 操作系统。

10) 此时 Machine-2 的 DomU guest 进行 ICMP 响应,它从 Machine-2 传送到 Machine-1,其经过与步骤 1 ~ 9 相似。

ICMP 请求/响应的往返时延分别包括步骤 2、3、7、8 各两次,与 Passthrough 方式相比,UDP 封装方式多余的时延(0.321 - 0.18 = 0.141 ms)主要产生在这几个步骤中。部分多余的时延还来自于步骤 1,ICMP 请求从 DomU guest 操作系统到 VNET/P 核心所消耗时间,比 Passthrough 方式中 Dom0 guest 发送 ICMP 请求到 Passthrough 网卡的时间要长,这依赖于 Palacios 和 guest 的性能。

4.2.2 通量

使用 ttcp 测试 TCP 包时,发送和接收窗口均设为 64 KB,而测试 UDP 时,UDP 包设为 1 400 B。从图 5 中可知,无论测试 UDP 还是 TCP 包,其通量按照 Native、Passthrough、直接桥、UDP 封装方式递减,但其间的差距

均很小,这是因为传输时间的消耗主要集中在对数据的拷贝、发送和接收上,而不是在 VNET/P 对包的路由、拆封和封装等操作上,这对大数据量传输来说是很重要的特性。由于 VNET/P 的目的在于局域网或集群中的高性能计算,所以我们更重视直接桥,这种不用进行 UDP 封装的方式,它的带宽很接近 Native。而且,在高性能计算中考虑更多的是大数据量的交换,直接桥方式接近Native的带宽能更好地满足要求,即使小数据量的延迟与Native相差较大,对高性能计算的影响可忽略。

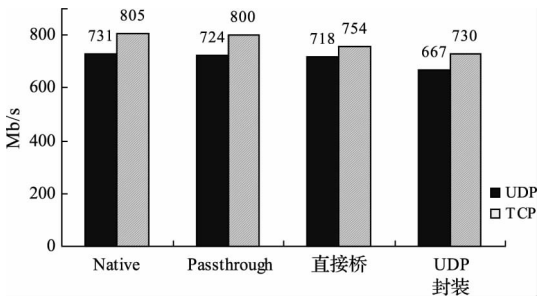


图 5　VNET/P 端到端通量

Fig. 5 End to end throughput of VNET/P

5 结　论

本文设计和实现了一种新的虚拟网络 VNET/P。VNET/P 一方面屏蔽广域自治网络中的异构问题,使用户可以使用局域网的所有技术来管理分布在广域网上的虚拟机;另一方面,由于 VNET/P 工作于 VMM 中,因此相对于用户层虚拟网络具有更小的系统开销和更高的性能。实验表明,VNET/P 具有接近于硬件的性能,并能达到高性能计算对网络带宽的要求。进一步的研究将增强 VNET/P 控制部分的功能,可以监视虚拟网络流量,随着网络性能、计算任务的变化,迁移虚拟机和动态调整网络拓扑。

参考文献

[1] MERGEN M F, UHLIG V, KRIEGER O, et al. Virtualization for high-performance computing[J]. Operating Systems Review,2006,40(2):8-11.

[2] JIANG X X,XU D Y. Violin: Virtual internetworking on overlay infrastructure[R]. Technical Report CSD TR 03-027, Purdue University, July 2003.

[3] GANGULY A, AGRAWAL A, BOYKIN P O, et al. IP over P2P: Enabling self-configuring virtual ip networks for grid computing[A]. Proceedings of the International Parallel and Distributed Processing Symposium [C]. 2006:1-10.

[4] SHOYKHET A, LANGE J, DINDA P. Virtuoso: A system for virtual machine marketplaces[R]. Technical Report NWU-CS-04-39, Northwestern University,2004.

[5] SUNDARARAJ A, DINDA P. Towards virtual networks for virtual machine grid computing[A]. Proceedings of the third USENIX Virtual Machine Research and Technology Symposium[C]. 2004:177-190.

[6] SUNDARARAJ A, GUPTA A, DINDA P. Increasing application performance in virtual environments through runtime inference and adaptation[A]. Proceedings of the 14th IEEE International Symposium on High Performance Distributed Computing[C]. 2005:47-58.

[7] LANGE J, DINDA P. Transparent network services via a virtual traffic layer for virtual machines[A]. Proceedings of the 16th IEEE International Symposium on High Performance Distributed Computing[C]. 2007:23-32.

[8] LANGE J, PEDRETTI K, HUDSON T, et al. Palacios and kitten: New high performance operating systems for scalable virtualized and native supercomputing[A]. Proceedings of the 24th IEEE International Parallel and Distributed Processing Symposium[C]. 2010:1-12.

[9] LANGE J, DINDA P. An introduction to the palacios virtual machine monitor—Release 1.0[R]. Technical Report NWU-EECS-08-11,Northwestern University, 2008.

[10] XIA L, LANGE J, DINDA P, et al. Investigating virtual passthrough I/O on commodity devices[J]. Operating Systems Review,2009,43(3):83-94.

[11] 罗秋凤,肖前贵,杨柳庆. 无人机自动检测系统的设计与实现[J]. 仪器仪表学报, 2011, 32(1):126-131.
LUO Q F, XIAO Q G, YANG L Q. Design and implement of automatic test system for UAV[J]. Chinese Journal of Scientific Instrument, 2011,32(1): 126-131.

[12] 曲良东,刘衍珩,余雪岗,等. 基于虚拟设备的车载异构网络互联模型[J]. 仪器仪表学报, 2010, 31(8): 1904-1909.
QU L D, LIU Y H, YU X G, et al. In-vehicle heterogeneous network interconnection model based on virtual devices[J]. Chinese Journal of Scientific Instrument, 2010,31(8):1904-1909.

[13] 张利,路园,张建军,等. OSEK 网络管理在汽车 CAN 系统中研究与实现[J]. 电子测量与仪器学报, 2011, 25(6):522-527.
ZHANG L, LU Y, ZHANG J J, et al. Research and implementation of OSEK network management in automotive CAN systems[J]. Journal of Electronic Measurement and Instrument, 2011, 25(6): 522-527.

[14] 张衡,陈东义,凌健中. 核电站设备无线监测路由与通信资源分配算法[J]. 电子测量与仪器学报,2010, 24

(12):1101-1106.
ZHANG H, CHEN D Y, LING J ZH. Routing and communication resources scheduling algorithm in nuclear devices wireless monitoring[J]. Journal of Electronic Measurement and Instrument, 2010, 24(12):1101-1106.

[15] GANGULY A, AGRAWAL A, BOYKIN P O, et al. WOW: Self-organizing wide area overlay networks of virtual workstations[A]. Proceedings of the Fourteenth International Symposium on High Performance Distributed Computing[C]. 2006:30-42.

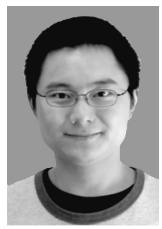
[16] TSUGAWA M, FORTES J A B. A virtual network (ViNe) architecture for grid computing[A]. Proceedings of the International Parallel and Distributed Processing Symposium[C]. 2006: 1212-1221.

[17] RUSTY R. Virtio: Towards a de-facto standard for virtual I/O devices[J]. Operating Systems Review, 2008, 42(5):95-103.

作者简介

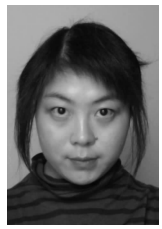
唐源,1999 年于河北金融学院毕业,2007 年于昆明理工大学获得硕士学位,现为电子科技大学在读博士研究生、美国西北大学联合培养博士生,主要研究方向为操作系统、虚拟技术和云计算。
E-mail: ytang2222@uestc.edu.cn

Tang Yuan graduated from Hebei Finance University in 1999, and got his master degree from Kunming University of Science and Technology in 2007. Now he is a Ph. D. student in University of Electronic Science and Technology of China, and a visiting scholar in Northwestern University, USA. His research has mainly focused on operating systems, virtualization and cloud computing.



夏磊,2004 年于南京大学获得学士学位,2007 年于南京大学获得硕士学位,现为美国西北大学在读博士研究生,主要研究方向为操作系统、高性能计算和云计算。
E-mail: lxia@northwestern.edu

Xia Lei got his bachelor and master degrees both from Nanjing University in 2004 and 2007, respectively. Now he is a Ph. D. student in Northwestern University, USA. His research has mainly focused on operating systems, high performance computing and cloud computing.



崔峥,2003 年于郑州大学获得学士学位,现为美国新墨西哥大学计算机系在读博士生,主要研究方向为虚拟技术、高性能计算和云计算。
E-mail: cuizheng@cs.unm.edu

Cui Zheng got her bachelor degree from Zhengzhuo University in 2003. Now she is a Ph. D. student in University of New Mexico, USA. Her research has mainly focused on virtualization, high performance computing and cloud computing.

李建平,1998 年于重庆大学获得博士学位,现为电子科技大学教授、博士生导师,主要研究方向为小波分析、智能计算和虚拟技术。
E-mail: jpli2222@uestc.edu.cn

Li Jianping got his doctor degree from Chongqing University in 1998. He is a professor and Ph. D. supervisor in University of Electronic Science and Technology of China now. His main research interests include wavelet analysis, intelligent computing and virtualization.