doi:10.3969/j.issn.1000-436x.2013.z2.017

基于虚拟化的网络测量平台

曹争[1,2]，何建斌[1,2]
（1 东南大学 计算机科学与工程学院，江苏 南京 211189；2.计算机网络与信息集成教育部重点实验室，江苏 南京 211189）

摘 要：随着网络测量研究内容的扩展，网络测量的设施在提高性能的同时必须支持测量的可扩展性以适应不同网络环境和添加新测量研究的需要。提出了一种基于虚拟化技术的网络测量平台。讨论了平台涉及的关键问题，设计了虚拟平台及其运行机制。通过一个组播测量实例表明，和现有测量平台相比，该虚拟平台具有并发性、可扩展性、可定制性、可重构性的特点。

关键词：虚拟化；网络测量；测量平台；虚拟平台
中图分类号：TP301 **文献标识码**：A **文章编号**：1000-436X(2013)Z2-0084-06

Virtualization based network measurement platform

CAO Zheng[1, 2]，HE Jian-bin[1, 2]
(1. School of Computer Science and Engineering, Southeast University, Nanjing 211189, China;
2. Key Laboratory of Computer Network and Information Integration, Ministry of Education, Nanjing 211189, China)

Abstract: With the expansion of network measurement research, the network measurement infrastructures must support the extensible to different network environment and services while improving performance at the same time. A network measurement platform based on virtualization technology was presented, by discussing the key issues of the platform, design of the virtual platform and its operation mechanism. The instance of a multicast measurement shows that compared with the existing measurement platforms, the virtualization platform has concurrency, scalability, customization and reconfigurability characteristics.

Key words: virtualization; network measurement; measurement platform; virtualization platform

1 引言

随着网络结构的日益复杂，Internet 规模的迅速扩大、网络传输内容的多样化、恶意攻击行为的增多，对网络的流量特征、性能特征、可靠性及安全性特征等网络行为模型越来越缺乏精确描述，严重影响到网络发展及更有效的应用。不管是网络运营商还是终端用户对网络性能状况的关心程度也都在提高，因此需要一个网络测量平台，能提供网络运行状态规律，分析定位网络问题，以逐步建立自动化网络。

现有的网络测量平台有各自的设计特点和目的，面向不完全相同的应用领域。如 perf SONAR[1]面向了一个跨域的环境，并通过抽象屏蔽域间差异，提供一个端到端的视角。PlanetLab[2,3]采用了虚拟化的实现技术，为计算机组网与分布式系统研究提供基础设施，测量节点遍布全球，因此所部署的环境是一个不可控的网络。ETOMIC[4,5]的设计目的是从复杂系统的角度理解网络。Emulab[6]与其他平台相比规模较小，但提供了一个共同的用户接口统一了所有的服务。GENI[7]主要面向下一代网络的研究与设计，提出了一个三层结构模型，使参与的国家与结构在保持自治的前提下，充分分享资源。上述网络测量平台主要提供了一种基础设施，其上运行各种测量工具，通过采用各种技术和测量方法，通过测量平台的调度来达到测量网络、研究网络的目的，同时通过引入抽象层或虚拟化等各种手段达到有效地利用底层硬件资源的目的。为了有效地进行网络测量，一个通用的网络测量平台应至少具备以下一个或多个特点[8~11]。

收稿日期：2013-09-05

1) 并发性：可同时开展多个测量任务的研究，多个测量任务可同时运行在平台内的一台或多台主机上，彼此之间不互相干扰，即使多个测量任务同时占据平台内同一个测量节点。一个测量任务造成的错误不能波及其他测量任务。

2) 可扩展性：平台的规模可以随着不断加入的物理设备而线性增长，同时对加入的各测量节点实行良好的管理以及资源的有效利用。

3) 可定制性：平台所备的测量功能应尽可能满足用户需求，同时应适合用户需求的变化。随着网络中新业务的应用，允许用户定制或加入新的测量功能。

4) 安全性：网络测量平台需保证运行过程的安全性，体现在以下几方面：只有平台授权用户才能使用测量平台进行任务测量；测量数据或控制数据传递时不能被修改或窃取；测量平台需具备可以经受一定程度的 DoS 攻击的能力。

5) 平稳性：测量平台所提供的测量任务的意义在于反映被测网络运行状态，其本身绝不能影响被测网络的性能。除非是出于测量目地，否则测量平台在运行过程中产生数据流量应尽量少且平缓。

6) 顽健性：网络测量平台构建在被测网络之上，本身应具备比被测网络更高的顽健性，同时测量平台应具备一定的容错能力与恢复机制，不能被其上运行测量任务的错误所影响。

近年来，虚拟化技术在业界得到大量运用，虚拟出来的多个系统运行环境是隔离的，互相不干扰，因此提高了安全性，并满足了用户独立拥有系统环境的需求。在上述未引入虚拟化技术的平台中，各个测量任务都存在于同一个操作系统中，各个任务所占用的资源全靠程序员自律，可能会有意或无意地对其他任务造成影响。同时，个别平台存在任务固定、不可扩展等问题[9]。本文把虚拟化技术应用到网络测量平台的开发中，提出一种基于虚拟化的网络测量平台方案。

2　平台关键问题

2.1　虚拟化技术的选择

本平台采用了基于半虚拟化技术的Xen作用虚拟化的选择。Xen[12]是一个开放源代码的 x86 虚拟机监控器，可在其上运行不同的客户操作系统，但客户操作系统内核必须经过修改，用户程序不受此影响，通过采用半虚拟化技术，Xen 无论在主机还是客户机上均展现出很高的性能。

在 PlanetLab 中，采用了 Linux vserver 作为虚拟化的实现方案。Linux vserver 是一种操作系统级的虚拟化技术。相关文献表明，Linux vserver 在 5 到 7 个 VM 时，表现出了更好的性能，在 VM 数量更高的情况下，Xen 性能表现更为优异[13]。在 Linux vserver 这种操作系统级的虚拟化方案中，所有 Guest OS 使用同一份内核，隔离程度小于 Xen，Xen 中各个 Guest OS 甚至可以运行不同的操作系统。结合并发性的需求，因此本平台最终选择了 Xen。

2.2　测量节点的 QoS 设计

在平台的每个测量节点中，对网络有以下 4 种基本需求。

1) 二层的网络隔离。不同的虚拟平台在各节点生成的虚构机对彼此是完全不可见的。

2) QoS 配置。在同一个测量点的 VM 共享同一个物理网卡，对于每台 VM 都应该有带宽限制，以免影响其他 VM。

3) 流量监控、Netflow、sFlow。通过 xxFlow 技术对数据分组采样，进而实现网络监控、网络规划、安全分析、会计和结算等操作。

4) bonding 支持，即链路层的网卡聚合。在一些测量点上，物理机拥有多个网卡，若不能全都利用起来会造成资源的浪费。如果简单地用多个网卡将外部交换机和内部虚拟交换机相连，在有 STP(spanning tree protocol)存在的情况下，STP 会阻断多余的路径，只保持一条可用，若没有启用 STP，将会引起广播风暴。所以需要 bonding 机制，通过将链路层的网卡聚合，达到资源利用的最大化。

在此，本平台的每个测量节点上均部署了 Open vSwitch，通过可编程扩展，可以实现大规模网络的自动化(配置、管理、维护)。它支持现有标准管理接口和协议(比如 netFlow、sFlow、SPAN、RSPAN、CLI、LACP、802.1ag)等，较好地解决了上述 4 个需求。

2.3　平台的可靠性

对于单个节点，定义基本指标 t_{mtbf}(平均无故障运行时间，单位为小时)。文献表明 t_{mtbf} 与可靠性的关系如下[14]：

$reliability = \exp(- t / t_{mtbf})$，　t 是运行时间

按照此公式，当 $t=t_{mtbf}$ 时，系统的可靠性为 36.8%，若一台服务器的 t_{mtbf} 是 100 000 h，连续运

行一年出现故障的概率是 8.4%。在 100 台服务器规模的分布式系统中，每个月出现硬件故障的概率是 51.3%，1 000 台服务器的系统中，每周出现故障的概率是 81.4%[15]。因此在开发分布式服务软件时，要求单个程序连续稳定运行一年是做无用功。分布式平台的高可用性关键不在于做到不停机，而是要做到能随时重启任何一个进程或服务。通过容错策略让系统保持整体可用。因此本平台在设计时，贯彻了以下 4 个原则。

1) 不对任何服务做过高的可靠性假设，每个服务均要求在依赖的服务不可用时，当前服务不受依赖服务错误的影响而出错。

2) 任何服务均要求能随时重启，自动恢复与之相关其他服务的通信，并恢复到崩溃前的状态，不影响平台的一致性。

3) 在节点上细分各个服务，将功能不同的模块拆分放入不同的服务中，使用单独进程实现，错误被隔离在一个进程内。各服务可单独升级，单个节点的升级不必停止此节点所有服务。

4) 进程内及不同节点间采用远程过程调用协议 RPC 作为通信方式。RPC 的消息编码做到向下兼容。已升级的节点可以与未升级节点正常交互，藉此做到平台级的热升级。

3 基于虚拟化的测量平台

3.1 平台的构成

本平台由部署在可控网络中多个关键节点上的测量主机与测量服务器构成。在每个测量节点中，又划分出管理主机、服务主机、虚拟平台主机等多个虚拟实例。称所有的测量主机、服务器及运行在其上的管理系统为物理平台。在此管理系统内，根据用户需求在不同测量点上生成的多个虚拟机实例与虚拟机管理系统称为虚拟平台。

如图 1 所示，若物理平台共有 9 个节点，在其上建立了 2 个虚拟平台，其中虚拟平台 1 选取了 5 个节点，虚拟平台 2 选取了 5 个节点，这 2 个虚拟平台彼此间相互独立，完全感知不到对方存在。各自运行各自的测量任务。

3.2 物理平台的构成

物理平台的管理服务器内共分为物理平台管理主机与各虚拟平台对应的管理主机。如图 2 所示，物理平台管理主机共分为 7 个模块。

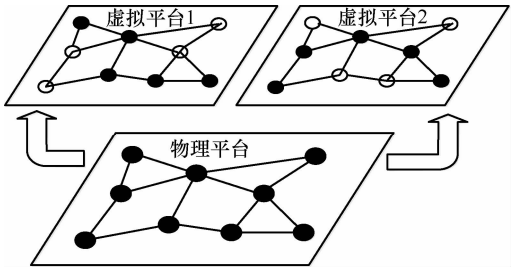


图 1 平台结构

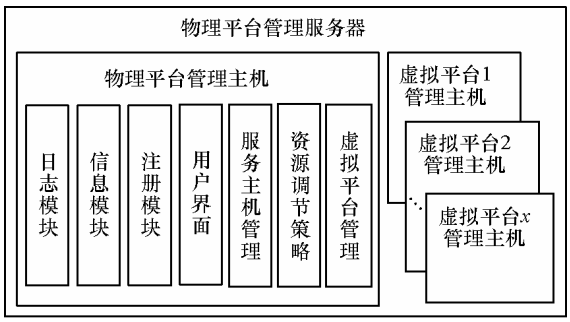


图 2 物理平台服务器架构

日志模块：收集本节点及各测量节点运行日志。

注册模块：新加入本测量平台的测量节点注册服务，并监听测量节点的心跳服务。

信息模块：负责统计平台各节点的资源使用情况。

服务主机管理：管理各测量节点上的服务主机，对服务主机内置工具集进行更新以及分发测量任务、收集运行结果。

虚拟平台管理：管理已生成的虚拟平台，如启动、暂停、状态查询等。提供生成新虚拟平台的接口。

用户界面：供管理员使用的界面，可观测到平台的运行状态、系统日志、活跃的节点等相关信息。通过本节点上其他模块提供的调用接口，对其他模块进行调用，对运行状态进行显示。

资源调节策略：负责对测量节点端的资源调节模块发送资源调节的策略。

测量节点连接在被控网络中指定的路由器或交换机上。如图 3 所示，各测量节点主要由 3 部分组成：节点管理主机、服务主机、各虚拟平台在本测量节点上生成的主机。其中节点管理主机由以下 5 个模块组成。

日志服务：为本机其他服务提供日志功能，同时将收集到的日志发送给平台管理主机中的日志模块。

资源调节模块：监控本节点上虚拟主机的资源消耗情况，根据服务器端配置的资源调节策略进行调节。

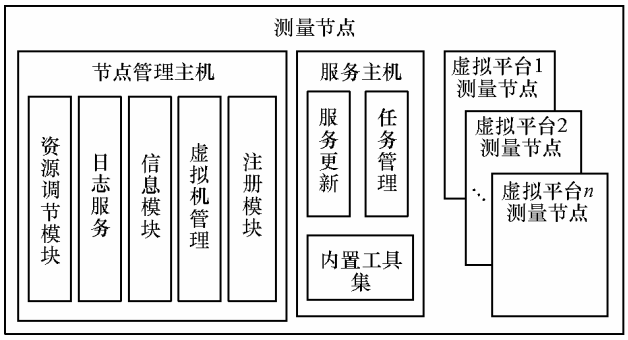


图 3　物理平台测量节点架构

虚拟机管理：管理本节点上运行的虚拟机，并根据服务器指令生成、删除、启动、暂停虚拟机。

信息模块：统计本节点运行状态及资源分配情况并发送给服务器。

注册模块：向平台服务器注册本节点，并与服务器发送心跳连接。

服务主机内置了经过验证可靠的测量工具集，目的是快速地提供简单任务的测量而无需分配新的虚拟测量平台，主要包含服务更新与任务管理 2 个模块。

3.3　虚拟平台

3.3.1　虚拟平台的构造

与物理平台类似，虚拟平台也分为服务器端与客户端 2 个部分。用户通过物理平台的管理界面选择需要生成虚拟机的节点，并在每个节点上指明需要生成虚拟机的配置，主要包括 CPU 核数、内存大小、磁盘大小、swap 空间大小、带宽、发行版等参数。物理平台将在指定节点生成虚拟机，并在服务器上为该虚拟平台生成一个虚拟服务器。该虚拟平台的管理员通过此虚拟服务器管理其所拥有的虚拟机，且此虚拟服务器同时作为此虚拟平台的 VPN 网关。

3.3.2　虚拟平台的运行机制

虚拟平台架构如图 4 所示，虚线部分为控制指令通信通道，主要通过 VPN 实现，可以保证信息的安全，并且做到与测量数据的隔离；实线为测量程序自身产生的数据。虚拟平台的服务器主要由以下 5 个模块构成。

1) 日志服务：节点运行日志并收集虚拟节点的运行日志。

2) 监控信息收集服务：收集虚拟平台各个节点的运行状态。

3) 注册模块：接收虚拟测量节点的注册，且通过心跳机制监控虚拟测量节点是否超时。

4) 程序分发服务：用户可以界面上提交程序，由平台传送到指定节点，并运行。

5) 管理界面：供管理员使用的界面，可观测到平台的运行状态、系统日志、活跃的节点等相关信息。通过本节点上其他模块提供的调用接口，对其他模块进行调用，对运行状态进行显示。

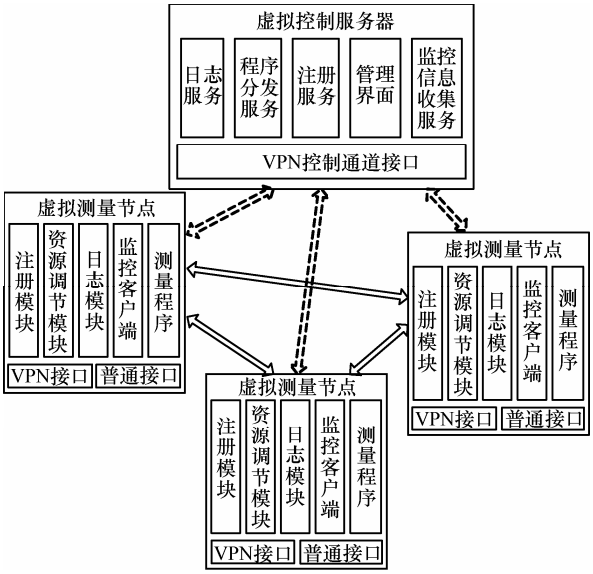


图 4　虚拟平台架构

虚拟测量节点由以下 5 个模块构成。

1) 注册模块：向服务器注册本节点，并运行心跳服务。只有注册过且未超时的节点才被认为是活跃可运行任务的节点。

2) 资源调节模块：接受测量节点 Hypervisor 内的资源调节模块的指令，对本虚拟机资源进行调节。

3) 日志模块：记录本节点运行日志，并发送给服务器端日志模块。

4) 监控客户端：监控本节点运行状态，并发送给服务器。

5) 测量程序：用户上传的测量程序。

3.4　虚拟器件的定制

虚拟器件是一个预配置的软件栈，包括一个或多个虚拟机。其中每个虚拟机都是可以自运行的，而且自带操作系统和相关应用，并明确其所需要的虚拟资源。通过虚拟器件，定制出不同类型的虚拟机，供用户选择，在部署时做到快速部署，而不用每次分配完虚拟机后，再在虚拟机中进行操作系统的安装。

本平台主要提供了 32 bit 和 64 bit 的 ubuntu 和 centOS。内置了平台管理系统如日志服务、监控服

务、平台更新服务以及常用的 libpcap、libnet、libdnet 等软件分组。

4 多播测量实例

作为一个实例实现的组播测量系统可以让管理员更直观地监测网络中组播的性能状况，并且能够定位到性能瓶颈点，为管理员做优化配置提供更准确的信息，对组播服务质量进行有效的评估和控制[16]。且此实例需要众多节点协同配合，可以有效地利用测量平台所提供的资源对平台全面测试，是个非常合适的用例。

4.1 系统架构

系统主要包括 MSCS（测量统计控制服务器）、测量点和用户 3 个模块。其中 MSCS 是整个系统的核心，测量点是实际测量工作的执行组件，用户是系统的使用者。整个系统的设计主要分成 2 个部分：以 MSCS 为核心的系统管理部分和由测量点参与的测量过程部分[17]。具体内容如下。

MSCS 以服务器的方式启动，接收用户和测量点的消息。

用户只能与 MSCS 交互，用户通过任务模板向 MSCS 提交测量请求，MSCS 处理用户任务，然后将测量结果按照标准格式返回给用户。

MSCS 管理所有的测量点。每个测量点完成初始化过程后，必须向 MSCS 报告其活跃状态。MSCS 将该测量点加入测量点管理列表，并记录其信息。

测量点可以以组播源和组播接收者 2 种方式工作，组播源构造探测数据分组，并按照测量过程的要求发送出去，组播接收者根据测量过程的要求捕获探测数据分组，然后进行统计计算，最后将结果返回给 MSCS[18]。

4.2 多播测量虚拟平台的建立

在本实验中，根据测量的需求，设计了由 1 个 MSCS、4 个测量点、4 个用户构成的虚拟平台。在平台管理界面选择好对应的位置，进行各位置上机器的配置，选择上传程序，点击生成虚拟平台。如果所有虚拟机都正常生成，且内置的平台管理程序均已向生成的虚拟服务器注册，则返回用户，此用户即成为虚拟平台的管理员。

4.3 平台的运行及结果验证

在本组播平台中，测量过程完成实际测量工作，由 MSCS 调度测量点执行，其过程如图 5 所示。

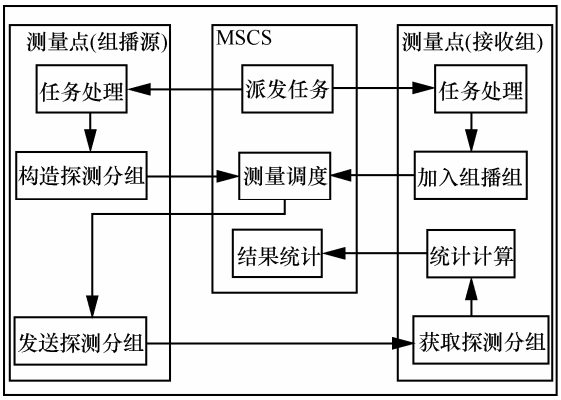


图 5 系统的运行流程

测量调度根据测量工作类型分为 2 种，一种是为完成常规测量对测量点进行的周期性的调度；另一种是为完成用户提交的任务而进行的调度，由任务驱动。

周期性调度是对所有的测量点进行轮询，选择其中一个测量点作为组播源，其他的为接收者，进行一次 30 s 时长的标准测量工作，启动计时器，5 min 后再次进行常规测量。

任务驱动测量是在用户提交任务后，MSCS 对任务进行解析，然后以派发任务的方式调度所需要的测量系统资源完成测量任务。

每个测量点根据任务要求捕获探测分组，统计计算被测性能参数。任务结束后，将本地测量结果汇报给 MSCS，MSCS 收集所有任务参与者汇报的结果，然后进行整理后，按照统一的格式显示给用户。

经过测试，此组播测量虚拟平台可以正常运行，结果与部署在物理测量节点平台上效果一致，且部署在虚拟化平台上时，花费的时间更为简短，明显减少了管理员的工作。

5 结束语

设计并实现了一个基于虚拟化技术的网络测量平台，具有多重通用性特点，对部署在网络中的测量节点形成有效管理，能监控每个加入平台节点的运行状态资源使用情况，平台支持随时加入新的测量节点；利用虚拟化技术，抽象测量节点的底层硬件资源，根据用户的需要，实时生成新的虚拟测量平台供用户使用。利用该平台实现了一个组播测量系统，对此平台进行了验证和功能扩展。在未来，本平台将专注于以下 2 点：一是基于相对低廉的 ARM 嵌入式设备实现测量平台，便于大规模地部署；二是对节点资源研究粒度更细更有效的管理策略。

参考文献：

[1] HANEMANN A, BOOTE J W, BOYD E L, *et al*. Perfsonar: a service oriented architecture for multi-domain network monitoring[A]. Service-Oriented Computing-ICSOC 2005[C]. Springer Berlin Heidelberg, 2005. 241-254.

[2] CHUN B, CULLER D, ROSCOE T, *et al*. Planetlab: an overlay testbed for broad-coverage services[J]. ACM SIGCOMM Computer Communication Review, 2003, 33(3): 3-12.

[3] PlanetLab:an open platform for developing, deploying and accessing planetary-scale services[EB/OL]. http://www.planet-lab.org/, 2012.

[4] ETOMIC: european traffic observatory measurement infrastructure [EB/OL]. http://www.etomic.org/, 2012

[5] MORATO D, MAGANA E, IZAL M, *et al*. The european traffic observatory measurement infrastructure(etomic): a testbed for universal active and passive measurements[A]. Testbeds and Research Infrastructures for the Development of Networks and Communities, 2005 Tridentcom[C]. 2005. 283-289.

[6] EIDE E, STOLLER L, STACK T, *et al*. Integrated scientific workflow management for the Emulab network testbed[A]. Appeared in Proceedings of the USENIX Annual Technical Conference[C]. Boston, MA, USA, 2006. 363-368.

[7] 姜典言, 张丽茹, 李艳. GENI 的设计与分析[J]. 无线电通信技术, 2012, 38(3): 35-38.
JIANG D Y, ZHANG L R, LI Y. Designing and analysis of GENI[J]. Radio Communications Technology, 2012, 38(3): 35-38.

[8] 陈中林, 金跃辉, 牛志升, *et al*. 网络性能测量平台的研究与实现[J]. 电信科学, 2005, 21(11): 63-68.
CHEN Z L, JIN Y H, NIU Z S, *et al*. Research and implementation of network performance platform[J]. Telecommunication Science, 2005, 21(11): 63-68.

[9] 李霞丽, 袁杰, 温辉敏. 一种基于 Plugin 技术的网络测量平台[J], 微电子学与计算机，2007, 24(3): 62-65.
LI X L, YUAN J, WEN H M. A plugin based network measurement platform[J]. Microelectronics & Computer, 2007, 24(3): 62-65.

[10] 朱畅华, 裴昌幸, 李建东. 网络测量及其关键技术[J]. 西安电子科技大学学报，2002, 29(6): 813-818.
ZHU C H, PEI C X, LI J D. Network measurement and its key technologies[J]. Journal of Xidian University, 2002, 29(6): 813-818.

[11] 朱畅华, 裴昌幸, 李建东. 分布式网络测量和分析基础架构研究与实现[J]. 北京邮电大学学报，2004, 27(Z1): 25-31.
ZHU C H, PEI C X, LI J D. Research on distributed network measurement and analysis infrastructure[J]. Journal of Beijing University of Posts and Telecommunications, 2004, 27(Z1): 25-31.

[12] BARHAM P, DRAGOVIC B, FRASER K, *et al*. Xen and the art of virtualization [J]. ACM SIGOPS Operating Systems Review, 2003, 37(5): 164-177.

[13] BENOIT DES LEGNERIS. Comparison of Open Source Virtualization Technology[EB/OL]. http://revolutionLinux.com, 2012.

[14] 陈硕. Linux 多线程服务端编程[M]. 北京: 电子工业出版社, 2012.
CHEN S. Programming of Multithreaded Server-side for Linux[M]. Beijing: Electronic Industry Press, 2012.

[15] Reliability engineering[EB/OL]. http://en.wikipedia.org/ wiki/reliability_ engineering, 2013.

[16] 于乐怡, 曹争. 一种组播视频业务的服务质量监控方案[J]. 中南大学学报，2010, 41(10): 275-277.
YU L Y, CAO Z. Service quality monitoring scheme for multicast video business[J]. Journal of Central South University, 2010, 41(10): 275-277.

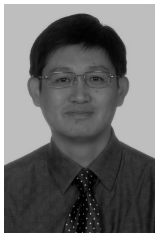
[17] ANDERSON T, CROVELLA M, DIOT C. Internet measurements: past, present and future[EB/OL]. https://www.cs.bu.edu/faculty/crovella/ internet-msmt-impact.ps, 2012.

[18] ADAMS A, BU T, CACERES R, *et a1*. The use of end-to-end multicast measurements for characterizing internal network behavior[J]. IEEE communication Magazine, 2004, 38(2): 341-349.

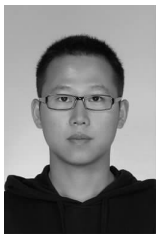
[19] NEVIL B, CHRIS L. Fundamentals of Internet measurement: a tutorial[J]. First published in the CMG Journal of Computer Resource Management, 2001, (102).

[20] 王佳玮, 田斌, 裴昌幸. 分布式网络测量探针关键技术研究[J]，现代电子技术，2007, 30(11):65-67.
WANG J W, TIAN B, BEI C X, *et al*. Research on key technologies of distributed network measurement probe[J]. Modern Electronics Technique, 2007, 30(11): 65-67.

作者简介：



曹争（1958-），男，江苏武进人，硕士，东南大学副教授，主要研究方向为网络体系结构、网络管理、下一代网络技术。



何建斌（1987-），男，江苏泰州人，东南大学硕士生，主要研究方向为网络体系结构、网络管理、下一代网络技术。