



基于虚拟化的网络测量平台

何建斌^{1,2}, 曹争^{1,2}

(1. 东南大学计算机科学与工程学院, 南京, 210000; 2. 计算机网络与信息集成教育部重点实验室, 南京, 211189)

摘要: 由于多媒体等新兴服务占用越来越大的带宽, 一个没有经过仔细规划的网络很容易遭遇到网络拥塞、服务性能瓶颈、用户数量限制等问题。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。本文提出了一种基于虚拟化技术的网络测量平台, 和现有测量平台相比, 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;

A virtualization based network measurement platform

He Jianbin^{1,2}, Cao Zheng^{1,2}

(1. School of Computer Science and Engineering, Southeast University, Nanjing, 211189;

2. Key Laboratory of Computer Network and Information Integration, Ministry of Education, Nanjing, 211189)

Abstract: since the multimedia and other emerging services occupy more and more bandwidth, a network without carefully planning is vulnerable to network congestion, the services bottleneck, restrictions of the number of users and other problems. With the expansion of network measurement research, when the network measurement infrastructures are improving performance, at the same time they must support the extensible to different network environment and services. This paper presents a network measurement platform based on virtualization technology, compared with the existing measurement platforms, it has some advantages as concurrency, scalability, can be customized and reconfigurable.

Key words: virtualization; network measurement; measurement platform;

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, 以及对网络的流量特征、性能特征、可靠性及安全性特征及网络行为模型等缺乏精确描述, 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序也都在提高, 因此需要一个平台, 能提供网络运行状态规律, 分析定位网络问题, 以便于更好的研究网络。近年来, 随着硬件技术的不断提升, 虚拟化技术成为学术界和工业界的研究热点, 并在业界得到大量运用。虚拟化技术通过在一台计算机上虚拟出多个操作系统, 共享底层硬件资源, 提高了硬件的利用率。由于虚拟出来的多个系统运行环境是隔离的, 互不干扰, 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。本文把虚拟化技术应用到网络测量平台的开发中, 根据现有网络测量平台的不足,

作者简介: 何建斌, (1987-), 男, 硕士研究生, E-mail: jbh@njnet.edu.cn; 曹争, (1958-), 男, 副教授, E-mail: zcao@njnet.edu.cn。

提出一种基于虚拟化的网络测量平台方案。

1 研究现状

1.1 现有的网络测量平台

ETOMIC^[8] (European Traffic Observatory Measurement Infrastructure 欧洲流量观测基础设施) 是 EVERGROW^[17] 的子项目, 该项目在欧洲的不同位置部署了 50 个测量节点, 计划在 2025 年之前实现从复杂系统的角度理解网络。

LOBSTER^[18] (large-scale monitoring of broadband internet infrastructures, 大规模宽带互连网络基础设施监测), 为实现精确的网络流量监测而设计和部署的跨越整个欧洲的项目。该项目目前已经终止, 但是部署的测量网络依旧保持运行状态。

AMP (active measurement project) 和 PMA (passive measurement and analysis project) 是由 NLANR^[3] 领导的研究项目, AMP 采用主动测量的



方式对网络的进行测量和分析,在全球共部署了 150 多个测量设备;PMA 在全球部署了 20 多个测量节点,通过被动测量的方式,对网络流量进行分析,监测网络的性能。

PlanetLab^[10] 是一个网络服务的实验床,具有开放共享体系架构,方便各机构组织提供机器加入平台。为支持不同网络应用的研究,PlanetLab 在设计上提出了切片(slice)的概念,即每个应用都是在 PlanetLab 网络中隔离的“切片”中运行,切片之间通过虚拟机共享着系统底层硬件资源。PlanetLab 采用了操作系统级的虚拟化技术 Linux vserver 做为虚拟化的实现,在最近又转向了 LXC,一种基于容器技术的 linux 内核虚拟化。

现有平台的不足:

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^{[2][3][4]}:

并行性: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了平台内的一个测量节点。

可扩展性:平台的规模可以随着不断加入的主机而线性增长,同时对加入的主机实行良好的管理和有效地利用。

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,定制新的测量功能。

可重构性:能够适应将来的测量需求,针对未知的测量需求,提供相应的测量节点,重构出虚拟的测量平台。

在未引入虚拟化技术的平台中,各个测量任务都存在同一个操作系统中,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。

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 作为虚拟化的选择。Xen^[9]是一个开放源代码的 x86 虚拟机监控器,可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。

PlanetLab 采用了 Linux vserver 做为虚拟化的实现方案。Linux vserver 是操作系统级的虚拟化技术。文献表明 5 到 7 个 VM 时, Linux vserver 表现出了更好的性能, VM 数量更高时 Xen 性能表现更为优异^[15]。在 Linux vserver 这类操作系统级的虚拟化方案中, Guest OS 使用同一份内核,隔离程度小于 Xen, Xen 中各个 Guest OS 可以运行不同的操作系统。结合并发性的需求,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,我们对网络有以下几种基本需求:

需求一:二层的网络隔离。不同的虚拟平台在各节点生成的虚拟机对彼此是完全不可见的;

需求二: 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。

需求三:流量监控, Netflow, sFlow。通过 xxFlow 技术对数据包采样,进而实现网络监控、网络规划、安全分析、会计和结算等操作。

需求四: 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 协议,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。Open vSwitch 是做一个开源的具有产品级质量的多层虚拟交换机。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。通过引入 open vSwitch,较好地解决了上述四个需求。

2.3 平台的可靠性

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

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

按照此公式,当 $t=t_{mbf}$ 时,系统的可靠性为 36.8%,若一台服务器的 t_{mbf} 是 10 万小时,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%，1000 台服务器的系统中，每周出现故障的概率是 81.4%^[16]。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，贯彻了以下几个原则：

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，本服务不受此错误的影响而同样崩溃；
2. 任何服务均要求能随时重启，并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. 在节点上细分服务，服务均使用单独进程实现，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 RPC 做为通信方式。RPC 的消息编码做到向下兼容。已升级的节点以与未升级节点正常交互。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。我们称所有的测量主机，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。

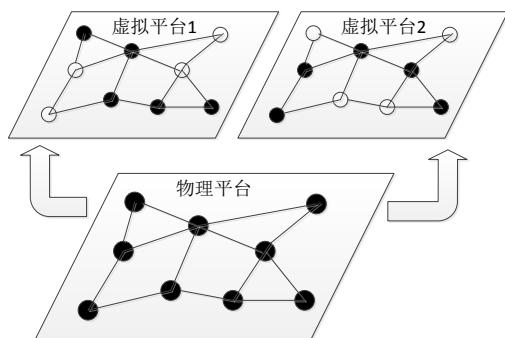


图1 平台结构图

如图 1，物理平台共有 9 个节点，在其上建立了两个虚拟平台，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，这两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。

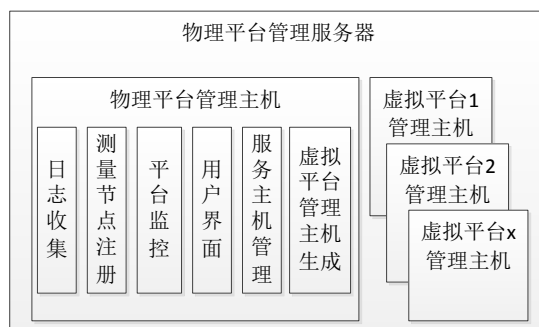


图2 物理平台服务器架构

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六个模块：

日志收集：收集平台运行日志及节点日志

测量节点注册：新加入本测量平台的测量节点注册服务，并提供心跳机制

平台监控：负责收集平台内每一台物理主机的监控服务返回的监控信息。

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。

服务主机管理：管理各测量节点上的服务主机，并进行任务的分配。

虚拟平台管理主机生成：为用户定制的虚拟平台生成对应的虚拟平台管理主机。

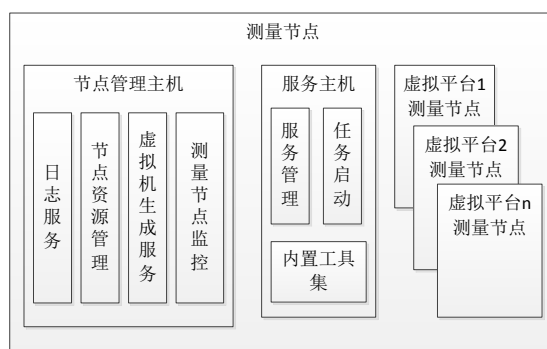


图3 物理平台测量节点架构

各测量节点主要由三部分组成：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 4 个模块组成：



日志服务：收集本机运行日志，并提供给平台管理主机。

节点资源管理：管理本机的资源，包括 CPU, RAM, 磁盘，带宽等。

虚拟机生成服务：接受服务器指令为虚拟平台生成相应的虚拟机。

测量节点监控：监控平节点运行信息，并反馈服务器。

服务主机内置目前已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。接收平台管理主机的指令，运行相应的服务。

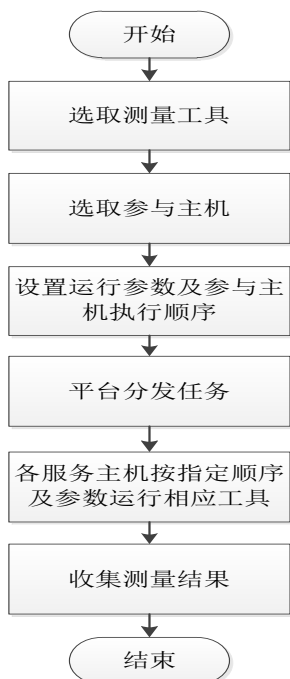


图4 服务主机运行流程图

3.3 虚拟平台

3.3.1 虚拟平台的构造

虚拟平台的构造过程如图5所示。用户在管理界面选择自己需要监控的节点，并且指明在每个节点上需要的机器配置（内存，CPU核数，磁盘大小，带宽大小，操作系统版本），上传平台使用的公钥文件。在平台生成后，用户可使用私钥登陆平台内的任何虚拟机。最后点击生成，管理系统自动地完成生成过程，并实时将生成状态反馈给用户。

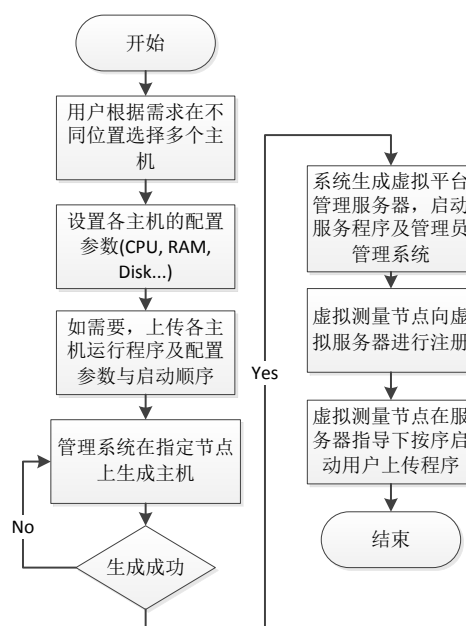


图5 虚拟平台构造流程图

3.3.2 虚拟平台的运行机制

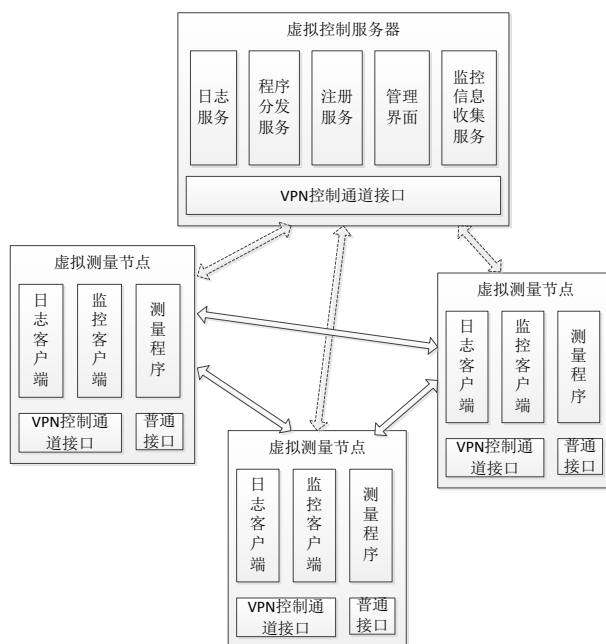


图6 虚拟平台架构图

图中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。

日志服务：在各客户端的运行日志客户端，收集日志，在服务器端需要的时候发送指定级别日志给服务器。

监控服务：提供给虚拟平台管理员的监控服务，以



便于对平台的实时运行情况进行掌控,由服务器端监控收集程序和各客户端监控服务构成。

3.5 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,我们定制出不同类型的虚拟机,供用户选择,在部署的时候做到快速部署,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 位和 64 位的 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 等软件包。

4 组播测量实例

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^[12]。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,对平台的全面测试,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 (包括 MSS) 和用户三个模块。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两个部分: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^[13]。具体内容如下:

- 1、MSCS 以服务器的方式启动,接收用户和测量点的消息。
- 2、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。
- 3、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。
- 4、测量点可以以组播源和组播接收者两种方式工作,组播源构造探测数据包,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

探测数据包,然后进行统计计算,最后将结果返回给 MSCS^[13]。

4.2 组播测量虚拟平台的建立

在本实验中,根据测量的需求,我们设计了由 1 个 MSCS, 2 个 MSS, 4 个用户构成的虚拟平台。其中 MSCS 配置为双核, 512M RAM, 5GB 硬盘, 发行版选择 ubuntu, MSS 配置与 MSCS 相同, 用户机器配置为单核, 512M RAM, 5GB 硬盘, 发行版 ubuntu。在平台管理界面选择好对应的位置, 以及各位位置上机器的配置。选择上传程序, 点击生成虚拟平台。等待返回, 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。其过程如下图

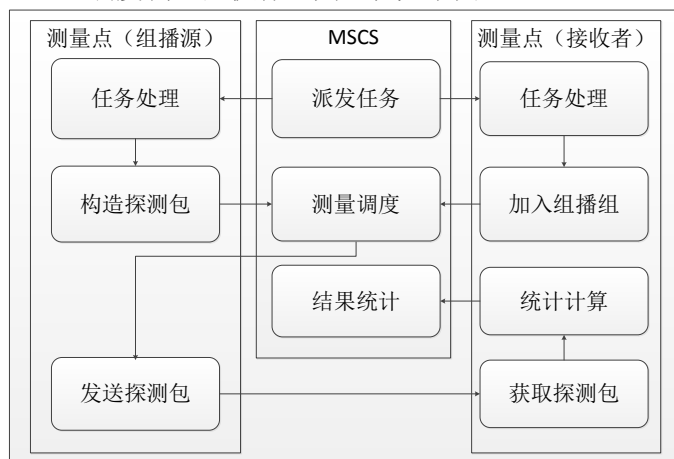


图 7 系统的运行流程

测量调度根据测量工作类型分为两种, 一种是为完成常规测量对 MSS 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 MSS 进行轮询, 选择其中一个 MSS 作为组播源, 其他的为接收者, 进行一次 30 秒时长的标准测量工作, 启动计时器, 5 分钟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包, 并根据任务要求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, 此基本虚拟化的组播平台的可以正常运行, 我们将此平台基于物理机进行部署, 结果与部署在虚拟化平台上效果一致, 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, 平台具有良好的并发性, 扩展性, 可定制性和可重构性, 并在该平台上实现了一个组播测量平台对此平台进行了验证和功能扩展。

参考文献

[1] 徐加羚, 龚俭. 可扩展高速网络流量被动测量平台的设计与实现. 《计算机技术与发展》 2006 年 9 期

[2] 李霞丽, 袁杰, 温辉敏, 于绍娜, 胥桂仙. 一种基于 Plugin 技术的网络测量平台. 《微电子学与计算机》 2007 年 3 期

[3] NLANR <http://www.nlanr.net> 2013

[4] 朱畅华, 裴昌幸, 李建东, 彭爱华, 肖海云 分布式网络测量和分析基础架构研究与实现. 《北京邮电大学学报》 2004 年 z1 期

[5] Chris Loosley, CMG (Computer Measurement Group). Fundamentals of Internet Measurement: A Tutorial Nevil Brownlee, CAIDA (Cooperative Association for Internet Data Analysis). CMG Journal of Computer Resource Management,

Issue 102, Spring 2001

[6] Tom Anderson, Mark Crovella, Christophe Diot. Internet Measurement: Past, Present and Future. November 1, 2004

[7] 王佳玮, 田斌, 裴昌幸, 曾宪玮 分布式网络测量探针关键技术研究. 《现代电子技术》 2007 年 11 期

[8] Surveyor <http://www.etomic.org> 2013

[9] xen <http://www.xen.org> 2013

[10] PlanetLab <http://www.planet-lab.com> 2013

[11] 曾宪玮, 裴昌幸, 朱畅华, 易运晖 基于 ARM 的分布式网络测量探针的设计与实现 《仪器仪表学报》 2007 年 5 期

[12] 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 2010.10.

[13] Adams A, Bu T, Caceres R, et al. The use of end-to-end multicast measurements for characterizing internal network behavior[J]. IEEE communication Magazine, 2004, 38(2): 341.349.

[14] 陈硕 《Linux 多线程服务端编程》第 9 节 分布式系统的可靠性浅说 电子工业出版社 2012 年

[15] Benoit des Legneris. Comparison of Open Source Virtualization Technology. RevolutionLinux.com

[16] Reliability engineering. http://en.wikipedia.org/wiki/-Reliability_engineering

[17] evergrow <http://www.evergrow.org> 2013

[18] LOBSTER <http://www.ist-lobster.org/> 2013

[19] 朱畅华, 裴昌幸, 李建东, 金旗 网络测量及其关键技术. 《西安电子科技大学学报(自然科学版)》 2002 年 6 期