

虚拟计算环境中的分布式虚拟网管技术

占旻, 李沁

(北京航空航天大学计算机学院, 北京 100191)

摘要:通过连接分布在不同网络中不同主机上的多个虚拟机构建虚拟网络, 并提出相应的管理方案。对于一对一节点间的通信采用 P2P 模式, 而一对多节点间的通信采用组播模式, 使节点间的通信无需其他节点转发, 从而提高虚拟网络中数据的传输效率。对该方案进行实验分析, 结果验证了其可行性。

关键词:虚拟机网络; P2P 模式; 传输效率

Distributed Virtual Network Management Technology in Virtual Computing Environment

ZHAN Min, LI Qin

(School of Computer, Beihang University, Beijing 100191)

【Abstract】By connecting multiple virtual machines in different hosts in different networks, the virtual network is set up. The relevant management scheme is proposed. P2P communication mode is used between one-to-one nodes, while multicast communication mode is used among one-to-many nodes, which need not other node to transmit the communication, and promote the efficiency of data transmission in virtual network. This scheme is tested and analyzed, and the results show its feasibility.

【Key words】virtual machine network; P2P mode; transmission efficiency

1 概述

复杂的网络应用需要多机协同, 要求将不同的资源连接成网络来满足应用需求。但通常情况下, 利用虚拟机封装的虚拟资源会部署在不同的主机上, 且这些主机可能分布在不同网络中, 因此, 需要将分布在不同网络的虚拟机连接起来形成独立的虚拟网络。这种技术称为面向虚拟计算环境的分布式虚拟网络管理技术。

虚拟网络由虚拟节点组成, 这些虚拟节点对应于一个个运行的虚拟机。相对于传统的物理节点, 它的动态性更强, 虚拟节点动态加入或退出更方便、更频繁。虚拟网络管理技术要确保在这些节点动态变化时, 网络仍能持续稳定运行。同时这些虚拟节点通常承载密集的网络计算任务, 对于虚拟网络间的数据传输率等网络性能有较高的要求。因此, 本文根据虚拟机网络的特点, 研究面向虚拟计算环境的分布式虚拟网络管理技术。

2 相关工作

N2N 可以构建一个 2 层的 P2P 的虚拟局域网^[1], 但 N2N 中的节点一经注册并加入社区后, 将保持一种永久关系, 不能适应网络的动态变化, 且它采取集中的包转发机制, 使网络中的数据传输效率不高。VDE 是种自适应的虚拟以太网技术^[2]。该虚拟网络对节点管理采用 C/S 结构, 在节点加入或者退出时网络不能动态地做出相应调整, 且在包转发的时候需要利用专门的端口进行转发, 数据传输效率也较低。IPOP 基于 P2P 中的 DHT 来构建虚拟网络^[3], 利用 DHT 的分布式结构完成虚拟网络中链路层地址的查询和更新, 且数据传输中没有采用组播方式, 致使网络中的数据效率不高。Open VPN 采用 C/S 模式来构建虚拟网络, 提供一对多的服务。但

这种结构对服务器端过分依赖, 频繁访问服务器端单一资源将造成严重的瓶颈问题, 甚至发生因服务器端出现故障而造成的网络瘫痪, 引发单点失效问题, 这就无法继续保证整个网络的稳定运行。

针对以上问题, 本文提出面向虚拟环境的分布式虚拟网络管理技术, 改进 C/S 模式, 建立基于 P2P 的虚拟网络, 从而保证整个网络的动态性和灵活性, 并对用户提供透明、稳定的虚拟机网络, 且在网络的数据传输部分采用分组组播的方式, 提高整个网络数据传输效率。

3 系统设计

由于 C/S 架构不能解决虚拟机网络中节点在不同主机迁移时所带来的网络动态性以及 Server 节点可能成为效率瓶颈等问题, 因此采用 P2P 的架构方式, 根据虚拟机网络环境中常常会发生变化, 虚拟机网络可能发生动态启动停止的情形, 提出一种信息的管理和协调机制, 对虚拟网络的节点及其状态变化等信息进行管理。为获得较高的网络数据传输效率, 采用单播与组播相结合的方式转发网络中的数据。

3.1 系统整体结构

系统整体层次结构如图 1 所示。其中, 最底层是资源层, 包含实际的物理资源和封装的虚拟机资源, 主要是构建虚拟网络节点的实际物理网络节点。在资源层之上是基础设施层,

基金项目:国家自然科学基金资助项目(90412011); 国家“973”计划基金资助项目(2005CB321803); 国家“863”计划基金资助项目(2005AA119010); 国家杰出青年基金资助项目(60525209)

作者简介:占旻(1985-), 女, 硕士, 主研方向: 网络安全, 网络计算, 虚拟机技术; 李沁, 博士

收稿日期:2009-04-20 **E-mail:** minminzhan@163.com

该层由 2 个部分组成 :VM-VPN 引擎和公告牌模块。VM-VPN 引擎实现利用 P2P 的机制进行构架网络的功能,并向上提供虚拟网络接口;而公告牌模块则主要负责对虚拟机网络中虚拟机节点信息进行管理和协调。交互层向用户或者上层的应用使用基础设施层功能的客户端工具。用户可以通过 Client 端命令行工具或者 Portal UI 来调用下层提供的功能。最上层是应用层,例如虚拟集群(Vcluster)、虚拟实验室(Vlab)以及其他应用程序,它们通过下层提供的服务来实现自身更复杂的网络功能。

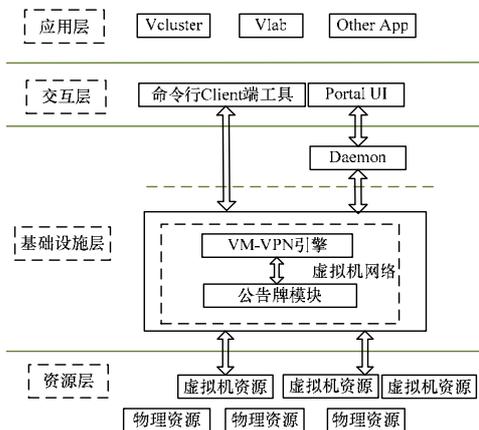


图 1 整体层次结构

3.2 VM-VPN 引擎

VM-VPN 引擎主要是对网络进行拓扑构造。针对前文提到的 C/S 模式带来的虚拟机网络的动态性以及效率瓶颈问题,在网络的拓扑架构中采用 P2P 方式,无需特殊的服务器参与通信的中转,这样不仅可以避免单点失效问题,而且解决了 C/S 模式中的服务瓶颈问题,从而提高通信效率。

VM-VPN 引擎安装在每个 VMM 中(VMM 安装在宿主机上),它向上提供一个虚拟网络接口,这个网络接口可以为节点分配虚拟专用网范围内的 IP,并从虚拟网络中获取数据包,也可以向网络中输入数据。

根据虚拟机节点之间 P2P 的架构思想,VM-VPN 引擎的各个模块以及模块之间的关系如图 2 所示。其中,VM-VPN 引擎核心由以下 6 个模块组成:

(1)Membership 模块

主要是负责向公告牌发送节点注册或注销消息,并同时负责对虚拟网卡进行相应的添加或者删除操作来提供对多个网卡的控制。

(2)虚拟网络节点查询模块

定期从公告牌模块获得邻居节点的 port、IP、子网掩码等信息,并维护虚拟网络节点数据库。

(3)虚拟网络节点数据库

该数据库记录网段中邻居节点的 IP、port、子网掩码等信息。该数据库中的信息主要在分段组播时使用。

(4)转发模块

主要是根据转发模块的原理对报文进行接收或者发送。

(5)MAC 数据库

该数据库记录接收到的报文的源主机的 IP、port、虚拟机的 MAC 地址等信息。

(6)MAC 数据库更新模块

对 MAC 数据库进行更新维护等操作,如对一定长时间

没有查询过的 MAC 地址进行删除,同时更新 MAC 数据库。

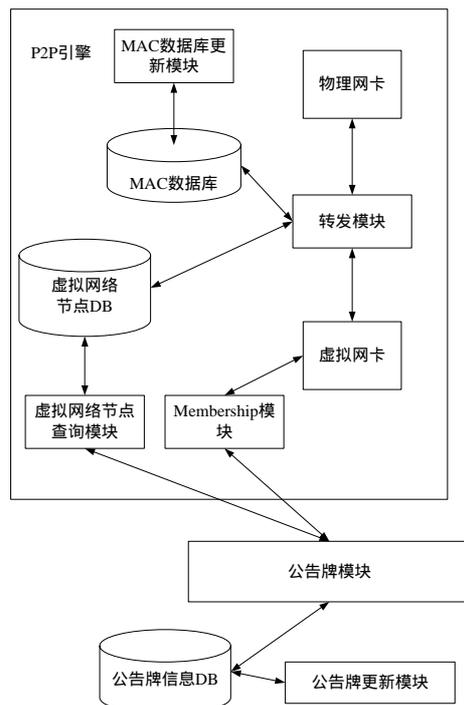


图 2 模块结构关系

3.3 公告牌

公告牌主要负责对不同的虚拟网段的虚拟机网络进行管理和协调。这部分是整个虚拟机网络的中心服务器,提供虚拟机网络中的所有虚拟机节点的信息,但不参与报文的转发。所有运行的虚拟机节点都将在此处注册,并标明所处的网段、域名、IP 地址以及相应的子网掩码。该公告牌记录了当前虚拟机网络的情况以及虚拟机节点的情况,其主要内容如表 1 所示。

表 1 公告牌内容

域名	虚拟网段	该网段宿主机 IP	该网段宿主机 IP 掩码
Lan1	192.168.1.0/24	192.168.2.1	255.255.255.0
Lan1	192.168.1.0/24	192.168.2.4	255.255.255.0
Lan1	192.168.1.0/24	192.168.3.3	255.255.255.0

如某台虚拟机节点启动后,将在该部分进行注册,根据自己的 IP 加入具体网络名称的相应网段,而一旦虚拟机节点退出该虚拟机网络,该部分公示的内容也要做相应修改。

公告牌更新模块主要是对一段时间内未更新的在公告牌模块中注册的节点信息进行删除,并发送删除通知消息给同一个虚拟网络中其他节点。

3.4 虚拟网络中的数据转发

这部分设计主要考虑虚拟网络中的数据转发方式。将一个节点的信号传送到多个节点,如果采用点对点通信方式,则会严重浪费网络带宽;如果采用广播方式,将会造成无关节点的处理负担。因此,在这些情况下采用组播的方式能提高效率,节省网络带宽。另外,对于单点之间的报文传送,多播的效率比不上单播,并会浪费资源。因此,在本文的网络中节点间的传输中考虑单播和组播相结合的传播方式,以达到较高的数据传输效率。

在本文中,节点之间一对一的通信同样采用单播的方式,

如果是一对多的通信,则采用组播方式。在一个 VPN 中的广播实现方式如图 3 所示。

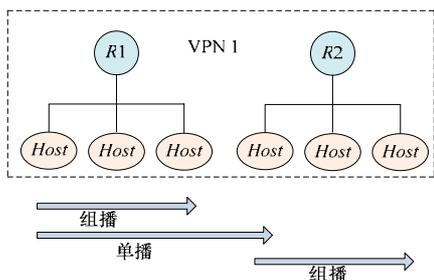


图 3 VPN 中的广播

在图 3 中,同一个网段中的节点之间一对多的通信,采取在该子网组播的通信方式。选择一个虚拟机节点作为向其他子网转发报文的转发节点,采用单播的方式向其他子网转发节点发送报文。其他子网的某虚拟机节点收到报文,采用组播的方式在本网段转发。这就可以实现在整个 VPN 中的报文广播。

4 系统实现

4.1 转发机制的实现

在一对一节点间的通信中,单播无疑是最好最有效的通信方式。下面以 2 台虚拟机节点之间如何 ping 通为例说明 MAC 数据库在报文转发中的工作原理。在图 4 中,宿主机 192.168.1.1/24 上的虚拟机节点 192.168.100.1/24 ping 通宿主机 192.168.1.2/24 上的虚拟机节点 192.168.100.2/24。

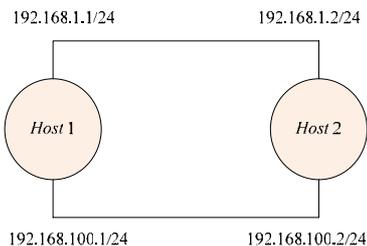


图 4 简单的网络连通示意图

工作原理如下(以下 IP 地址均省略 192.168.):

- (1)100.1 发出 broadcast ARP, 请求 100.2 的 MAC;
- (2)1.1 隧道 ARP, 通过组播地址进行广播;
- (3)1.2 解包,接收 ARP 报文,同时获得{MAC1,100.1,port}的信息,并存储到 MAC 数据库中;
- (4)100.2 发出 ARP Reply, 回复 100.2 的 MAC;
- (5)1.2 根据 MAC 数据库中的内容查询发送到 1.1;
- (6)1.1 解包,获得{MAC2,100.2,port}的信息,并存储到 MAC 数据库中;
- (7)100.1 ping 100.2, 报头为 [MAC1, MAC2][100.1, 100.2];
- (8)1.1 根据 MAC 数据库隧道发给 1.2, 报头为[1.1,1.2][MAC1,MAC2][100.1,100.2];
- (9)1.2 解包;
- (10)100.2 pong 100.1, 报头为[MAC2, MAC1] [100.2, 100.1];
- (11)1.2 根据 MAC 数据库直接隧道给 1.1,报头为[1.2,1.1][MAC2, MAC1][100.2,100.1];
- (12)1.1 解包。

从上面的步骤可以看出,通过 MAC 数据库的参与,记

录宿主机节点的 IP, port 以及虚拟机节点的 MAC 地址,通过查询 MAC 数据库就可以唯一地匹配 MAC 地址,发向相应的 IP, port, 使单播的通信方式得以实现。

4.2 虚拟网络的实现

4.2.1 虚拟网卡的实现

本文在 Linux 的环境中使用 Tap 设备模拟虚拟网卡驱动。Tap 驱动程序的数据接收和发送并不直接与真实网卡交互,而是通过用户态驱动进程转交。考虑设备驱动本身就是核心态和用户态的一个接口,并且访问设备文件会调用设备驱动相应的例程,因此,本文利用设备文件实现用户态和核心态的数据交互。

4.2.2 虚拟网络的构建和执行方式

下面通过上述提到的多个 Tap 设备构建虚拟网络,如图 5 所示。

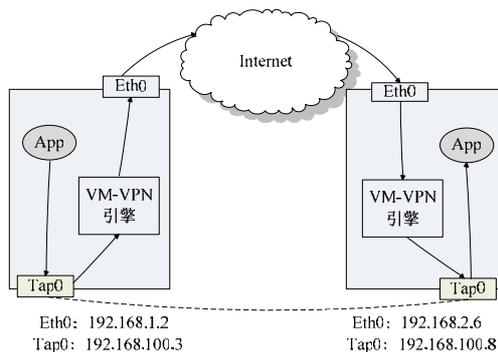


图 5 虚拟网络

为组建虚拟网络,这 2 个节点必须安装 VM-VPN 引擎,并分配 IP。当本地应用进程向虚拟专用网范围内的 IP 发送数据包时,这些数据都会发送到该 Tap 设备。一旦监听到应用程序发出的网络数据包就将其从 Tap 设备中读取。再将这个数据包封装后,经由物理通信链路发送到虚拟专用网中的对方节点。对方节点拆包后,将这个数据包写入该节点的 Tap 虚拟设备,此时该节点上的远程进程就可以接收到这个数据包。按照同样的方式来连接虚拟机节点构建虚拟机网络。

目前系统为用户执行虚拟网络中的应用提供 2 种方式。除了直接使用命令行控制台输入参数来执行应用之外,还提供 Portal UI 的方式。通过 Portal UI 界面连接命令行控制台并由用户输入命令来执行,使得用户可以直接控制程序的执行。

5 应用场景

本系统的设计初衷是提供一个面向虚拟计算环境的高效率、灵活、动态的虚拟网络。本节将介绍 VM-VPN 在虚拟计算环境下的一些应用。

(1)Virtual cluster: 类似真实环境中的集群,由多个虚拟机节点构成。在虚拟机节点中可以预装软件和操作系统,而且通过虚拟网络将虚拟机节点链接起来,所形成的虚拟网络与物理网络逻辑上相互独立。在虚拟集群中可以进行和真实集群中同样的操作,例如通过 PBS 提交作业,通过 Ganglia 监控集群运行状态。并且对于用户来说,可以定制集群中节点的个数和一些软件,这样可以更加灵活方便地满足用户的需求。

(2)Virtual Lab: 由多个虚拟机节点组成具有一定拓扑结构的虚拟网络,称为 Virtual Lab。该虚拟网络拓扑与其运行的物理机之间的拓扑逻辑上相互独立。由于虚拟机又具有各自的 CPU、内存等硬件设备、独立的 OS 和独立的 IP 协议栈,

因此通过这样一个虚拟的网络平台,完全可以满足仿真实验和应用开发的需求,例如开发网络层协议 IPsec, Mobile IP 或针对用户自己定制的虚拟机的软件和拓扑结构,测试在特定的网络拓扑下软件的功能和性能等。

6 实验分析

6.1 实验描述

本实验的主要目的是测试依据本文构建的系统网络执行环境和执行应用的性能。一对一的节点间通信采用单播方式,不需要其他节点的参与转发;一对多的节点通信采用组播的方式。由于系统通过 2 种方式组合的方法提高效率,因此实验将从这 2 个角度加以说明,从而证明本方案的可行性和良好的性能。

实验在局域网环境中进行,实验环境是 2 台 CPU 为 Intel PD 3.4 GHz、内存为 1 GB 的实验用机,操作系统安装的为 Debian etch,虚拟监控器采用 Xen 3.0.3-1-i386-pae 版本。

实验 1 对 3 种 CASE 测试在不同连接方式下的工作负载,以确定使用 Tap 设备连接网络的性能,其中,CASE1 直接测试以太网的网络性能;CASE2 是本系统采用的 Tap 设备来连接虚拟网络测试网络性能;CASE3 是对 2 台宿主机采用 Openvpn 方式架构网络,测试 C/S 模式下的网络性能。

实验 2 分别通过测试本系统中的 P2P 方式与 OpenVPN 架构网络中的 C/S 方式比较网络中 2 组不同节点的并发通信效率。

实验 3 分别通过测试本系统中的组播传输方式与 OpenVPN 架构网络中的 C/S 方式比较网络通信效率。

6.2 实验结果分析

实验 1 使用 tbench 工具测试使用 Tap 设备架构的虚拟机网络、以太网以及 OpenVPN 的架构方式下虚拟机网络 3 种情况的进程负载,在 60 s 的时间里,并发 10 个进程的负载情况如图 6 所示。

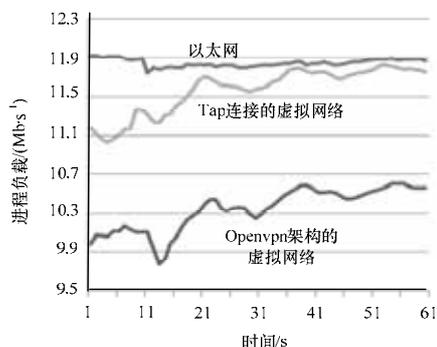


图 6 不同连接方式的网络进程负载

从图 6 可以看出,以太网的网络性能最好,使用 Tap 设备连接的虚拟网络性能比以太网稍差,但其性能损失在可以接受的范围,且 OpenVPN 方式的性能比 P2P 的网络性能低 10%,说明 P2P 方式下网络性能比 C/S 模式下网路性能高,从而证明了采用 Tap 设备连接的虚拟机网络性能的稳定以及效率较高,该方案是可行的。

实验 2 采用实验 1 的测试方法,考虑到 OpenVPN 架构下的网络中的 C/S 方式,此处架构成一个 Server 多个 Client,并选择 OpenVPN 架构的网络中的 2 组节点,其传输方式分别为 C-S(A 组)和 C-C(B 组),并对于同样的这 2 组节点采用本系统的网络架构方式。实验结果如表 2 所示。

表 2 网络中节点并发通信效率

连接方式	A 组吞吐量/(Mb·s ⁻¹)	B 组吞吐量/(Mb·s ⁻¹)
Tap	11.711 80	11.385 70
OpenVPN	6.275 50	4.416 68

从表 2 可以看出,本系统 Tap 采用单播的方式实现网络中 2 个单独节点之间的通信,它们的通信互不影响。但由于 OpenVPN 方式中只取决于中心点的处理能力,因此本系统中采用的 P2P 方式的吞吐量近似 OpenVPN 方式中节点传输方式的 2~3 倍。

实验 3 采用实验 2 的测试方法,选择 OpenVPN 架构下的网络中 2 组节点的通信方式为 C-S 以及 C-C。实验结果如表 3 所示。

表 3 不同传输方式下的网络通信效率

连接方式	吞吐量/(Mb·s ⁻¹)	性能比/(%)
Tap	11.510 00	100.0
OpenVPN(C-S)	10.469 00	91.0
OpenVPN(C-C)	6.417 18	55.8

从表 3 可以看出,C-C 方式中需要 Server 的参与,且组播方式可以节省网络带宽。

7 结束语

本文利用虚拟机技术作为虚拟环境的底层支持,并借助 P2P 思想实现一个稳定的虚拟机网络。该虚拟机网络可以对网络的重新部署提供良好的动态性、灵活性,提高了网络的数据传输效率,为上层应用提供良好性能。通过一系列实验分析,验证了虚拟网络通信时良好的运行效率。当虚拟网络中单独的 2 个节点间进行通信时,采用单播方式,其性能比 C/S 架构的虚拟网络性能高 10%。而对于虚拟网络中大量无关的两两节点间的通信时,其性能将是 C/S 架构的虚拟网络的 2~3 倍,且在虚拟网络中一对多节点进行通信时,系统采用的组播方式将提高网络效率。今后的工作将增加运行维护模块的功能,并实现对虚拟机网络的安全访问。

参考文献

- [1] Tom S. N2N: a Layer Two P2P VPN[Z]. (2007-08-10). <http://www.ntop.org/n2n/>.
- [2] Shammash G. Virtual Distributed Ethernet[Z]. (2007-08-23). <http://vde.sourceforge.net/>.
- [3] Ganguly A. Enabling Self-configuring Virtual IP Networks for Grid Computing[C]//Proc. of the IEEE Int'l Conf. on Parallel and Distributed Processing. Rhodes, Greece: [s. n.], 2006.

编辑 陈文