第 36 卷 第 8 期 计 算 机 工 程 2010 年 4 月
Vol.36 No.8 Computer Engineering April 2010
·网络与通信· 文章编号：1000—3428(2010)08—0113—04 文献标识码：A 中图分类号：TP393.04

# 分布式 VMM 通信架构的研究与原型实现

宋忠雷，肖利民
(北京航空航天大学计算机学院，北京 100191)

**摘 要：**针对分布式虚拟机监控器(DVMM)的通信需求，研究并实现一套 DVMM 的通信方案，利用精简可靠数据协议为分布于多台物理主机之上的 VMM 提供可靠、有序、高效的通信服务。通信方案测试表明，与现行 TCP/IP 协议栈相比，尽管该方案的带宽并无提升，但是对通信延迟却减小了 45%左右，显示了该方案的可行性和优越性。

**关键词：**分布式虚拟机监控器；通信模块；通信协议；精简可靠数据协议

# Research and Prototype Implementation of Communication Architecture in Distributed VMM

SONG Zhong-lei, XIAO Li-min
(School of Computer Science and Engineering, Beijing University of Aeronautics and Astronautics, Beijing 100191)

**【Abstract】** This paper presents Distributed Virtual Machine Monitor(DVMM) and because of the uniqueness of requirement, a communication solution is adopted and implemented. Reduced Reliable Data Protocol(RRDP) provides a highly efficient, orderly and reliable communication service for DVMM that runs on multiple hosts. Test results in DVMM compared with the current communication protocol stack show that, although the bandwidth is not improved, the latency is reduced by 45%, so it is feasible and advantageous.

**【Key words】** Distributed Virtual Machine Monitor(DVMM); communication module; communication protocol; Reduced Reliable Data Protocol(RRDP)

## 1 概述

随着大规模集成电路的发展，现代计算机处理能力越来越强大，这使得现代计算机具有足够的能力来利用虚拟化技术支持多个虚拟机(Virtual Machine, VM)，提供给每个操作系统一套独立的虚拟物理资源。虚拟化技术指的是对计算机资源的抽象化[1]，为提供更高的资源利用率和灵活性而将物理硬件与操作系统分开的一种技术。目前 X86 架构下的 VMM 主要有 EMC 公司的 VMWare、微软公司的 Virtual Server 和剑桥大学的 Xen[2]，这种技术由 IBM 公司于 20 世纪 60 年代发明[3]，在当时用于用户尽可能充分地利用大型机资源。

然而，近年来，随着集群的出现以及高性能计算的需要，一些特殊应用需要更强大的物理资源，单宿主机已经满足不了应用需求，将多个主机物理资源有效地进行整合并提供给上层应用成为必然趋势。分布式虚拟机通过分布式虚拟监控器(Distributed Virtual Machine Monitor, DVMM)对硬件资源进行管理，并通过彼此间的协作提供给客户操作系统统一的物理资源，从而满足高端服务需求。

DVMM 技术将多台物理机资源进行整合，为运行在上层的 GOS(Guest Operation System)提供统一的资源，从而充分利用物理资源，满足应用需求。Xen 是一个优秀的开源虚拟机，本文讨论的 DVMM 是基于 Xen 实现的，它由各物理机上运行的 VMM 组成，各个 VMM 包括了系统启动模块、指令虚拟模块、I/O 虚拟化模块、中断虚拟化模块、内存虚拟化模块和通信模块。各个模块除了独立对硬件资源进行虚拟化外，还需要通过通信连接起来。VMM 之间同样需要通信进行连接整合，为 GOS 呈现出一台“物理主机”资源，如图 1 所示。

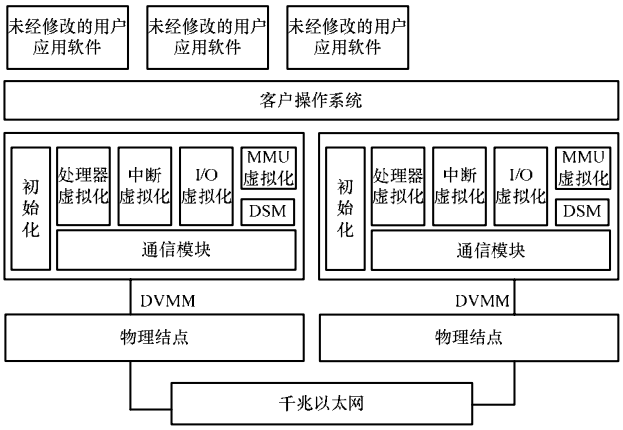


**图 1 DVMM 各模块视图及通信系统所处的位置**

实现 DVMM 的关键之一是为运行在不同主机上的 VMM 之间提供高效可靠的通信。由图 1 可知，通信模块是分布式 VMM 的基础，它为不同 VMM 的各个模块之间提供了协作的平台。各物理主机的 VMM 通过底层通信软件传递控制信息和数据信息，从而完成分布式 VMM 的各项功能，如：为操作系统实现单一内存空间以及跨节点的处理器统一调度等。

**基金项目：**国家“863”计划基金资助项目“新型服务器体系结构及其关键技术”(2006AA01Z108)
**作者简介：**宋忠雷(1984－)，男，硕士研究生，主研方向：计算机系统结构，虚拟化技术，网络通信；肖利民，教授、博士生导师
**收稿日期：**2009-11-30 **E-mail：**szl_84@sina.com

## 2 需求分析

通信模块作为 DVMM 系统的一部分，其主要作用是为分布在不同物理主机上的 VMM 提供可靠高效的通信，从而使硬件资源得到整合，为上层的操作系统呈现出统一的资源映像，使 GOS 拥有强大的硬件资源。

通信是整个系统实现分布式功能的基础，为 DVMM 中的其他功能模块提供通信支持。其他的模块通过通信实现跨节点资源的整合，实现集群的目的，从而可以为上层的操作系统提供更多的硬件资源。由于其他的所有功能模块都需要用到通信，因此通信在 DVMM 中占有重要的地位，是实现其他功能的基础。通信性能的好坏在很大程度上决定着整个 DVMM 系统的性能，要求通信模块能提供正确、有序和高效的服务。

通信模块的功能主要是实现 DVMM 之间的通信，为各物理机器之间资源的感知与整合提供通信基础；使各物理机器的内存能够统一编址，实现单一存储空间，跨物理机的各处理器之间能够相互通信、协调，进而实现单一进程空间；同时也使跨物理机操纵外设成为可能。

涉及到跨节点的通信种类主要有以下几种：

(1)中断虚拟化。

(2)分布式共享存储(Distribute Shared Memory, DSM)页面交换。

(3)远程 I/O 操作。

(4)BIOS 信息交换。

分布在各个物理机上的 VMM 之间通信有以下几个特点与要求：

(1)可靠高效性：VMM 之间的通信要求可靠、高效。

(2)数据固定大小：通过分析得知 VMM 之间每次通信的数据量都比较小，最大是一页数据，即 4 KB。这个特点是选择协议时重要的考虑因素。

(3)专有性：这里的通信不涉及到客户操作系统，对 GOS 来说，整个 DVMM 系统都是透明的。

(4)精简性：通信协议栈只为 DVMM 服务，应尽可能简洁、实用。

总之，DVMM 系统要求可靠、高效的通信，为 DVMM 系统中其他功能的实现提供基础支持。考虑到它的专用性，协议应尽可能精简，以便于实现。

通过上面的分析可知，DVMM 系统中的通信有其独特性，它要求一种简洁但是却可靠、高效的通信传输层协议。而在 TCP/IP 协议栈中，TCP 协议为面向连接的字节流传输，可以提供可靠的通信，但是它的繁杂性使其不适合 DVMM 通信，而 UDP 协议为高效的通信协议，但是它是一种无连接的通信协议，它的不可靠性使其更不适合 DVMM 通信。RDP 是一种基于数据包的面向连接的可靠的数据传输 ，它既提供了可靠的通信，又能高效地传输数据，易于设计实现。本文对该协议进行了研究，并改进和优化，最终实现了一种精简、可靠、有序和高效的通信协议，称之为精简可靠数据传输协议(Reduced Reliable Data Protocol, RRDP)。

## 3 通信架构

通过对 DVMM 的研究与通信的实现要求的分析，通信整体架构如图 2 所示，DVMM 由 VMM 和 DM 这 2 个部分组成， DM 模块为虚拟机监控器的设备模拟部分，通过对其扩展实现 DVMM 的应用层。而在 Dom0 中，基于以太网实现与 TCP/IP 协议栈对等的传输层协议。应用层为 VMM 中各模块发送接收信息所遵循的协议层，依据应用层头部来区分各种传输数据，如 IPI 数据、中断、内存传输、指令传输等；可靠传输层为本文所重点研究并实现的传输层，它主要为 VMM 应用层提供可靠、高效的通信服务；设备抽象和驱动层使用内核的网卡驱动。

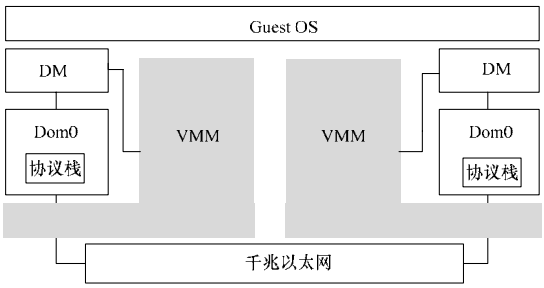


**图 2 通信整体架构**

DVMM 系统通信的最重要部分为传输层通信协议部分的设计。基于以上分析，本文研究并设计实现了 RRDP 协议。

RRDP 工作在链路层协议之上，应用层协议之下，是一个面向连接的、端到端的、可靠并且高效的传输层协议。

### 3.1 RRDP 的目标

RRDP 的目标如下：

(1)提供全双工的通信。

(2)通过正确的确认以及重传机制来提供可靠传输。

(3)避免不必要的重传，一个数据包丢失时只重传这一个或者若干个。这样保证它的高效性。

(4)利用序列号和检验保证数据包的正确性。协议允许用户指定一个连接并对数据包按序发送。

(5)协议不是面向复杂的字节流，而是面向数据包，简洁高效地为 DVMM 提供专有服务。

### 3.2 RRDP 的状态及头部

如图 3 所示，RRDP 共有 6 种链路状态。RRDP 为面向连接的可靠传输协议，在传输数据前进行类 TCP 的 3 次握手建立连接，并以 C/S 方式进行通信，一端为客户端发起连接，另一端为服务器端监听并接受连接请求。在数据传送结束后采用简洁的发送连接请求关闭以及和定时器(在若干长时间内无数据传输)结合的方式来释放连接。通信双方在 OPEN 状态下才能传输数据。

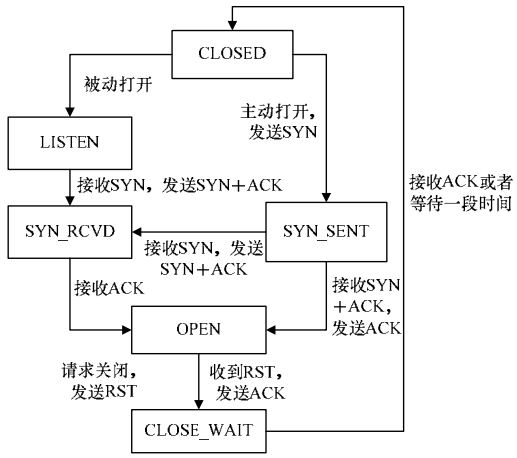


**图 3 RRDP 状态转换**

图 4 为 RRDP 协议的首部域，它们各自的功能及意义为：

(1)Control Flags，第 1 个字节

1)SYN，建立链接和同步序号

2)ACK，确认域有效

3)EACK，乱序确认

4)RST，重新开始建立链接

5)NUL，表明该帧没有有效数字

6)第 5 位没有使用

7)第 6 位和第 7 位是协议的版本号

(2)Header Length，协议头部的长度

(3)Source and Destination Ports，源端口和目的端口号

(4)Data Length，有效数据的长度

(5)Sequence Number，包的序号

(6)Acknowledgement Number，确认序号

(7)Checksum，校验和

在以上各字段的共同作用下，RRDP 完成连接建立与关闭，以及正确传输数据的功能。

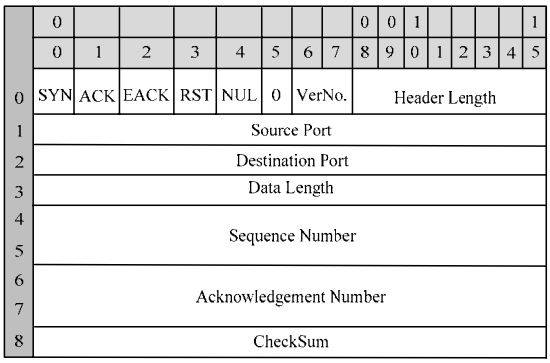


**图 4　可靠传输 RRDP 协议首部**

### 3.3　协议实现

根据上文设计，本文实现了 RRDP 协议，并加入了 DVMM 系统中，使其可以为 DVMM 系统提供可靠、有序、高效的通信服务。

图 5 为 RRDP 协议实现层次视图，与 TCP 和 UDP 协议栈相比，RRDP 协议更加简洁，由于是局域网内的协议，去除了 IP 层，体现了其专用性。而且，在协议实现过程中，为提高数据传输的效率，使用精简的数据结构和优化的算法。在通信协议中，无论是发送数据还是接收数据的过程都会涉及到很多通信数据结构和通信控制结构，通过精简这些结构和优化处理这些结构的算法能适当提高协议的传输效率。

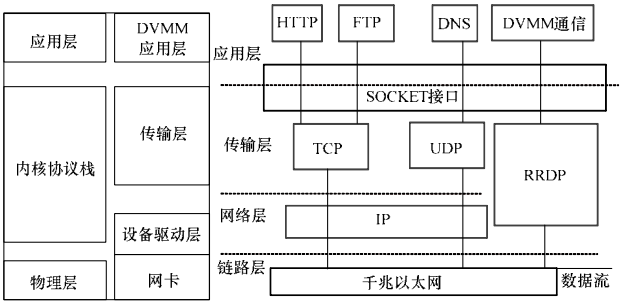


**图 5　RRDP 协议层次及实现结构**

其次，减少数据在主机中的拷贝次数，对于 TCP 或者 UDP，数据首先从用户态拷贝到内核态，在 IP 层对数据块进行分片时会再一次进行数据拷贝。而在 RRDP 协议的实现中，通过用户态与内核态共享数据的方法实现用户到内核的零拷贝，在去除了 IP 层之后，传输层通过 DMA 技术将数据直接交付给网卡处理从而达到了真正的零拷贝。这样极大地提高了协议性能。

### 3.4　地址转换层

为保证与标准 socket 接口兼容，RRDP 模块必须提供简单的地址转换协议，它负责将 DVMM 系统中各个物理结点的结点号与它们的 MAC 地址对应起来。转换模块负责维护 MAC 表(保存物理结点号与 MAC 地址之间的对应关系)，在通信的建立连接过程中，各个 VMM 使用它们各自的结点号，这样提供了高效的地址转换方式。

## 4　性能测试与结果分析

本文针对 DVMM 系统通信应用进行了性能测试，环境为 2 个具有千兆以太网卡的服务器运行 DVMM 系统以及 Dom0。dm 为设备管理模块，协议以模块形式加载到 Dom0 中，测试程序模仿 DVMM 应用环境编程。测试模型如图 6 所示。测试内容主要包括点到点的单向通信延迟和通信带宽，并与 TCP/IP 协议栈的相应性能进行了比较。

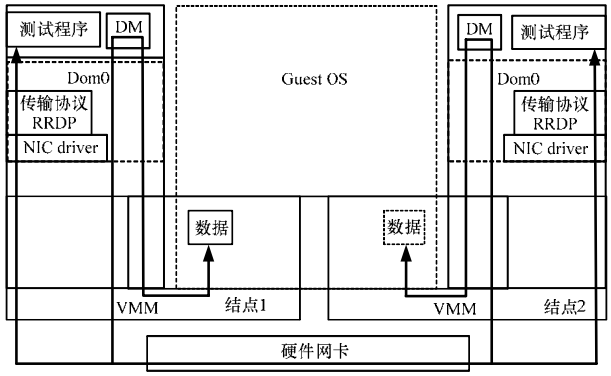


**图 6　通信测试模型**

### 4.1　性能测试

#### 4.1.1　通信延迟测试

图 7 显示了 DVMM 通信协议 RRDP 延迟和传输数据块大小的关系，并与 TCP 延迟进行比较。

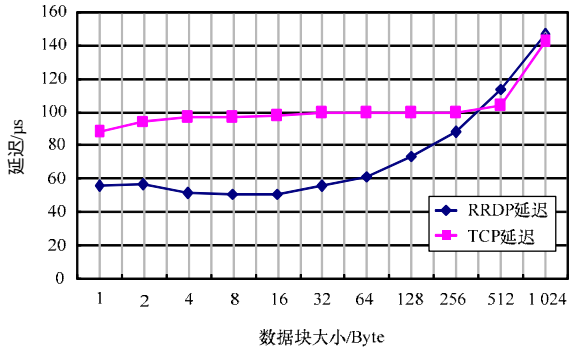


**图 7　RRDP 与 TCP 协议延迟比较**

从图 7 可以看出，随着数据块不断增大，通信协议传输延迟逐渐增大。而对于小数据块(32 Byte 以下)增加不太显著，这主要是因为网络传输的小数据帧中头部占得较大(RRDP 头部和 MAC 头部)，而随着数据所占部分增大延迟明显大幅增加。与 TCP 协议相比，RRDP 协议延迟大概减小 45%左右。当数据块大于 512 Byte 时，RRDP 较 TCP 协议延迟减小不太明显。因为数据帧较大时，TCP 和 RRDP 的延迟大部分都消耗在了用户层到内核层之间的数据拷贝，使得两者差别不大。

#### 4.1.2　通信带宽测试

图 8 给出 DVMM 通信协议 RRDP 带宽和传输数据块大小的关系，并与 TCP 进行了对比。由于 RRDP 无分片功能，因此在千兆网卡大帧支持下最大数据块为 9 000 Byte。

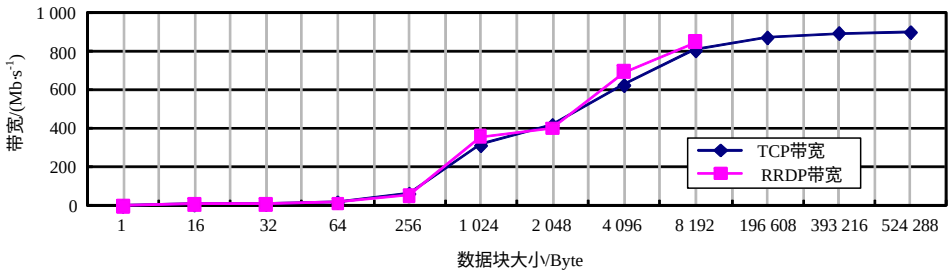


**图 8　RRDP 与 TCP 协议带宽比较**

在通信带宽测试中，同一环境下，RRDP 和 TCP 协议带宽均逼近于 900 Mb/s，两者差异不太明显。在 DVMM 通信中，提高通信协议质量关键在于提高延迟，因此，对于带宽，意义并不是太大。

### 4.2　结果分析

分析可知，实验中在数据字节数较小时延迟有很大提高，这对 DVMM 系统有很大意义。因为在 DVMM 系统通信中，多数的信息交换都是小数据块传输，如中断的发送、BIOS 信息交换、远程 I/O 请求等。

而对于大数据块可以通过零拷贝技术进一步减小延迟。虽然带宽无显著增大，但通信方案最主要的目的——低延迟已经达到。作为简洁的三层协议栈，具有一定的潜力，但是通用性不强，可考虑在加入网络层的基础上使其成为与 TCP, UDP 对等的因特网协议。

## 5　相关工作

### 5.1　MM 间的通信

日本东京大学的 Virtual Multiprocessor 项目也是实现分布式虚拟机。在这个项目中，在 VMM 和硬件层之间，运行了一个简化的操作系统，称为 Host OS。VMM 间的通信使用自己的应用协议，它调用 Host OS 的 TCP/IP 协议栈来完成数据的发送与接收。

目前常见的局域网通信协议主要有 NetBEUI, IPX/SPX, NWLink, TCP/IP。在这几种协议中用得最多、最为复杂的是 TCP/IP 协议。这些协议的实现都离不开操作系统的支持。以最常见的 TCP/IP 协议为例，其实现复杂，不仅可以实现局域网通信，而且支持“路由”功能。因此，不适合 VMM 对通信的要求。

### 5.2　其他几种典型的通信

VMMC 是一种基于虚拟内存映射的通信机制，它支持从发送方虚拟内存到接收方虚拟内存的数据直接传送。

Active Message 是一种异步通信机制，采用与传统的通信机制完全不同的设计思路，更直接地使用通信硬件提供的功能。

Fast Sockets 是一种面向局域网的通信软件，它采用新的高效通信协议在用户空间实现了一般 Unix 操作系统均提供的 Berkeley Sockets API。该软件试图通过修改现有的网络通信协议，使得通信过程中软件开销最小，同时又能保持 Fast Sockets 与现有应用程序以及广域网通信协议的兼容性。

## 6　结束语

本文提出并实现了分布式 VMM 间的通信架构，为实现分布式 VMM 提供了基础。它的性能是影响分布式 VMM 性能的关键因素之一，因此，研究如何进一步提高底层通信的性能就显得十分重要[4]。通过分析得知，由用户态到内核态的数据拷贝影响了通信的性能，后续将实现通信部分零拷贝技术来提高它的性能，采用 Infininband 高速网代替千兆以太网等策略来进一步优化分布式 VMM 底层通信的性能。协议直接运行在网卡驱动层之上，通过修改驱动也能从一定程度上提高它的性能。

## 参考文献

[1] 董耀祖, 周正伟. 基于 X86 架构的系统虚拟机技术与应用[J]. 计算机工程, 2006, 32(13): 71-73.

[2] Barham P, Dragovic B, Fraser K, et al. Xen and the Art of Virtualization[C]//Proc. of the 19th ACM Symposium on Operating Systems Principles. Bolton Landing, NY, USA: Virtual Machine Monitors, 2003: 164-177.

[3] Creasy R J. The Origin of the VM/370 Time-sharing System[J]. IBM Journal of Research and Development, 1981, 25(5): 483-490.

[4] 付赛平, 任国林. XEN 网络 I/O 完全虚拟化机制的可扩展性研究[J]. 计算机工程, 2008, 34(23): 102-104.

编辑　顾逸斐

(上接第 112 页)

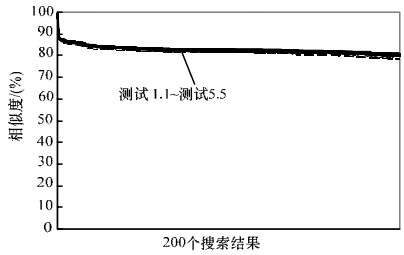


**图 5　扰动环境下 SWAD 的仿真结果**

## 6　结束语

本文对 SWIM 算法进行了抗扰动改进，实现了基于多媒体特征的抗扰动 P2P 搜索。在发布文件并与邻居文件交换更新信息的过程中，如何减少重复数据的发送，提高通信效率，仍有很大的改进空间。

## 参考文献

[1] Novak D, Zezula P. M-Chord: A Scalable Distributed Similarity Search Structure[C]//Proceedings of INFOSCALE'06. Hong Kong, China: [s. n.], 2006.

[2] Androutsos P, Androutsos D, Venetsanopoulos A N. A Distributed Fault-tolerant MPEG-7 Retrieval Scheme Based on Small World Theory[J]. IEEE Transactions on Multimedia, 2006, l8(2): 278-288.

[3] Maymounkov P, Mazières D. Kademlia: A Peer-to-Peer Information System Based on the XOR Metric[C]//Proceedings of the 1st International Workshop on Peer-to-Peer Systems. London, UK: [s. n.], 2002.

[4] ISO/IEC. ISO/IEC 15938-6-2003 Information Technology — — Multimedia Content Description Interface-Part 6: Reference Software[S]. 2003.

编辑　索书志