跨域虚拟网络环境下虚拟机 Live 的迁移机制

魏晓辉, 蒋 娜, 郭庆南, 李洪亮
(吉林大学 计算机科学与技术学院, 长春 130012)

摘要: 针对跨域虚拟网络环境下虚拟机的在线迁移机制, 设计了支持虚拟机跨域通信的虚拟网络路由协议和控制虚拟机在线迁移的路由更新协议, 以支持虚拟机的跨域动态管理, 并基于流行虚拟化环境 Xen, 完成了原型实现, 从而解决了虚拟机跨域的在线迁移问题, 实现了利用广域网络环境搭建动态虚拟化平台.

关键词: 虚拟网络; 虚拟机; 跨域在线迁移; 路由协议; 虚拟机迁移协议
中图分类号: TP391 **文献标志码**: A **文章编号**: 1671-5489(2011)03-0481-06

Mechanism of Virtual Machine Live Migration in Cross-Domain Virtual Networks

WEI Xiao-hui, JIANG Na, GUO Qing-nan, LI Hong-liang
(*College of Computer Science and Technology*, *Jilin University*, *Changchun* 130012, *China*)

Abstract: The authors proposed a novel approach of virtual machine live migration in cross-domain virtual networks. A virtual network routing protocol was presented to support cross-domain communication between VMs. A VM migration protocol controlling the migration process was designed to enable dynamic management of virtual infrastructure. We implemented a prototype based on a typical virtualization tool, Xen. VM Live migration method presented in this work enables dynamic virtual infrastructures extending to wide-area network.

Key words: virtual network; virtual machine; cross-domain live migration; routing protocol; VM migration protocol

0 引 言

虚拟机(VM)[1]和虚拟机群技术[2]有效解决了高性能计算环境对硬件平台的依赖性问题, 为进一步解决计算环境的动态管理, 如负载均衡、容错容侵以及节能管理等问题提供了新途径.

虚拟机迁移[3]是指将一台主机(源主机)上运行的虚拟机迁移到另一台主机(目的主机)上运行. 虚拟机在线迁移(live migration)是指在整个迁移过程中, 虚拟机的暂停时间(downtime)非常短, 在虚拟机上运行的服务始终能响应用户的请求, 以保证虚拟机环境对用户的透明性.

现阶段的虚拟机迁移通常仅限于一个局域网络. 如 Xen[4]和 VMWare[5]等均基于网络文件服务器实现虚拟机镜像文件的共享, 虚拟机的迁移只是迁移 CPU 和内存状态. 这样的迁移方案只能解决同一

收稿日期: 2010-01-22.

作者简介: 魏晓辉(1972—), 男, 汉族, 博士, 教授, 博士生导师, 从事网格计算与网络安全的研究, E-mail: weixh@jlu.edu.cn. 通讯作者: 蒋 娜(1985—), 女, 汉族, 硕士研究生, 从事网格计算与网络安全的研究, E-mail: jiangnajlu@sohu.com.

基金项目: 国家自然科学基金(批准号: 60703024)、吉林省科技发展计划项目(批准号: 20070122; 20060532)和新世纪优秀人才支持计划项目(批准号: NCET-09-0428).

局域网内的虚拟机动态管理问题. 随着虚拟计算环境的发展和普及, 单一局域网管理机器的数量和网络环境将限制虚拟计算环境的可扩展性; 如何在多个管理域间进行虚拟计算资源共享成为研究热点. 因此, 基于跨域虚拟网络的虚拟机在线迁移受到研究者的广泛关注.

文献[3]给出了当虚拟机在局域网内迁移后, 它的IP地址等网络配置无需改变, 虚拟机可以按照同样的方式与外部网络通信的方法: 采用目标节点在迁移完成后广播一个未被恳求的ARP应答的方法, 通知其他节点其IP地址与新MAC地址的绑定, 但存在两方面的问题: 1) ARP响应报文机制不适用于广域网; 2) 在迁移过程的Downtime时间内会有少部分数据包丢失, 不适于并行计算等应用领域.

不同子网之间在线迁移的网络连接解决方案中, 文献[6]提出了一种在不同子网间的虚拟机在线迁移方案. 当虚拟机迁移时和迁移后, 保持源虚拟机不停止运行, 继续向外提供服务, IP地址保持不变, 不向外提供新的服务; 而目标虚拟机迁移到另外一个网络后, 配置新的IP地址和网络配置, 然后启动虚拟机. 当新的服务开始运行时, 服务对外的IP地址即为新设置的值. 源虚拟机和目标虚拟机需要共同运行一段时间, 当源虚拟机中的服务都运行完毕后, 停止该虚拟机, 此后所有的服务由目标虚拟机提供. 但迁移后虚拟机的网络配置需要改变, 对虚拟机不具有透明性.

本文使用IP隧道[7]、Overlay网络[8]等网络通信技术, 结合虚拟化技术对跨域虚拟网络环境下虚拟机在线迁移机制进行研究, 从而实现透明、可靠的虚拟机广域网在线迁移. 首先, 需要解决迁移前后通信地址的一致性问题, 即迁移到不同网络的虚拟机仍然使用原有的IP和端口地址与其他虚拟机上的应用程序进行通信; 其次, 为支持并行计算等应用, 必须保证迁移过程中没有任何的消息丢失和乱序. 目前本文的工作只关注减少应用程序的冻结时间, 而不是严格意义上的在线迁移.

本文提出的迁移机制采用虚拟网络路由更新协议和分布式异步数据缓存协议相结合的方法, 保证迁移过程中虚拟机IP地址不变和Downtime时间内数据包不丢失, 从而实现透明、可靠的虚拟机在线迁移.

1 虚拟网络通信机制与虚拟机迁移协议

1.1 虚拟网络通信机制

本文基于以下几点提出虚拟网络原型: 1) 确保迁移对虚拟机(VM)透明, 即迁移时VM的IP地址保持不变; 2) 确保虚拟网络的动态性. 虚拟网络通信代理(Agent, 负责虚拟机群中跨局域网通信时数据的路由、存储和转发)能根据虚拟机群的要求进行动态创建, 在VM都迁移出后被撤销; 3) 确保虚拟网络路由管理的灵活性, 即Agent可以根据VM的物理位置调整路由信息.

本文设计的虚拟网络原型架构如图1所示. 底层是物理网络, 包括路由器、交换机、主机(HOST)和虚拟机(VM), 在每个主机上运行一个通信代理Agent. Overlay[8]层中实体只包含Agent和VM. 每个虚拟私有网络根据需求由若干Agent和VM动态组成, 位于不同网络的VM通过Agent之间建立的专用通信隧道连接. 其中, VN1和VN2内的VM分别构成各自独立的虚拟机群.

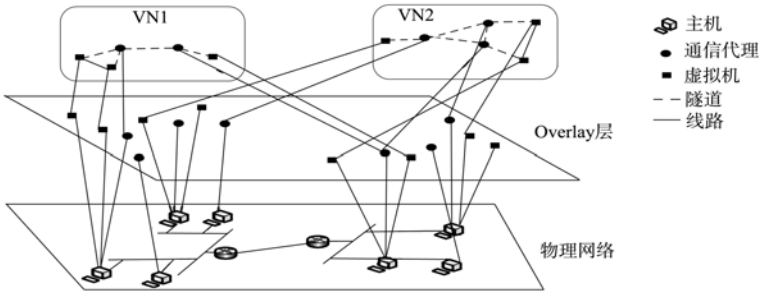

图1 虚拟网络原型

Fig. 1 Prototype of virtual network

虚拟网络通信过程如图2所示, 由路由管理模块、带过滤捕获模块、封装模块、隧道转发模块、解封装模块和注入模块组成. 各模块的作用如下:

1）路由管理模块：负责路由表的管理，为捕获模块制定过滤规则及由虚拟机迁移引起的路由更新；

2）带过滤捕获模块：根据过滤规则捕获与相应虚拟机属于同一虚拟机群且跨局域网通信的数据包，其中虚拟机和捕获模块一一对应；

3）封装模块：把从捕获模块发来的数据包当作新数据包的数据部分，在包头填入控制信息（源主机 IP 地址 Src_host_ip，目的虚拟机 IP 地址 Dst_vm_ip，虚拟机所在机群的群标志 Cluster_id，数据包的大小 Packet_size，序列号 Seq_num），组成新的数据包发送给隧道；

4）隧道转发模块：根据数据包包头的 Dst_vm_ip 和 Cluster_id 查找路由表，找到与 Dst_vm_ip 对应的 Dst_host_ip，从而借助以 Src_host_ip 和 Dst_host_ip 为端点建立的隧道进行数据传输；

5）解封装模块：对应于封装模块，去掉包头的控制信息，将数据包的纯数据部分作为新的数据包，交给注入模块；

6）注入模块：根据数据包内附加的信息，将数据包注入到当前网络.

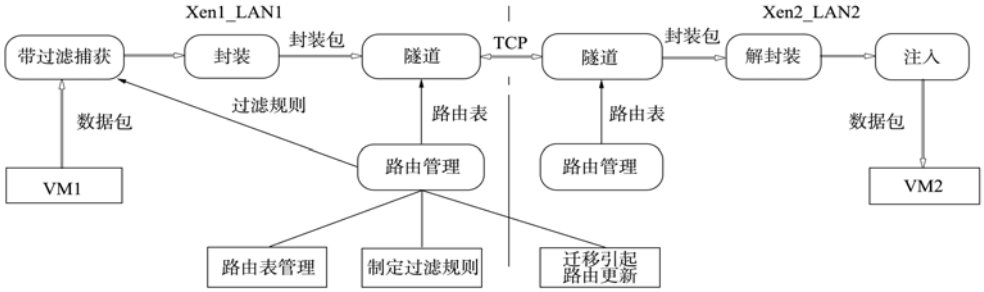

图 2　虚拟网络通信过程

Fig. 2　Communication process of virtual network

1.2　路由管理

虚拟网络通信需要借助路由表进行查询选路，路由表的结构如下：（Host_name，Host_ip，VM_name，VM_ip，Bcast，Cluster_id，Valid）. Cluster_id 是虚拟机所在机群的标识，便于对虚拟机进行动态管理. 子网广播地址 Bcast 辅助虚拟网络通信机制中的过滤规则，Valid 在虚拟机迁移协议中具有重要作用.

局域网内路由管理采用主从式（Master-Slave），如图 3 所示. 在每个局域网内，都有一个主控（Master）和一个备份主控（Backup Master），它们通过 Hello 协议选举产生.

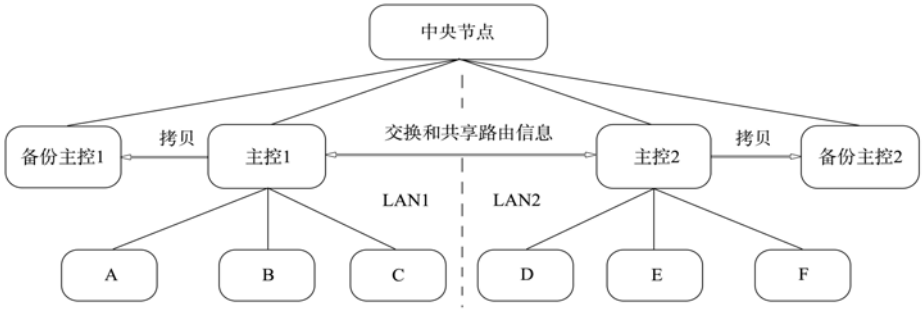

图 3　主从式路由管理

Fig. 3　Router management based on Master-Slave model

Hello 协议的运行通过中央节点完成，它辅助进行各局域网内 Master 的选举. 当某新成员加入局域网时，它需要向中央节点发送一条注册消息，中央节点把本局域网内第一个向它注册的成员任命为 Master，把第二个向它注册的成员任命为 Backup Master，然后把任命消息发送给注册机.

Master 与本网内的所有其他机器建立一种星型的邻接关系，这种邻接关系用于交换各机器的路由

信息，Master在路由信息的同步上具有核心作用. Backup Master保存有Master的一份实时数据拷贝，它的设立是为了保障当Master发生故障时尽快接替Master的工作，避免出现由于重新选举Master和重新构筑拓扑(路由信息)数据库而产生大范围的数据库震荡.

路由管理模块根据Cluster_id派生出多个子进程，维护相应虚拟机群内部的路由信息，使各虚拟机群内的路由信息管理相对隔离. 当某主机上启动、关闭或迁移VM时，路由管理模块需要将相关路由信息发送给本网的Master，Master再根据路由信息中的Cluster_id交给相关子进程，然后根据信息中的操作标识进行相应的插入、删除或更新操作，并将此信息发布给其他相关节点(其上拥有隶属于Cluster_id的虚拟机)和其他的Master，完成虚拟机群内部的路由信息同步.

1.3 支持虚拟机在线迁移的路由更新协议

虚拟机迁移是操作系统级的，迁移时把整个操作系统以及其上运行的所有应用程序作为一个单元处理. 虚拟机的迁移过程主要分为“挂起”、“虚拟机映像复制”和“重启”3个阶段. 迁移期间，其上运行的应用程序处于中断状态. 中断时间越短，对其他程序和网络用户的影响越小. 目前，虚拟机在线迁移通常使用“Pre-copy”技术[3]减少虚拟机迁移过程中的停机时间.

假设Xen1和Xen2位于同一局域网内，Xen1上的VM1与Xen2上的VM3通信，当Xen2上的VM3开始向另一局域网的Xen4迁移时，应在源主机的角度分析虚拟机迁移的逻辑步骤和虚拟网络路由更新协议以保证迁移过程中IP地址不变，进而保证迁移的透明性.

图4为虚拟机迁移协议时间轴，说明如下：

1）预迁移阶段. 预先选择一个满足虚拟机资源需求的目的主机，其中的资源需求即为CPU数、内存资源和某些特定的服务.

2）资源预约. 源主机Xen2向目的主机Xen4发送虚拟机迁移请求，预约VM container.

3）虚拟机迁移. 迁移阶段主要进行路由更新和迭代拷贝.

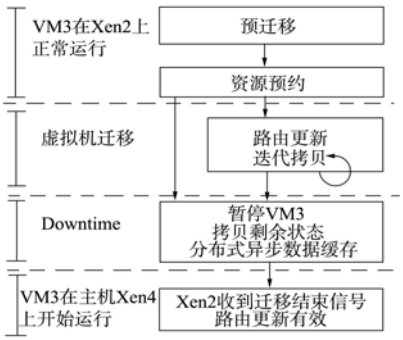

图4 虚拟机迁移协议时间轴

Fig. 4 Timeline of VM live migration protocol

Xen2上控制模块捕获到迁移信号，获得预迁移目的机的IP地址等信息，控制模块通知路由模块进行路由更新. 如果源主机是本网内普通节点，则执行下述步骤①～⑤；如果源主机是本网的Master，则执行下述步骤③～⑤.

① 源主机的路由更新模块将更新信息发送给本网的Master；

② 源主机等待本网Master发来的回复确认，若超时仍未收到回复，则源主机重传更新信息；

③ Master的路由管理模块进入相关子进程，将此路由表项添加到路由表，并发布给相关节点和其他的Master；

④ 本网的Master等待本网内相关节点和其他Master发来的回复确认，如果超时仍未收到回复，则重传更新信息；

⑤ 本网相关成员节点与Master一样更新路由表；其他局域网的Master采用超时重传机制进行③的处理.

更新完路由表后，此时与VM3对应的路由表项有两个：对应于源主机Xen2和对应于目的机Xen4. 把后者的Valid设置为0，标志此表项暂时不可用. 因为目的机上的虚拟机并未启动，暂时还不能进行数据处理. 截止到Downtime，数据传输仍然使用对应于Xen2的路由表项.

使用“Pre-Copy”技术通过多次迭代将VM3内存空间从源主机迭代拷贝至目的主机，保持虚拟机运行状态的一致性.

4）Downtime. 暂停源主机上的VM3，把VM3的剩余状态拷贝到目的机.

从最后一次迭代开始到VM3在目的机正常启动恢复工作时截止，网络中没有VM3的活动副本.

即 Downtime 时间段内，与 VM3 建立的通信无法被响应，数据包在超时情况下将被丢弃，从而导致各虚拟机之间无法保证状态的一致性. 基于分布式异步数据缓存协议，发送端和接收端同时缓存，保证迁移期间报文传递的可靠性和有序性，最终支持虚拟机的在线迁移.

5）迁移结束，路由更新有效. 源主机控制模块捕获到迁移结束信号，虚拟机已在目的机上启动，可以进行数据处理. 使 3）中的路由更新有效.

如果源主机是本网内普通节点，则执行下述步骤① ~ ⑤；如果源主机是本网的 Master，则执行下述步骤③ ~ ⑤.

① 源主机将虚拟机可用消息发送给本网的 Master；

② 源主机等待本网 Master 发来的回复确认，若超时仍未收到回复，则源节点重发此信息；

③ Master 的路由管理模块进入相关子进程，将对应于虚拟机是 VM3 且依附主机是 Xen4 的路由表项的 Valid 设置为 1，删除对应于 VM3 且依附主机是 Xen2 的路由表项，发布给相关节点和其他Master；

④ 本网 Master 等待本网相关节点和其他 Master 发来的回复确认，如果超时仍未收到回复，则重发消息；

⑤ 本网相关成员节点使相应路由表项生效；其他局域网的 Master 采用超时重传机制进行③的处理.

更新完毕，新的路由表项生效，分布式异步缓存的数据可借助新的隧道进行传输. 缓存数据传输完毕后的虚拟机间通信即是正常情况下跨域虚拟网络的通信. 本文设计的路由更新协议将路由更新和数据传输分离，减少了迁移完成后恢复数据通信的时间，提高了通信效率，保证了服务的可靠性.

跨域环境下虚拟机迁移协议将路由更新协议和分布式异步数据缓存协议相结合，保证迁移过程中 IP 地址不变，解决了 Downtime 时间段内的丢包问题，保证迁移期间报文传递的可靠性和有序性，最终实现透明、可靠的虚拟机在线迁移.

2　实　验

下面由两方面验证本文工作：1）检验虚拟网络路由协议的可靠性，协议实现采用 C/S 架构并基于传输层协议 TCP；2）验证虚拟机发生迁移时，路由更新协议和数据缓存协议能够有效结合，实现透明、可靠的虚拟机在线迁移.

实验环境：局域网 LAN1（主机 Master1，Backup Master1，A，B，C，D），LAN2（主机 Master2，Backup Master2，E，F，G，H），一个中央节点 N.

首先验证虚拟网络路由协议的可靠性. 实验条件：在主机 A 上启动一台虚拟机 VM，过一段时间后，Master1 发生故障. 测得的相关时间数据列于表 1.

表 1　虚拟网络路由协议的可靠性

Table 1　Reliability of virtual network routing protocol

实验次数	共享路由时间/μs	恢复路由系统时间/μs
1	2 861	880
2	3 056	869
3	3 200	872
4	2 806	865
5	3 178	870
平均	3 020	871

由表 1 可见：1）主机 A 上启动了一台虚拟机，将它的相关路由信息扩散到广域网相关主机上的平均时间是 3 020 μs，非常短暂，从 CUP 进程调度的角度看，路由管理所用时间越短，越有更充裕的时间处理数据；2）当 Master1 发生故障后，Backup Master1 通过 871 μs 接替它的工作并恢复正常运行，采用容错原理[9]中的冗余机制虽然增加了备份成本，但在故障发生时，可以快速恢复路由系统. 而且在主控变更的过程中，各 Agent 上的路由表完好无损，所以变更不会影响虚拟机群的内部通信.

其次验证在虚拟机 VM 发生迁移的情况下，路由更新协议和数据缓存协议能有效结合，保证迁移

期间 IP 地址不变和报文传递的可靠性和有序性，相关实验数据列于表 2.

表 2 虚拟机迁移相关数据

Table 2 Relevant data of VM live migration

实验次数	路由更新时间/μs	Downtime/ms	路由生效时间/μs	迁移时间/s	丢包/个
1	2 784	525	2 071	22	0
2	3 047	510	2 556	24	0
3	2 827	568	2 694	23	0
4	2 978	535	2 816	25	0
5	3 115	517	3 005	20	0
平均	2 950	531	2 628	23	0

由表 2 可见，虚拟机迁移的平均时间是 23 s，Downtime 是 531 ms. 迁移时路由更新时间比迁移完成后使路由生效时间稍长，因为前者传输的数据较多，如目的主机 IP、子网广播地址等相关信息，后者只需要有相关虚拟机名和 IP 地址即可. 由表 2 还可见，虚拟机迁移虽然存在 Downtime 时间，但将路由更新协议和数据缓存协议相结合后，并不存在丢包现象，从而保证迁移期间 IP 地址不变和报文传递的可靠性与有序性，最终支持虚拟机在线迁移.

综上所述，本文设计了支持虚拟机跨域通信的虚拟网络路由协议和控制虚拟机在线迁移的路由更新协议，并结合分布式异步数据缓存协议[10]实现了可靠、透明的虚拟机在线迁移.

参 考 文 献

[1] Garfinkel T. Virtual Machine Monitors: Current Technology and Future Trends [M]. Washington DC: IEEE Computer Society, 2005: 39-47.

[2] Ruth P, McGachey P, XU Dong-yan. VioCluster: Virtualization for Dynamic Computational Domains [C]//Proceedings of IEEE Cluster Computing. Burlington: MA, 2005: 1-10.

[3] Clark C, Fraser K, Hand S, et al. Live Migration of Virtual Machines [C]//NSDI'05 Proceedings of the 2nd Conference on Symposium on Networked Systems Design and Implementation. [S. l.]: USENIX Association Berkeley, 2005: 273-286.

[4] Barham P, Dragovic B, Fraser K, et al. Xen and the Art of Virtualization [C]//SOSP'03 Proceedings of the Nineteenth ACM Symposium on Operating Systems. New York: ACM, 2003.

[5] Vmware Inc. VMware-Virtualization Software [C/OL]. [2007-11-04]. http://www.Vmware.com/.

[6] Snoeren A C, Balakrishnan H. An End-to-End Approach to Host Mobility [C]//Proc of the 6th Annual International Conference on Mobile Computing and Networking. New York: ACM, 2000: 155-166.

[7] Nordmark. RFC 4213, Basic Ipv6 Transition Mechanisms [C/OL]. [2009-08-15]. http://tools.ietf.org/html/rfc4213.

[8] JIANG Xu-xian, XU Dong-yan. VIOLIN: Virtual Internetworking on Overlay INfrastructure [R]. Department of Computer Sciences Technical, Purdue University. Berlin: Springer, 2003.

[9] Tenzekhti F, Day K, Ould-Khaoua M. On Fault-Tolerant Data Replication in Distributed Systems [J]. Microprocessors and Microsystems, 2002, 26(7): 301-309.

[10] GUO Qing-nan. The Data Caching Protocol Supporting the Communication of Virtual Machine Live-Migration in WAN [D]: [Master's Degree Thesis]. Changchun: College of Computer Science and Technology, Jilin University, 2010. (郭庆南. 支持广域网虚拟机在线迁移通信的数据缓存协议 [D]: [硕士学位论文]. 长春: 吉林大学计算机科学与技术学院, 2010.)