



(12) 发明专利申请

(10) 申请公布号 CN 103201721 A

(43) 申请公布日 2013. 07. 10

(21) 申请号 201280002397. 7

G06F 9/455(2006. 01)

(22) 申请日 2012. 08. 29

(85) PCT申请进入国家阶段日  
2013. 03. 01

(86) PCT申请的申请数据  
PCT/CN2012/080703 2012. 08. 29

(71) 申请人 华为技术有限公司  
地址 518129 广东省深圳市龙岗区坂田华为  
总部办公楼

(72) 发明人 董益兼

(74) 专利代理机构 北京龙双利达知识产权代理  
有限公司 11329  
代理人 王君 肖鹏

(51) Int. Cl.  
G06F 9/48(2006. 01)

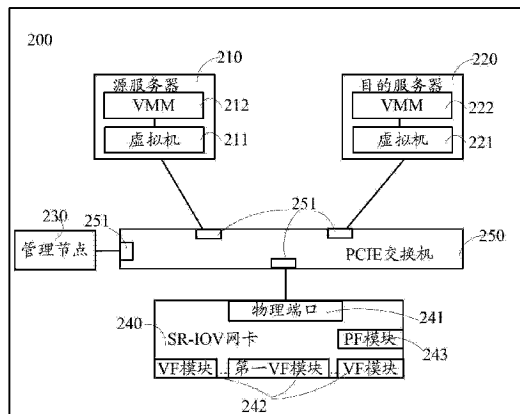
权利要求书3页 说明书11页 附图5页

(54) 发明名称

虚拟机热迁移的系统和方法

(57) 摘要

虚拟机热迁移的系统,包括管理节点、源服务器、目的服务器、PCIE交换机和SR-IOV网卡,源服务器包括热迁移前的VM;目的服务器包括热迁移后的VM;源服务器、目的服务器、管理节点以及SR-IOV网卡分别连接到PCIE交换机的物理端口以进行通信;管理节点用于通过PCIE交换机,将热迁移前的VM使用的VF模块与源服务器的连接关系配置为VF模块与目的服务器的连接关系;目的服务器通过PCIE交换机,根据管理节点配置的与VF模块的连接关系,使用VF模块,以完成虚拟机热迁移。该系统通过切换上述连接关系从而保证收发数据包业务不丢失,保证网络业务的连续性。



1. 一种虚拟机 VM 热迁移的系统,其特征在于,包括管理节点、源服务器、目的服务器、高速外围部件互连 PCIE 交换机和单根模式的 I/O 设备虚拟化 SR-IOV 网卡;

所述源服务器、所述目的服务器、所述管理节点以及所述 SR-IOV 网卡分别连接到所述 PCIE 交换机的物理端口以进行通信;

所述源服务器包括热迁移前的虚拟机,所述热迁移前的虚拟机为所述虚拟机热迁移的源端;

所述 SR-IOV 网卡包括输入输出的物理端口以及虚拟功能 VF 模块,其中,所述输入输出的物理端口用于与所述 PCIE 交换机的一个物理端口连接以进行通信;

所述管理节点用于通过所述 PCIE 交换机,将所述热迁移前的虚拟机使用的所述 VF 模块与所述源服务器的连接关系配置为所述 VF 模块与所述目的服务器的连接关系;

所述目的服务器包括热迁移后的虚拟机,所述热迁移后的虚拟机为所述虚拟机热迁移的目的端,所述目的服务器的虚拟机监控器,用于通过所述 PCIE 交换机,根据所述管理节点配置的与所述 VF 模块的连接关系,使用所述 VF 模块,以完成虚拟机热迁移。

2. 根据权利要求 1 所述的系统,其特征在于:

所述 VF 模块与所述源服务器的连接关系为所述 VF 模块通过所述 PCIE 交换机的一个物理端口与所述源服务器的连接关系,所述 VF 模块与所述目的服务器的连接关系为所述 VF 模块通过所述 PCIE 交换机的一个物理端口与所述目的服务器的连接关系。

3. 根据权利要求 1 或 2 所述的系统,其特征在于:

所述管理节点,还用于向所述源服务器的虚拟机监控器发起迁移请求;以及将所述 VF 模块与所述源服务器的连接关系配置为所述 VF 模块与所述目的服务器的连接关系后,通知所述目的服务器的虚拟机监控器;

则,所述源服务器的虚拟机监控器,用于根据所述迁移请求暂停所述热迁移前的虚拟机,并获取所述热迁移前的虚拟机的数据和所述 VF 模块的配置信息,发送到所述目的服务器的虚拟机监控器;

所述目的服务器的虚拟机监控器用于根据所述管理节点配置的与所述 VF 模块的连接关系,使用所述 VF 模块,具体为:所述目的服务器的虚拟机监控器,用于接收所述管理节点的所述通知,将所述 VF 模块配置给所述热迁移后的虚拟机,并根据从所述源服务器的虚拟机监控器接收的所述热迁移前虚拟机的数据和所述 VF 模块的配置信息配置所述 VF 模块,以使所述热迁移后的虚拟机使用所述 VF 模块在所述目的服务器上恢复运行。

4. 根据权利要求 3 所述的系统,其特征在于,所述管理节点将所述 VF 模块与所述源服务器的连接关系配置为所述 VF 模块与所述目的服务器的连接关系,包括:

将所述 PCIE 交换机的一个物理端口与所述源服务器的硬件的地址映射配置为所述 PCIE 交换机的一个物理端口与所述目的服务器的硬件的地址映射,所述物理端口为所述 VF 模块所使用的。

5. 根据权利要求 3 或 4 所述的系统,其特征在于,所述目的服务器的虚拟机监控器将所述 VF 模块配置给所述热迁移后的虚拟机包括:

改变所述目的服务器的输入输出内存管理单元 IOMMU 硬件的地址映射信息,以便所述 VF 模块根据所述地址映射信息访问所述热迁移后的虚拟机。

6. 根据权利要求 3 至 5 任一所述的系统,其特征在于:

所述源服务器的虚拟机监控器,还用于在所述管理节点将所述 VF 模块与所述源服务器的连接关系配置为所述 VF 模块与所述目的服务器的连接关系前,如果所述热迁移前的虚拟机还有待发送的数据包,将所述待发送的数据包发送到所述目的服务器的虚拟机监控器;

所述热迁移后的虚拟机,还用于在所述目的服务器上恢复运行后,通过所述 VF 模块发送所述目的服务器的虚拟机监控器接收的所述待发送的数据包。

7. 根据权利要求 3 至 6 任一所述的系统,其特征在于:

所述 VF 模块,用于,所述管理节点将所述 VF 模块与所述源服务器的连接关系配置为所述 VF 模块与所述目的服务器的连接关系后,如果所述热迁移前的虚拟机还有待接收的数据包时,执行直接内存存取 DMA,将所述待接收的数据包发送到所述目的服务器的虚拟机监控器;并向所述热迁移后的虚拟机发起中断请求,由所述目的服务器的虚拟机监控器接收并记录所述中断请求;

则,所述目的服务器的虚拟机监控器,还用于,在所述热迁移后的虚拟机在所述目的服务器上恢复运行后,将所记录的所述中断请求发送到所述热迁移后的虚拟机,

所述热迁移后的虚拟机,还用于处理所述目的服务器的虚拟机监控器接收的所述待接收的数据包。

8. 一种虚拟机 VM 热迁移的方法,其特征在于,包括:

管理节点通过高速外围部件互连 PCIE 交换机,将虚拟功能 VF 模块与源服务器的连接关系配置为所述 VF 模块与目的服务器的连接关系,以使所述目的服务器的虚拟机监控器根据所述管理节点配置的与所述 VF 模块的连接关系使用所述 VF 模块完成虚拟机热迁移;

其中,所述 VF 模块是配置给所述热迁移前的虚拟机使用的单根模式的 I/O 设备虚拟化 SR-IOV 网卡上的 VF 模块;

其中,所述源服务器、所述目的服务器、所述管理节点以及所述 SR-IOV 网卡分别连接到所述 PCIE 交换机的物理端口以进行通信;所述源服务器为热迁移前的虚拟机所在的服务器,所述热迁移前的所述虚拟机为所述虚拟机热迁移的源端;所述目的服务器为热迁移后的虚拟机所在的服务器,所述热迁移前的所述虚拟机为所述虚拟机热迁移的目的端;所述 SR-IOV 网卡包括输入输出的物理端口以及 VF 模块,其中,所述输入输出的物理端口用于与所述 PCIE 交换机的一个物理端口连接以进行通信。

9. 根据权利要求 8 所述的方法,其特征在于:

所述 VF 模块与源服务器的连接关系为所述 VF 模块通过所述 PCIE 交换机的一个物理端口与所述源服务器的连接关系,所述 VF 模块与所述目的服务器的连接关系为所述 VF 模块通过所述 PCIE 交换机的一个物理端口与所述目的服务器的连接关系。

10. 根据权利要求 8 或 9 所述的方法,其特征在于,所述方法还包括:

所述管理节点向所述源服务器的虚拟机监控器发起迁移请求;

所述源服务器的虚拟机监控器 VMM 根据所述迁移请求暂停所述热迁移前的虚拟机,并获取所述热迁移前的虚拟机的数据和所述 VF 模块的配置信息,发送到所述目的服务器的虚拟机监控器;

所述管理节点将所述 VF 模块与所述源服务器的连接关系配置为所述 VF 模块与所述目的服务器连接关系后,通知所述目的服务器的虚拟机监控器;

所述目的服务器的虚拟机监控器接收所述管理节点的所述通知,将所述 VF 模块配置给所述热迁移后的虚拟机,并根据从所述源服务器的虚拟机监控器接收的所述热迁移前虚拟机的数据和所述 VF 模块的配置信息配置所述 VF 模块;

所述热迁移后的虚拟机使用所述 VF 模块在所述目的服务器恢复运行。

11. 根据权利要求 10 所述的方法,其特征在于,所述管理节点将所述 VF 模块与所述源服务器的连接关系配置为所述 VF 模块与所述目的服务器的连接关系包括:

所述管理节点将所述 PCIE 交换机的一个物理端口与所述源服务器的硬件的地址映射配置为所述 PCIE 交换机的一个物理端口与所述目的服务器的硬件的地址映射,所述物理端口为所述 VF 模块所使用的。

12. 根据权利要求 10 或 11 所述的方法,其特征在于,所述目的服务器的虚拟机监控器将所述 VF 模块配置给所述热迁移后的虚拟机包括:

改变所述目的服务器的输入输出内存管理单元 IOMMU 硬件的地址映射信息,以便所述 VF 模块根据所述地址映射信息访问所述热迁移后的虚拟机。

13. 根据权利要求 10 至 12 任一所述的方法,其特征在于,所述方法还包括:

在所述管理节点将所述 VF 模块与所述源服务器的连接关系配置为所述 VF 模块与所述目的服务器的连接关系前,如果所述热迁移前的虚拟机还有待发送的数据包,所述源服务器的虚拟机监控器将所述待发送的数据包发送到所述目的服务器的虚拟机监控器;

所述热迁移后的虚拟机在所述目的服务器上恢复运行后,通过所述 VF 模块发送所述目的服务器的虚拟机监控器接收的所述待发送的数据包。

14. 根据权利要求 10 至 13 任一所述的方法,其特征在于,所述方法还包括:

在所述管理节点将所述 VF 模块与所述源服务器的连接关系配置为所述 VF 模块所述目的服务器的连接关系后,如果所述热迁移前的虚拟机还有待接收的数据包,所述 VF 模块还执行直接内存存取 DMA,将所述待接收的数据包发送到所述目的服务器的虚拟机监控器;

并向所述热迁移后的虚拟机发起中断请求,由所述目的服务器的虚拟机监控器接收并记录所述中断请求;

在所述热迁移后的虚拟机在所述目的服务器上恢复运行后,所述目的服务器的虚拟机监控器还将所记录的所述中断请求发送到所述热迁移后的虚拟机;

所述热迁移后的虚拟机处理所述目的服务器的虚拟机监控器接收的所述待接收的数据包。

## 虚拟机热迁移的系统和方法

### 技术领域

[0001] 本发明涉及服务器领域,并且更具体地,涉及虚拟机热迁移的系统和方法。

### 背景技术

[0002] 服务器虚拟化技术是当前的热点,而虚拟机(英文为Virtual Machine,缩写为VM)的“热迁移(英文为Live Migration)”技术则是虚拟化技术当中的热点。虚拟机热迁移技术主要应用于保持服务器的负载均衡或虚拟机的容灾、备份等场景。

[0003] 虚拟机热迁移技术是指将整个虚拟机的运行状态完整保存下来,同时可以快速地恢复到另一个物理服务器上。理想的情况,恢复以后,虚拟机仍旧平滑运行,用户不会察觉到任何差异。

[0004] 现有技术中,虚拟机热迁移根据系统的不同,存在多种应用方案。然而,虚拟机热迁移的过程中经常出现接收数据包和 / 或发送数据包丢失,无法保证网络通信的连续性。

### 发明内容

[0005] 有鉴于此,本发明实施例提供一种虚拟机热迁移的系统和方法,以保证虚拟机热迁移过程中网络通信的连续性。

[0006] 第一方面,提供了一种虚拟机 VM 热迁移的系统,包括管理节点、源服务器、目的服务器、PCIE 交换机和 SR-I/OV 网卡,其中,源服务器、目的服务器、管理节点以及 SR-I/OV 网卡分别连接到 PCIE 交换机的物理端口以进行通信;源服务器包括热迁移前的虚拟机,热迁移前的虚拟机为虚拟机热迁移的源端;SR-I/OV 网卡包括输入输出的物理端口以及虚拟功能 VF 模块,其中,所述输入输出的物理端口用于与 PCIE 交换机的一个物理端口连接以进行通信;管理节点用于通过 PCIE 交换机,将热迁移前的虚拟机使用的 VF 模块与源服务器的连接关系配置为 VF 模块与目的服务器的连接关系;目的服务器包括热迁移后的虚拟机,热迁移后的虚拟机为虚拟机热迁移的目的端;目的服务器的虚拟机监控器用于通过 PCIE 交换机,根据管理节点配置的与 VF 模块的连接关系,使用 VF 模块,以完成虚拟机热迁移。

[0007] 在第一种可能的实现方式中, VF 模块与源服务器的连接关系为 VF 模块通过 PCIE 交换机的一个物理端口与源服务器的连接关系, VF 模块与目的服务器的连接关系为 VF 模块通过 PCIE 交换机的一个物理端口与目的服务器的连接关系。

[0008] 在第二种可能的实现方式中或结合第一方面的第一种可能的实现方式,在第三种可能的实现方式中,管理节点,还用于向源服务器的虚拟机监控器发起迁移请求;以及将 VF 模块与源服务器的连接关系配置为 VF 模块与目的服务器的连接关系后,通知所述目的服务器的虚拟机监控器;则,源服务器的虚拟机监控器 VMM,还用于根据迁移请求暂停热迁移前的虚拟机,并获取热迁移前的虚拟机的数据和 VF 模块的配置信息,发送到目的服务器的虚拟机监控器;目的服务器的虚拟机监控器,用于根据管理节点配置的与 VF 模块的连接关系,使用 VF 模块,具体为:目的服务器的虚拟机监控器,用于接收管理节点的通知,将 VF 模块配置给热迁移后的虚拟机,并根据从源服务器的虚拟机监控器接收的热迁移前虚拟机

的数据和 VF 模块的配置信息配置 VF 模块,以使热迁移后的虚拟机使用 VF 模块在目的服务器上恢复运行。

[0009] 结合第一方面的第三种可能的实现方式,在第四种可能的实现方式中,管理节点将 VF 模块与源服务器的连接关系配置为 VF 模块与目的服务器的连接关系,包括:将 PCIE 交换机的一个物理端口与源服务器的硬件的地址映射配置为 PCIE 交换机的一个物理端口与目的服务器的硬件的地址映射,该物理端口为 VF 模块所使用的。

[0010] 结合第一方面的第三种可能的实现方式或第四种可能的实现方式,在第五种可能的实现方式中,目的服务器的虚拟机监控器将 VF 模块配置给热迁移后的虚拟机包括:改变目的服务器的输入输出内存管理单元(IOMMU)硬件的地址映射信息,以便 VF 模块根据地址映射信息访问热迁移后的虚拟机。

[0011] 结合第一方面的第三种至第五种任一可能的实现方式,在第六种可能的实现方式中,源服务器的虚拟机监控器,还用于在管理节点将 VF 模块与源服务器的连接关系配置为 VF 模块与目的服务器的连接关系前,如果热迁移前的虚拟机还有待发送的数据包,将待发送的数据包发送到目的服务器的虚拟机监控器;热迁移后的虚拟机,还用于在目的服务器上恢复运行后,通过 VF 模块发送目的服务器的虚拟机监控器接收的待发送的数据包。

[0012] 结合第一方面的第三种至第六种任一可能的实现方式,在第七种可能的实现方式中,VF 模块,用于,在管理节点将 VF 模块与源服务器的连接关系配置为 VF 模块与目的服务器的连接关系后,如果热迁移前的虚拟机还有待接收的数据包时,执行直接内存存取 DMA,将待接收的数据包发送到目的服务器的虚拟机监控器;并向热迁移后的虚拟机发起中断请求,由目的服务器的虚拟机监控器接收并记录中断请求;目的服务器的虚拟机监控器,还用于在热迁移后的虚拟机在目的服务器上恢复运行后,将所记录的中断请求发送到热迁移后的虚拟机,热迁移后的虚拟机还用于处理目的服务器的虚拟机监控器接收的待接收的数据包。

[0013] 第二方面,提供了一种虚拟机 VM 热迁移的方法,包括管理节点通过 PCIE 交换机,将 VF 模块与源服务器的连接关系配置为 VF 模块与目的服务器的连接关系,以使目的服务器的虚拟机监控器根据管理节点配置的与 VF 模块的连接关系,使用 VF 模块,以完成虚拟机热迁移;其中, VF 模块是配置给热迁移前的虚拟机使用的 SR-IOV 网卡上的 VF 模块;源服务器、目的服务器、管理节点以及 SR-IOV 网卡分别连接到 PCIE 交换机的物理端口以进行通信;源服务器为热迁移前的虚拟机所在的服务器,热迁移前的虚拟机为虚拟机热迁移的源端;目的服务器为热迁移后的虚拟机所在的服务器,热迁移前的虚拟机为虚拟机热迁移的目的端;SR-IOV 网卡包括输入输出的物理端口以及 VF 模块,其中,输入输出的物理端口用于与 PCIE 交换机的一个物理端口连接以进行通信。

[0014] 在第一种可能的实现方式中, VF 模块与源服务器的连接关系为 VF 模块通过 PCIE 交换机的一个物理端口与源服务器的连接关系, VF 模块与目的服务器的连接关系为 VF 模块通过 PCIE 交换机的一个物理端口与目的服务器的连接关系。

[0015] 在第二种可能的实现方式中或结合第二方面的第一种可能的实现方式,在第三种可能的实现方式中,管理节点向源服务器的虚拟机监控器发起迁移请求;源服务器的虚拟机监控器 VMM 根据迁移请求暂停热迁移前的虚拟机后,并获取热迁移前的虚拟机的数据和 VF 模块的配置信息,发送到目的服务器的虚拟机监控器;管理节点将 VF 模块与源服务器的

连接关系配置为 VF 模块与目的服务器的连接关系后,通知目的服务器的虚拟机监控器;目的服务器的虚拟机监控器接收管理节点的通知,将 VF 模块配置给热迁移后的虚拟机,并根据从源服务器的虚拟机监控器所接收的热迁移前虚拟机的数据和 VF 模块的配置信息配置 VF 模块;热迁移后的虚拟机使用 VF 模块在目的服务器恢复运行。

[0016] 结合第二方面的第三种可能的实现方式,在第四种可能的实现方式中,管理节点将 VF 模块与源服务器的连接关系配置为 VF 模块与目的服务器的连接关系包括:管理节点将 PCIE 交换机的一个物理端口与源服务器的硬件的地址映射配置为 PCIE 交换机的一个物理端口与目的服务器的硬件的地址映射,该物理端口为 VF 模块所使用的。

[0017] 结合第二方面的第三种可能的实现方式或第四种可能的实现方式,在第五种可能的实现方式中,目的服务器的虚拟机监控器将 VF 模块配置给热迁移后的虚拟机包括:改变目的服务器的输入输出内存管理单元(IOMMU)硬件的地址映射信息,以便 VF 模块根据地址映射信息访问热迁移后的虚拟机。

[0018] 结合第二方面的第三种至第五种任一可能的实现方式,在第六种可能的实现方式中,在管理节点将 VF 模块与源服务器的连接关系配置为 VF 模块与目的服务器的连接关系前,如果热迁移前的虚拟机还有待发送的数据包,源服务器的虚拟机监控器将待发送的数据包发送到目的服务器的虚拟机监控器;热迁移后的虚拟机在目的服务器上恢复运行后,通过 VF 模块发送目的服务器的虚拟机监控器接收的待发送的数据包。

[0019] 结合第二方面的第三种至第六种任一可能的实现方式,在第七种可能的实现方式中,在管理节点将 VF 模块与源服务器的连接关系配置为 VF 模块目的服务器的连接关系后,如果热迁移前的虚拟机还有待接收的数据包, VF 模块还执行直接内存存取 DMA,将待接收的数据包发送到目的服务器的虚拟机监控器;并向热迁移后的虚拟机发起中断请求,由目的服务器的虚拟机监控器接收并记录中断请求;在热迁移后的虚拟机在目的服务器上恢复运行后,目的服务器的虚拟机监控器还将所记录的中断请求发送到热迁移后的虚拟机,热迁移后的虚拟机处理目的服务器的虚拟机监控器接收的待接收的数据包。

[0020] 通过上述技术方案,可以在虚拟机热迁移过程中,通过 PCIE 交换机切换 SR-IOV 网卡上的 VF 模块的连接关系以保持热迁移前后虚拟机所连接的 SR-IOV 网卡的 VF 模块不变从而保证收发数据包业务不丢失,保证网络业务的连续性。

## 附图说明

[0021] 为了更清楚地说明本发明实施例的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0022] 图 1 是相关技术中虚拟机热迁移的一种迁移架构的示意图。

[0023] 图 2 是本发明实施例的虚拟机热迁移的系统的示意图。

[0024] 图 3 是本发明实施例的虚拟机热迁移前的系统的示意架构。

[0025] 图 4 是本发明实施例的虚拟机热迁移后的系统的示意架构。

[0026] 图 5 是本发明实施例的虚拟机热迁移过程中发送数据包处理方法的示意图。

[0027] 图 6 是本发明实施例的虚拟机热迁移过程中接收数据包处理方法的示意图。

## 具体实施方式

[0028] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0029] 本文中术语“和/或”,仅仅是一种描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。另外,本文中字符“/”,一般表示前后关联对象是一种“或”的关系。

[0030] 虚拟机热迁移主要包括以下几个技术点:外部存储设备迁移、内存状态迁移和网络设备迁移。

[0031] 外部存储设备迁移目前常用的方法包括通过网络共享外部存储设备,如SAN(英文为Storage Area Network,中文为存储局域网)或NAS(英文为Network-Attached Storage,中文为网络连接存储)之类的集中式共享外存设备。

[0032] 内存状态迁移目前常用的方法包括内存预拷贝技术等。

[0033] 网络设备迁移目前常用的方法大多适用于虚拟网络设备的迁移,虚拟网络设备也称为虚拟网卡。

[0034] 图1是相关技术中虚拟机热迁移的一种迁移架构的示意图。迁移架构中包括虚拟机迁移前的源服务器110和虚拟机迁移后的目的服务器150。源服务器110包括:源服务器的物理设备111、源服务器的虚拟机监控器(英文为Virtual Machine Monitor,缩写为VMM)112和热迁移前的虚拟机160。目的服务器150包括:目的服务器的物理设备151、目的服务器的VMM152和热迁移后的虚拟机165。

[0035] 源服务器的物理设备111和目的服务器的物理设备151通过互连网络180彼此连接。如常规服务器一样,源服务器110和目的服务器150的IT(英文为Information Technology,中文为信息技术)基础结构中包括各种物理设备,例如处理器、内存、网卡、磁盘和其他设备等。

[0036] VMM用于在服务器的物理设备上创建和管理一台或多台虚拟机,实现了物理和虚拟的IT基础结构的集中管理。

[0037] 虚拟机可以模拟完整硬件系统功能,可以运行操作系统、应用程序,通过VMM分配的虚拟网络设备,经服务器的物理设备的网卡可以访问网络资源。现有技术中,虚拟网络设备在虚拟机迁移中会造成接收数据包和/或发送数据包业务中断,无法保证网络通信的连续性。此外,由于虚拟网络设备的性能较低,VMM必须直接参与每项I/O(英文为Input/Output,中文为输入/输出)。这不仅会减缓数据传输速度,还会由于更频繁的VMM活动而增大服务器的处理器的负载,从而导致占用了较多的服务器的物理资源,不利于提高服务器性能。

[0038] 针对上述迁移架构存在的问题,本发明实施例提供了一种虚拟机热迁移的系统,可以保证迁移过程网络的连续性,使得收发数据包业务不中断。

[0039] 图2是本发明实施例的虚拟机热迁移的系统200的示意框图。

[0040] 系统200包括源服务器210、目的服务器220、管理节点230、SR-IOV(英文为Single



Root I/O Virtualization,中文为单根模式的 I/O 设备虚拟化)网卡 240 和 PCIE (英文为 Peripheral Component Interconnect Express,中文为高速外围部件互连)交换机 250 等多个物理节点。其中源服务器 210、目的服务器 220、管理节点 230 和 SR-IOV 网卡 240 分别连接到 PCIE 交换机 250 的不同物理端口 251 以进行通信。

[0041] 源服务器 210 包括 VMM212 和热迁移前的虚拟机 211,其中虚拟机 211 为虚拟机热迁移的源端。

[0042] 目的服务器 220 包括 VMM222 和热迁移后的虚拟机 221,其中虚拟机 221 为虚拟机热迁移的目的端。

[0043] PCIE 交换机 250,英文为 PCIE Switch,包括提供 PCIE 总线的多个 P2P (英文为 Peer to Peer,中文为端到端)的物理端口 251,用于连接不同的设备。

[0044] PCIE 总线使用 P2P 的连接方式,在一条 PCIE 链路(英文为 link)的两端只能各连接一个设备,这两个设备互为数据发送端和数据接收端。PCIE 总线的物理层为 PCIE 链路两端的设备间的数据通信提供了传送介质,为数据传送提供了可靠的物理环境。PCIE 交换机扩展了 PCIE 链路,可以连接多个物理设备,可以支持 SR-IOV 设备共享。PCIE 交换机支持 SR-IOV 设备共享是指通过 PCIE 交换机将 SR-IOV 设备共享给不同的设备。其中,SR-IOV 设备就是指支持 SR-IOV 功能的设备,例如,常用的 SR-IOV 网卡。

[0045] SR-IOV 网卡 240 包括一个输入输出的物理端口 241,物理端口 241 用于与 PCIE 交换机的一个物理端口连接以进行通信。

[0046] SR-IOV 网卡 240 还包括与物理端口 241 存在一对多映射关系的多个 VF (英文为 Virtual Function,中文为虚拟功能)模块 242,以及一个 PF (英文为 Physical Function,中文为物理功能)模块 243。

[0047] SR-IOV 网卡包括两种功能模块:PF 模块支持 SR-IOV 扩展功能的 PCIE 功能,被用于配置和管理 SR-IOV 功能特性;VF 模块实现简化的 PCIE 功能,包括数据迁移必需的资源,以及经过精简的配置资源集。SR-IOV 网卡可以配置成多个 VF 模块形式呈现在 PCI(英文为 Peripheral Component Interconnect,中文为外围部件互连)配置空间中,每个 VF 模块都有它们自己的 PCI 配置空间。每个 VF 模块可以支持针对 PCIE 层中 I/O 相关功能的一个特有且单独的数据路径(英文为 lane)。在网络设备中使用 SR-IOV 功能可使一个端口(功能)的带宽分成更小的块,然后通过一个标准接口分配到特定虚拟机中。这种资源的共享提高了在一个支持 SR-IOV 设备共享的 PCIE 交换机上任何特定资源的总利用率,从而降低了虚拟系统的成本。也就是说,SR-IOV 网卡包括的多个 VF 模块中的一个可以由 VMM 通过模拟配置空间来分给虚拟机。这样,一个 SR-IOV 网卡对多个虚拟机来说,具有呈现为多个独立的物理网卡的 I/O 共享特性。通过上述直通技术,可以使得虚拟机直接操作硬件,此处硬件为 SR-IOV 网卡,从而在带宽、延时等方面的性能都接近物理机直接使用设备的性能。而采用虚拟网络设备的情况,由于 VMM 必须直接参与每项 I/O,从而减缓数据传输速度,且由于更频繁的 VMM 活动而增大服务器的处理器的负载,影响服务器的性能,由此说 SR-IOV 网卡拥有高性能。

[0048] 管理节点 230 通过 PCIE 交换机 250,配置所述 SR-IOV 网卡 240 的 VF 模块与源服务器 210 之间的连接关系,即 VF 模块通过 PCIE 交换机的一个物理端口与源服务器的连接关系,例如,配置第一 VF 模块 242 通过 SR-IOV 网卡的物理端口 241,连接到 PCIE 交换机的

物理端口 251, 并通过 PCIE 交换机连接到源服务器 210, 以及配置 SR-IOV 网卡 240 的 VF 模块与目的服务器 220 之间的连接关系, 通过切换所述 VF 模块与源服务器和目的服务器之间的连接关系以管理虚拟机热迁移。管理节点是管理 PCIE 交换机的一个物理节点, 例如可以是一台主机。

[0049] 作为一种虚拟机热迁移的实现过程, 可选的, 管理节点向源服务器的虚拟机监控器发起迁移请求; 并在将热迁移前的虚拟机使用的 VF 模块通过 PCIE 交换机配置给目的服务器, 即将 VF 模块与源服务器的连接关系配置为 VF 模块与目的服务器的连接关系后, 通知目的服务器的虚拟机监控器; 则, 源服务器的虚拟机监控器根据迁移请求暂停热迁移前的虚拟机后, 获取热迁移前的虚拟机的数据和 VF 模块的配置信息, 发送到目的服务器的虚拟机监控器; 目的服务器的虚拟机监控器接收管理节点的通知, 将该 VF 模块配置给热迁移后的虚拟机, 并根据从源服务器的虚拟机监控器所接收的热迁移前虚拟机的数据和 VF 模块的配置信息配置该目的服务器的虚拟机的 VF 模块; 热迁移后的虚拟机使用 VF 模块在目的服务器恢复运行。

[0050] 本发明实施例提供了一种虚拟机热迁移的系统, 可以实现虚拟机热迁移, 且通过迁移前的虚拟机和迁移后的虚拟机利用同一物理 SR-IOV 网卡上的同一 VF 模块可以保证接收数据包业务不丢失, 从而保证网络业务的连续性。

[0051] 可选的, 管理节点将热迁移前的虚拟机使用的 VF 模块与源服务器的连接关系配置为 VF 模块与目的服务器的连接关系包括: 将 PCIE 交换机的一个物理端口与源服务器的硬件的地址映射配置为 PCIE 交换机的一个物理端口与目的服务器的硬件的地址映射, 该物理端口为该 VF 模块所使用的。

[0052] 可选的, 目的服务器的虚拟机监控器将 VF 模块配置给目的服务器的虚拟机包括: 改变目的服务器的输入输出内存管理单元 (英文为 Input Output Memory Management Unit, 缩写为 IOMMU) 硬件的地址映射信息, 以便 VF 模块根据地址映射信息访问热迁移后的虚拟机。

[0053] 可选的, 在管理节点将热迁移前的虚拟机使用的 SR-IOV 网卡的 VF 模块与源服务器的连接关系配置为 VF 模块与目的服务器的连接关系前, 如果热迁移前的虚拟机还有待发送的数据包, 源服务器的虚拟机监控器还用于将待发送的数据包发送到目的服务器的虚拟机监控器; 热迁移后的虚拟机在目的服务器恢复运行后, 通过 VF 模块发送目的服务器的虚拟机监控器从源服务器接收的待发送的数据包。

[0054] 可选的, VF 模块还用于, 在管理节点将 VF 模块与源服务器的连接关系配置为 VF 模块与目的服务器的连接关系后, 如果热迁移前的虚拟机还有待接收的数据包, 执行 DMA (英文为 Direct Memory Access, 中文为直接内存存取), 将待接收的数据包发送到目的服务器的虚拟机监控器, 并向热迁移后的虚拟机发起中断请求, 由目的服务器的虚拟机监控器接收并记录中断请求; 热迁移后的虚拟机在目的服务器恢复后, 目的服务器的虚拟机监控器, 还用于将所记录的中断请求发送到热迁移后的虚拟机, 热迁移后的虚拟机还用于处理待接收的数据包。

[0055] 图 3 是本发明实施例的虚拟机热迁移前的系统的示意架构。如图 3 所示, 虚拟机热切换前, 管理节点 230 通过配置 PCIE 交换机 250, 建立源服务器 210 和 SR-IOV 网卡 240 之间的连接关系。此时源服务器 210 的 VMM211 将 SR-IOV 网卡 240 的多个 VF 模块 242 中

的一个,如图中所示的第一 VF 模块 242 配置给热迁移前的虚拟机 212,即第一 VF 模块 242 通过 SR-IOV 网卡的物理端口 241 连接 PCIE 交换机的物理端口 251,并进一步通过 PCIE 交换机实现与源服务器的连接,虚拟机 212 通过第一 VF 模块 242 实现与其他设备的互连网络通信,例如收发数据包。同时,目的服务器 220 连接到 PCIE 交换机 250 的一个的物理端口 251。与 PCIE 交换机 250 连接的 SR-IOV 网卡 240 可以有多个。通过配置 PCIE 交换机 250,管理节点 230 建立了目的服务器 220 与 SR-IOV 网卡 240 中的一个 VF 模块 242 的连接关系。如果目的服务器 220 的虚拟机 222 已经创建,VMM221 可以从与目的服务器 220 建立连接关系的 SR-IOV 网卡 240 中选择一个 VF 模块 242,并将所选择的 SR-IOV 网卡中的一个 VF 模块 242 配置给虚拟机 222,被配置的 VF 模块是除第一 VF 模块以外的其他 VF 模块。可能的情况中包括,目的服务器 210 和源服务器 220 都连接到同一个 SR-IOV 网卡,但配置给虚拟机 212 和虚拟机 222 的 VF 模块是不同的。可选的,还可以以现有技术的方式为虚拟机 222 配置有虚拟网卡。这样,虚拟机 222 可以通过配置的 VF 模块或虚拟网卡与其他设备进行互连网络通信。可选的,VMM221 暂未对虚拟机 222 进行任何网络设备的配置工作。在本发明实施例中,为突出本发明实施例中连接关系的改变,在图 3 中未示出目的服务器的虚拟机 222 与 VF 模块的连接关系。

[0056] 图 4 是本发明实施例的虚拟机热迁移后的系统的示意架构。如图 4 所示,虚拟机热迁移后,原来配置给热迁移前的虚拟机 212 的第一 VF 模块 242 被目的服务器 220 的 VMM221 配置给热迁移后的虚拟机 222,且通过配置 PCIE 交换机 250,管理节点 230 建立了目的服务器 220 与最初源服务器 210 所使用的 SR-IOV 网卡 240 的连接关系,即第一 VF 模块 242 通过 SR-IOV 网卡的物理端口 241 连接 PCIE 交换机的物理端口 251,并进一步通过 PCIE 交换机实现与目的服务器的连接。此时,热迁移前的虚拟机 212 与第一 VF 模块之间的配置关系已经不存在,例如虚拟机 212 暂停(英文为 suspend),或直至虚拟机 212 被删除。

[0057] 本发明实施例提供的虚拟机热迁移的系统,通过切换上述图 3 和图 4 中的连接关系,实现了虚拟机 SR-IOV 网卡的热迁移,特别地,通过迁移前和迁移后利用同一 SR-IOV 网卡的同一 VF 模块可以保证接收数据包业务不丢失,从而保证网络业务的连续性。此外,由于虚拟机使用 VF 模块直接通过 SR-IOV 网卡进行网络通信,减少了 VMM 的参与,提高了源服务器或目的服务器的资源的利用率,从而提高了源服务器或目的服务器的性能,降低了虚拟系统的成本。进一步地,当源服务器和目的服务器通过 PCIE 交换机进行数据迁移时,比起使用外部网络通信进行的数据迁移,则效率高且稳定。

[0058] 图 5 是本发明实施例 50 的虚拟机热迁移过程中发送数据包处理方法 50 的示意图。

[0059] S51,管理节点向源服务器的 VMM 发起迁移请求。

[0060] 此时,与现有技术相同,源服务器和目的服务器已经存在通信连接。当源服务器收到迁移请求后,源服务器和目的服务器建立迁移的通道,也即源服务器发迁移请求到目的服务器;目的服务器接到迁移请求后,在目的服务器上构建和源服务器上的虚拟机配置相同的虚拟机。

[0061] 当前,管理节点已经通过配置 PCIE 交换机的端口与源服务器的硬件的地址的映射,建立了 SR-IOV 网卡与源服务器的连接关系,如图 3 所示。同时,热迁移前的虚拟机通过 SR-IOV 网卡的第一 VF 模块与其他设备进行网络通信。

[0062] S52,源服务器的VMM接收迁移请求,将虚拟机暂停,同时虚拟机的第一VF模块驱动停止运行。

[0063] 此处的虚拟机为虚拟机热迁移前的虚拟机,如图3中的212。

[0064] S53,源服务器的VMM将数据、第一VF模块的配置信息发送到目的服务器的VMM。

[0065] 该过程同现有技术,数据包括虚拟内存数据、磁盘数据、虚拟机的CPU(英文为Central Processing Unit,中文为中央处理器)的上下文等,如果此时迁移前的虚拟机还有未完成的需要发送的数据包,则数据中还包括待发送的数据包。可选的,该步骤可以通过源服务器和目的服务器的外部的互连网络。当通过本发明实施例的PCIE交换机交互数据时,较之通过外部的互连网络更为高效和可靠。

[0066] S54,管理节点将热迁移前的虚拟机使用的第一VF模块与源服务器的连接关系配置为第一VF模块与目的服务器连接关系,即将第一VF模块配置给目的服务器,并通知目的服务器的VMM。

[0067] 管理节点通过将PCIE交换机的一个物理端口与源服务器的硬件的地址映射配置为PCIE交换机的一个物理端口与目的服务器的硬件的地址映射,该物理端口为第一VF模块所使用的,从而建立SR-IOV网卡与目的服务器的连接关系,如图4所示。

[0068] S55,目的服务器的VMM根据从源服务器的VMM所接收的源服务器VM的第一VF模块的配置信息和热迁移前虚拟机的数据,配置目的服务器的虚拟机的VF模块。

[0069] 当目的服务器的VMM接收到管理节点的通知时,则根据从源服务器的VMM获知的第一VF模块的配置信息和热迁移前虚拟机的数据,将所述第一VF模块配置作为热迁移后的虚拟机的VF模块。目的服务器的VMM配置目的服务器的VF模块包括在内存中改变目的服务器的IOMMU硬件的地址映射信息,以便VF模块根据地址映射信息访问热迁移后的虚拟机。此后,目的服务器的VMM配置后的第一VF模块成为热迁移后的虚拟机的虚拟网络设备。

[0070] IOMMU是一种I/O硬件辅助虚拟化技术中的组件,用于维护虚拟机的内存和物理内存的地址映射关系。当外部设备例如SR-IOV网卡的VF模块访问虚拟机内存地址时,通过IOMMU转换,可以访问与虚拟机内存地址相映射的物理内存上存储的信息。I/O硬件辅助虚拟化技术例如包括VT-d(英文为Virtual Technology of Directed I/O,中文为直接的输入输出的虚拟技术)或AMD-Vi(英文为Advanced Micro Devices-Virtual I/O,中文为超微半导体设备-虚拟输入输出技术)。

[0071] S56,目的服务器的VMM恢复(resume)虚拟机。

[0072] S57,恢复后的虚拟机重启第一VF模块的通信,如果有迁移前未完成的需要发送的数据包,则还执行未完成数据包的发送流程。

[0073] 至此,热迁移后的虚拟机可以通过SR-IOV网卡的第一VF模块与其他设备进行网络通信。虚拟机完成从源服务器到目的服务器的迁移。

[0074] 由于在上述过程中,发送数据包的数据、VF模块的驱动状态信息、VF模块的配置信息都没有改动,只有PCIE交换机中第一VF模块的配置空间地址映射信息以及IOMMU硬件的地址映射信息被改变,IOMMU硬件的地址映射信息也就是虚拟机的内存与物理内存的地址映射信息。上层的虚拟机对此没有察觉,无需做任何改动,因此对发送数据包流程不会有任何影响。与其他虚拟机热迁移技术中需要更改虚拟机内核的方法比较,该热迁移的过程对上层软件透明,实现简便,且保证发送数据包业务不丢失,从而保证网络业务的连续

性。

[0075] 图 6 是本发明实施例 60 的虚拟机热迁移过程中接收数据包处理方法 60 的示意图。

[0076] S61,管理节点向源服务器的 VMM 发起迁移请求。

[0077] 当前,管理节点已经通过配置 PCIE 的端口与源服务器的硬件的地址的映射,建立了 SR-IOV 网卡与源服务器的连接关系,如图 3 所示。同时,热迁移前的虚拟机通过 SR-IOV 网卡的第一 VF 模块与其他设备进行网络通信。

[0078] S62,源服务器的 VMM 接收迁移请求,将待热迁移的虚拟机暂停(英文为 suspend),同时第一 VF 模块驱动停止运行。

[0079] S63,源服务器的 VMM 将数据、第一 VF 模块的配置信息发送到目的服务器的 VMM。

[0080] 该过程同现有技术。数据包括虚拟内存数据、磁盘数据、虚拟机的 CPU 的上下文等,如果此时迁移前的虚拟机还有未完成的需要发送的数据包,则数据中还包括待发送的数据包。可选的,可以通过源服务器和目的服务器的外部的互连网络。当通过本发明实施例的 PCIE 交换机交互数据时,较之通过外部的互连网络更为高效和可靠。

[0081] S64,管理节点将原先热迁移前的虚拟机使用的第一 VF 模块与源服务器的连接关系配置为第一 VF 模块与目的服务器连接关系,即将第一 VF 模块配置给目的服务器,并通知目的服务器的 VMM。

[0082] 管理节点通过通过将 PCIE 交换机的一个物理端口与源服务器的硬件的地址映射配置为 PCIE 交换机的一个物理端口与目的服务器的硬件的地址映射,该物理端口为第一 VF 模块所使用的,从而建立 SR-IOV 网卡与目的服务器的连接关系,如图 4 所示。

[0083] S65,目的服务器的 VMM 根据从源服务器的 VMM 所接收的源服务器 VM 的第一 VF 模块的配置信息和热迁移前虚拟机的数据,配置热迁移后的虚拟机的 VF 模块。

[0084] 目的服务器的 VMM 接收到管理节点的通知时,则根据从源服务器的 VMM 获知的第一 VF 模块的配置信息和热迁移前虚拟机的数据,将所述第一 VF 模块配置作为热迁移后的虚拟机的 VF 模块。目的服务器的 VMM 配置目的服务器的 VF 模块包括在内存中改变目的服务器的 IOMMU 硬件的地址映射信息,以便 VF 模块根据地址映射信息访问热迁移后的虚拟机。此后,目的服务器的 VMM 配置后的第一 VF 模块成为热迁移后的虚拟机的虚拟网络设备。

[0085] IOMMU 是一种 I/O 硬件辅助虚拟化技术中的组件,用于维护虚拟机的内存和物理内存的地址映射关系。当外部设备例如 SR-IOV 网卡的 VF 模块访问虚拟机内存地址时,通过 IOMMU 转换,可以访问与虚拟机内存地址相映射的物理内存上存储的信息。

[0086] S66,在迁移过程中,如果第一 VF 模块还在继续接收数据包,第一 VF 模块执行 DMA,将获得的数据包发送到热迁移后的虚拟机的内存中。具体的,第一 VF 模块通过 SR-IOV 网卡的描述符表寄存器(英文为 Descriptor Table Register)获得接收的数据包数据。

[0087] 需要注意的是,第一 VF 模块执行 DMA 时,由于管理节点和目的服务器的 VMM 已经配置了 SR-IOV 网卡的第一 VF 模块及 PCIE Switch 中的硬件地址的映射关系,并根据源服务器和目的服务器的 VM 虚拟内存分布信息重新配置 IOMMU,因此 DMA 过程中,SR-IOV 网卡的第一 VF 模块仍然可以把数据传送到目的服务器的 VM 内存中。源服务器和目的服务器的 VM 虚拟内存分布信息为虚拟机的内存到物理内存的映射信息,可以包括在虚拟机迁移所发送的数据中。

- [0088] S67, DMA 完成后,第一 VF 模块向目的服务器的 VM 发起一个中断请求。
- [0089] S68,由于中断映射到目的服务器的 VMM,中断不会直接到 VM (此时虚拟机尚未恢复),由 VMM 接收该中断请求并记录。
- [0090] S69,目的服务器的 VMM 恢复虚拟机后,将中断请求重新发送到恢复后的 VM,该 VM 完成被中断的接收数据包的处理过程。
- [0091] 同时,结合本发明实施例 50 可以还完成迁移过程中未处理的数据包处理流程。
- [0092] 本发明实施例提供的虚拟机热迁移的系统和方法,实现了虚拟机热迁移,且通过迁移前和迁移后利用同一 SR-IOV 网卡的同一 VF 模块可以保证接收数据包业务不丢失,从而保证网络业务的连续性。
- [0093] 此外,在实施例 50 和 60 中,由于虚拟机使用 VF 模块直接通过 SR-IOV 网卡进行网络通信,减少了 VMM 的参与,提高了源服务器或目的服务器的资源的利用率,从而提高了源服务器或目的服务器的性能,降低了虚拟系统的成本。进一步地,当源服务器和目的服务器通过 PCIE 交换机进行数据迁移时,比起使用外部网络通信进行的数据迁移,则效率高且稳定。
- [0094] 本领域普通技术人员可以意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本发明的范围。
- [0095] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统、装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。
- [0096] 在本申请所提供的几个实施例中,应该理解到,所揭露的系统、装置和方法,可以通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,所述单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。
- [0097] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。
- [0098] 另外,在本发明各个实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。
- [0099] 所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用,可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U 盘、移动硬盘、只读存储器(ROM, Read-Only Memory)、随机存取

存储器(RAM, Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0100] 以上所述, 仅为本发明的具体实施方式, 但本发明的保护范围并不局限于此, 任何熟悉本技术领域的技术人员在本发明揭露的技术范围内, 可轻易想到变化或替换, 都应涵盖在本发明的保护范围之内。因此, 本发明的保护范围应所述以权利要求的保护范围为准。

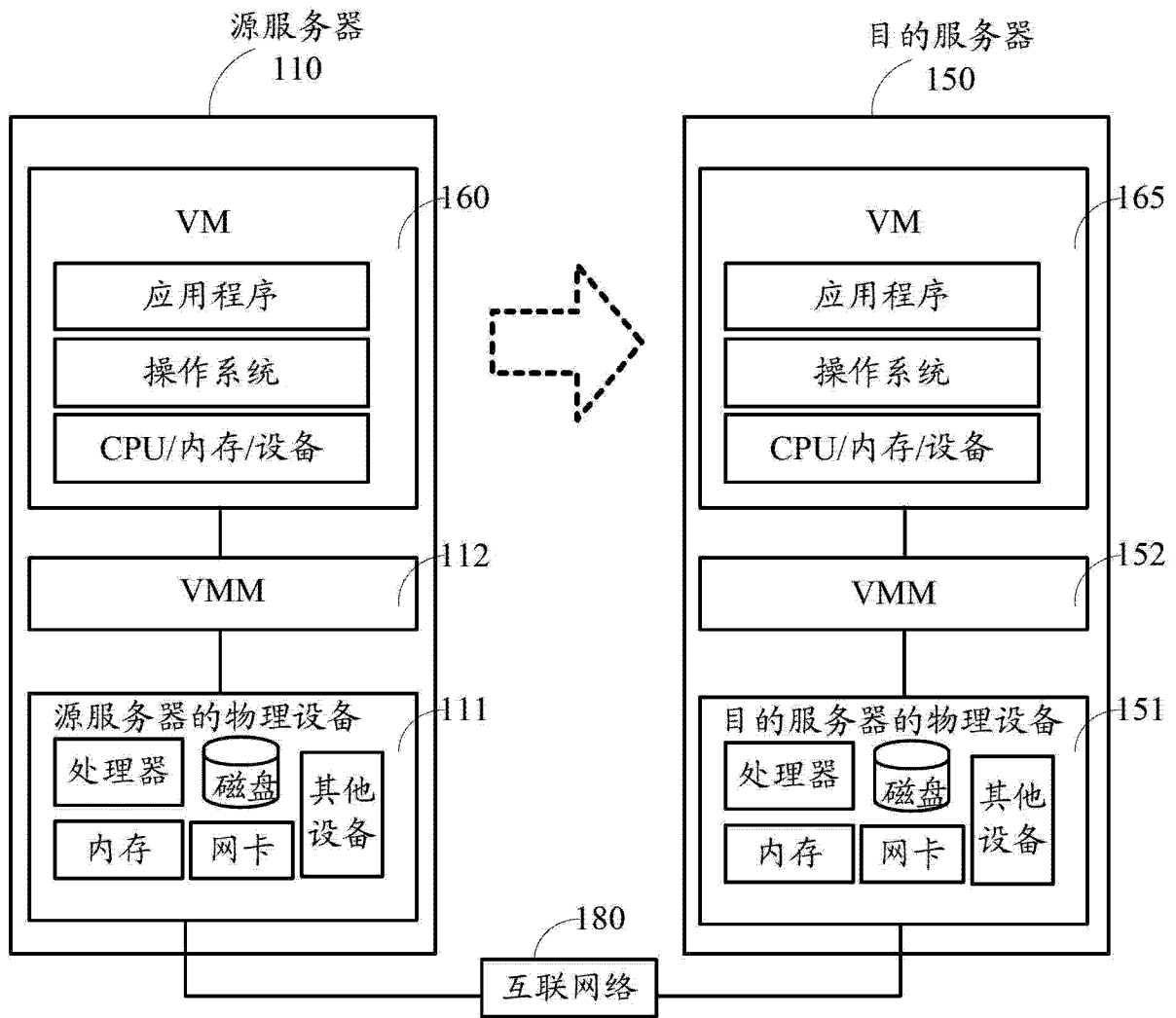


图 1



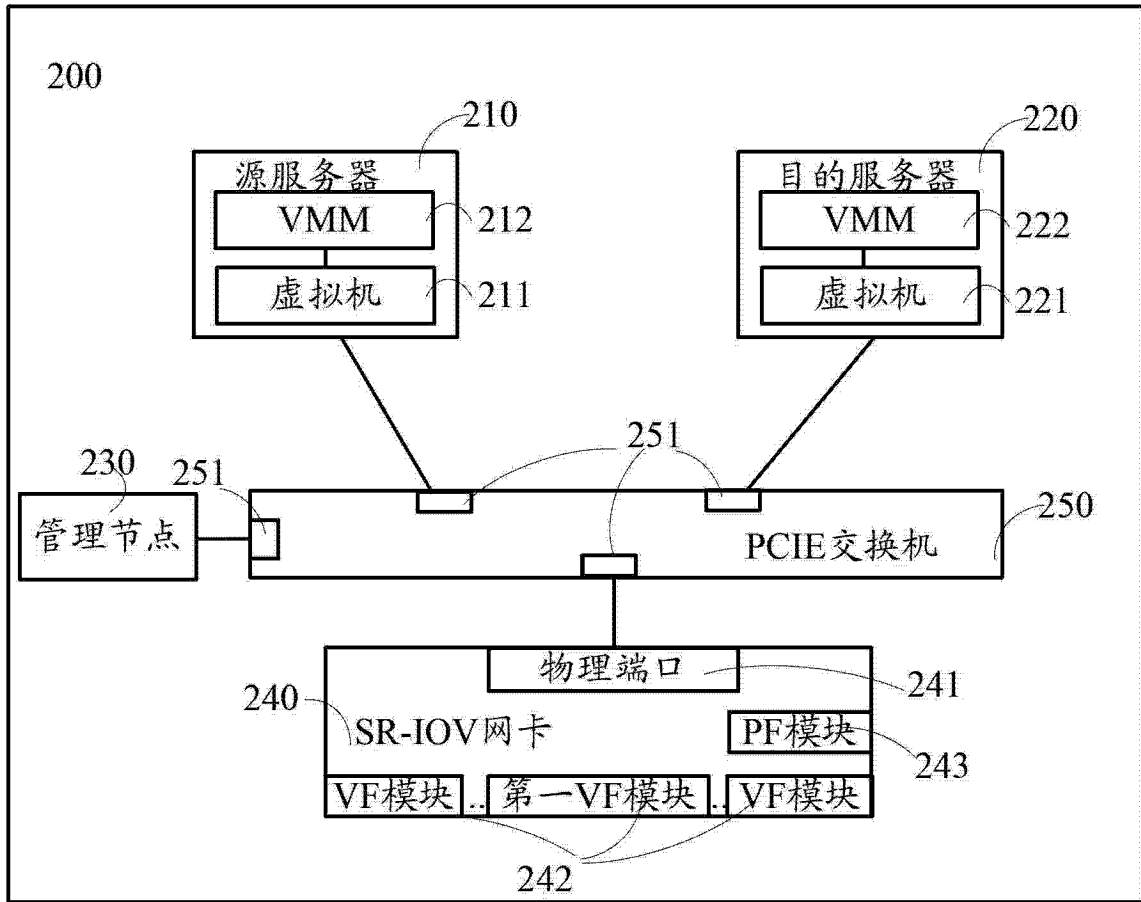


图 2

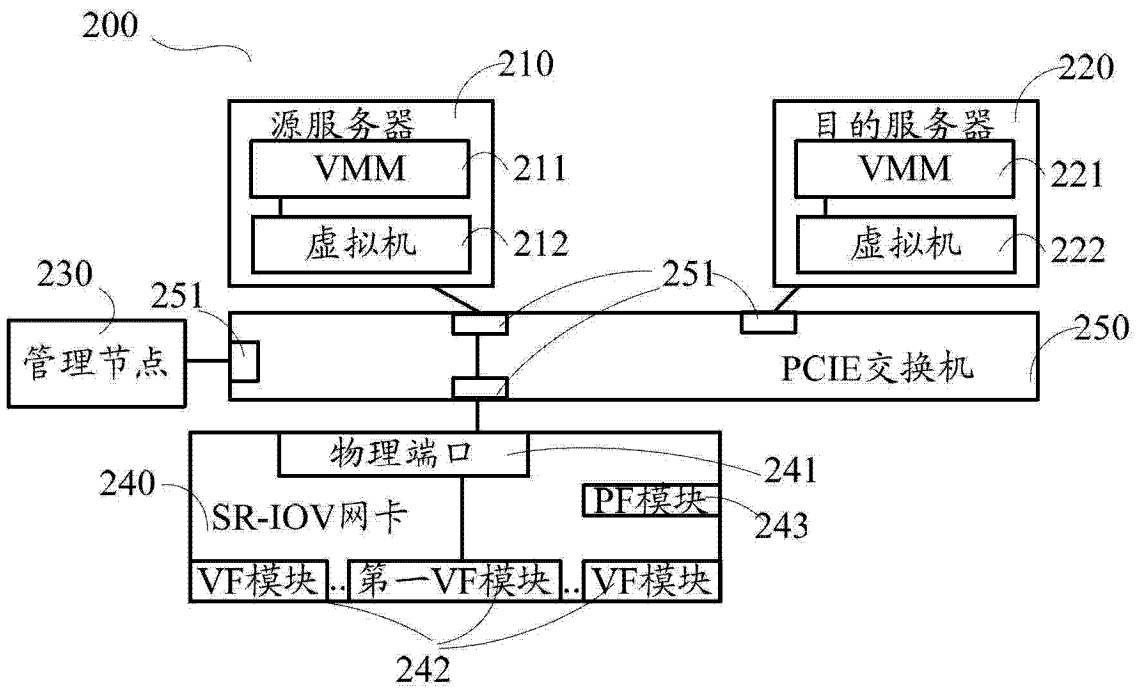


图 3

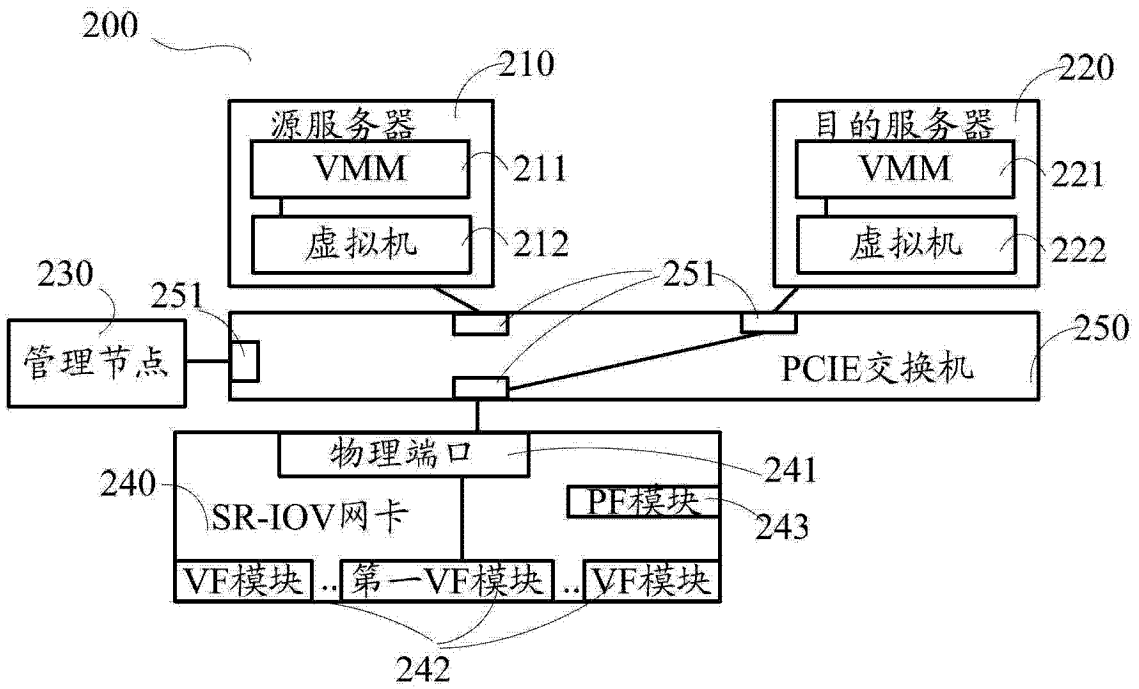


图 4

50

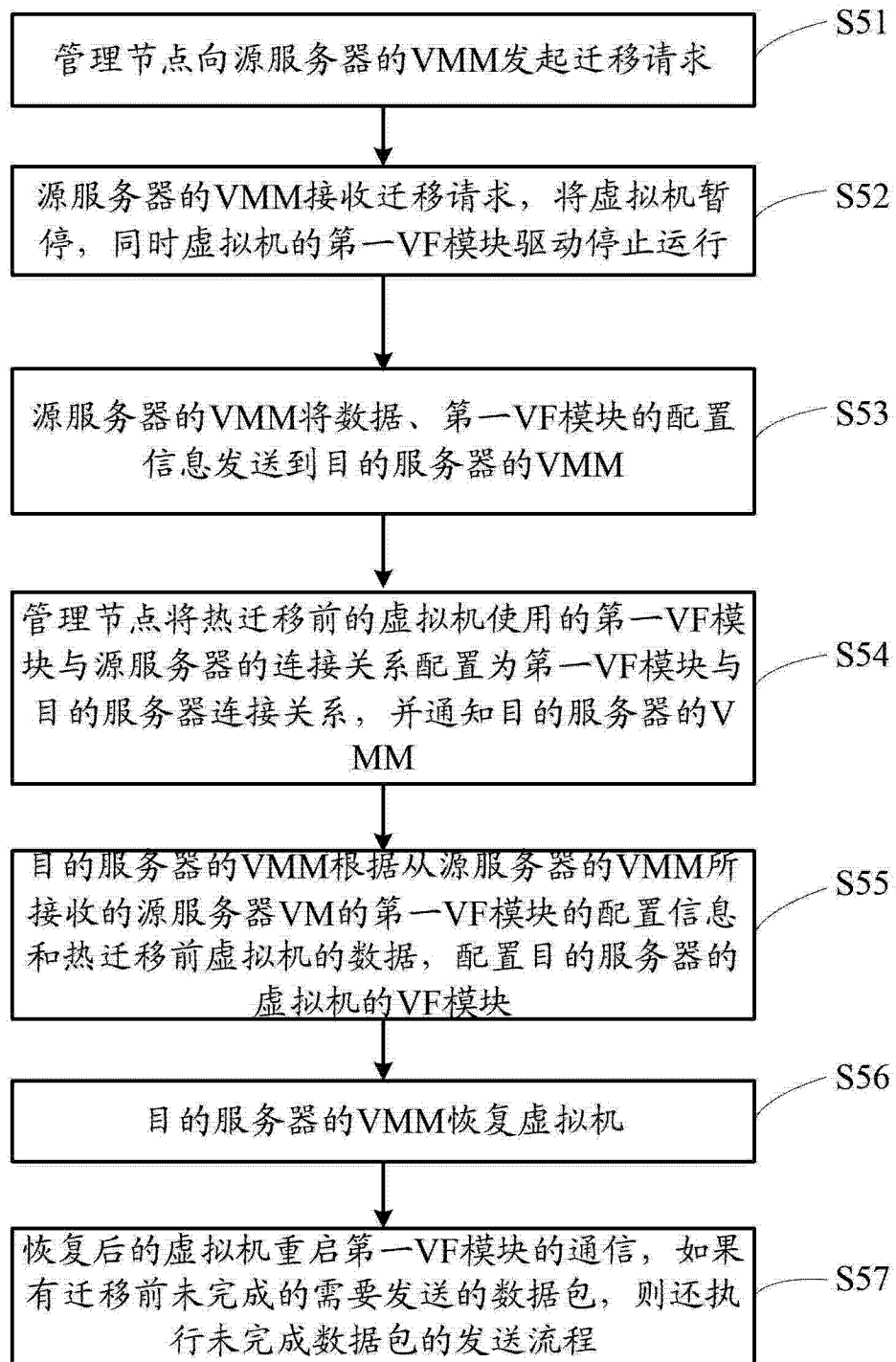


图 5

60

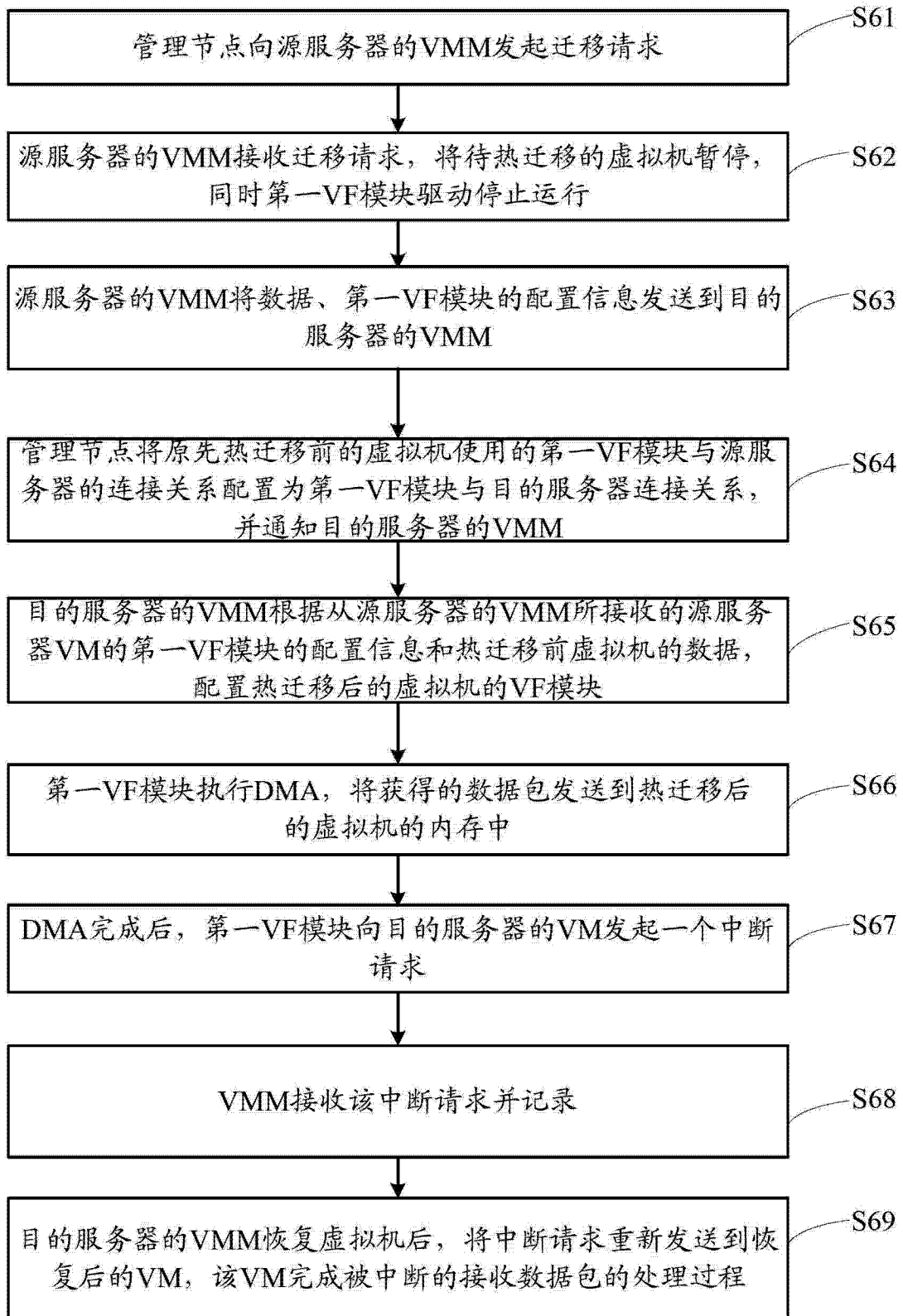


图 6