



(12)发明专利

(10)授权公告号 CN 103761145 B

(45)授权公告日 2017.06.09

(21)申请号 201310753146.9

(22)申请日 2010.01.26

(65)同一申请的已公布的文献号
申请公布号 CN 103761145 A

(43)申请公布日 2014.04.30

(30)优先权数据
61/147,269 2009.01.26 US
12/611,595 2009.11.03 US
12/611,607 2009.11.03 US
12/611,622 2009.11.03 US

(62)分案原申请数据
201080009842.3 2010.01.26

(73)专利权人 超威半导体公司
地址 美国加利福尼亚州

(72)发明人 本杰明·C·塞雷布林
唐纳德·W·麦考利
约翰·F·维德伊恩
伊丽莎白·M·库珀
马克·D·胡梅尔

(74)专利代理机构 上海胜康律师事务所 31263
代理人 李献忠

(51)Int.Cl.

G06F 9/48(2006.01)

(56)对比文件

CN 101128807 A, 2008.02.20,
CN 1297893 C, 2007.01.31,
US 7130949 B2, 2006.10.31,
US 7209994 B1, 2007.04.24,
Antonino Tumeo等. An Interrupt
Controller for FPGA-based
Multiprocessors.《2007 International
Conference on Embedded Computer Systems:
Architectures, Modeling and Simulation》
.2007, 第82-87页.

丁一等. 高级可编程中断控制器系统的研
究.《计算机工程与科学》.2005, 第97-100页.

(续)

审查员 杨龙

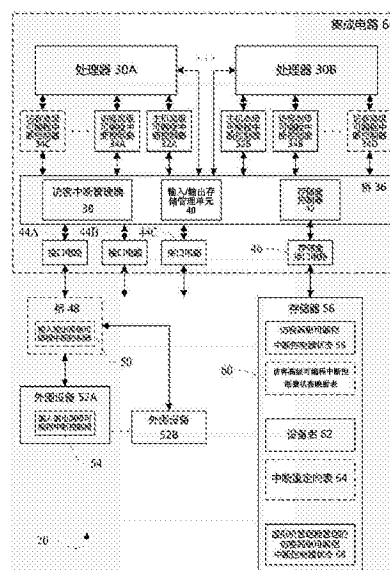
(54)发明名称

用于各个处理器的辅助中断虚拟化的访客
中断控制器

(57)摘要

在一实施方式中,系统包括处理器、与该处理器耦合的第一中断控制器和第二中断控制器。该第一中断控制器被配置为用信号通知处理器中断以对接收第一中断消息作出响应,该第一中断消息传送指向系统中的主机的第一中断。第二中断控制器被配置为用信号通知处理器中断以对接收第二中断消息作出响应,该第二中断消息传送指向访客的第二中断,该访客由主机控制并且在处理器上可执行。

权利要求书3页 说明书20页 附图12页



[转续页]

[接上页]

(56)对比文件

马玉英等.PCI中断机制分析.《青海师范大学学报(自然科学版)》.2005,(第2期),第51-53页.

Luis E. Leyva-del-foyo等.Real-Time

Scheduling of Interrupt Requests over Conventional PC Hardware.《2006 Seventh Mexican International Conference on Computer Science》.2006,第27-36页.

1. 一种中断控制方法,其包括:

在输入/输出存储管理单元中接收中断消息,其中该中断消息传送中断,其中所述中断源自外围设备,所述外围设备被分配给系统中多个访客中的第一访客;

通过所述输入/输出存储管理单元存取在存储器中的一个或多个第一数据结构以将所述中断消息映射到所述第一访客,其中所述一个或多个第一数据结构还包括指向另一个存储位置的指针;

响应于所述指针,通过访客中断管理器定位存储器中的一个或多个第二数据结构,其中所述一个或多个第二数据结构包括所述第一访客的中断控制器状态;

在所述一个或多个第二数据结构中记录所述中断,允许所述中断在所述第一访客执行时被传送到所述第一访客;以及

除了记录所述中断,响应于确定所述第一访客在所述中断消息被接收时执行,将所述中断传送到第一中断控制器,所述第一中断控制器被分配给所述第一访客并且是所述第一访客中的所述中断的目的地。

2. 根据权利要求1所述的方法,其中所述系统包括多个可分配给多个访客的中断控制器,并且其中所述第一中断控制器是所述多个中断控制器之一,并且其中所述传送包括将所述中断消息传送至所述多个中断控制器中的每一个。

3. 根据权利要求2所述的方法,其中所述多个中断控制器中的每一个包括访客标识符,该访客标识符标识被分配了该中断控制器的访客,并且其中传送所述中断包括传送用所述中断标识所述第一访客的第一访客标识符,并且其中所述方法还包括:

每个中断控制器将所述访客标识符与所述第一访客标识符进行比较;以及

所述第一中断控制器记录所述中断以响应在所述第一中断控制器中的所述访客标识符与所述第一访客标识符的匹配。

4. 根据权利要求3所述的方法,其中所述多个中断控制器中的每一个包括至少一个在被分配了该中断控制器的所述访客内的目的地标识符,并且其中传送所述中断包括传送与所述中断相关的第一目的地标识符,并且其中所述方法还包括:

每个中断控制器将所述目的地标识符与所述第一目的地标识符进行比较;以及

所述第一中断控制器记录所述中断以响应在所述第一中断控制器中的所述目的地标识符与所述第一目的地标识符的匹配。

5. 根据权利要求1所述的方法,其中所述中断控制器状态包括所述中断控制器中的中断请求寄存器的状态,且其中所述中断包括中断向量,且其中所述中断请求寄存器包括与所述向量相关的位位置,且其中记录所述中断包括在数据结构的所述位位置中设置所述位。

6. 根据权利要求5所述的方法,其中设置所述位是原子的。

7. 根据权利要求5所述的方法,其中设置所述位包括原子地将置位或运算到所述位位置中。

8. 根据权利要求5所述的方法,其中使用比较和交换操作来执行设置所述位。

9. 一种中断控制系统,其包括:

存储系统;以及

访客中断管理器,其被配置为接收中断消息,所述中断消息对应于指向在所述系统上

可执行的访客的中断,其中所述访客中断管理器被配置为将所述中断记录于所述存储系统中的一个或多个第一数据结构中,以确保即使所述访客在所述系统中在所述中断消息被接收时不是活动的,所述中断依然可被传送到所述访客;

外围设备,其被配置来初始化所述中断;

输入/输出存储管理单元,其被配置为接收来自所述外围设备的所述中断,其中所述输入/输出存储管理单元被配置为将所述中断与所述访客关联以响应所述存储系统中的一个或多个第二数据结构中的数据,且其中所述输入/输出存储管理单元被配置为传送包含有标识所述访客的访客标识符的所述中断消息以响应与所述访客相关的所述中断,且其中所述一个或多个第二数据结构还包括指向另一个存储位置的指针,其中所述访客中断管理器被配置成定位所述存储系统中的所述一个或多个第一数据结构以响应所述指针;以及

可分配给所述访客的中断控制器,其中所述访客中断管理器被配置成响应于在所述中断消息被接收时第一访客的执行,发送所述中断消息至所述中断控制器以捕获所述中断控制器中的所述中断。

10. 根据权利要求9所述的系统,其中所述中断控制器被配置为存储所述访客的所述访客标识符,且其中所述中断控制器被配置为将来自接收到的中断消息的访客标识符与存储在所述中断控制器中的访客标识符进行比较,且其中所述中断控制器被配置为接受所述中断以响应所述访客标识符比较过程中的匹配。

11. 根据权利要求9所述的系统,其中所述一个或多个第一数据结构包括多个条目,每个条目被配置为为给定访客存储中断控制器的状态的至少一部分,且其中所述条目中的所述状态包括中断请求寄存器的状态,且其中所述访客中断管理器被配置来更新所述中断请求寄存器的所述状态以记录所述中断。

12. 根据权利要求11所述的系统,其中所述中断请求寄存器包括与所述中断控制器所支持的每一个中断向量相对应的位,且其中所述中断消息包括所述中断的所述中断向量,且其中所述访客中断管理器被配置来更新与来自所述中断消息的所述中断向量相对应的所述位。

13. 根据权利要求12所述的系统,其中所述更新是原子的。

14. 根据权利要求13所述的系统,其中所述更新是原子或运算。

15. 一种中断控制方法,包括:

编程中断控制器以接受第一访客的中断;

在编程所述中断控制器之后读取由访客中断管理器维持的数据结构以捕获与所述第一访客相对应的中断状态,其中所述数据结构包括多个条目,每个条目被配置为为给定访客存储中断控制器的状态的至少一部分,且其中所述状态包括中断请求寄存器的状态;以及

用读取自所述数据结构的所述中断状态更新所述中断控制器的状态,其中所述中断控制器包括所述中断请求寄存器并且该中断请求寄存器被与所述第一访客对应的、来自所述数据结构中的所述条目的状态更新。

16. 根据权利要求15所述的方法,其中,更新所述中断控制器的状态包括将读取自所述数据结构的一组中断请求位或运算入所述中断控制器内的中断请求寄存器。

17. 根据权利要求15所述的方法,进一步包括:

编程所述中断控制器以停止接受所述第一访客的中断；
从所述中断控制器读取中断控制器状态；以及
将该中断控制器状态写入所述数据结构。

用于各个处理器的辅助中断虚拟化的访客中断控制器

[0001] 本申请是申请号为201080009842.3、申请日为2010年1月26日、发明名称为“用于各个处理器的辅助中断虚拟化的访客中断控制器”的发明专利申请的分案申请。

技术领域

[0002] 本发明涉及处理器和虚拟化,尤其涉及至虚拟机访客(guest)的传送中断。

背景技术

[0003] 出于各种不同的目的,虚拟化已经被应用在计算机系统中。例如,虚拟化可以被用来在“容器”中执行特权软件(privileged software)以防止特权软件未经控制虚拟机的虚拟机管理器(VMM)许可就直接存取和/或改变至少一些物理机状态。这样的容器能够防止“有臭虫的(buggy)”或恶意的软件在物理机上产生问题。除此之外,利用虚拟化可以允许两个或两个以上的特权程序在同一个物理机上并发执行。由于控制了对物理机的访问,因此可以防止特权程序彼此干扰。特权程序可以包括操作系统,也可以包括其他当其在硬件上执行时能够对硬件实现完全控制的软件。在另一实施例中,可利用虚拟化以在不同于特权程序所预期的硬件上执行该特权程序。

[0004] 通常地,处理器或计算机系统的虚拟化可以包括提供一个或多个对虚拟机(上面提到的容器)进行存取的特权程序,且该特权程序能够完全控制虚拟机,但是物理机的控制被VMM保留。上述虚拟机可以包括一个处理器(或多个处理器)、存储器以及特权程序预期在虚拟机中找到并能在其上执行特权程序的各种外围设备。虚拟机的元件可由VMM分配给(至少临时地分配给)虚拟机的硬件来实现,和/或由软件仿真来实现。每一个特权程序(以及某些情况下的相关软件,诸如在操作系统上执行的应用)此处可以被称作访客。虚拟化可以以软件(例如上面提到的VMM)实现,而不需要在其上执行VMM及其虚拟机的物理机中的任何具体的硬件虚拟化支持。然而,如果提供一些硬件支持,那么虚拟化就可以简化和/或达到更高的性能。

[0005] 虚拟化所引发的一个问题是中断传送的等待时间(latency)。如上所述,外围设备可以被虚拟机分配使用(来作为虚拟机中的虚拟外围设备)。这样的外围设备能够生成有待虚拟机中的软件来处理的中断。在非虚拟化环境中,中断处理的等待时间相对较短。在虚拟化环境中,利用某种软件机制,中断通常可被VMM拦截、处理并传送给目标虚拟机。但是中断处理的等待时间会显著变长(例如,大约增加了100倍)。

发明内容

[0006] 在一实施方式中,系统包括处理器、与该处理器耦合的第一中断控制器和第二中断控制器。该第一中断控制器被配置为用信号通知处理器中断以对接收第一中断消息作出响应,该第一中断消息传送指向系统中的主机的第一中断。第二中断控制器被配置为用信号通知处理器中断以对接收第二中断消息作出响应,该第二中断消息传送指向访客的第二中断,该访客由主机控制并且在处理器上可执行。

附图说明

- [0007] 下面参照附图来详细描述本发明,现在先简要说明附图。
- [0008] 图1是实行虚拟化的计算机系统的一实施方式的方块图。
- [0009] 图2是图1所示的主机硬件的一实施方式的方块图。
- [0010] 图3是说明被传送给访客的中断的一实施方式的方块图。
- [0011] 图4是说明访客高级可编程中断控制器 (APIC) 的一实施方式的方框图。
- [0012] 图5是说明访客APIC状态数据结构中的访客APIC状态条目 (entry) 的一实施方式的方块图。
- [0013] 图6是说明图2所示的访客中断管理器对接收指向访客的中断作出响应的一实施方式的操作流程图。
- [0014] 图7是说明访客APIC对接收中断消息作出响应的一实施方式的操作流程图。
- [0015] 图8是说明虚拟机监视器将访客APIC状态由一个访客改变为另一个访客的一实施方式的操作流程图。
- [0016] 图9是说明在访客APIC状态条目中排列中断状态的一实施方式的方块图。
- [0017] 图10是说明为中断定位访客APIC状态条目的一实施方式的方块图。
- [0018] 图11是说明为中断定位访客APIC状态条目的另一实施方式的方块图。
- [0019] 图12是说明为中断定位访客APIC状态条目的又一实施方式的方块图。
- [0020] 图13是说明为中断定位访客APIC状态条目的再一实施方式的方块图。
- [0021] 图14是图1所示的主机硬件的另一实施方式的方块图。
- [0022] 图15是存储VMM一实施方式的计算机可存取存储介质的一实施方式的方块图。
- [0023] 本发明的具体实施方式通过附图中的实施例示出并在本文中进行详细描述,同时本发明允许种种修正和替代形式。但应该明白的是,此处的附图及相关的详细描述并不试图将本发明限定在所公开的特定形式上,相反地,本发明应涵盖由所附权利要求书所定义的所有落于本发明精神与范围内的修正、等同方式与替代形式。此处使用的标题仅用于组织目的而不意味着限制说明书的范围。本申请中所使用的“可/可以”用来表示许可的意思(也就是具有可能性)而不是强制的意思(也就是必须)。同样的,“包括/包含”意味着包括在其中,但不是仅限于包括在其中的内容。
- [0024] 各种元件、电路或者其他组件可以被描述为“被配置为/用于(configured to)”执行一个或多个任务。在这样的上下文中,“被配置为/用于”是对通常意味着在操作过程中“具有电路系统”来执行一个或多个任务的结构之广义描述。同样地,上述元件/电路/组件可被配置为/用于执行任务即使是在该元件/电路/组件当前未打开时。通常来说,符合“被配置为/用于”的结构之电路系统可包括硬件电路来执行所述操作。类似的,为了便于描述,各种元件/电路/组件可被描述为执行一个或多个任务。这样的描述应该被解释为包括短语“被配置为/用于”。列举被配置为/用于执行一个或多个任务的元件/电路/组件并不特别意指使用35U.S.C. §112第6段中对元件/电路/组件的解释。

具体实施方式

- [0025] 在一实施方式中,计算机系统包括至少一个主机中断控制器和至少一个访客中断

控制器。主机中断控制器可管理由主机(例如,虚拟化环境中的虚拟机管理器(VMM))带来的中断。这样的中断可包括诸如从计算机系统上的未被分配给执行于该系统上的访客的设备而来的中断、VMM不希望暴露给访客的系统级中断,等等。访客中断控制器可管理由访客带来的中断。这样的中断可包括诸如由分配给访客以提供访客的虚拟机中的设备之功能性(functionality)的设备所分发的中断。

[0026] 该系统中的硬件可发送中断给主机中断控制器或访客中断控制器。可选地,中断也可以被发送给上述两个中断控制器,并且其中一个中断控制器可接收该中断,具体由哪个接收基于该中断是主机中断还是访客中断。每一个中断控制器都可与处理器耦合并且都可与处理器通信以传送中断。在一些实施方式中,提供访客中断控制器以传送中断到访客可以使得访客中断的传送具有与主机中断大致相等的等待时间。

[0027] 在一实施方式中,访客中断控制器可以是主机中断控制器的复本(duplicate)。也就是说,访客中断控制器与主机中断控制器可包含相同的硬件。在另一实施方式中,访客中断控制器可以仅重复(duplicate)一部分的主机中断控制器。与主机中断控制器不同的部分可以经由VMM拦截及仿真被VMM虚拟化。

[0028] 虚拟化概述

[0029] 图1示出了实行虚拟化的计算机系统5的一实施方式的方块图。在图1的实施方式中,出现了多个访客10A-10N。访客10A包括访客操作系统(OS) 12和一个或多个运行在访客操作系统上的应用14A-14N。访客10N包括特权代码16。访客10A-10N由虚拟机管理器(VMM) 18来管理。VMM18和访客10A-10N在主机硬件20上执行,该主机硬件20可以包括包含于计算机系统5中的物理硬件。在一实施方式中,VMM18可保有一组虚拟机控制块(VMCB) 22。每一个访客10A-10N都可对应一个VMCB22。尽管图1中VMCB22是作为VMM18的一部分来图示的,但是所述VMCB22可被存储在存储器和/或非易失性介质中,例如主机硬件20中的磁盘驱动器。

[0030] 主机硬件20通常包括包含在计算机系统5中的所有硬件。在各种实施方式中,主机硬件20可以包括一个或多个处理器、存储器、外围设备以及其他用以耦接前述组件的电路。特别是,主机硬件20可包括一个或多个主机中断控制器、一个或多个访客中断控制器和/或一个或多个访客中断管理器。例如,个人计算机(PC)式的系统可包含耦接处理器的北桥、存储器以及使用高级图形端口(AGP)接口的图形设备。此外,北桥可耦接到诸如外设组件接口(PCI)总线的外设总线,各种外设组件可直接或间接地耦接至外设总线。该PC式系统也可以包括耦接至PCI总线的南桥,以提供传统(legacy)功能性和/或耦接至传统(legacy)硬件。在各种实施方式中,访客中断管理器可在北桥、南桥或者在上述接口之一上面的设备中实施。主机中断控制器和访客中断控制器可实施于每一个处理器中,或者在一组处理器之间共享。在其他实施方式中,其他电路可用于连接各种硬件组件。例如,可使用超级传输(HT)链接来连接节点,每一个节点都可包括一个或多个处理器、主桥以及存储控制器。每一个节点也可包括可以包含有访客中断管理器和/或主机中断控制器及访客中断控制器的北桥。此外,主桥可包含访客中断管理器和/或主机中断控制器及访客中断控制器。该主桥可被用来以菊花链(daisy chain)的形式经由HT链接耦接至外围设备。任何期望的电路系统/主机硬件结构均可被使用。

[0031] VMM18可用于为每一个访客10A-10N提供虚拟化并且可以控制访客10A-10N对主机硬件20的访问。VMM18也可负责调度访客10A-10N以在主机硬件20上执行。VMM18可用于将主

机硬件20中提供的硬件支持用于虚拟化。例如,处理器可为虚拟化提供硬件支持,包括用以拦截事件并将访客退出到VMM18以进行处理的硬件。访客中断管理器和/或访客中断控制器也可以是配备来支持虚拟化的硬件。

[0032] 在一些实施方式中,VMM18可以作为在主机硬件20上实施的并且为访客10A-10N提供虚拟化的“薄(thin)”的独立软件程序来执行。这样的VMM实施方式有时可被称为“管理程序(hypervisor)”。在其他实施方式中,VMM18可被集成到主机OS中或者执行于主机OS上。在这样的实施方式中,VMM18可依赖主机OS,包括该主机OS中的任何驱动器、由系统BIOS提供的平台系统管理模式(SMM)代码等等。因此,主机OS组件(以及诸如平台SMM代码的各种较低层级的组件)直接执行于主机硬件20上且未被VMM18虚拟化。在一实施方式中,VMM18和主机OS(如果包括的话)可合起来称为主机。通常来说,主机可包括在使用过程中直接控制主机硬件20的任何代码。例如,主机可以是VMM18、结合主机OS的VMM18或者单单是主机OS(例如在非虚拟化环境中)。

[0033] 在各种实施方式中,VMM18可支持全虚拟化、半虚拟化或者两者都支持。另外,在一些实施方式中,VMM18可并发地执行半虚拟化的访客和全虚拟化的访客。

[0034] 全虚拟化时,访客10A-10N不知道虚拟化正在发生。每一个访客10A-10N可在其虚拟机中具有相邻的零基(zero based)存储器,并且VMM18可使用影子页表或嵌套页表来控制对主机物理地址空间的访问。影子页表可从访客虚拟地址重新映射到主机物理地址(有效地将访客10A-10N中的存储器管理软件所分配的访客“物理地址”重新映射到主机物理地址),而嵌套页表则可接收访客物理地址作为输入并映射到主机物理地址。通过对每一个访客10A-10N使用影子页表或嵌套页表,VMM18可以确保访客不会访问主机硬件20中的其他访客的物理存储器。

[0035] 半虚拟化时,访客10A-10N可以至少部分地感知虚拟机。这样的访客10A-10N可与VMM18协商存储页面,因此可以不需要将访客物理地址重新映射到主机物理地址。在半虚拟化的一实施方式中,可允许访客10A-10N直接与主机硬件20中的外围设备交互。在任何给定的时间,外围设备可被访客10A-10N中的一个或多个所“拥有”。例如,在一实施方式中,外围设备可通过一个或多个当前拥有此外围设备的访客10A-10N被映射到保护域中。只有拥有外围设备的访客可直接与之交互。还可以有保护机制来防止保护域中的设备读/写分配给其他保护域中的访客的页面。

[0036] 如前所述,VMM18可以为每一个访客10A-10N保留一个VMCB22。所述VMCB22通常可包括存储在由VMM18为相应的访客10A-10N分配的存储区域中的数据结构。在一实施方式中,VMCB22可包含存储页,尽管其他实施方式可以利用更大或更小的存储区域和/或在其他介质上的存储器(例如非易失性存储器)。在一实施方式中,VMCB22可包括访客的处理器状态,当调度该访客来执行时,该状态可被载入主机硬件20的处理器中,而当该访客退出时(要么由于其预定时间的结束,要么由于处理器检测的用于退出访客的一个或多个拦截),该状态可被存回VMCB22。在一些实施方式中,仅一部分的处理器状态经由将控制转移给对应于该VMCB22的访客的指令(“虚拟机运行(VMRUN)”指令)进行载入,其他的期望状态可在执行该VMRUN指令前由VMM18载入。同样地,在这样的实施方式中,仅一部分的处理器状态可在访客退出时由该处理器存储至VMCB22,且VMM18可在需要时负责存储任何其他状态。在其他实施方式中,VMCB22可以包括指向存储处理器状态的另一存储区域的指针。另外,在一实

施方式中,可以定义两种或两种以上的退出机制。在一实施方式中,所存储的状态的数量以及载入状态的位置可以根据所选择的退出机制而变。

[0037] 在一实施方式中,VMM18也可具有被分配来存储对应于该VMM18的处理器状态的存储器的区域。当执行VMRUN时,对应于所述VMM18的处理器状态可被保存在该区域中。当访客退出到VMM18时,该区域中的处理器状态可重新从该区域载入以让VMM18继续执行。例如,在一实施方式中,处理器可具备寄存器(例如模式特定寄存器(MSR))以存储VMM18保存区的地址。

[0038] 此外,该VMCB22可包括用以识别使能访客的拦截事件的拦截配置,以及如果检测到使能的拦截事件时用于退出访客的机制。在一实施方式中,拦截配置可包括一组拦截指示,其中每一个拦截指示用来指示处理器支持的每一个拦截事件。拦截指示可指示处理器是否要拦截相应事件(或者,从另一角度来看,该拦截是否可用)。如本文中所使用的,如果该处理器退出访客以处理事件,其中该事件应该发生在该访客中,则该事件在该访客内被“拦截”。在一实施方式中,拦截配置可包括第二组指示以指示使用两个退出机制中的哪一个。其他实施方式也可定义多于两个的退出机制。在另一实施方式中,拦截配置可包括一组拦截指示,每个事件一个指示,指示第一退出机制是否应用于该事件;以及第二组拦截指示,每个事件一个指示,指示第二退出机制是否应用于该事件。

[0039] 通常来说,退出机制可定义由处理器进行的退出访客执行的操作(一般是以重新启动的方式),并且开始执行其他代码。在一实施方式中,退出机制可包括存储少量的处理器状态并为微程序(Minivisor)载入状态。该微程序可在访客物理地址空间内执行并执行相对简单的拦截处理。另一退出机制可退出至该VMM,保存大量的处理器状态并载入VMM的处理器状态。因此,拦截事件可根据该事件由不同的指令代码来处理。此外,相对简单的拦截处理可通过耗费较少时间执行的“较轻权重(lighter weight)”退出机制来处理,在一些实施方式中这样的处理能够提高性能。在使用“较重权重(heavier weight)”机制以退出之后,可在该VMM中执行较复杂的处理。因此,在本实施方式中,该VMM18可配置处理器去拦截那些VMM18不希望访客10A-10N在内部处理的事件,也可以配置处理器使用哪种退出机制。事件可包含指令(即,拦截指令而不是执行该指令)、中断、异常和/或在访客执行过程中可能发生的其他事件。

[0040] 在一实施方式中,该VMCB22可进一步包括其他控制位,所述控制位可使得处理器在载入该VMCB22时执行某些动作。例如,控制位可包括刷新(flush)处理器中的TLB的指示。其他控制位可为访客指定其执行的环境(例如,中断处理模式,访客的地址空间标识符,等等)。还有一些其他控制位可用来传送描述访客为何退出的退出码,等等。

[0041] 通常来说,一个“访客”可包括任意一个或多个将被虚拟化以在计算机系统5中执行的软件程序。一个访客可包括至少一些执行在特权模式中的代码,并且因此希望可以完全控制执行该访客的计算机系统。如前所述,访客10A是一个包含有访客OS12的实施例。所述访客OS12可以是任何OS,诸如微软公司(Redmond,WA)的任何Windows OS、任何UNIX类型的操作系统(例如Linux)、IBM公司(Armonk,NY)的AIX、太阳微系统公司(Santa Clara,CA)的Solaris、惠普公司(Palo Alto,CA)的HP-UX,等等。访客10N是包含有非操作系统特权代码16的访客实施例。

[0042] 需要注意的是字母“N”,当此处使用诸如10N的参考数字时,一般是指带有参考数

字的元件编号(例如,访客10A-10N中的任何编号均包括一个访客)。此外,使用字母“N”的不同参考数字(例如,10N和14N)并不意指不同的元件具有相同的数量(例如,访客10A-10N的数量与应用14A-14N的数量不同),除非另有说明。

[0043] 主机硬件

[0044] 现在翻到图2,该图是说明主机硬件20的一实施方式的方块图。在所示的实施方式中,主机硬件20包括多个处理器30A-30B、各自的主机高级可编程中断控制器(hAPIC)32A-32B、各自的访客APIC(gAPIC)34A-34B以及可选的附加gAPIC34C-34D、桥36(其包括访客中断管理器38、输入/输出(I/O)存储管理单元(IOMMU)40以及存储器控制器42)、多个接口电路(IF)44A-44C、存储器接口电路(MIF)46、可选的桥48(其可包括IOAPIC50)、外围设备52A-52B(其中一些可包含诸如所述IOAPIC54的IOAPIC)以及存储器56。处理器30A-30B与桥36耦接并与图2所示的各自的hAPIC32A-32B和gAPIC34A-34D耦接。该hAPIC32A-32B和gAPIC34A-34D与桥36耦接,该桥36与接口电路44A-44C和存储器接口电路46耦接。该存储器接口电路46与存储器56耦接,接口电路44A与桥48耦接,桥48与外围设备52A-52B耦接。

[0045] 在所示的实施方式中,每一个处理器30A-30B具有相关的hAPIC32A-32B以及至少一个相关的gAPIC34A-34D。在一实施方式中,可根据由Intel公司(Santa Clara,CA)制定的APIC规范在主机硬件20中传送中断。在上述规范中,每一个处理器具有相关的能够接收中断的本地APIC(中断都来自所述处理器自身、其他处理器、内部APIC中断源以及与外围设备设备相关的IOAPIC)。本地APIC对未决的中断按照优先级进行排列,并且如果一个中断的优先级高于处理器中正在执行的另一中断的优先级和/或该中断的优先级高于处理器当前任务的优先级,则将该中断传送到处理器。

[0046] 在图2的实施方式中,hAPIC32A-32B可以是针对处理器的主机中断的本地APIC(即,中断由主机处理)并且gAPIC34A-34D可以是针对处理器的访客中断的本地APIC(即,中断由在各自的处理器30A-30B上处于活动状态的访客来执行)。如果访客当前正在处理器上执行(例如,VMM18指令已经在对访客的处理器上被执行并且访客退出还未发生),或者如果访客已经退出且VMM18正在执行,但是期待该访客再次在该处理器上被执行,则该访客在该处理器中是活动的。

[0047] 当VMM18调度处理器30A-30B上的访客时,VMM18可载入处理器30A-30B的gAPIC34A-34D及其对应于访客的gAPIC状态。尤其是,给定的访客可以具有多个虚拟CPU(vCPU)。该VMM18可调度访客的vCPU以在处理器30A-30B上执行,并且可载入gAPIC34A-34D以及该访客虚拟机中的vCPU的中断状态。此外,任何指向访客(以及vCPU)的中断都可由gAPIC34A-34D来捕获,所述访客在活动状态时被信号通知中断。如上述所述,gAPIC34A-34D可根据的APIC规范来中断访客。

[0048] 给定处理器30A-30B的hAPIC32A-32B和gAPIC34A-34D可具有到处理器的任何接口。例如,应用在本地APIC和他们各自的处理器之间的接口可是任何接口。每一个APIC可用来独立地用信号通知处理器中断将要被传送过来以供服务。如果处理器正在执行访客并且用信号通知了访客中断,处理器可用来中断访客代码并且开始执行访客虚拟机中的正确的中断句柄(handler)。这样的话,在一实施方式中,传送访客中断与传送主机中断可具有同样的等待时间。如果所述处理器正在执行访客并且该hAPIC用信号通知了中断,则处理器可用来退出访客到VMM18以处理主机中断。如果处理器未在执行访客,则由gAPIC用信号通知

的中断可被处理器屏蔽直到该访客被再次执行。如果处理器未在执行访客并且hAPIC用信号通知了中断,则处理器可用于中断主机执行并且分行到该主机中断句柄。

[0049] 在一实施方式中,每一个处理器30A-30B可包括不止一个gAPIC34A-34D。每一个gAPIC34A-34D可存储对应于不同访客/vCPU的APIC状态。在这样的实施方式中,当用信号通知访客中断给处理器时,每一个gAPIC34A-34D可用来标识其所对应的访客(或者处理器30A-30B可具有内部寄存器用来标识当前被分配给每一个gAPIC34A-34D的访客)。如果一个不同的访客当前正在执行,则处理器可屏蔽该访客中断,其与当VMM18未在执行时屏蔽该访客中断相类似。可选地,当相应的访客被调度时,每一个gAPIC34A-34D可以包括可被VMM18设置为活动状态的主动指示,并且当相应的访客处于活动状态时,所述gAPIC34A-34D可用来仅通知自身的访客中断。随着时间的推移,当多个访客被调度以在处理器上执行时,每一个处理器30A-30B包括多于一个的gAPIC34A-34D可减少gAPIC状态变化的数量。例如,如果每一个处理器30A-30B具有N个gAPIC34A-34D(其中N是大于0的整数),在gAPIC状态需要为任何一个访客保存之前,多达N个不同的访客可被调度来执行。在实现每一个处理器30A-30B具有多于一个的gAPIC34A-34D的一些实施方式中,该gAPIC34A-34D可包括附加状态来确保该中断消息可被正确地接受和/或记录(log)。例如,该gAPIC34A-34D可包括“当前运行”的指示用来标识相应的虚拟机当前是否运行在相应的处理器30A-30B上(相对于因VMM的执行而暂停或者于另一虚拟机正在执行时的暂停)。如果当前运行标识表明该虚拟机正在运行,则gAPIC可接受该中断消息。如果当前运行标识表明该虚拟机未在运行,则gAPIC可用信号通知不接受该中断。可选地,gAPIC可包括附加指示用来标识该gAPIC是否用信号通知不接受中断。在这样的实施方式中,该gAPIC可用信号通知中断,即使当前运行标识指出当前未运行,并且不接受(not-accepted)标识指示该gAPIC用信号通知不被接受的中断。这样的功能可用来检测针对未运行访客的中断是否被接收,这可用来调度该中断所指向的访客。

[0050] gAPIC34A-34D可包括包含在hAPIC32A-32B中的至少部分硬件,也可包括其所有硬件(例如,可以是hAPIC32A-32B的副本)。除了APIC状态之外,gAPIC34A-34D还可用访客标识符(ID)编程以确定gAPIC34A-34D被分配给哪一个访客。如果所述访客包括多个vCPU,那么物理APIC ID和逻辑APIC ID可标识该访客内部的vCPU。在一实施方式中,访客ID可以与由IOMMU40支持的外围设备的域ID相同。可选地,访客ID可以是被单独管理的资源。在另一个例子中,VMM18可分配访客ID给访客并可确保gAPIC34A-34D针对每个访客而进行恰当地编程。vCPU和/或gAPIC和/或这两者在此处可更简洁地被称为访客中的中断的目的地。目的地最终可以是中断服务的vCPU,但是由于gAPIC与相应的处理器相关并且也记录中断,因此相应的gAPIC也可被看作是目的地。

[0051] 该gAPIC34A-34D和hAPIC32A-32B被耦接到桥36上以接收中断。可使用任意的接口来传送中断到gAPIC34A-34D和hAPIC32A-32B。例如,可使用任意的为APIC中断传送而设置的接口。在一实施方式中,可使用与用来传送其他操作给处理器30A-30B/从处理器30A-30B传送其他操作(诸如由处理器30A-30B发起的存储器读/写操作、缓存一致性维护的探测,等等)的相同的传送机制来传送中断消息。从另一角度看,可将gAPIC34A-34D与hAPIC32A-32B的耦接与处理器30A-30B到桥36的耦接共享。可选地,处理器30A-30B可具有到桥36的独立路径,比如在gAPIC34A-34D和hAPIC32A-32B采用APIC“3线接口”时。中断消息可以是在任意接口上识别正被传送的中断和中断目的地的任意通信。例如,中断可具有相关的中断向量,

且该中断向量可以是中断消息的一部分。中断消息也可以包括访客ID和目的地ID(例如,逻辑APIC ID或物理APIC ID)。

[0052] 该hAPIC32A-32B可与本地APIC相同。例如,由于该hAPIC为主机中断所使用,因此该hAPIC32A-32B不包括用于访客识别的额外硬件。可选地,hAPIC32A-32B可包括额外的硬件但是该额外的硬件可被编程以表明该hAPIC32A-32B是用于主机中断的。被桥36传送到hAPIC32A-32B和gAPIC34A-34D的中断消息可识别访客中断以及与之相对的主机中断,并可包括访客中断的访客ID(或者可使用预留的访客ID,诸如零或全二进制以指示主机中断)。hAPIC32A-32B可用于接受被标识为主机中断的中断(如果主机中断的物理APIC ID或逻辑APIC ID与相应的hAPIC ID相匹配),并且gAPIC34A-34D可用于接受其各自访客的访客中断(如果访客ID相匹配,以及如果访客中断的物理APIC ID或者逻辑APIC ID与相应的gAPIC ID相匹配)。

[0053] 当gAPIC34A-34D可为活动访客管理中断时,一些访客可以是非活动的(或者可具有被访客中断指向的非活动的vCPU)。在一实施方式中,访客中断管理器38可用于为非活动的访客维持访客中断状态并确保活动访客的gAPIC接收它们的中断。

[0054] 尤其是,在一实施方式中,访客中断管理器38可利用分布式中断传送方案,其中该访客中断管理器38可用于记录桥36中所接收的每一个访客中断,也可用于传送该访客中断到每一个gAPIC34A-34D。如果gAPIC34A-34D接受该中断,那么访客中断所指向的访客是活动的。如果没有gAPIC34A-34D接受该中断,那么该访客中断所指向的访客是非活动的。

[0055] 在所示的实施方式中,访客中断管理器38可用于为访客维持gAPIC的状态,所述访客被定义于系统5的存储器56中的gAPIC状态数据结构58中。该gAPIC状态数据结构58可包括在系统5中所定义的每一个gAPIC的gAPIC状态条目(例如,该系统中每一个访客10A-10N的每一个vCPU的一个条目)。如果gAPIC与系统中活动或非活动的访客相关联的话,则gAPIC可被定义在该系统中。据此,该访客中断管理器38可被配置为为中断所指向的访客/vCPU更新gAPIC状态数据结构58中的gAPIC状态以对接受访客中断作出响应。在一实施方式中,访客中断管理器38可用于更新gAPIC状态而不管访客是否是活动的。为了组播和广播具有多于一个目标的中断,访客中断管理器38可用于为每一个中断目的地更新gAPIC状态数据结构中的gAPIC状态。可选地,访客中断管理器38可被配置为为这些多目的地中断而依赖VMM18。在这样的实施方式中,访客中断管理器38可用于在VMM18可存取的存储位置中记录该中断,且可用于用信号通知VMM18去处理该消息。

[0056] 在一些实施方式中,该访客中断管理器38可用于直接定位gAPIC状态数据结构58中的gAPIC状态条目以响应访客中断消息中的访客ID和/或其他信息。在其他实施方式中,为了提供gAPIC状态数据结构58中的灵活性和/或为了节省存储器空间,访客中断管理器38可被配置为采用gAPIC状态映射表60来定位gAPIC状态数据结构58中的gAPIC状态条目。gAPIC状态数据结构58和映射表60(对某些实施方式)的不同实施方式被图示于图10-13中并在下面进行更为详细的讨论。据此,访客中断管理器38可用于通过gAPIC状态映射表60来定位gAPIC状态条目并且用于更新该gAPIC状态条目以记录该中断以响应访客中断。

[0057] 在一实施方式中,gAPIC状态数据结构58可存储gAPIC状态的子集。该子集可以是被硬件20(例如访客中断管理器38与IOMMU40的联合)追踪的gAPIC状态。更特别的是,该子集可以gAPIC状态的一部分,其中当相应的访客是非活动的时,该部分gAPIC状态可随之改

变。例如,在一实施方式中,当相应的访客是非活动的时,外围设备52A-52B可用信号通知中断,这将导致相应的中断请求被gAPIC捕获。该中断请求可在gAPIC状态数据结构58中被追踪。其他gAPIC状态可通过处理器追踪服务中的中断、处理器的任务优先级,等等。只有当访客处于活动状态时,这些值才可以改变。在一实施方式中,当访客处于非活动状态时不可改变的gAPIC状态可由VMM18追踪,其利用在图2中显示为VMM管理gAPIC状态数据结构68的一个或多个其他数据结构来进行追踪。在该系统中,VMM18可在VMM管理的状态68和作为部分的激活访客与关闭访客的gAPIC34A-34D之间传送状态。

[0058] 虽然gAPIC状态映射表60和gAPIC状态数据结构58在所示的实施方式中都显示为存储在存储器56中,但其中之一者或两者的部分可被访客中断管理器38和/或桥36可存取的缓存存储起来。此外或可替换的,一个或多个gAPIC状态条目的专用存储器可被实现在桥36中。该专用存储器可存储一组可被快速转入和转出gAPIC34A-34D的“快速”gAPIC状态。而其他的gAPIC状态则可在存储器56中以较慢的速度来存取。在一些实施方式中,该快速gAPIC状态开关可由访客中断管理器38操纵而较慢的gAPIC状态开关则可由VMM18操纵。

[0059] 在APIC中断机制中,每一个处理器(通过其本地APIC)可具有物理APIC ID和逻辑APIC ID。物理APIC ID存储在APIC ID寄存器中。物理APIC ID在一对一的基础上与由物理传送模式中中断所指示的物理APIC ID相匹配。逻辑APIC ID作为逻辑目的地寄存器被存储在本地APIC中。逻辑APIC ID具有簇ID和本地APIC ID,其中本地APIC ID是独热(one-hot)向量。逻辑传送模式中中断可包括独热向量中的任意置位(set bit)以传送中断到簇中的一个或多个本地APIC。据此,逻辑APIC ID的匹配可包括簇ID的比较以及本地APIC ID向量中的置位的检测,该置位与本地APIC中独热向量的置位具有相同的位置。从另一个角度看,在逻辑传送模式中中断中的本地APIC ID向量可与本地APIC的本地APIC ID向量进行逻辑与操作(AND),并且如果结果为非零且簇ID能够匹配,则本地APIC是逻辑中断的目标。此处的逻辑APIC ID可更简洁地被称为逻辑ID,类似的,物理APIC ID可更简洁的被称为物理ID。与中断相关的给定ID(物理的或逻辑的)可被称为中断的目的地ID。中断的相应传送模式可识别中断的目的地ID。

[0060] gAPIC34A-34D也可以支持物理传送和逻辑传送两种模式。除了根据上面突显的模式在中断消息中匹配APIC ID之外,gAPIC34A-34D可将中断消息中的访客ID匹配到gAPIC中的访客ID。

[0061] IOMMU40可用于对I/O启动的存储操作(例如,源自外围设备52A-52B的存储器读/写操作或由代表外围设备52A-52B的DMA控制器执行的存储器读/写操作)执行虚拟地址至物理地址的映射。作为转换操作的一部分,IOMMU40可用于存取设备表62且可选地存取访问中断重定向表64。设备表62可包括对每一个外围设备52A-52B的条目(并可包括一个外围设备的多个条目,该条目包括与外围设备耦接的外围设备接口上的不止一个标识符)。设备表62可包括用于转换存储器读/写操作(未图示)的存储地址的指向I/O页表的页表指针并可包括指向中断重定向表64的指针。在一些实施方式中,设备表62可以为分配给访客的外围设备存储访客ID。在一实施方式中,访客ID可以与IOMMU40中用于设备访问保护的域ID相同。可选地,访客ID可被独立分配。在一实施方式中,设备表62也可存储指向gAPIC状态映射表60(如果使用的话)的指针,或者指向gAPIC状态数据结构58的指针。在另一实施方式中,访客ID和/或指向表60/数据结构58的指针可被存储在中断重定向表64中。该中断重定向表

64可用于将中断从其初始的目的地和/或中断向量重定向到新的目的地和/或中断向量。为了简化本发明的其他部分,将采用这样一种实施方式,在此实施方式中,访客ID是来自设备表62的域ID且指向映射表60和/或gAPIC状态数据结构58的指针被存储在设备表62中。然而,本发明剩余部分的实施方式通常可按照上述讨论的内容进行修正。

[0062] 在其他实施方式中,可以不提供访客中断管理器38。例如,可能存在这样的配置,如果在访客从一个处理器30A-30B转移到另一个时由VMM18更新设备表62和/或中断重定向表64,以及如果处理器30A-30B专用于代表非活动访客接收中断(以按期望更新存储器56中的gAPIC状态数据结构58和/或为中断服务),则不需要访客中断管理器38。

[0063] 存储器控制器42可以被耦合以接收由处理器30A-30B发出的存储操作(例如,提取指令、存取载入/存储数据、存取用于转换的处理器页表,等等)、来自访客中断管理器38的存储操作(例如,读取/更新gAPIC状态数据结构58和/或gAPIC状态映射表60)、来自IOMMU40的存储操作(例如,访问I/O页表、设备表62以及中断重定向表64),以及接收来自接口电路44A-44C的存储操作(在某些实施方式中)。存储器控制器42可用于对存储操作进行排序并与存储器56通信以执行存储器操作。存储器接口电路46可执行到存储器56的物理层访问。

[0064] 存储器56可包括任何类型的存储器。例如,存储器56可包括诸如同步DRAM(SDRAM)的动态随机存取存储器(DRAM)以及双倍数据速率(DDR、DDR2、DDR3等)SDRAM、RAMBUS DRAM、静态RAM,等等。存储器56可包括一个或多个包含有多存储芯片的存储模块,诸如单列直插式存储模块(SIMM)、双列直插式存储模块(DIMM),等等。

[0065] 在此实施方式中,除了包括访客中断管理器38、IOMMU40以及存储器控制器42之外,桥36也可包括其他通信功能以在处理器30A-30B、hAPIC32A-32B、gAPIC34A-34D以及耦接到接口电路44A-44D的设备之间进行通信。例如,在所示的实施方式中,另一个桥48可被耦接到接口电路44A上,并可用于在接口电路44A所采用的协议和外围设备52A-52B所采用的协议之间进行桥通信。在一实施方式中,接口电路44A-44C可实现诸如上面提到的HT接口,并且桥48可从HT桥接到诸如PCI Express(PCIe)接口的另一接口。在这一实施方式中,外围设备52A-52B可以是PCIe设备。桥48也可用于桥接到其他接口,或另一个桥可被耦合到桥48以桥接到其他接口。可使用任何一个或多个外围设备接口。此外,外围设备52A-52B可包括被配置为直接耦接到HT接口的HT外围设备。这种外围设备可以不需要桥48。

[0066] 在一实施方式中,桥48和/或一个或多个外围设备52A-52B可包括IOAPIC(图2中的50和54)。IOAPIC可负责接收来自外围设备的中断请求并形成中断消息以传送中断请求给hAPIC32A-32B和中断管理器38(用于到gAPIC34A-34D的传送和/或记录在存储器中的gAPIC状态数据结构58中)。

[0067] 如上所述,在一实施方式中,接口电路44A-44C可用于HT接口上的通信。该接口电路44A-44C可用于通过HT与外围设备/桥进行通信。此外,在一些实施方式中,接口电路44A-44C可用于耦接到其他具有处理器、hAPIC和gAPIC等的节点。在这样的实施方式中,桥36可包括除了之前描述的电路系统之外的一致性管理电路系统(coherence management circuitry)。

[0068] 处理器30A-30B可实施任意指令集架构,并可用于执行定义在指令集架构中的指令。处理器30A-30B可包括任意微架构,诸如超流水线架构、超标量架构和/或二者的结合;按顺序或不按顺序的执行;预测的执行;等等。根据需要,处理器30A-30B可以实施或不实施

微编码技术。

[0069] 外围设备52A-52B可包括任意类型的外围设备。外围设备52A-52B可包括存储设备,比如磁盘驱动器、固态磁盘驱动器、光盘驱动器或者诸如闪存之类的非易失性存储设备,等等。外围设备52A-52B可包括I/O设备,比如用户I/O设备(键盘、鼠标、显示器、语音输入等等)、网络设备、诸如通用串行总线(USB)或火线的外部接口设备,等等。

[0070] 在所示的实施方式中,处理器30A-30B、桥36、hAPIC32A-32B、gAPIC34A-34D、接口电路44A-44C以及存储器接口电路46可被集成在一个单独的半导体基底上以形成集成电路66。其他实施方式可根据需要设置不同数量的集成电路系统和离散电路系统。值得注意的是,尽管图2示出了不同数量的组件,诸如处理器、hAPIC、gAPIC、接口电路、外围设备、桥,等等,但其他实施方式也可根据需要设置任何数量的一个或多个上述每一个组件。

[0071] 在其他实施方式中,IOMMU40的位置和访客中断管理器38的位置是可变的。例如,其中一个或所有两个可位于桥48中、外围设备52A-52B中、耦接到该桥的另一个桥中,等等。

[0072] 在所示的实施方式中,每一个gAPIC34A-34D和每一个hAPIC32A-32B与图2所示的特定处理器30A-30B相关联。这样,给定的中断控制器可专用于该实施方式中相应的处理器30A-30B。更具体地,在图2中,hAPIC32A、gAPIC34A和34C专用于处理器30A,hAPIC32B、gAPIC34B和34D专用于处理器30B。中断控制器可以用任何方式将中断用信号通知给其相应的处理器。通常来说,该通知可表示需要中断。该通知可包括中断向量,或者中断向量可被中断传送之后执行的软件读取。在一实施方式中,传送中断可以指用信号通知处理器以及处理器接受该中断。为中断服务可以指执行与中断向量相关的中断服务程序以执行中断设备所需要的操作。

[0073] 现在翻到图3,图3所示为在一实施方式中,中断从外围设备进到gAPIC的方块图。类似于图3,来自其他处理器的中断(内处理器中断(IPI))可也被传送到访客中断管理器38且可从那一点向前进行处理。可选地,从初始化IPI的处理器接收IPI的gAPIC可传送更新到访客中断管理器38(来更新接收访客的gAPIC状态,如果访客是非活动的)并且也可传送IPI(包括访客ID)到其他gAPIC。

[0074] 在所示的实施方式中,外围设备52A确定需要中断。外围设备52A中的IOAPIC54(参见图2)可生成响应外围设备52A的操作的中断消息。特别是,IOAPIC54可生成对应于预期中断(例如,基于外围设备52A所需要的服务、用信号通知中断的特定功能(如果外围设备52A执行多种功能),等等)的中断向量。中断向量是中断通信的一部分,且可通过软件被用于识别中断源、区分中断的优先顺序,等等。在一些实施例中,中断向量可由IOMMU40重新映射,这样该中断向量在图3中可标示为“初始向量”。外围设备52A可将中断消息传送到IOMMU40A(箭头A)。在此实施方式中,中断可按照消息通知中断(Message-Signalled Interrupt, MSI)的方式来传送,例如在PCIe规范中所定义的。在其他实施方式中,中断可以以任意的期望方式来传送。通常来说,传送设备可识别中断、中断的传送模式(例如,逻辑的或物理的)以及中断的目的地ID(DestID)。

[0075] IOMMU40可接收MSI。MSI包括外围设备的标识符。例如,实施PCI编程模型的接口可利用总线号以及在该总线上的设备号识别每一个设备(允许多个PCI接口以分层形式和/或并行形式存在于系统中)。设备可具有多种“功能”,其可以是物理设备上的独立虚拟设备,或者是设备上的操作分区。标识符还可包括功能号。这样,在此实施方式中,标识符可被称

为总线设备功能(BDF)。IOMMU40可利用BDF来索引设备表62(箭头B),并可识别对应于外围设备52A的设备表条目。在一些实施方式中,该条目可包括访客ID和指向gAPIC状态映射表60或gAPIC状态数据结构58的指针(箭头C)。在此实施方式中,设备表条目也可包括中断重定向表指针(IRTP),该指针可确定对应于设备的中断重定向表64(箭头C1)。中断重定向表64可被初始中断向量来索引,并可为该中断提供输出向量和目的地ID(DestID,例如,逻辑APIC ID或物理APIC ID)(箭头C2)。

[0076] 在图3所示的实施例中,MSI被重新映射到向量42和访客ID99。该重新映射可包括添加该访客ID,在使用中断重定向表64的情况下,该向量也可被改变。否则,在该中断消息中提供来自MSI的初始中断向量。描绘了使用中断向量42和访客ID99的特定实施例的图3中的要点在图中被括在方括号中,例如[]。

[0077] IOMMU40可传送包括访客ID(例如,在这个实施例中是99)的中断消息到访客中断管理器38。中断消息也可包括中断向量(例如,在这个实施例中是42)和目的地ID。中断消息还可包括指向gAPIC状态映射表60或gAPIC状态数据结构58的指针(箭头D)。

[0078] 在具有gAPIC状态映射表60的实施方式中,访客中断管理器38可利用诸如访客ID和/或目的地ID的指针以及潜在的其他信息来定位gAPIC状态映射表60中的gAPIC状态指针(箭头E1,以及如箭头E2所示的到访客中断管理器38的返回指针)。gAPIC状态指针可确定gAPIC状态数据结构58中的gAPIC状态条目,并且访客中断管理器38可利用该gAPIC状态指针来执行gAPIC状态数据结构58中的gAPIC状态更新(箭头E)。在这个实施例中,gAPIC状态更新可设定对应于向量42的中断请求寄存器中的位。中断请求寄存器(IRR)参照图4在下面进行更为详细的描述。

[0079] 在一实施方式中,gAPIC状态58的更新是原子的。例如,访客中断管理器38可生成原子或运算(OR)事务,该事务原子地将正在设置的中断请求位或运算到gAPIC状态条目中的中断请求寄存器的当前状态中。一个原子操作可以是作为一个单位来有效执行的操作,即使该操作由多个步骤完成。试图访问被原子地更新的位置的观察者(observer)要么接收原子更新之前的值,要么接收更新之后的值,但不会接收中间的值。试图更新被原子地更新的位置的观察者要么在该原子操作之前执行其更新,要么在之后执行其更新,但不会在操作过程中进行更新。尽管此实施方式可执行原子或运算,其他实施方式也可执行更加通用的原子更新操作。例如,原子更新可包括识别不会被修改的目标的位的与屏蔽(AND mask),以及识别会被进行或运算的位的或屏蔽(OR mask)。其他实施方式也是可能的。例如,可使用比较和交换实现,其中来自存储位置的初始值被读取并且执行比较和交换操作以用经过或运算得到的新值替代初始值。如果比较操作失败,该过程可被重复(读取新的初始值,然后执行比较和交换)。如果需要,可使用后退和/或暂停机制来退出循环。

[0080] 访客中断管理器38也可向gAPIC34A-34D广播包括中断向量、访客ID和目的地ID的中断消息(箭头F)。gAPIC中的一个(图3中的gAPIC34A)可具有访客ID99以及与目的地ID相匹配的逻辑APIC ID或物理APIC ID,因此gAPIC34A可用确认信号(Ack)来响应中断消息,该确认信号表示gAPIC34A已经接受该中断消息(箭头G)。在该实施例中,gAPIC34A也可更新其中断请求寄存器以设置对应于向量42的位。如果该中断的优先级比任何当前正在进行的中断(如果有的话)和/或处理器的任务的优先级都要高的话,则gAPIC34A也可将该中断用信号通知处理器30A。其他gAPIC34B-34D也可响应广播的中断消息,但是由于它们不是该中断

的目标,因此它们不会发出确认接受的信号(箭头H)。对逻辑中断而言,如果其确定有多个目标的话,则可能有不止一个确认信号。

[0081] 使用上述机制,访客中断管理器38不需要知道哪一个gAPIC34A-34D被分配给了哪一个访客。其他实施方式也是可预期的,其中访客中断管理器38追踪哪个gAPIC34A-34D被分配给了哪个访客以及哪个访客仅传送中断到目标gAPIC。访客中断管理器38可自动追踪gAPIC,或者每次一个gAPIC被重新分配给另一个访客时由VMM18进行编程。在这样的实施方式中,访客中断管理器38可仅传送中断消息到目标gAPIC。

[0082] 到hAPIC32A-32B的中断的传送可按常用的APIC方式进行。特别地,在一实施方式中,中断可能未由访客中断管理器38操作在其上,但在其他方面可与图3中的操作相同。

[0083] 需要注意的是,尽管访客中断管理器38在此处图示为一个方块,但实现访客中断管理器38的电路系统可以是分布式的。例如,在一实施方式中,接收指针、可选择地处理gAPIC状态映射表 60并生成gAPIC状态数据结构58的更新的那部分访客中断管理器38可被包括在IOMMU40中,因此该IOMMU40为gAPIC状态数据结构58和被传送到gAPIC34A-34D的中断消息传送原子的或运算。访客中断管理器38在一个或多个物理位置中的任何实现方式均可采用。

[0084] 现在翻到图4,该图示出了gAPIC34A的一实施方式的方块图。其他的gAPIC34B-34D可以类似。在图4的实施方式中,gAPIC34A包括中断请求寄存器(IRR)70、中断服务寄存器(ISR)72、触发模式寄存器(TMR)74、任务优先级寄存器(TPR)76、控制单元78、物理ID寄存器80、逻辑ID寄存器82、访客ID寄存器84以及其他可选的APIC状态86。控制单元78与IRR70、ISR72、TMR74、TPR76、物理ID寄存器80、逻辑ID寄存器82、访客ID寄存器84和其他APIC状态86相耦合。此外,控制单元78与访客中断管理器38耦合而通信以接收中断,并与处理器接口相耦合以与处理器30A通信。

[0085] 为了对从访客中断管理器38接收中断消息作出响应,如果该中断指向与gAPIC34A相应的访客,则控制单元78可用于将该中断写入IRR70中。IRR中的中断请求的位置与中断向量相对应。IRR可追踪“固定的(fixed)”中断。其他中断类型可包括不可屏蔽中断(NMI)、系统管理中断(SMI)、传统外部中断(extINT)等等。这些中断可被作为其他APIC状态86的一部分处理。

[0086] 在一实施方式中,中断消息也可为每一个中断包括一个触发模式(电平的或边缘的)。TMR74可存储哪一个触发模式应用于该中断的指示。例如,边缘触发中断可由TMR74中的二进制0来表示而电平触发中断则可由二进制1来表示。在其他实施方式中,在gAPIC34A中仅支持边缘触发中断,则可除去TMR74(及其在 gAPIC状态数据结构58中的副本)。在另一实施方式中,可重新使用TMR74以允许VMM18记录虚拟的电平敏感中断。VMM18可在TMR74中设置多个位以指示如下情况,如果处理器30A用信号通知中断的结束以作相应的中断向量,则处理器30A退出到该VMM18。

[0087] 对于固定中断,gAPIC34A可用于对中断请求以及服务中(in-service)中断的优先级进行排序以判断中断请求是否应被传送给处理器。通常来说,如果最高优先级中断请求的优先级高于服务中中断的最高优先级(其中如果处理器已中断其软件执行来处理与一中断对应的中断句柄,则该中断为服务中中断),控制单元78可用于将所请求的中断传送给处理器30A。此外,TPR76可被软件编程以建立处理器30A所接受的中断的最低优先级。如果最

高优先级中断请求的优先级高于服务中中断的最高优先级并且高于TPR76中所指明的优先级,则控制单元78可用于传送该最高优先级中断请求。

[0088] 当处理器30A接收到该中断时,该处理器可用中断确认命令响应gAPIC34A。控制单元78可用于从IRR70中移出最高优先级中断请求并且将该中断作为服务中中断记录在ISR72中。相应于该ISR中的中断的服务中指示的位置可对应于该中断的中断向量。处理器30A可执行中断服务程序来为该中断服务。中断服务程序可利用给gAPIC的结束中断(EOI)命令来结束以用信号通知该中断服务完成。控制单元78可用于从ISR72中移出该最高优先级的服务中中断以响应该EOI命令。

[0089] 每一个IRR70、ISR72和TMR74都包括与gAPIC34A支持的每一个中断向量相对应的位置。在所示的实施方式中,支持从0到255的向量。中断向量号也可指示其与其他中断的相对优先级(例如,在其他实施方式中,较高向量号的优先级高于较低向量号优先级,反之亦然)。对每一个中断向量而言,IRR70存储的中断请求位指示在该中断向量处是否有中断被请求。例如,该指示可以是指示当设置时有请求以及当清空时没有请求的位。类似的,对每一个中断向量而言,ISR72存储的服务中(in-service)位为该中断向量指示是否有中断在服务中(例如,当设置时表示有服务中中断,当清空时表示无服务中中断)。对每一个中断向量而言,TMR74存储触发模式。对每一个IRR70、ISR72和TMR74而言,寄存器中的位位置等同于与中断相应的中断向量号。

[0090] 在所示的实施方式中,中断被逻辑地分配到组中,这些组都被指定优先级等级以确定未决的中断请求是否被传送给处理器。例如,将优先级等级0分配给中断向量0到15,将优先级等级1分配给中断向量16到31,等等,直到将优先级等级15分配给中断向量244到255。在此实施方式中,提高优先级等级的号码表示更高的优先级等级。控制单元78可计算请求的优先级等级,如果是最高优先级等级,则IRR70中至少有一个中断请求处于未决状态。控制单元78也可计算服务中的优先级等级,如果是最高优先级等级,则ISR72中至少有一个中断被指示为在服务中。如果请求的优先级等级超过了服务中的优先级等级并且也超过了TPR76中所表明的优先级等级,则控制单元78可传送该中断。应该注意的是,尽管在所示的实施方式中256个中断向量由16个优先级等级组别来支持,然而在其他实施方式中也可以采用更多或更少的中断向量和/或更多或更少的优先级等级组别。

[0091] 物理ID寄存器80和逻辑ID寄存器82可分别存储分配给gAPIC34A的物理APIC ID和逻辑APIC ID。访客ID寄存器84可存储分配给gAPIC34A的访客ID。据此,如果中断的访客ID与访客ID寄存器84中的访客ID相匹配,或者物理中断中的APIC ID与物理ID寄存器80中的物理ID相匹配,亦或者逻辑中断中的APIC ID与逻辑ID寄存器82中的逻辑ID相匹配,则控制单元78可用于接受来自访客中断管理器38的中断。

[0092] 其他APIC状态86可包括内部生成的中断、计时器、本地向量表等等。在各种实施方式中,一些或所有的其他APIC状态86可被包括在gAPIC34A中或通过到VMM18的拦截以及该状态的VMM18仿真虚拟化。

[0093] 除了不包含访客ID寄存器外,hAPIC32A-32B可以与gAPIC34A类似。可选地,hAPIC32A-32B和gAPIC34A-34D可以是具有相同硬件的例子(如果gAPIC34A-34D实现所有的APIC状态),并且访客ID寄存器可包含指示该访客ID是否有效的使能位,或者访客ID寄存器可被编程为调0以指示hAPIC。

[0094] 现在翻到图5,该图示出了gAPIC状态条目90的一实施方式和VMM管理的gAPIC状态条目92的一实施方式的方块图。图5中所示的是状态的逻辑视图。在存储器中,状态的实际排列可按照图9、12或13中的一些实施方式而进行变化。

[0095] 通常来说,gAPIC状态条目90可包括至少如下的gAPIC状态,当对应于该gAPIC状态的访客是非活动的时,该gAPIC状态是可变的。在当前的实施方式中,外围设备可将中断用信号通知访客,这可改变IRR的状态。但是,ISR状态仅当访客中的vCPU接受该中断时才可改变,当访客是非活动时,ISR状态不会发生改变。类似的,可以通过vCPU来改变TPR,并且当访客是非活动时,TPR不改变。VMM18可在VMM18管理的gAPIC状态条目92中管理这些状态的存储和恢复。

[0096] 相应地,对类似于图4的APIC34A的一实施方式而言,gAPIC状态条目90可包含IRR70的状态。VMM管理的gAPIC状态条目92可包括ISR72、TMR74、TPR76以及其他各种APIC状态86的状态。VMM管理的gAPIC状态条目92也可存储访客ID、逻辑ID以及物理ID,或者这些内容在选择条目92中是固有的(也就是说,VMM18可使用这些值来从数据结构68中选择条目92)。

[0097] 现在翻到图6,该图示出了响应接收来自IOMMU40的访客中断消息的访客中断管理器38的一实施方式的操作流程图。尽管为了便于理解而将方块按特定顺序显示,但其他顺序也可以使用。可按照访客中断管理器38中的组合逻辑并行地执行各个方块。方块、方块组合和/或流程图在多个时钟周期期间以流水线方式处理。通常来说,访客中断管理器38可用于执行如图6所示的操作。

[0098] 在一些实施方式中,中断消息的处理根据该中断是逻辑的还是物理的可变(也就是说,根据中断的传送模式是逻辑的还是物理的)。例如,在图11的实施方式中,逻辑中断和物理中断读取不同的表。在图12和13中,逻辑表和物理表在存储器中可以是相邻的,但是可对基址指针增加偏移以定位逻辑中断的逻辑表而物理中断则不需要增加任何偏移。据此,访客中断管理器38可用于确定中断是逻辑的还是物理的(判定框100)。其他实施方式不可基于传送模式而变换,于是判定框100可以除去(且下面要讨论的广播或多个目的地的检查可被并入一个检查中)。

[0099] 如果中断是逻辑的(判定框100,“是”支路),访客中断管理器38可用于确定从逻辑中断到gAPIC状态数据结构58中相应的gAPIC状态条目90的映射(方块102)。不同的实施方式可执行不同的映射,如图10-13所示,因此上述的确定是可变的。访客中断管理器38可用于设置对应于中断向量的位,该中断向量位于表示在gAPIC状态条目90中的IRR中(方块104)。逻辑中断可具有多个目的地(例如,在簇中的目的地是位向量,其可以具有多于一个的置位)。如果逻辑中断包含多个目的地(判定框106,“是”支路),则访客中断管理器38可用于为每一个附加的目的地重复方块102和104。可选地,在图12的实施方式中,逻辑目的地向量在一个操作中可被写入gAPIC状态条目中,如下详述。访客中断管理器38可用于传送中断消息到gAPIC34A-34D(方块108)。

[0100] 如果中断是物理的(判定框100,“否”支路),访客中断管理器38可用于确定从物理中断到gAPIC状态数据结构58中相应的gAPIC状态条目90的映射(方块110)。不同的实施方式可执行不同的映射,如图10-13所示,因此上述的确定是可变的。访客中断管理器38可用于设置对应于中断向量的位,该中断向量位于表示在gAPIC状态条目90中的IRR中(方块

112)。物理中断可以是广播的或者是单一的目的地。如果物理中断是广播的(判定框114,“是”支路),则访客中断管理器38可用于为访客虚拟机(例如,每一个vCPU)中的每一个目的地重复方块110和112。可选地,在图12的实施方式中,该广播可在一个操作中被记录在gAPIC状态条目中,如下详述。访客中断管理器可用于传送中断消息到gAPIC34A-34D(方块108)。

[0101] 被表示在gAPIC状态条目90中的IRR中的位的设置可作为原子的或运算操作来执行,其中该置位被或运算到该存储位置中的其他IRR位中。原子或运算操作的实际执行情况是可以变化,从锁定的读/修改/写操作到定义来执行或运算的专用电路系统作为一个操作。如上所述,在其他实施方式中可执行比较和交换操作。

[0102] 在另一实施方式中,具有多于一个目的地的逻辑中断和被广播的物理中断可由访客中断管理器38通过在VMM18可存取的数据结构中记录该中断来处理(例如,事件队列)。访客中断管理器38可也用于用信号通知VMM18(例如,引起从处理器30A-30B之一上的虚拟机退出)来将事件通知给VMM18。可选地,访客中断管理器38可仅周期性地用信号通知VMM18(例如,每N毫秒一次和/或在事件队列中处于高水位时),并且VMM18也可周期性地检查事件队列以比可支持的信号通知更快地为任一事件服务。在一实施方式中,事件队列可由IOMMU40而不是访客中断管理器38来管理。

[0103] 现在翻到图7,该图所示为说明gAPIC34A-34D对接收来自访客中断管理器38的中断消息的作出响应的一实施方式的操作流程图。尽管为了便于理解而将方块按照特定顺序显示,但其他顺序也可以使用。可按照gAPIC中的组合逻辑并行地执行各个方块。方块、方块组合和/或流程图在多个时钟周期期间以流水线方式处理。通常来说,gAPIC可用于执行如图7所示的操作。

[0104] 在一实施方式中,通过设置gAPIC的访客ID(参见图4,在访客ID寄存器84中)为0可使得该gAPIC非激活。因此,为了对接收中断消息作出响应,如果该gAPIC的访客ID为0(判定框120,“是”支路),则该gAPIC是非活动的并且不可处理中断。其他实施方式可通过其他方式(例如,在寄存器中的活动位)来使得gAPIC非激活并且判定框120可根据gAPIC的活动/非活动状态的检查来相应地进行修改。

[0105] 如果gAPIC的访客ID不为0,则该gAPIC可被配置为既比较该访客ID和接收到的中断的访客ID,又比较接收到的目的地ID和分别位于寄存器80与82中的逻辑ID和物理ID(参见图4)。如果gAPIC的访客ID与接收到的访客ID不匹配(判定框122,“否”支路),则该gAPIC当前被分配给不同的访客并且该gAPIC因而不 再被该中断指向。该gAPIC可被配置为用中断的非确认信号进行响应(方块124)。该非确认信号可表示该gAPIC已经接收到该中断但是已经确定该中断不是相应的处理器所指向的,因此该中断未被接受。类似的,如果gAPIC的访客ID与接收到的访客ID相匹配,但是该中断是逻辑的并且与该gAPIC的逻辑ID不匹配或者该中断是物理的、单目的地的并且与该gAPIC的物理ID不匹配(判定框126和128,“否”支路),则该gAPIC可被配置为用中断的非确认信号进行响应(方块124)。

[0106] 匹配逻辑中断通常包括比较逻辑ID的簇ID部分的相同之处,并且检测gAPIC逻辑ID寄存器中的置位也被设置在从该中断接收的逻辑ID的目的地部分中。如果有不止一个逻辑目的地的话,则该中断逻辑ID的目的地部分中的其他位也可被设置。物理ID可被比较相同之处,除了广播的物理中断可被看作只要访客ID匹配就匹配。

[0107] 如果中断是逻辑的并且与逻辑ID相匹配(判定框126,“是”支路)或者中断是物理的并且与物理ID相匹配或者中断是广播的(判定框128,“是”支路),则该gAPIC可用于用表示该gAPIC正在接受该中断的确认信号来响应访客中断管理器38以报告给相应的处理器30A-30B(方块130)。该gAPIC也可用于更新IRR寄存器70、设置对应于中断消息中的中断向量的中断请求位(方块132)。该gAPIC可用于重新评估与任何服务中中断和/或任务优先级寄存器有关的中断的优先级(方块134),并可用于基于重新评估的结果将中断用信号通知给处理器(方块136)。也就是说,如果该中断的优先级高于服务中中断和任务优先级寄存器的优先级的话,则该gAPIC可被配置为用信号通知该中断。

[0108] 现在翻到图8,该图示出了一个流程图以说明VMM18将gAPIC状态从一个访客改变到另一个访客的一实施方式的操作。也就是说,图8的流程图可表示将gAPIC34A-34D从一个访客/vCPU重新分配给其他访客或者同一访客中的其他vCPU。尽管为了便于理解,方块以特定的顺序来显示,但也可以用其他顺序。通常来说,VMM18可包括一些指令,当这些指令在系统5上执行时可实现图8所示的操作。

[0109] VMM18可确定相应于“旧访客”(来自gAPIC的被关闭的访客)的位于gAPIC状态数据结构58中的gAPIC状态条目90的位置(方块140)。gAPIC状态条目90中的数据被认为是“过时的”,因为它可能已经被gAPIC修改过。例如,IRR位可以已被重置以对传送中断到相应的处理器作出响应。据此,VMM18可将gAPIC状态条目90中的IRR调0(方块142)。VMM18可清空访客ID寄存器84、逻辑ID寄存器82以及物理ID寄存器80(方块144)。由于访客ID、逻辑ID和物理ID不会与任何中断消息相匹配,因此清空操作可阻止gAPIC接受任何额外的中断。在寄存器80-84被清空(方块144)之后、IRR状态被写入gAPIC状态条目90之前,中断可能已经被传送。据此,为了防止中断状态的丢失,VMM18可原子地将IRR状态从IRR70或运算(OR)到gAPIC状态条目90中(方块146)。VMM18也可将其他gAPIC状态写入与旧访客相关的VMM管理的gAPIC状态条目92中(方块148)。

[0110] VMM18可清空gAPIC的IRR、ISR和TMR寄存器70、72和74以移除旧访客的中断状态(方块150)。VMM18可将分配给gAPIC的访客的新访客ID、逻辑ID以及物理ID各自写入到访客ID寄存器84、逻辑ID寄存器82和物理ID寄存器80(方块152)。一旦方块152被执行,则gAPIC可开始接受访客中断。VMM18可为“新访客”(在gAPIC中激活的访客)确定gAPIC状态条目90(方块154),并可从gAPIC状态条目90中读取IRR状态(方块156)。既然对寄存器80-84的编程可使得gAPIC开始接受中断,因此可能的情况是gAPIC已经接受IRR中的中断,该中断在VMM18读取gAPIC状态条目90之后被记录于该条目中。据此,VMM18可原子将IRR状态或运算到IRR寄存器70中。也就是说,gAPIC可支持IRR寄存器70上的原子或运算操作(方块158)。VMM18可从VMM管理的gAPIC状态条目92中为新的访客读取其他状态(方块160)并可将状态写入gAPIC中(方块162)。值得注意的是,方块160和162也可在方块150之后的任意其他点执行。

[0111] 通常来说,方块140-148可表示关闭来自gAPIC的访客的操作,同时方块150-162通常可表示激活gAPIC中的访客的操作。据此,如图8中的水平虚线所示,如果VMM18仅希望关闭gAPIC中的访客,则水平虚线以上的方块可被执行。如果VMM仅希望激活gAPIC中的访客,则水平虚线以下的方块可被执行。

[0112] 现在翻到图9,图中所示的方块图说明在一实施方式中在gAPIC状态条目90中的

gAPIC状态的示例性排列170。在图9的实施方式中,IRR的每一个位都被存储在不同字节中。例如,图9中的IRR位0(IRR0)被存储在存储器中一组连续字节的字节0中;IRR1被存储在字节1中;等等,直到IRR255被存储在字节255中。尽管字节中的任何位位置(bit position)均可使用,但在所示的实施方式中IRR位被存储在字节的0位处。在所示的实施方式中,字节中的其他位是不必关心的位(DC)。通过将每一个位存储在分离的字节(存储访问的最小单位)中,每一个位可分别写入而不影响其他的位。这样,一个位可通过到字节的写入操作被设置,这是一个原子操作。通过将置位写入到字节中的IRR位的位置且不更新其他字节,结果可以是对IRR位的原子或运算。在其他实施方式中,原子的或运算可以以其他方式完成,且IRR状态的位可以以其他方式存储。

[0113] 现在翻到图10,该图示出了定位gAPIC状态条目90的一实施方式的方块图。在所示的实施方式中显示了设备表62和中断重定向表64,还有gAPIC状态映射表60的一实施方式。在该实施方式中,传送中断的外围设备的BDF作为到设备表62的索引使用,且该条目可包括将BDF分配给其的访客的访客ID。此外,在这个实施例,该条目包括中断重定向表指针(IRTP),其指向中断重定向表64的基址。到中断重定向表64的索引是中断的中断标识符。中断标识符可包括中断向量,也可包括物理或逻辑的中断传送模式(Delmode)。选定的条目可包括新的向量以及目的地ID(DestID)。在不使用中断重定向表64的实施方式中,由外围设备提供的中断向量和目的地ID可直接用来索引gAPIC状态映射表60。

[0114] gAPIC状态映射表60可通过gAPIC状态映射表的基址被定位在存储器中。在不同实施方式中,基址可以对所有访客相同,也可以是不同访客不同,或者可被存储在设备表62中。在图10中,基址标识一组分层表的最高层级(L3),其可存储指向较低层级表的指针(例如,L2以及类似的由来自L3但不指向L2的指针标识的表)。L2表可存储指向更低层级表(L1)的指针,其可存储指向gAPIC状态数据结构58中的gAPIC状态条目90的指针。其他实施方式可在层级系统中使用任意的层级数,包括多于或少于图10所示的3层。

[0115] 在gAPIC状态映射表60中到每一层L3-L1的索引可以是下述值的一部分,该值由来自设备表62的访客ID、来自外围设备或来自中断重定向表64的中断向量以及来自外围设备或来自中断重定向表64的目的地ID串联而形成。到层L3-L1的索引可耗尽该串联值的所有位,这样在gAPIC状态映射表60中访客ID、向量和目的地ID的每一个组合都具有其自身的唯一指针。但是一些指针可指向同一个gAPIC状态条目90(例如,在一实施方式中,同一gAPIC的逻辑ID和物理ID可具有指向同一个gAPIC状态条目90的指针)。

[0116] 现在翻到图11,该图示出了定位gAPIC状态条目90的另一实施方式的方块图。在所示的实施方式中显示了设备表62和中断重定向表64,还有gAPIC状态映射表60的一实施方式。在该实施方式中,传送中断的外围设备的BDF作为到设备表62的索引使用,且该条目可包括将BDF分配给其的访客的访客ID。此外,在这个实施例,该条目包括中断重定向表指针(IRTP),其指向中断重定向表64的基址。设备62可进一步包括一个或多个指向gAPIC状态映射表60中的表的指针。尤其是,指向访客物理表的指针和指向访客逻辑表的指针均可被存储。访客物理表可将物理目的地ID映射到gAPI状态条目90。也就是说,访客物理表可通过目的地ID进行索引,且可存储指向gAPIC状态条目90的指针。类似的,访客逻辑表可将逻辑目的地ID映射到gAPIC状态条目90。

[0117] 到中断重定向表64的索引是中断的中断标识符。中断标识符可包括中断向量,也

可包括物理或逻辑的传送模式(Delmode)。选定的条目可包括新的向量和目的地ID(DestID)。在不使用中断重定向表64的实施方式中,外围设备提供的中断向量和目的地ID可直接用来索引gAPIC状态映射表60。

[0118] 现在翻到图12,该图示出了定位gAPIC状态条目90的又一实施方式的方块图。在这一实施方式中没有gAPIC状态映射表60。与图10-11中的实施方式类似的是,传送中断的外围设备的BDF 作为到设备表62的索引使用,且该条目可包括将BDF分配给其的访客的访客ID以及可选地包括中断重定向表指针(IRTP),该指针指向中断重定向表64的基址。设备表62可进一步包括至少一个指向gAPIC状态数据结构58中的表的基址的指针(Ptr)。在所示的实施方式中,该表包括访客物理片段(section)180和访客逻辑片段182。尽管为了绘图的清楚在图12中将片段180和182之间隔开一定的空间,但在存储器中片段180和182是相邻的。也就是说,访客物理片段180的顶部与访客逻辑片段182的底部相邻。设备表62的条目可进一步包括指示访客逻辑部分182的顶部的逻辑限制(LLim)域。在其他实施方式中,访客物理部分180和访客逻辑部分182可不相邻并且可在设备表62的条目中存储单独的指针以分别指示访客物理部分180和访客逻辑部分182。

[0119] 在图12的实施方式中,访客物理部分180可由中断向量(或者来自外围设备或者来自中断重定向表64)来索引。访客物理部分180中的每一个条目可包括对应于访客物理机所支持的目的地ID(例如,多达64个目的地,在图12中编号0到63)的位向量。为了响应物理中断,访客中断管理器38可用于为对应于目的地ID的中断向量设置条目中的位。对广播中断来说,访客中断管理器38可用于设置条目中的每一个对应于中断向量的位,直到虚拟机中的vCPU的数量。

[0120] 访客逻辑部分182可通过逻辑ID的簇部分以及向量来索引。簇部分是索引的最值得注意的位,因此访客逻辑部分182被分割成对应每一个逻辑簇的簇部分(图12中的簇0到簇N)。在每一个簇中,条目由中断向量来排列,每一个条目存储对应逻辑ID向量部分的位向量。在所示的实施方式中,一个簇中可包括多达16个目的地(例如,逻辑ID的位向量部分可以是16位)。为了响应 逻辑中断,访客中断管理器38可被配置为利用相应于中断向量的条目中的内容来逻辑地对逻辑ID的位向量部分进行或运算。

[0121] 相应地,图12的实施方式可支持广播物理中断的记录和逻辑中断的记录,所述逻辑中断在对gAPIC状态数据结构58的单个更新中具有多目的地。gAPIC的gAPIC状态条目可包括对应于gAPIC物理ID的访客物理部分180的列,其与来自被gAPIC逻辑ID标示的簇的列逻辑地进行或运算,该列被gAPIC逻辑ID的位向量部分中的置位所标识。响应使gAPIC中的访客不活动而更新gAPIC状态数据结构58包括将对应于访客的列中的一列调0并将IRR写到其他列。

[0122] 图13所示为定位gAPIC状态条目90的再一实施方式。图13中的实施方式与图12中的实施方式类似,除了在访客物理部分180和访客逻辑部分182中的数据排列不同之外。每一个条目对应一个IRR,因此每一个条目为每一个中断向量包括一个位。访客物理部分180由中断的物理ID来索引,同时访客逻辑部分182由访客的逻辑ID来索引。对应于中断向量的IRR位根据中断的传送模式被设置在逻辑部分182或物理部分180中。访客/vCPU的gAPIC状态是来自对应于分配给该访客/vCPU的物理ID的访客物理部分180的行与来自对应于分配该访客/vCPU的逻辑ID的访客逻辑部分182的行的或运算。

[0123] 现在翻到图14,该图所示为主机硬件20的另一实施方式的方块图。在所示的实施方式中,与图2中的集成电路66类似的包括两个集成电路66A-66B。因此,如图所示,每一个集成电路可包括gAPIC,例如,集成电路66A中的34A-34D,以及集成电路66B中的34E-34G。每一个集成电路66A-66B可包括各自的访客中断管理器28A-28B和IOMMU(图14中未图示)。至少一个集成电路66A-66B与存储器56A-56B相耦合,可选的是集成电路66A-66B二者都包括存储器。集成电路66A-66B通过所示实施方式中的接口电路44C和44D来耦合。在其他实施方式中,可提供不止两个的集成电路66A-66B并且不同的集成电路可以以任意想要的方式互连。

[0124] 在一实施方式中,每一个访客中断管理器28A-28B可被使能并可管理指向在同一集成电路中的gAPIC34A-34G的中断消息。因此,访客中断管理器28A-28B可对访客中断传送提供可扩展解决方案。访客中断管理器28A-28B所使用的数据结构可被存储在一个存储器中(例如,存储器58A),或者每一个访客中断管理器28A-28B可在其自身的存储器58A-58B中具有自己的数据结构。尽管可能存在一些对存取数据结构的争用,但在许多情况下外围设备被分配给特定的访客(其执行于位于集成电路66A-66B之一中的处理器上),这样实际争用的数量可相对较小。

[0125] 在另一实施方式中,访客中断管理器28A-28B中的一个可被使能并可为系统中的每一个gAPIC34A-34G执行访客中断传送。这样的实施方式在集成电路66A-66B之间的交互中可经受更多的与中断相关的传送,但也可可为访客中断管理提供中心点的概念简单性。

[0126] 现在翻到图15,该图所示为计算机可存取存储介质200的方框图。一般来说,计算机可存取存储介质可包含任何在用以对计算机提供指令和/或数据期间可由计算机存取的存储介质。例如,计算机可存取存储介质可包含诸如磁性介质或光学介质的存储介质,例如,磁盘(固定的或可移动的)、磁带、CD-ROM或DVD-ROM、CD-R、CD-RW、DVD-R、DVD-RW和/或蓝光光盘。存储介质可进一步包括易失性或非易失性存储介质诸如RAM(例如,同步动态RAM(SDRAM)、Rambus DRAM(RDRAM)、静态RAM(SRAM),等等)、ROM、闪存、可由诸如通用串行总线(USB)接口或任何其他接口的外围设备接口进行存取的易失性存储器(例如,闪存),等等。存储介质可包括微机电系统(MEMS)以及由诸如网络和/或无线链接的通信介质进行存取的存储介质。图15中的计算机可存取存储介质200可存储VMM18,VMM18可执行图8的流程图和/或任何其他分配给VMM18的功能。一般来说,该计算机可存取存储介质200可存储任何指令组,该指令组执行时实现图8所示的流程图的部分或全部。载体介质可包括计算机可存取存储介质以及诸如有线或无线传输的传输介质。

[0127] 一旦完全地了解上述公开的内容,本发明的多种变化和修正对本领域技术人员来说是显而易见的。其目的在于使下面的权利要求书解释为包括所有这样的变化和修正。

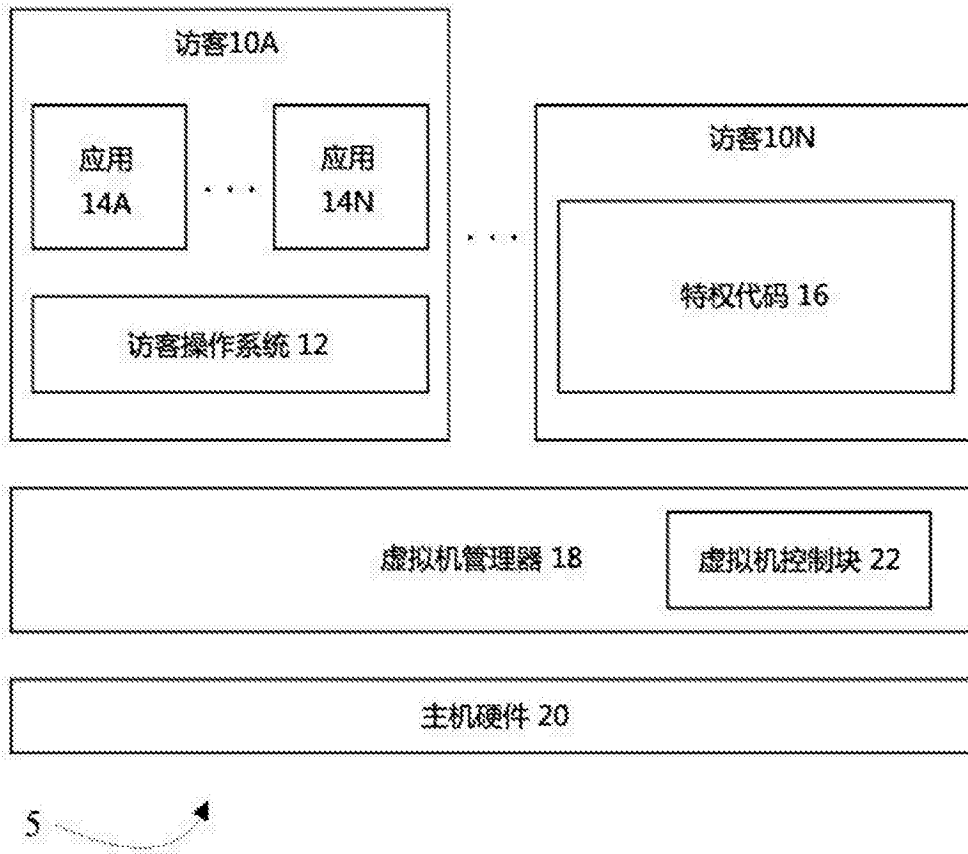


图1

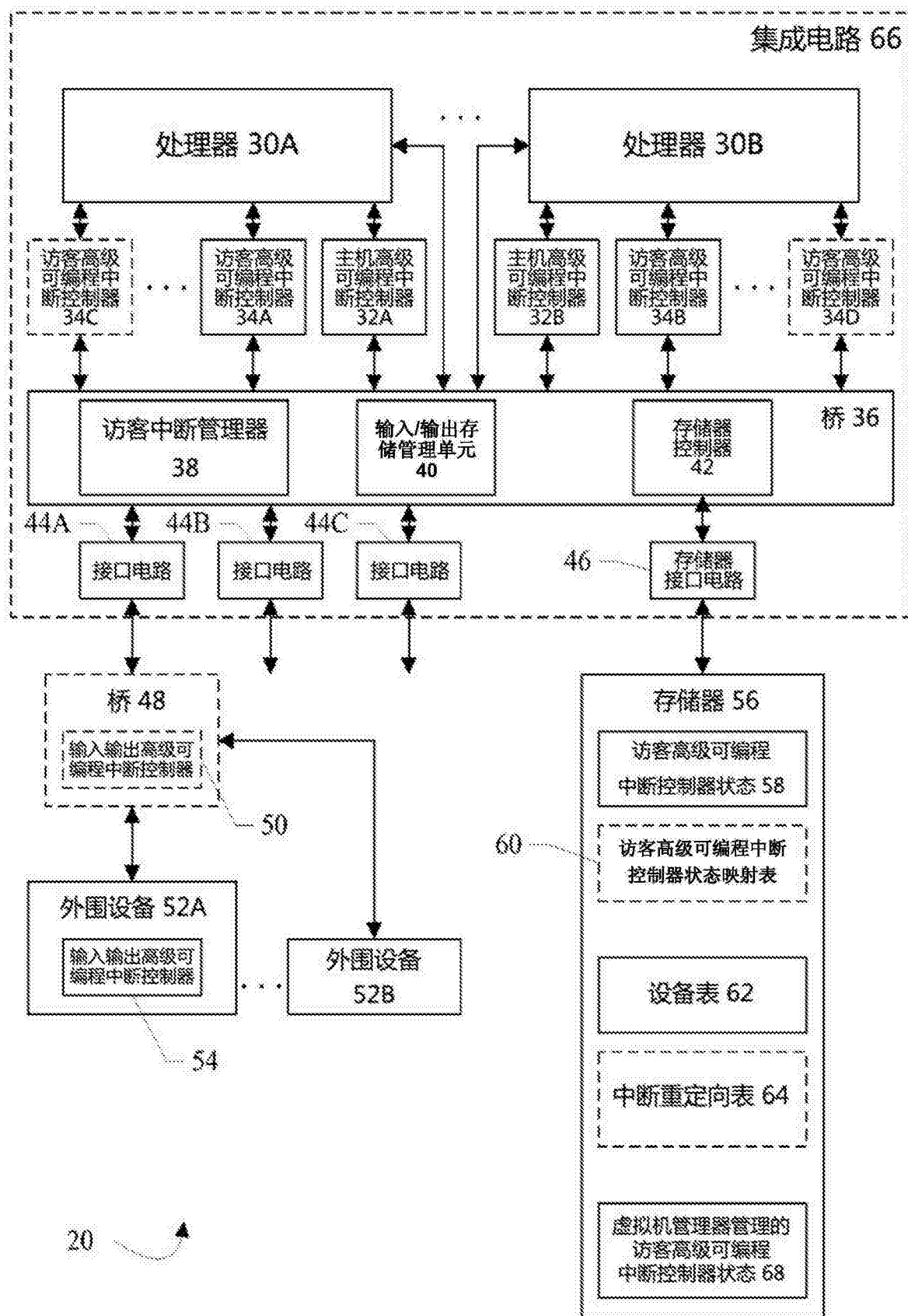


图2

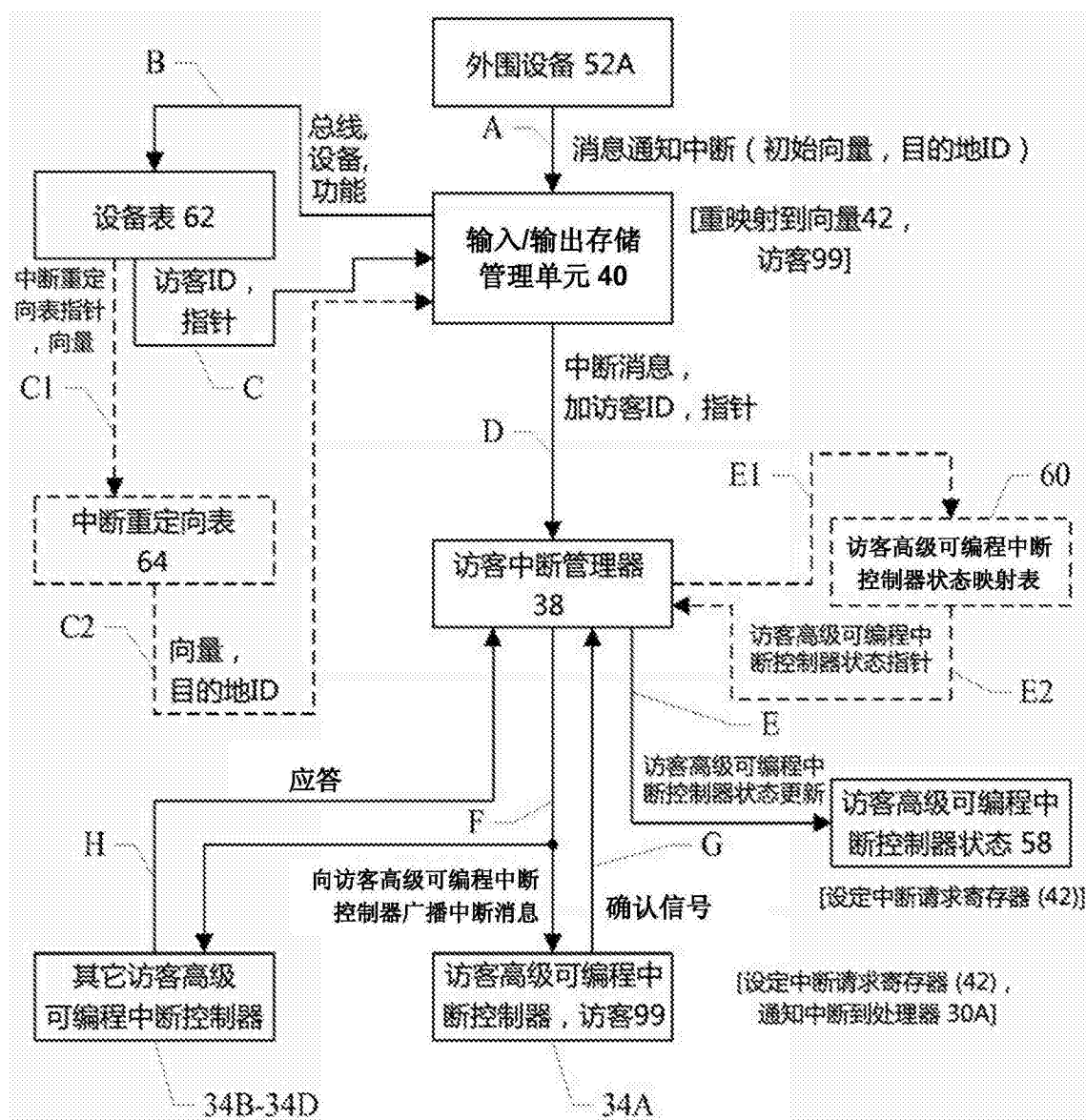


图3

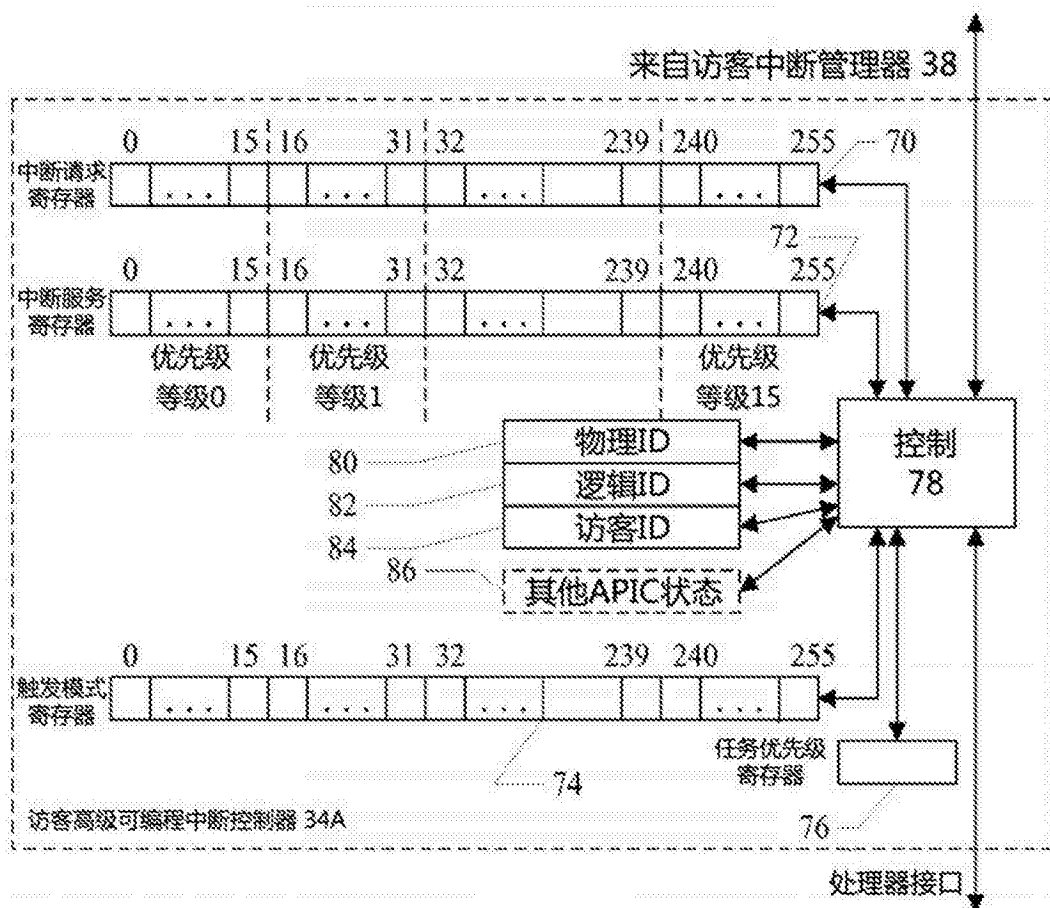


图4

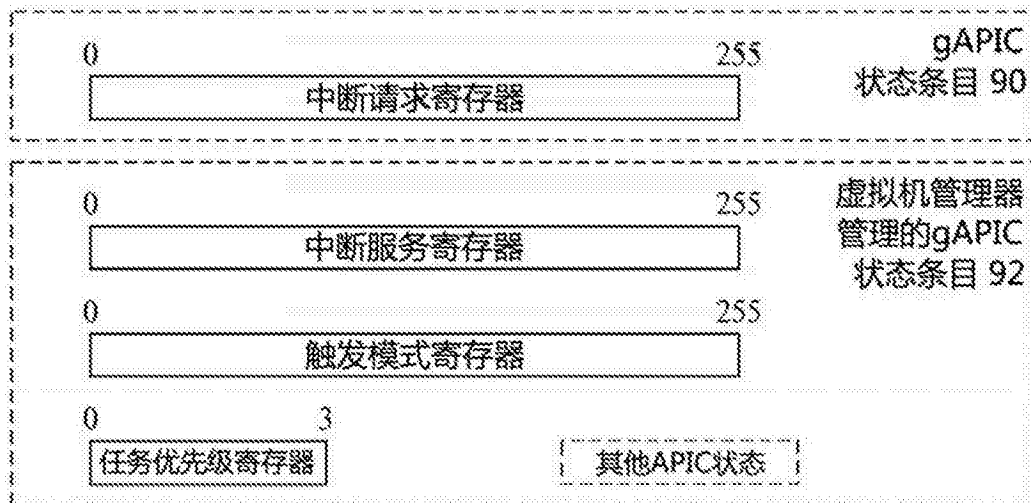


图5

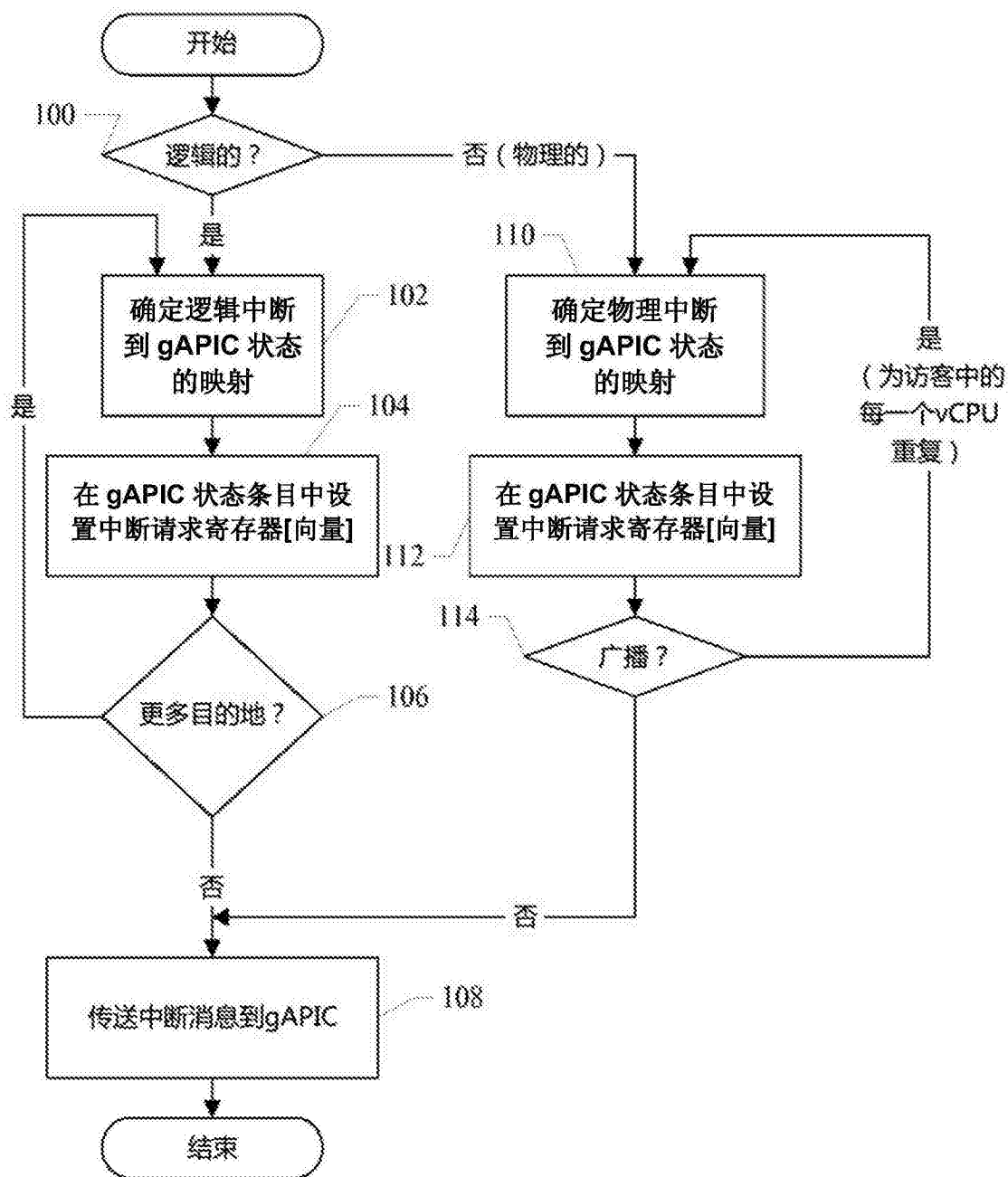


图6

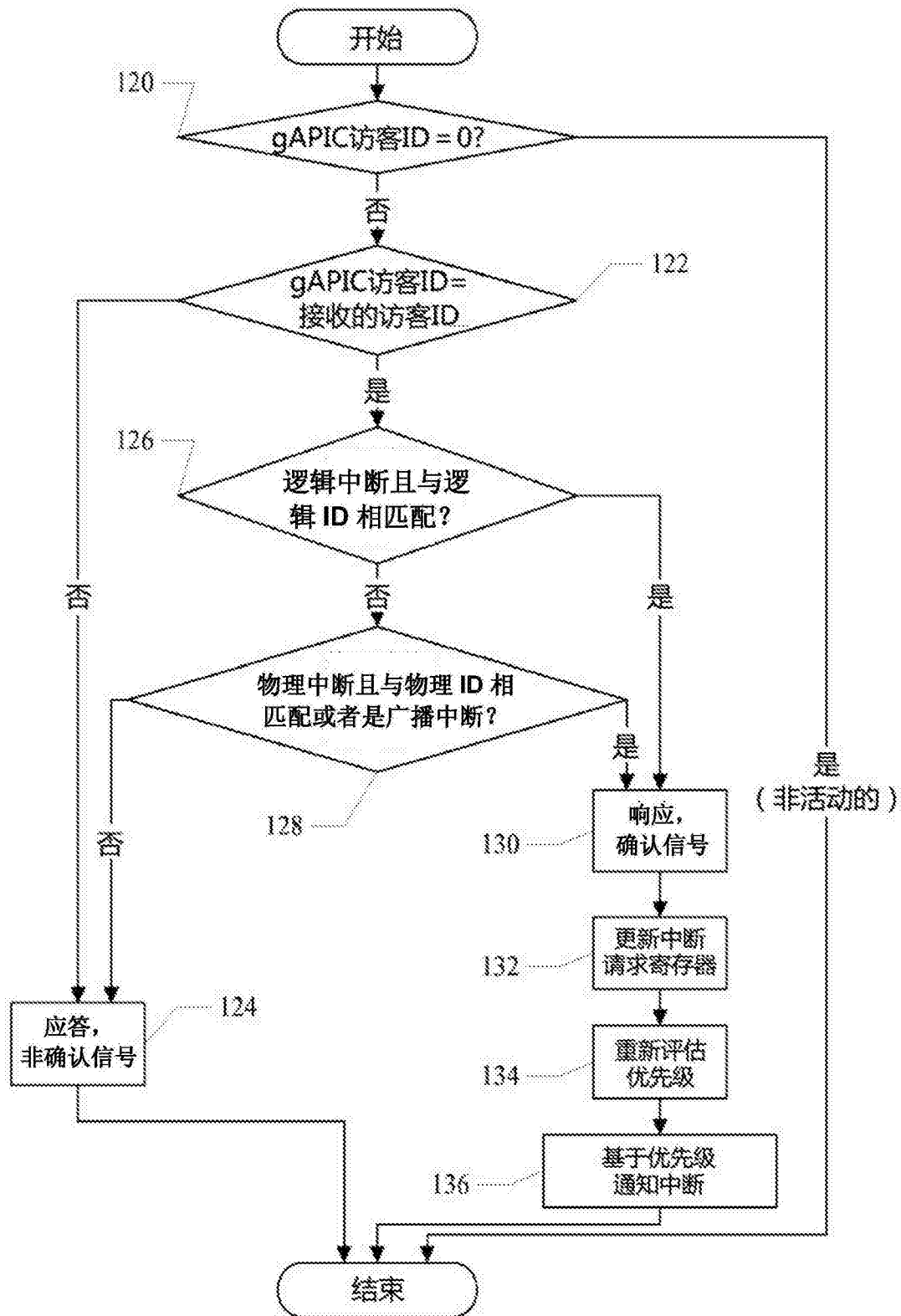


图7

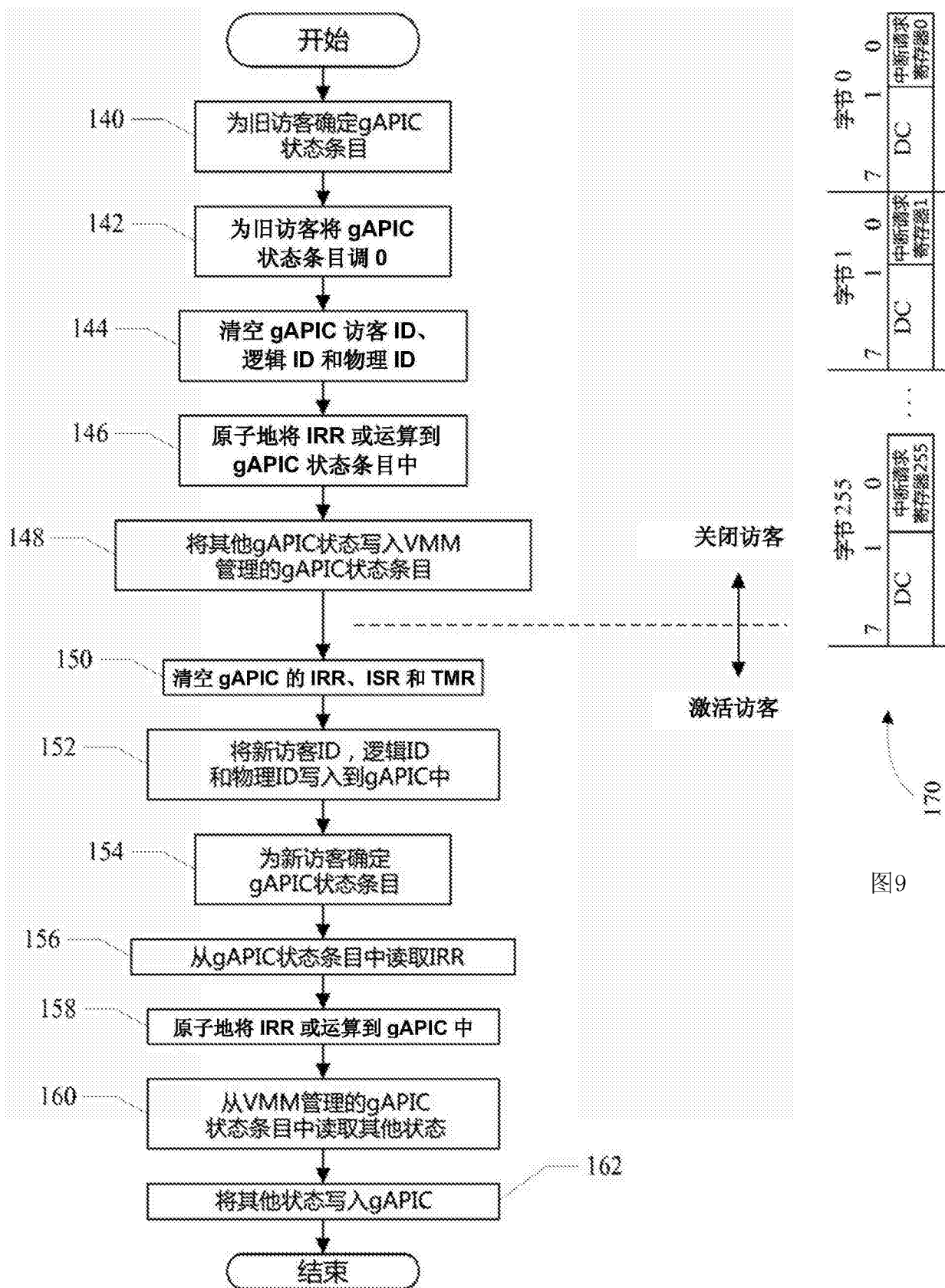


图9

图8

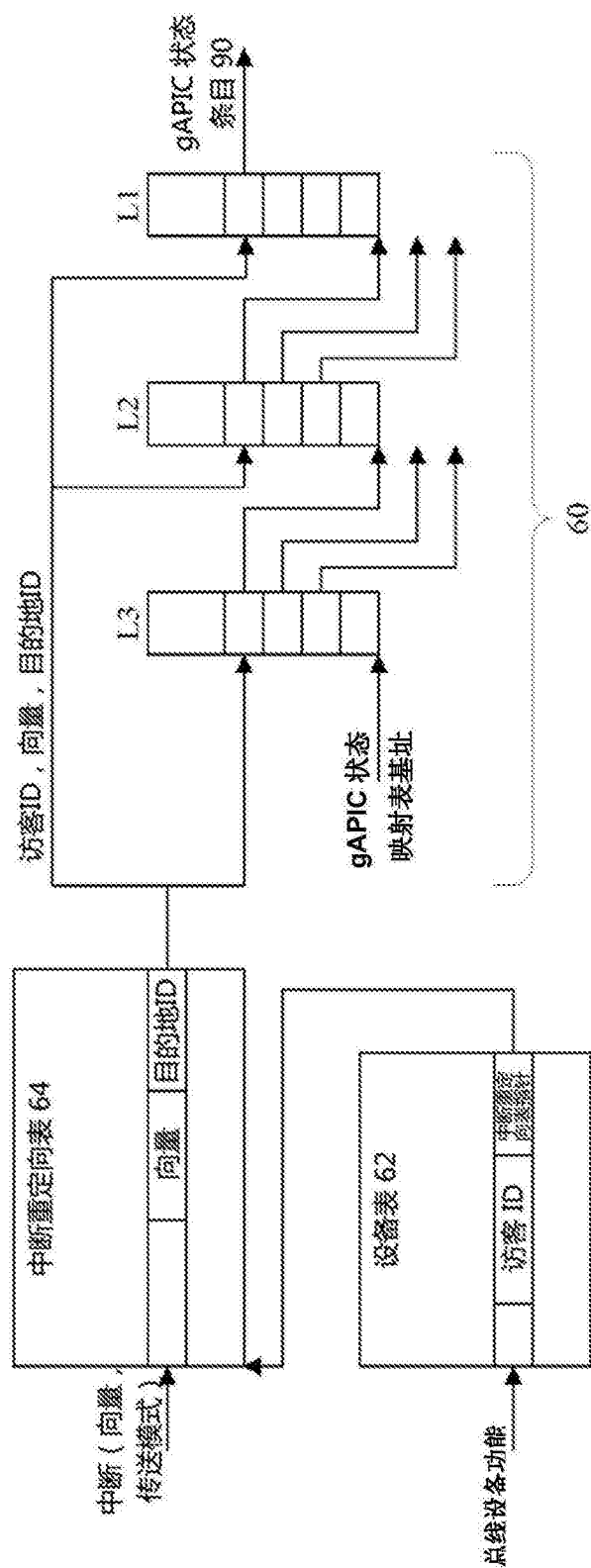


图10

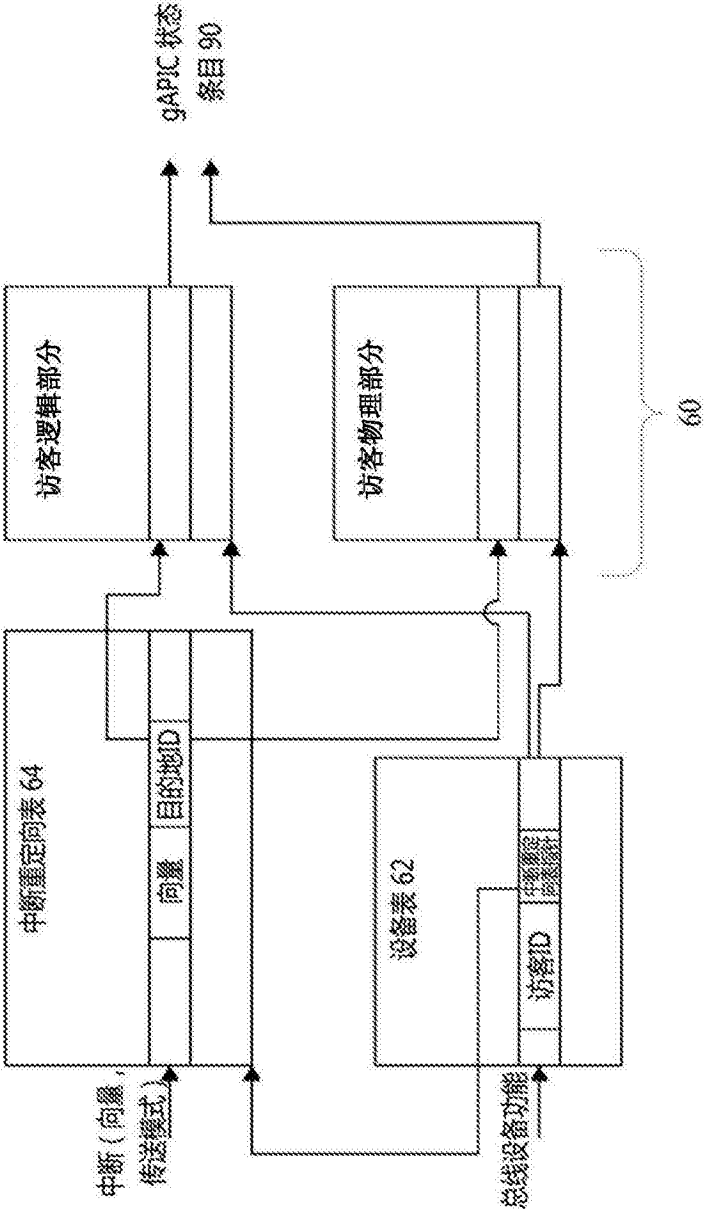


图11

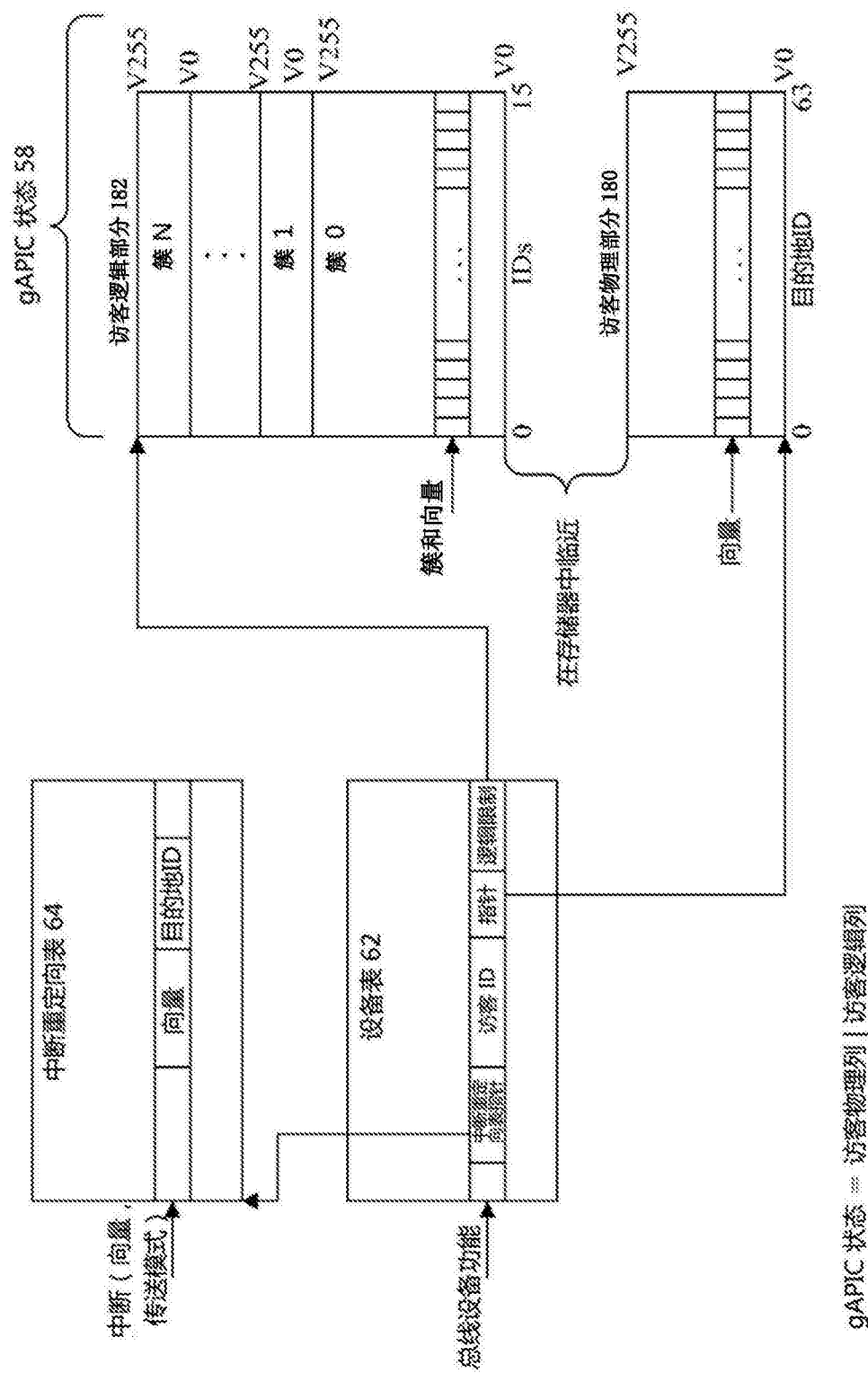


图12

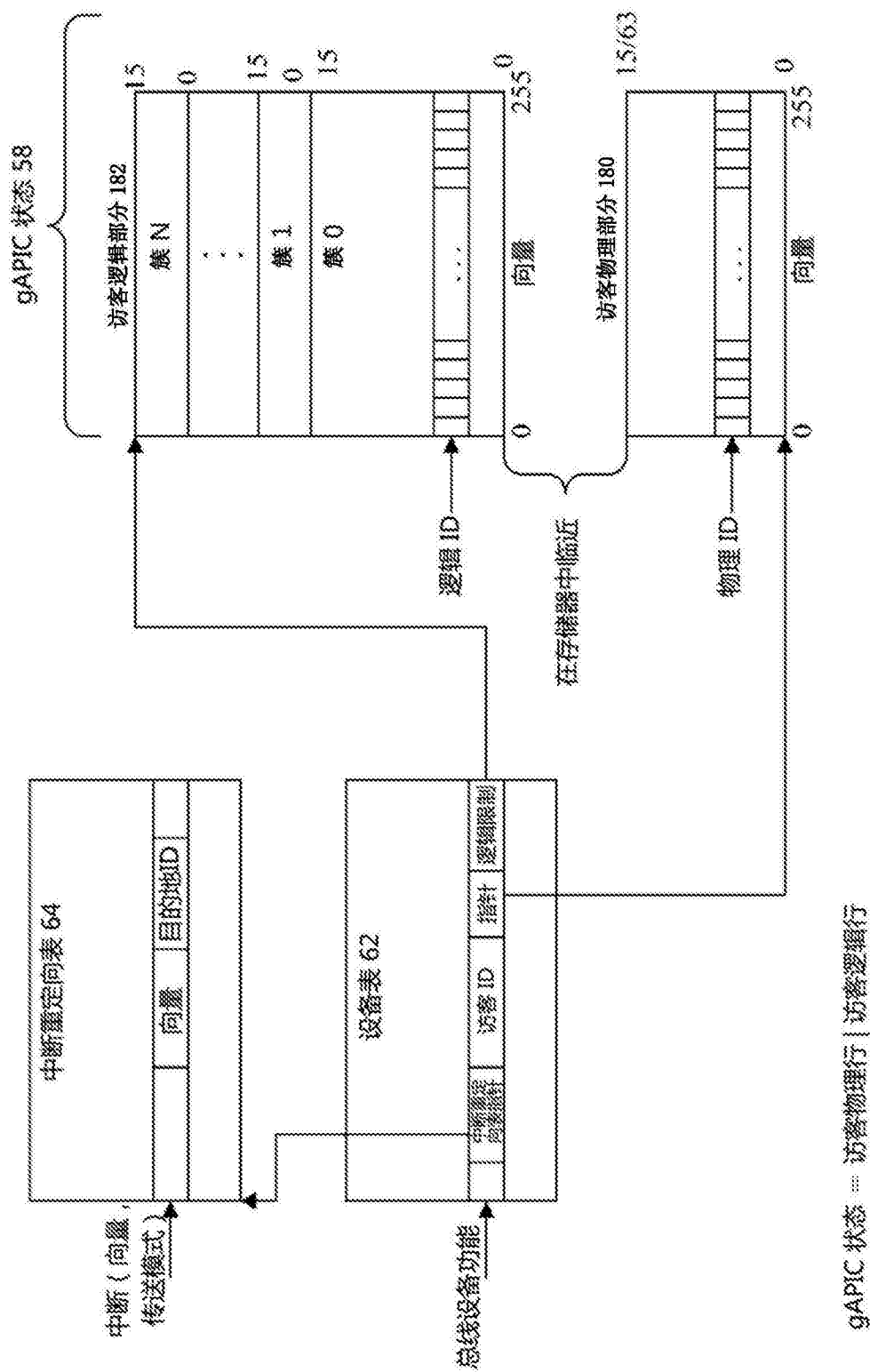


图13

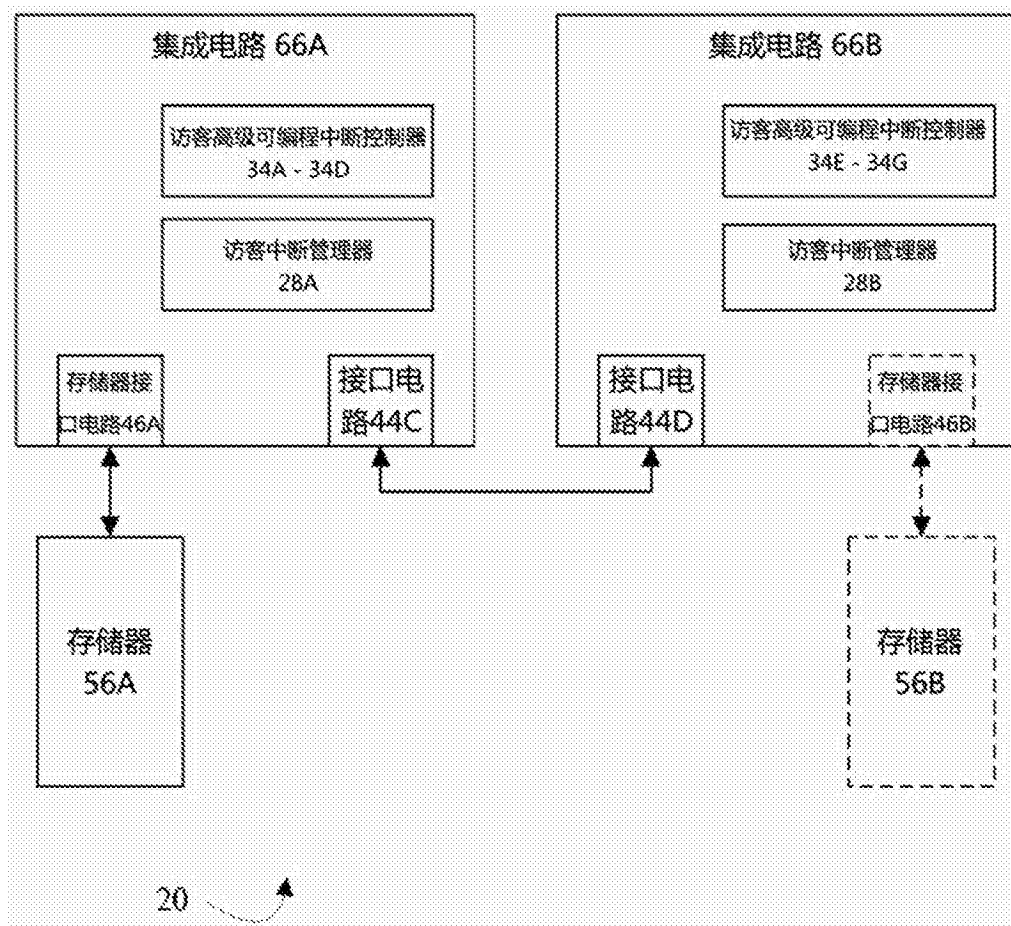


图14

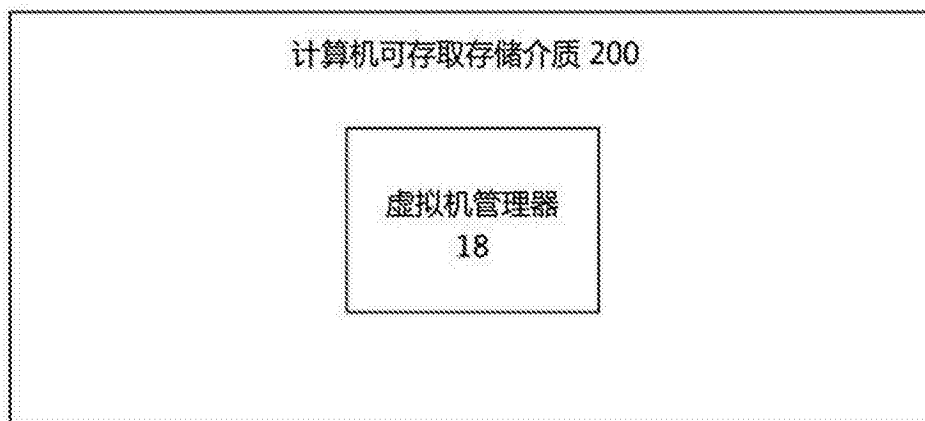


图15