[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 12/02 (2006.01)
G06F 9/46 (2006.01)

[12] 发明专利说明书

专利号 ZL 200610136100.2

[45] 授权公告日 2010 年 3 月 17 日

[11] 授权公告号 CN 100594481C

[22] 申请日 2006.10.19
[21] 申请号 200610136100.2
[30] 优先权
[32] 2005.10.20 [33] US [31] 11/255,393
[73] 专利权人 国际商业机器公司
地址 美国纽约
[72] 发明人 安托尼萨米·A.·拉杰德拉恩
[56] 参考文献
CN1222885C 2005.10.12
CN1536481A 2004.10.13
CN1637724A 2005.7.13
US6336995B1 2002.4.2
WO2004/053698A2 2004.6.24
审查员 刘 琳

[74] 专利代理机构 中国国际贸易促进委员会专利商标事务所
代理人 李镇江

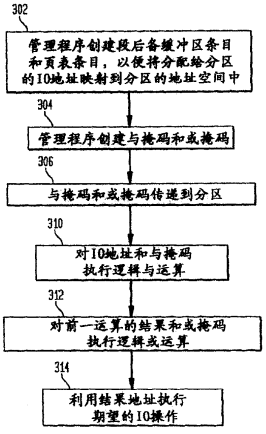
权利要求书 3 页 说明书 8 页 附图 3 页

[54] 发明名称

允许运行在逻辑分区上的程序访问资源的方法与系统

[57] 摘要

公开了用于使运行在逻辑分区的数据处理系统的逻辑分区上的程序直接访问数据处理系统的资源的方法与系统。 该方法包括步骤：所述程序将用于数据处理系统上特定类型资源的第一地址转换成分配给所述逻辑分区的地址空间中的第二地址；及所述程序使用所述第二地址空间访问分配给所述逻辑分区的所述特定类型的资源。 以这种方式，本发明可以用于使运行在分区地址空间中的程序直接访问 IO 设备，从而避免进行管理程序调用的开销。



1、一种使运行在逻辑分区的数据处理系统的逻辑分区上的程序直接访问数据处理系统的资源的方法，该方法包括步骤：

所述程序将用于数据处理系统上特定类型资源的第一地址转换成分配给所述逻辑分区的地址空间中的第二地址；及

所述程序使用所述第二地址空间访问分配给所述逻辑分区的所述特定类型的资源；

其中所述将所述第一地址转换成所述第二地址的步骤包括：

利用所述第一地址执行一系列逻辑运算以获得所述第二地址，该步骤进一步包括利用第一地址和给定的与掩码执行逻辑与运算的步骤。

2、如权利要求 1 所述的方法，其中数据处理系统包括资源集合和用于管理对所述资源集合的访问的管理程序，而且其中所述方法还包括如下步骤：所述程序将所述第一地址转换成所述第二地址并使用所述第二地址而不使用所述管理程序访问所述资源。

3、如权利要求 1 所述的方法，其中所述执行一系列逻辑运算的步骤还包括如下步骤：利用给定的或掩码执行逻辑或运算。

4、如权利要求 3 所述的方法，其中数据处理系统包括用于管理对系统资源的访问的管理程序，而且所述处理步骤还包括如下步骤：利用该管理程序创建所述与掩码和所述或掩码。

5、如权利要求 4 所述的方法，其中：

所述执行逻辑与运算的步骤包括利用第一地址和给定的与掩码执行逻辑与运算以便获得中间地址值的步骤；及

所述执行逻辑或运算的步骤包括利用所述中间地址值和给定的或掩码执行逻辑或运算以便获得所述第二地址的步骤。

6、如权利要求 1 所述的方法，其中第一地址在分配给所述逻辑分区的地址空间之外。

7、一种使运行在逻辑分区的数据处理系统的逻辑分区上的程序直接访问数据处理系统的资源的系统，该系统包括：

在所述程序中将用于数据处理系统上特定类型资源的第一地址转换成分配给所述逻辑分区的地址空间中的第二地址的装置；及

在所述程序中使用所述第二地址访问分配给所述逻辑分区的所述特定类型的资源的装置；

其中所述用于将所述第一地址转换成所述第二地址的装置包括：

用于利用所述第一地址执行一系列逻辑运算以获得所述第二地址的装置，该装置进一步包括利用第一地址和给定的与掩码执行逻辑与运算的装置。

8、如权利要求 7 所述的系统，其中数据处理系统包括资源集合和用于管理对所述资源集合的访问的管理程序，而且其中所述用于将所述第一地址转换成所述第二地址的装置和所述用于使用所述第二地址的装置不使用所述管理程序而运行。

9、如权利要求 7 所述的系统，其中所述用于利用所述第一地址执行一系列逻辑运算以获得所述第二地址的装置还包括利用给定的或掩码执行逻辑或运算的装置。

10、如权利要求 9 所述的系统，其中所述利用第一地址和给定的与掩码执行逻辑与运算的装置用于利用第一地址和给定的与掩码执行逻辑与运算以获得中间地址值；并且其中所述利用给定的或掩码执行逻辑或运算的装置用于利用所述中间地址值和给定的或掩码执行逻辑或运算以获得所述第二地址。

11、如权利要求 10 所述的系统，其中所述特定类型的所述资源是 IO 资源。

12、一种提供对分区成逻辑分区集合的数据处理系统中 IO 资源的访问的方法，每个所述逻辑分区都具有相应的地址空间和相应的资源集合，而且其中所述数据处理系统包括用于管理对数据处理系统资源的访问的管理程序，该方法包括步骤：

在所述逻辑分区中的第一逻辑分区上运行程序，所述程序具有用于分配给所述逻辑分区中的第二逻辑分区的 IO 资源的地址；

将所述第一地址转换成用于分配给所述第一逻辑分区的 IO 资源

的第二地址；

利用所述第二地址向程序提供对分配给所述第一逻辑分区的所述 IO 资源的访问；及

在不使用所述管理程序的情况下执行所述转换和使用步骤，以便将所述第一地址转换成所述第二地址或者向程序提供对分配给所述第一逻辑分区的所述 IO 资源的访问；

其中所述将所述第一地址转换成用于分配给所述第一逻辑分区的 IO 资源的第二地址的将所述第一地址转换成用于分配给所述第一逻辑分区的 IO 资源的第二地址步骤包括：

利用所述第一地址执行一系列逻辑运算以获得所述第二地址，该步骤进一步包括利用第一地址和给定的与掩码执行逻辑与运算的步骤。

13、如权利要求 12 所述的方法，其中所述利用所述第一地址执行一系列逻辑运算以获得所述第二地址的步骤进一步包括利用给定的或掩码执行逻辑或运算。

14、如权利要求 13 所述的方法，其中：

利用第一地址和与掩码执行逻辑与运算，以便获得中间地址值；及

利用所述中间地址值和或掩码执行逻辑或运算，以便获得所述第二地址。

15、如权利要求 14 所述的方法，还包括如下步骤：利用管理程序创建所述与掩码和所述或掩码。

允许运行在逻辑分区上的程序访问资源的方法与系统

技术领域

本发明总体上涉及管理逻辑分区的数据处理系统中多个操作系统之间的资源。更具体而言，本发明涉及用于管理这种数据处理系统中对例如 I/O 资源的资源的访问的方法与系统。

背景技术

数据处理系统（平台）中的逻辑分区选项（LPAR）允许单个操作系统（OS）的多个拷贝或多个异种操作系统同时运行在单个数据处理系统平台上。给操作系统图像在其中运行的分区指定不重叠的平台资源子集。这些平台可分配的资源包括一个或多个体系结构不同的处理器及其中断管理区域、系统存储器区域和 I/O 适配器总线槽。分区资源是由其自己的对 OS 图像的开放固件设备树表示的。

运行在平台中的每个不同 OS 或 OS 图像是彼此保护的，使得一个逻辑分区上的软件错误不会影响任何一个其它分区的正确运行。这是通过分配由每个 OS 图像直接管理的不相交的平台资源集合和通过提供用于确保各种图像不能控制还未分配给它的任何资源来提供的。此外，防止 OS 分配资源控制中的软件错误影响任何其它图像的资源。因此，OS 的每个图像（或每个不同的 OS）直接控制平台中可分配资源的不同集合。

为了控制和/或管理各种平台环境中的多个操作系统，通常使用可以称为管理程序的多个全局软件系统和/或固件组件。管理程序通常配置成管理和/或控制多个计算机硬件系统上每个操作系统可用资源的分配/使用。例如，除计算机系统的其它已知特征以外，管理程序可以控制用于整个计算机系统数据存储介质的资源访问与分配、对可用

系统 CPU 的访问和/或系统输入/输出（IO）设备适配器的任何一个。管理程序还可以配置成确保各个独立的分区不会注意到每个其它分区的存在并且不会干扰它们各自的运行。

通过向分区分配地址范围，然后将要分配的资源指定给该分区的所分配分区范围中的地址，来将资源指定给特定的分区。例如，在逻辑分区的系统中，除了其它资源，IO（输入/输出）资源分配给逻辑分区。这些 IO 资源中的许多在分区之间不是共享的，而是专用于一个分区。IO 地址空间落在分区允许的地址空间范围之外。因此，运行在分区中的例如设备驱动器的程序必须进行管理程序调用来访问 IO 设备。尽管管理程序有效地管理对 IO 资源的访问，但还有一定量与进行管理程序调用相关的开销。

发明内容

本发明的一个目的是提供用于管理对逻辑分区的数据处理系统中资源的访问的改进的过程。

本发明的另一目的是使运行在逻辑分区的数据处理系统中的逻辑分区地址空间中的程序能够直接访问 IO 设备，从而避免进行管理程序调用的开销。

本发明还有一个目的是将落在分区地址空间之外的存储器映射 IO 地址映射到分区的地址空间中，从而使运行在逻辑分区中的程序能够直接访问分配给它的存储器映射的 IO 资源。

这些和其它目的是利用使运行在逻辑分区的数据处理系统中的逻辑分区上的程序能够直接访问数据处理系统的资源的方法和系统来获得的。该方法包括步骤：所述程序将用于数据处理系统上特定类型资源的第一地址转换成分配给所述逻辑分区的地址空间中的第二地址；及所述程序利用所述第二地址访问分配给所述逻辑分区的所述特定类型资源。以这种方式，本发明可以用于例如使运行在分区地址空间中的程序能够直接访问 IO 设备，从而避免进行管理程序调用的开销。

参考指定并显示本发明优选实施方式的附图，本发明的更多好处与优点将从以下具体描述的考虑中变得显而易见。

附图说明

图 1 描述了可以用于实现本发明的数据处理系统的框图。

图 2 示出了其中本发明可以实现的逻辑分区平台的框图。

图 3 是说明用于实践本发明的示例处理的流程图。

具体实施方式

现在参考附图，尤其是参考图 1，描述了可以实现为逻辑分区的数据处理系统的数据处理系统的框图。数据处理系统 100 可以是包括连接到系统总线 106 的多个处理器 101、102、103 和 104 的对称微处理器（SMP）系统。例如，数据处理系统 100 可以是 IBM RS/6000，这是位于纽约 Armonk 的国际商用机器公司的产品。可选地，可以采用单处理器系统。连接到系统总线 106 的还有向多个本地存储器 160-163 提供接口的存储器控制器/高速缓冲存储器 108。I/O 总线桥 110 连接到系统总线 106 并向 I/O 总线 112 提供接口。如所描述的，存储器控制器/高速缓冲存储器 108 和 I/O 总线桥 110 可以集成。

数据处理系统 100 是逻辑分区的数据处理系统。因此，数据处理系统 100 可以具有同时运行的多个异种操作系统（或单个操作系统的多个实例）。这多个操作系统中的每一个都可以具有任意多个在其中执行的软件程序。数据处理系统 100 逻辑分区成使不同的 I/O 适配器 120-121、128-129、136-137 和 146-147 可以指定给不同的逻辑分区。

因此，例如，数据处理系统 100 可以分成三个逻辑分区，P1、P2 和 P3。I/O 适配器 120-121、128-129 和 136-137 中的每一个、处理器 101-104 中的每一个及本地存储器 160-164 中的每一个都指定给三个分区中的一个。例如，处理器 101、存储器 160 及 I/O 适配器 120、128 和 129 可以指定给逻辑分区 P1；处理器 102-103、存储器 161 及 I/O 适配器 121 和 137 可以指定给逻辑分区 P2；而处理器 104、存储器

162-163 及 I/O 适配器 136 和 146-147 可以指定给逻辑分区 P3。

在数据处理系统 100 中执行的每个操作系统都指定给不同的逻辑分区。因此，在数据处理系统 100 中执行的每个操作系统都只可以访问在其逻辑分区中的那些 I/O 单元。因此，例如，高级交互执行体（AIX）操作系统的一个实例可以在分区 P1 中执行，AIX 操作系统的第二实例（图像）可以在分区 P2 中执行，而 Windows 2000™ 操作系统可以在逻辑分区 P3 中执行。Windows 2000 是位于华盛顿 Redmond 的微软公司的产品与商标。

连接到 I/O 总线 112 的外围组件互连（PCI）主桥 114 提供与 PCI 本地总线 115 的接口。多个终端桥 116-117 可以连接到 PCI 总线 115。典型的 PCI 总线实现将支持四个用于提供扩展槽或内插式连接器的终端桥。每个终端桥 116-117 都通过 PCI 总线 118-119 连接到 PCI I/O 适配器 120-121。每个 I/O 适配器 120-121 提供数据处理系统 100 与例如作为服务器 100 的客户端的其它网络计算机的输入/输出设备之间的接口。只有单个 I/O 适配器 120-121 可以连接到每个终端桥 116-117。每个终端桥 116-117 配置成防止错误传播进入 PCI 主桥并进入数据处理系统 100 的更高层次。通过这样做，由任一终端桥 116-117 接收的错误都与可以在不同分区中的其它 I/O 适配器 121、128-129 和 136-137 的共享总线 115 和 112 隔离。其后，在一个分区中的 I/O 设备中发生的错误不会被另一分区的操作系统“看到”。因此，一个分区中操作系统的完整性不会受另一逻辑分区中发生的错误影响。没有这种错误的隔离，在一个分区的 I/O 设备中发生的错误可能造成另一分区的操作系统或应用程序停止运行或停止正确运行。

附加 PCI 主桥 122、130 和 140 提供用于附加 PCI 总线 123、131 和 141 的接口。每个附加 PCI 总线 123、131 和 141 都连接到分别通过 PCI 总线 126-127、134-135 和 144-145 连接到 PCI I/O 适配器 128-129、136-137 和 146-147 的多个终端桥 124-125、132-133 和 142-143。因此，可以通过 PCI I/O 适配器 128-129、136-137 和 146-147 中的每一个支持例如调制解调器或网络适配器的附加 I/O 设备。以这

种方式，服务器 100 允许连接到多个网络计算机。如所描述的，存储器映射的图形适配器 148 和硬盘 150 也可以直接或间接地连接到 I/O 总线 112。硬盘 150 可以在各分区之间逻辑分区，而不需要附加的硬盘。但是，如果期望，也可以使用附加的硬盘。

本领域普通技术人员将理解图 1 所描述的硬件可以变化。例如，除了或代替所描述的硬件，如光盘驱动器等的其它外围设备也可以使用。所描述的例子不是要暗示关于本发明的限制。

现在参考图 2，描述其中本发明可以实现的示例逻辑分区平台的框图。逻辑分区平台 200 中的硬件可以实现为例如图 1 中的数据处理系统 100。逻辑分区平台 200 包括分区的硬件 230、虚拟地址转换硬件 280、管理程序 210 和操作系统 202-208。操作系统 202-208 可以是同时运行在平台 200 上的单个操作系统的多个拷贝或者多个异种操作系统。

分区的硬件 230 包括多个处理器 232-238、多个系统存储器单元 240-246、多个输入/输出（I/O）适配器 248-262 及存储单元 270。处理器 232-238、系统存储器单元 240-246 和 I/O 适配器 248-262 中的每一个都可以指定给逻辑分区平台 200 中的多个分区中的一个，每个分区对应于操作系统 202-208 中的一个。

实现为固件的管理程序 210 创建并执行逻辑分区平台 200 的分区。固件是存储在不需要电能就可以保持其内容的存储器芯片中的软件，例如只读存储器（ROM）、可编程 ROM（PROM）、可擦除可编程 ROM（EPROM）、电可擦除可编程 ROM（EEPROM）及非易失随机存取存储器（非易失 RAM）。

虚拟地址转换硬件 280 提供用于将操作系统 202-208 中一个的资源的虚拟存储器地址页转换成对应于该资源的物理硬件地址的机制。虚拟存储器是模拟比实际存在的更多的存储器的方法，使得平台 200 可以运行更大的软件程序或者同时运行更多的程序。虚拟存储器将软件程序分成小段，称为页，并将尽可能多的页带到适合该软件程序保留区域的存储器 240-246 中。当需要附加页时，虚拟存储器通过利用

一个 I/O 适配器 248-260 交换当前在存储器中但磁盘存储器 270 或某种其它输入/输出设备不再需要的页来为它们腾出空间，从而释放用于该附加页的存储器。虚拟地址转换硬件 280 跟踪被修改的页，因此当又需要时，被修改的页可以检索。

在现有技术中，当运行在一个逻辑分区中的程序想访问 IO 设备时，该程序必须进行管理程序调用。管理程序检查被请求的 IO 是否已经分配给该逻辑分区。如果 IO 已经分配给该逻辑分区，则管理程序将请求的程序映射到 IO 资源。

本发明使运行在分区地址空间中的程序可以直接访问 IO 设备，从而避免进行管理程序调用的开销。

更具体而言，参考图 3，如在步骤 302 所表示的，当逻辑分区在资源分配后启动时，管理程序可以创建段后备缓冲区条目和页表条目，以便将分配给分区的 IO 地址映射到分区的地址空间中。然后，在步骤 304，管理程序创建与掩码和或掩码。这些掩码是以当应用到 IO 实际地址时将产生分区地址空间中有效地址的方式选择的。然后，在步骤 306，这些与掩码和或掩码可以传递到分区。

如在步骤 310 所表示的，当运行在逻辑分区中的程序想访问 IO 设备时，它对 IO 地址和上面提到的与掩码执行逻辑与运算。然后，在步骤 312，程序对前一运算的结果和或掩码执行逻辑或运算。结果将是分区地址空间中的地址。如在步骤 314 所表示的，它可以利用结果地址执行期望的 IO 操作。结果地址将在分区的地址空间中。

例如：

假定：

IO 设备实际地址空间： 0x3F2 0000 0000 到 0x3FD 87FF FFFF

分区的 RMOR： 0x0004

分区的 RMLR： 256GB

管理程序可以创建段后备缓冲区条目和页表条目，以便将有效地址 0x0004 0032 0000 0000 到 0x0004 003D 87FF FFFF 映射到实际地址 0x0000 03F2 0000 0000 到 0x0000 03FD 87FF FFFF。

有效地址－>实际地址

0x0004 0032 0000 0000－>0x0000 03F2 0000 0000

0x0004 003D 87FF FFFF－>0x0000 03FD 87FF FFFF

这种情况下的与掩码将是 0x0000 000F FFFF FFFF，而或掩码将是 0x0000 0030 0000 0000。

比方说，当分区想从存储器映射的 IO 地址 0x3F200000000 读时，它与 0xFFFFFFFFF 执行逻辑与。然后，对结果和 0x3000000000 执行逻辑或。结果地址将是 0x3200000000。

0x0000 03F2 0000 0000 实际地址

0x0000 000F FFFF FFFF 与掩码

---------- 与运算

0x0000 0002 0000 0000

0x0000 0030 0000 0000 或掩码

----------或运算

0x0000 0032 0000 0000

现在，分区可以在地址 0x32 0000 0000 执行读操作。然后，因为 RMOR 是 4，所以这个有效地址将变成 0x4 0032 0000 0000。然后，这个有效地址将由 PowerPC 地址转换机制转换成实际地址 0x3F2 0000 0000。

尽管本发明已经在完整功能数据处理系统的环境下进行了描述，但本领域普通技术人员将理解本发明的处理能够以指令的计算机可读介质的形式和多种形式分布，还应当理解不管实际用于执行分布的信号承载介质的特定形式是什么，本发明都同等地适用，指出这些是很重要的。计算机可读介质的例子包括如软盘、硬盘驱动器、RAM 和 CD-ROM 的记录类型介质，及如数字和模拟通信链路的发射类型介质。

本发明描述的提出是为了说明和描述，而不是穷尽的，或者要将本发明限定到所公开的形式。许多修改与变化对本领域普通技术人员都是显而易见的。实施方式的选择与描述是为了最好地解释本发明的

原理、实践应用，并使本领域其它普通技术人员能够理解本发明的具有适于所期望特定应用的各种修改的各种实施方式。

图1

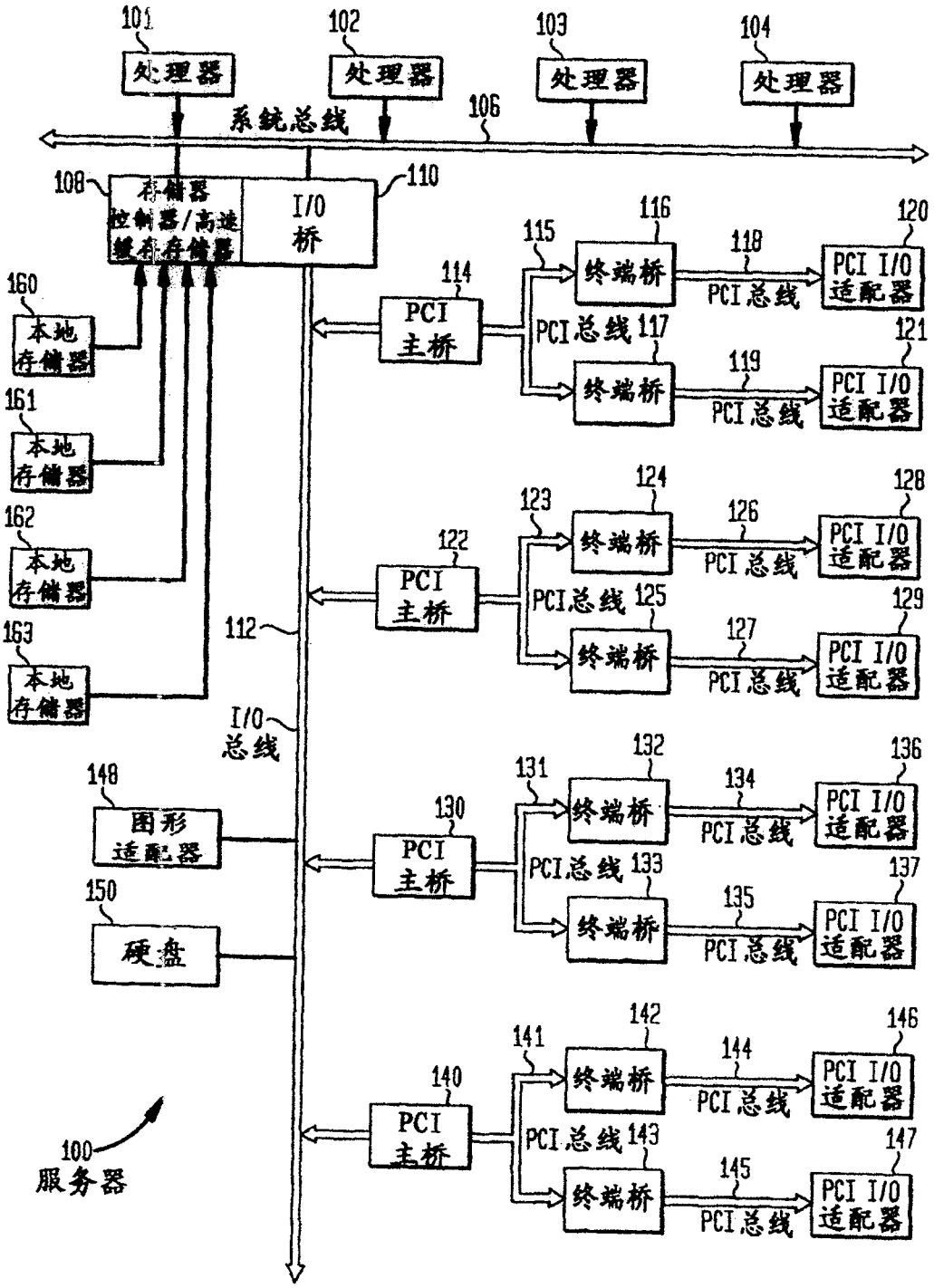

101
处理器
102
处理器
103
处理器
104
处理器
106
系统总线
108
存储器控制器/高速缓存存储器
110
I/O桥
160
本地存储器
161
本地存储器
162
本地存储器
163
本地存储器
112
I/O总线
114
PCI主桥
115
116
终端桥
118
PCI总线
120
PCI I/O适配器
PCI总线
117
终端桥
119
PCI总线
121
PCI I/O适配器
122
PCI主桥
123
124
终端桥
126
PCI总线
128
PCI I/O适配器
PCI总线
125
终端桥
127
PCI总线
129
PCI I/O适配器
130
PCI主桥
131
132
终端桥
134
PCI总线
136
PCI I/O适配器
PCI总线
133
终端桥
135
PCI总线
137
PCI I/O适配器
140
PCI主桥
141
142
终端桥
144
PCI总线
146
PCI I/O适配器
PCI总线
143
终端桥
145
PCI总线
147
PCI I/O适配器
148
图形适配器
150
硬盘
100
服务器

图2

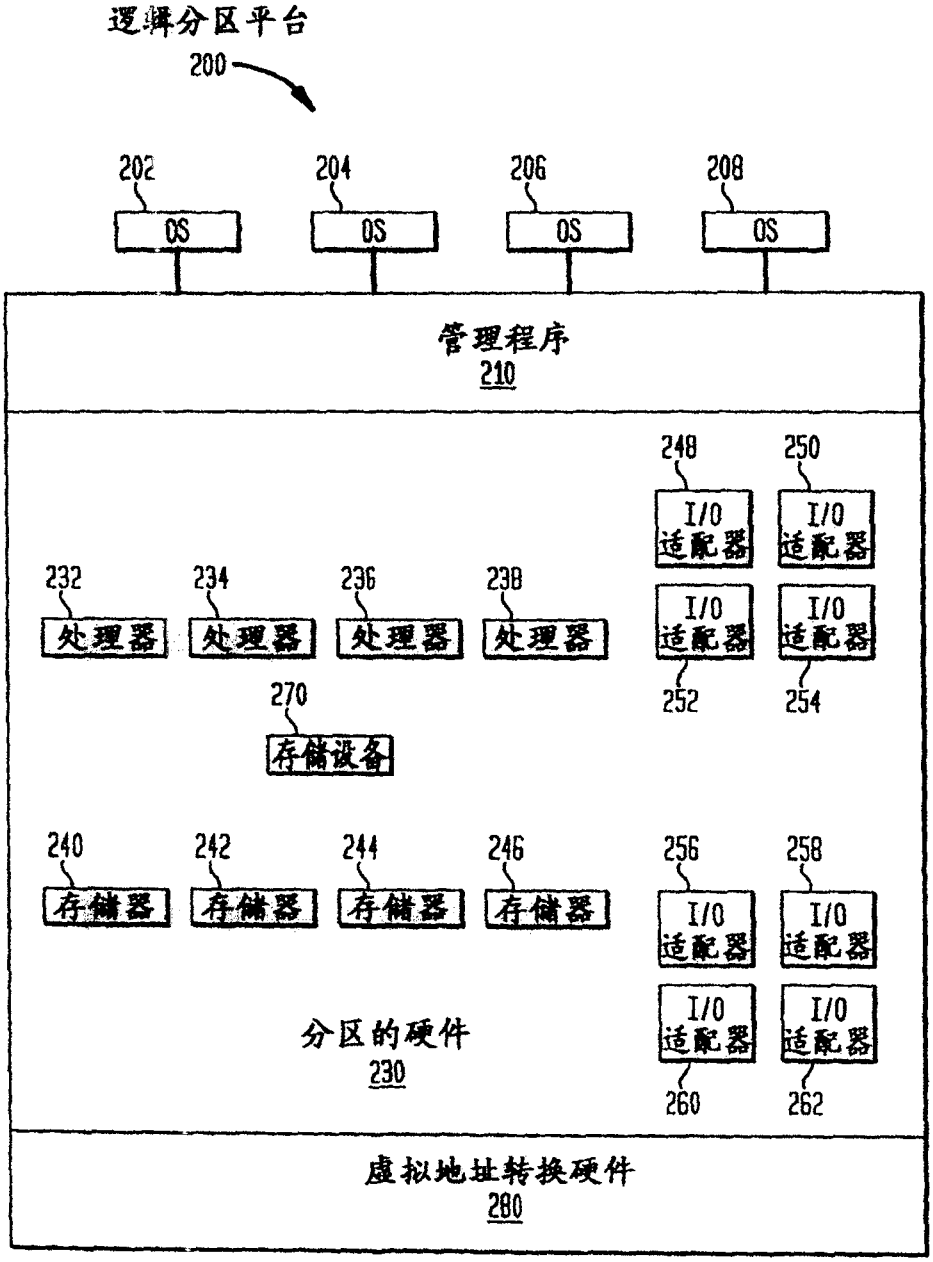
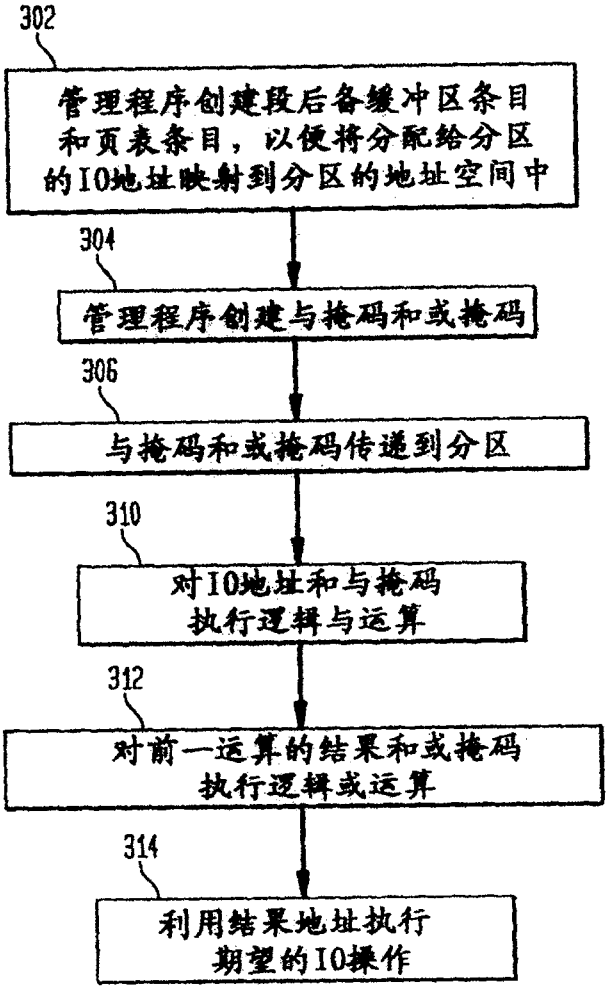

逻辑分区平台
200
202
204
206
208
OS
OS
OS
OS
管理程序
210
248
250
I/O 适配器
I/O 适配器
232
234
236
238
处理器
处理器
处理器
处理器
I/O 适配器
I/O 适配器
252
254
270
存储设备
240
242
244
246
256
258
存储器
存储器
存储器
存储器
I/O 适配器
I/O 适配器
I/O 适配器
I/O 适配器
分区的硬件
230
260
262
虚拟地址转换硬件
280

图 3


302
管理程序创建段后备缓冲区条目和页表条目，以便将分配给分区的IO地址映射到分区的地址空间中
304
管理程序创建与掩码和或掩码
306
与掩码和或掩码传递到分区
310
对IO地址和与掩码执行逻辑与运算
312
对前一运算的结果和或掩码执行逻辑或运算
314
利用结果地址执行期望的IO操作