



# [12] 发明专利申请公开说明书

[21] 申请号 200510070056.5

[43] 公开日 2005 年 11 月 9 日

[11] 公开号 CN 1694079A

[22] 申请日 2005.4.28

[21] 申请号 200510070056.5

[30] 优先权

[32] 2004.4.28 [33] US [31] 10/834,138

[71] 申请人 微软公司

地址 美国华盛顿州

[72] 发明人 K·W·斯特夫尔比姆

[74] 专利代理机构 上海专利商标事务所有限公司

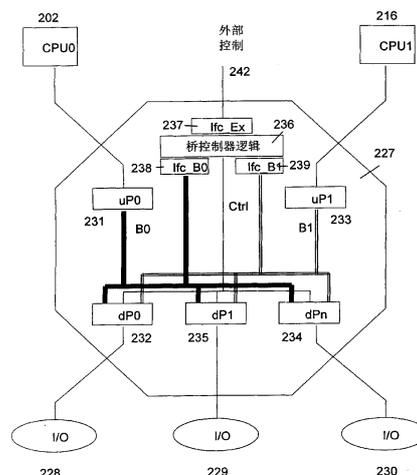
代理人 李 玲

权利要求书 3 页 说明书 6 页 附图 7 页

[54] 发明名称 可配置的 PCI Express 开关

[57] 摘要

一种使多个 CPU 能够通过单个开关连接到多个 I/O 设备的可配置开关。该开关可以被级联，以允许树中更多的 CPU 和/或更多的 I/O。配置对总线和端点设备的枚举是透明的。诸如 SMBus 或硬件跨接等简单管理输入用于设置对 CPU 的设备分配。使用管理器和 PCI Express 热插拔控制器寄存器在设备通过开关内的 PCI 总线在 CPU 之间切换时允许设备树的热插拔重新配置。



1. 一种可配置 PCI Express 开关, 其特征在于, 包括:  
多个上游 PCI 对 PCI 端口;  
5 多个下游 PCI 对 PCI 端口;  
与每一上游端口相关联的内部 PCI; 以及  
配置哪一上游端口与哪一下游端口通信的控制器。
2. 如权利要求 1 所述的开关, 其特征在于, 与所述上游端口的预定一个通信的 CPU 是可配置的, 以与所述下游端口的任一个通信。
- 10 3. 如权利要求 2 所述的开关, 其特征在于, 每一所述内部 PCI 总线连接到每一下游端口。
4. 如权利要求 2 所述的开关, 其特征在于, 所述 CPU 通过读取 PCI 配置空间寄存器发现设备, 并枚举在与所述 CPU 相关联的 PCI 总线上找到的设备, 其中, 所述设备连接到与所述上游端口的预定一个通信的下游端口。
- 15 5. 如权利要求 4 所述的开关, 其特征在于, 如果第二 CPU 请求对向所述 CPU 枚举的设备的访问, 则所述控制器启动对所述 CPU 的请求, 以释放所述设备所连接的下端口, 以及  
其中, 如果所述请求被准许, 则所述 CPU 默许所述设备, 并启动对所述控制器的准许, 以释放下游端口。
- 20 6. 如权利要求 5 所述的开关, 其特征在于, 所述控制器指令所述下游端口执行从与所述 CPU 相关联的所述总线的断开序列, 并且所述控制器指令下游端口执行向与所述第二 CPU 相关联的总线的连接序列。
7. 如权利要求 1 所述的开关, 其特征在于, 一内部配置控制寄存器用于定义所述下游端口的每一个响应于哪一内部总线。
- 25 8. 如权利要求 1 所述的开关, 其特征在于, 硬件跨接用于定义所述下游端口的每一响应于哪一内部总线。
9. 如权利要求 1 所述的开关, 其特征在于, 提供一外部配置管理接口, 用于一外部配置管理实体的配置控制。
10. 如权利要求 1 所述的开关, 其特征在于, 每一下游端口连接到每一内部  
30 PCI 总线, 并且其中, 每一下游端口仅响应于一个内部 PCI 总线。

11. 一种控制可配置 PCI Express 开关的方法，其特征在于，包括：  
读取 PCI 配置空间寄存器；  
发现多个下游 PCI 对 PCI 桥的一个；  
发现用于与所述多个上游 PCI 对 PCI 桥的所述一个相关联的总线的控制接口；  
5 以及  
枚举在所述总线上发现的设备。
12. 如权利要求 11 所述的方法，其特征在于，还包括：  
通过接口启动对控制器的发现请求；  
读取所述总线上每一设备的配置空间；以及  
10 通过所述接口响应，以返回所述发现请求所请求的信息。
13. 如权利要求 11 所述的方法，其特征在于，还包括：  
向多个 CPU 之一分配下游资源，每一所述 CPU 与所述多个上游 PCI 对 PCI  
桥的所述一个相关联。
14. 如权利要求 13 所述的方法，其特征在于，还包括查询一控制器，以确定  
15 哪些设备被分配给与所述多个上游 PCI 对 PCI 桥的所述一个相关联的所述总线。
15. 如权利要求 11 所述的方法，其特征在于，如果第一 CPU 请求对向第二  
CPU 枚举的设备的访问，则一控制器启动对所述第二 CPU 的请求，以释放所述设  
备所连接的下游端口，以及  
其中，如果所述请求被准许，则所述第二 CPU 默许所述设备，并启动对所述  
20 控制器的准许，以释放下游端口。
16. 如权利要求 15 所述的方法，其特征在于，所述控制器指令所述下游端口  
执行从与所述 CPU 相关联的所述总线的断开序列，并且所述控制器指令下游端口  
执行到与所述 CPU 相关联的总线的连接序列。
17. 一种连接多个 CPU 复合体的可配置 PCI Express 开关，其特征在于，包  
25 括：  
多个上游 PCI 对 PCI 端口，每一上游端口连接到所述 CPU 复合体的一个；  
多个下游 PCI 对 PCI 端口；  
多个内部 PCI 总线，每一内部 PCI 总线连接到一相应的上游端口；以及  
配置哪一上游端口与哪一下游端口通信的控制器，  
30 其中，每一下游端口连接到每一内部 PCI 总线，并且其中，每一下游端口仅  
响应于一个内部 PCI 总线，以及

其中，所述控制器通过一与每一 CPU 复合体相关联的接口接收发现请求。

18. 如权利要求 17 所述的开关，其特征在于，与所述上游端口的预定一个通信的 CPU 是可配置的，以与所述下游端口的任一个通信。

19. 如权利要求 17 所述的开关，其特征在于，如果 CPU 请求对向第二 CPU  
5 枚举的设备的访问，则所述控制器启动对所述第二 CPU 的请求，以释放所述设备所连接的下游端口，以及

其中，如果所述请求被准许，则所述第二 CPU 默许所述设备，并启动对所述控制器的准许，以释放下游端口。

20. 如权利要求 19 所述的开关，其特征在于，所述控制器指令所述下游组件  
10 执行从与所述第二 CPU 相关联的所述总线的断开序列，并且所述控制器指令下游组件执行到与所述 CPU 相关联的总线的连接序列。

## 可配置的 PCI Express 开关

## 5 技术领域

本发明一边涉及计算设备领域，尤其涉及用于 PCI Express 的可配置开关，以使多个上游端口能够连接到多个下游端口。

## 背景技术

10 在 1990 年代早期，引入了外围部件互连 (PCI) 标准。PCI 为连接的设备提供了对系统存储器的直接访问，但是使用桥来连接到前侧总线并连接到 CPU。PCI 能够连接多个组件。PCI 桥接芯片独立于 CPU 的速度调整了 PCI 总线的速度，以允许更高层次的可靠性，并确保 PCI 硬件制造商具有一致的设计约束。PCI 支持即插即用，它使设备或卡能够被插入到计算机中，并被自动识别和配置以对系统起作用。

15 用。

当今的软件应用程序更需要平台硬件，尤其是 I/O 子系统。来自各个视频和音频源的数据流现在在台式机和移动机器上是常见的。诸如视频点播和音频重分发等应用程序也在服务器上施加的实时约束。PCI 体系结构已经不再能够应付这些需求，并且提出了一种新标准，称为 PCI Express。

20 参考图 1，示出了一种可包括在计算设备中的 PCI Express 拓扑 100。该拓扑除 CPU 102 和存储器 103 之外，包含主机桥 (Host Bridge) 101 以及若干端点 104-109 (即，I/O 设备)。多个点对点连接由开关 110 来实现。开关 110 替换 PCI 使用的多点总线，并用于为 I/O 总线提供扇出 (fan-out)。开关 110 可提供不同端点 104-109 之间的对等通信，且如果该话务不涉及与高速缓存相干的存储器传输的话，它不需要被转发到主桥 101。开关 101 被示出为单独的逻辑元件，但是它可被集成到主桥

25 101 中。

尽管这是对于较旧的 PCI 体系结构的改进，然而它不提供在不同的计算设备间连接并共享端点的方法。由此，需要一种共享端点的系统和方法。这样的系统将很大程度上增强计算设备的灵活性，并提供降低功率消耗的方法。本发明提供了这样

30 一种解决方案。

### 发明内容

本发明允许多个 CPU 通过一个开关连接到多个 I/O 设备。开关可被级联，以在树中允许更多的 CPU 和/或更多的 I/O 设备。这一配置方法对于总线和端点设备的枚举是透明的。诸如 SMBus 或硬件跨接（strapping）等简单管理输入是设置向 CPU 的设备分配所需要的一切。

依照本发明的一个方面，提供了一种可配置的 PCI Express 开关，它包括多个上游 PCI 对 PCI 端口、多个下游 PCI 对 PCI 端口、唯一地与一上游端口相关联的内部 PCI 总线、以及配置哪一上游端口与哪一下游端口通信的控制器。

10 依照本发明的另一方面，提供了一种控制可配置 PCI Express 开关的方法。该方法包括读取 PCI 配置空间寄存器、发现多个下游 PCI 对 PCI 桥中的一个、发现与多个上游 PCI 对 PCI 桥之一相关联的总线相关联的控制接口、以及枚举在该总线上发现的设备。

依照本发明的又一方面，提供了一种连接多个 CPU 复合体（complex）的可配置 PCI Express 开关。该开关包括多个上游 PCI 对 PI 桥，其每一个唯一地连接到 CPU 复合体中的一个；多个下游 PCI 对 PCI 桥；多个内部 PCI 总线，其每一个连接到唯一的（或单个）上游端口；以及配置哪一上游端口与哪一下游端口通信的控制器。每一下游端口连接到每一内部 PCI 总线，且每一下游端口仅响应于一个内部 PCI 总线。并且，控制器通过与每一 CPU 复合体相关联的接口接收发现请求。

20 当结合附图阅读以下说明性实施例的详细描述时，可以清楚本发明的另外的特征和优点。

### 附图说明

25 当结合附图阅读时，可以更好地理解本发明的以上概述以及以下较佳实施例的详细描述。为说明本发明的目的，附图中示出了本发明的示例性构造；然而，本发明不限于所揭示的具体方法和手段。附图中：

图 1 所示是一常规个人计算机的框图；

图 2 所示是依照本发明使用可配置 PCI Express 开关共享组件的通用系统的框图；

30 图 3 所示是可配置 PCI Express 开关的框图；

图 4 所示是配置 PCI Express 开关的控制接口和命令逻辑的框图；

图 5 所示是依照本发明共享组件的示例性系统的框图；以及  
图 6-8 所示是使用可配置 PCI Express 开关的组件共享的若干实施例的框图。

### 具体实施方式

5 现在参考图 2，示出了用于共享组件的系统 200 的综述。当 PCIExpress 替代了 PCI，并且多个 CPU 变为计算设备中的一种标准实现时，标准系统组件的灵活配置将成为一种十分期望的特征。基于可用的硬件和应用程序要求动态地重新配置一组硬件资源是对于客户机台式 PC 的期望特征。本发明提供了如用户和应用程序所需要地配置系统配置的简单控制方法。然而，本发明不限于台式机设计，因为它  
10 可应用于采用 PCI Express 和类似的体系结构的服务器和其它计算设备。

图 2 示出了支持两个上游 CPU 拓朴的可配置开关设计，这两个拓朴被指定为 201 和 215。第一系统拓朴 201 被示出为典型的 PC 计算机，它可包括 CPU 202、图形卡 203、系统总线、存储器 204、芯片组（北桥 205 和南桥 206）、存储设备 207（例如，硬盘、闪存等）、通信设备 210（例如，MODEM、NIC 等）、以及  
15 连接到鼠标 210、键盘 211 和软盘驱动器 213 的超级 I/O 控制器 208。PCI Express 总线 214(1) 连接到可配置 PCI 开关 227。类似地，第二系统拓朴 215 包括 CPU 216、图形卡 217、系统总线、存储器 218、芯片组（北桥 219 和南桥 220）、存储设备 222、通信设备 221、以及连接到鼠标 224、键盘 225 和软盘驱动器 226 的超级 I/O 控制器 223。PCI Express 总线 214(2) 连接到 PCI 开关 227。PCI Express 开关连接到  
20 I/O 设备 228-230。

现在参考图 3 和 4，更详细地示出了可配置 PCI 开关 227。在图中，“u”表示上游端口；“P”表示 PCI 对 PCI (P2P)；“d”表示下游端口；“B0”、“B1”、“B2”表示与上游端口相关联的 PCI Express 内部 PCI 总线；“0”、“1”、“2”和“n”表示信号路径或端口。

25 如 PCI Express 规范中所定义的，PCI Express 开关被模型化为一组 PCI 对 PCI (P2P) 桥设备。上游 P2P 桥（连接到主机控制器或另一 PCI 总线）连接到公用的 PCI 总线，在该公用 PCI 上，该（内部）PCI 总线上找到的唯一设备是（下游）PCI 对 PCI 桥，它进而连接到输出上的 PCI 设备。因此，典型的 PCI Express 开关包括仅一个连接到 PCI/芯片组主机控制器的上游 P2P 桥、内部 PCI 总线以及一组下游  
30 P2P 桥。

本发明有利地实现了一组上游 PCI 对 PCI 桥，用于扩展 PCI Express 点对点体

系结构的扇出的目的。如图 3 所示,如由 uP0 231 和 uP1 233 所表示的 n 个上游 P2P 桥的每一个具有其自己的独立内部 PCI 总线,如由 B0 和 B1 所表示的,以及由 dP0 232、dP1 235 和 dPn 234 所表示的多个下游 P2P 桥。较佳的是,每一下游 P2P 桥连接到每一内部 PCI 总线。与常规的 PCI Express 开关不同,每一下游 P2P 桥是可配置的,以响应任一内部 PCI 总线 B0 或 B1 的枚举实行。

控制方法包括内部配置控制寄存器或外部硬件跨接或其它外部配置管理接口外部控制 242。控制方法定义了下游 P2P 桥 (232、234 和 235) 响应哪一总线。来自其它 PCI 总线的通信被忽略。例如,在加电序列的末端,任意方法向总线 B0 或 B1 分配资源 (I/O 和 dPx), 用于初始化配置的目的。由此,下游端口 (dPx) 响应于来自任一内部总线 B0 或 B1 的周期,但不是两者。存在物理连接,但是响应仅对总线 B0 或总线 B1 上的周期发生。

图 4 是桥控制逻辑 236 的详细图示,它具有用于 PCI 枚举和发现的其相关联的外部总线和配置接口 237,用于总线 0 (238) 的其内部 PCI 总线配置接口以及用于总线 1 (239) 的内部 PCI 总线配置接口。在设备枚举和配置过程中,运行在 CPU0 202 上的操作系统通过读取 PCI 配置空间寄存器内容发现设备。CPU0 202 将发现 uP0 231 中找到的 PCI 对 PCI 桥。操作系统将枚举在总线 B0 上找到的设备,并发现与开关内部总线 B0 相关联的控制接口 Ifc\_B0 238。该设备具有唯一的桥标识号,它将其标识为可配置的 PCI Express 开关。因此,接口 Ifc\_B0 238 与开关的内部总线 B0 相关联。它可以是 B0 上的配置和 I/O 周期的主设备或目标。CPU0 202 然后将枚举总线 B0 上发现的所有设备。

当完成之后,CPU0 202 将通过 Ifc\_B0 238 接口向开关控制器启动一发现请求。控制器然后可启动配置请求,并读取总线 B1 上每一设备的配置空间,或通过 Ifc\_B1 239 向 CPU 216 启动一请求,请求在总线 B1 上枚举的设备。在收集了 CPU0 202 请求的信息之后,开关控制器将通过 Ifc\_B0 238 向 CPU0 202 启动一响应,并返回所请求的信息。因此,该机制使 CPU0 202 和 CPU1 216 都能够确定在请求时哪些设备可用。

由外部控制 237 提供的外部控制接口使得以管理容量执行的总线管理器能够向 CPU0 202 或 CPU1 216 分配下游资源 (I/O)。外部控制 237 通过向桥控制器逻辑 236 询问有哪些设备可从配置开关的内部总线 B0 和 B1 得到来执行这一功能。当基于 CPU/ 操作系统责任来分配资源,并且当向每一上游服务器实体分配任务时,这一特征在服务器体系结构中是特别期望的。

当 CPU0 202 期望分配给 CPU1 216 的资源时，它将通过 Ifc-B0 238 启动一对当前下游（dPx）端口或端点（I/O）的请求。桥控制器逻辑 236 然后将启动一对 CPU1 216 的请求，以释放下游端口。如果请求被准许，则 CPU1 216 将默许该端点，并通过 Ifc\_B1 239 启动对桥控制器逻辑 236 的准许，以释放下游端口（dPx）。  
5 桥控制器逻辑 236 然后将指令下游端口（dPx）通过开关端口控制接口执行从 B1 开始的 PCI Express 断开序列。当断开时，桥控制器逻辑 236 将指令下游端口（dPx）通过开关端口控制接口执行到 B0 的连接序列。当连接时，CPU0 202 将接收一热插拔事件，如 PCI 体系结构规范中所定义的。当被通知该事件时，CPU0 202 将枚举设备并加载与其相关联的适当驱动程序，以完成转移。

10 如果 CPU1 216 拒绝准许请求，则 CPU1 216 通过 Ifc\_B1 239 向 CPU2 202 启动一消息，通知被拒绝的请求的始发者。桥控制器逻辑 236 通过其接口 Ifc\_B0 238 经由 B0 向 CPU0 202 启动一对 CPU0 202 的响应，由此完成拒绝序列。

现在参考图 5，示出了一个示例，其中，对接的膝上 PC（系统 201）和增强的对接站（系统 215）都共享通过可配置开关关联的资源。当应用程序被加载到膝上 PC，并且用户期望使用当前由增强对接站拓扑配置的扫描仪来获取照片时，膝上 PC 将请求扫描仪的所有者。当用户期望打印通过扫描仪获取的所扫描且操纵的照片时，膝上 PC 拓扑将请求与增强对接站相关联的照片质量打印机的所有者。当膝上 PC 脱离对接时，通过可配置开关与膝上 PC 相关联的所有资源被解除关联，并且该可配置开关然后将向增强对接站重新分配资源，以在该拓扑内使用。  
15

20 图 6 示出了具有多个 PCI Express 总线的 CPU 复合体如何可通过多个可配置 PCI Express 开关来配置以使用 I/O 设备。在此示例中，CPU1 216 与开关 SW0 227(1) 和 SW1 227(2)接口。任何连接到 SW0 和 SW1 的 I/O 设备然后可被分配到 CPU1 216。在此配置中，仅 SW0 内的下游 P2P 桥可被分配给 CPU0 202，并且仅 SW1 内的下游 P2P 桥可被分配给 CPU2 214。

25 图 7 是图 6 的修改。并非在多个开关之间共享带宽，图 7 相反示出了使用可从 CPU1 216 对两个开关可获得的全带宽。

图 8 是能够访问两个开关 SW0 和 SW1 内的所有资源的三个 CPU 复合体 202、216 和 241 的又一示例。尽管图 8 是开关可伸缩性的另一示例，然而它在示出添加上游 CPU 复合体的相关联的内部开关复合体中也是有用的。此外，尽管看似这一  
30 多个上游 P2P 桥的实现消耗了下游桥，然而这并不是下游 P2P 桥比上游桥更容易添加到设计的情况。

尽管结合各附图的较佳实施例描述了本发明，然而可以理解，在不脱离本发明的的情况下，可以使用其它相似的实施例，或可以向描述的实施例作出修改和添加，以执行本发明的相同功能。例如，本领域的技术人员将认识到，本申请中描述的本发明可应用于任一计算设备或环境，不论是有线还是无线的，并且可应用于通过通信网络连接并通过网络交互的任意数量的这类计算设备。此外，应当强调，考虑各种计算机平台，包括手持式设备操作系统和其它应用专用操作系统，尤其是当无线联网设备的数量持续增长的时候。再者，本发明可以在多个处理芯片或设备内或跨多个处理芯片或设备实现，并且存储可以类似地跨多个设备实现。因此，本发明不应当限于任何单个实施例，而是相反，应当依照所附权利要求书的宽度和范围来解释。

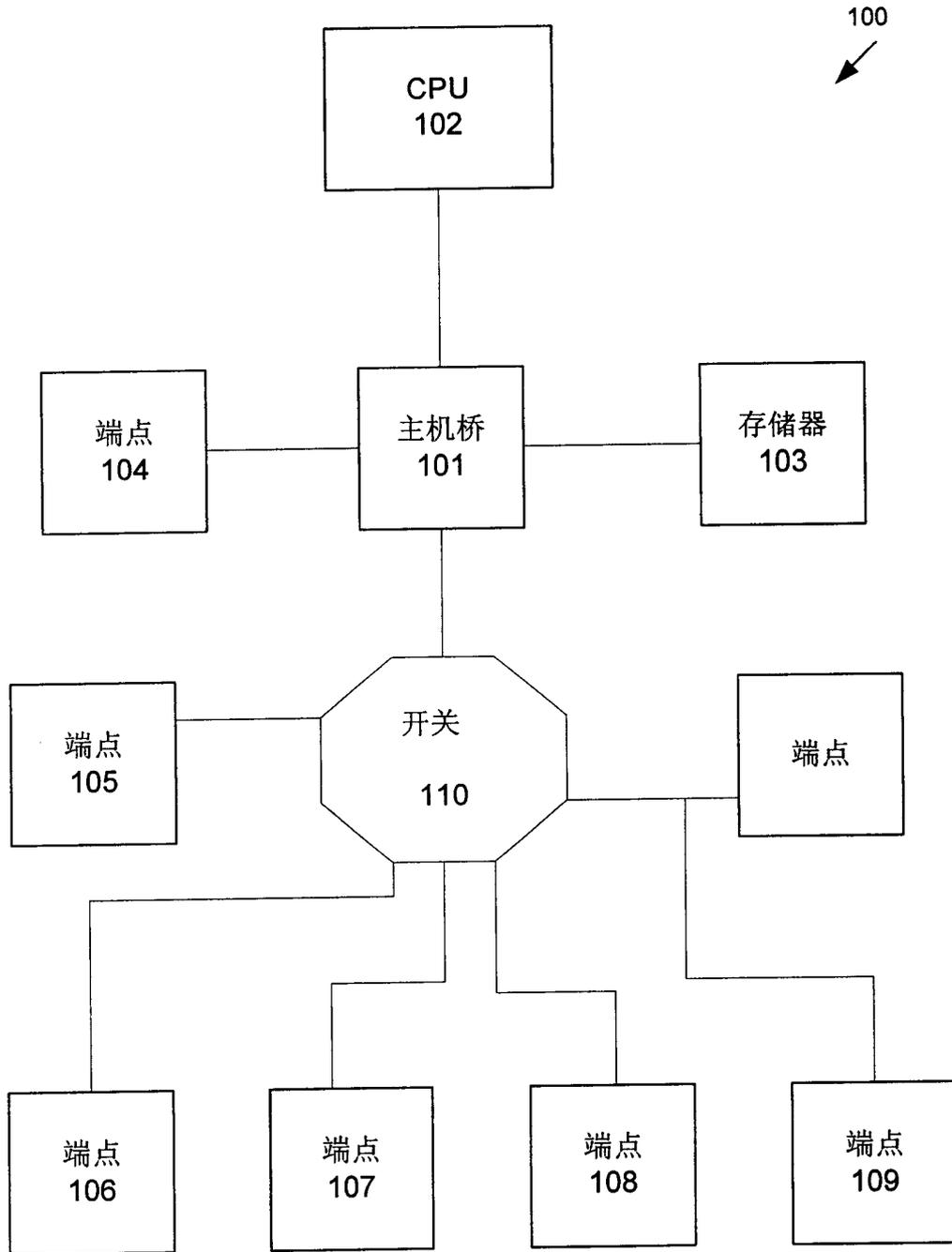


图 1

现有技术

200

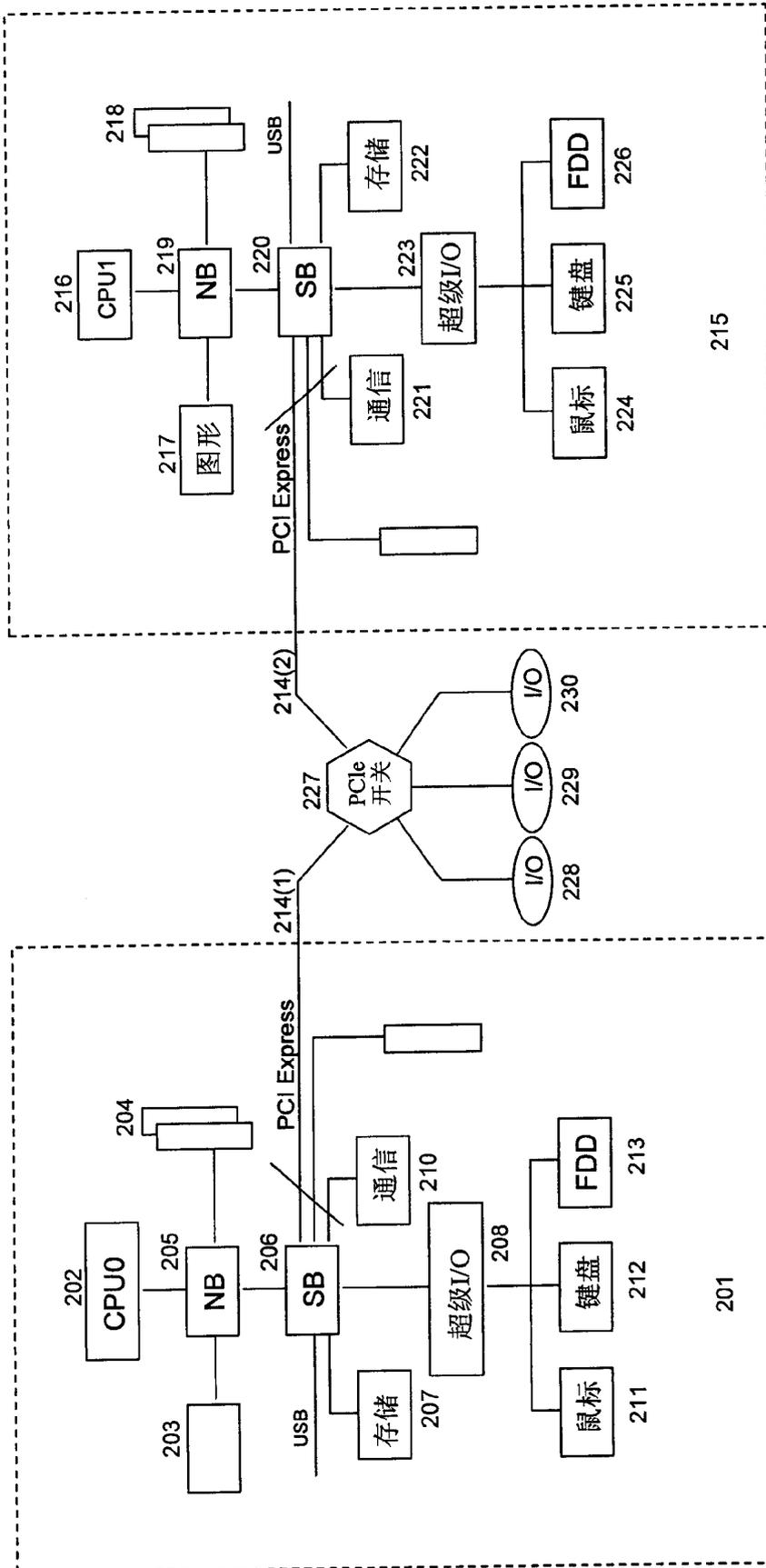


图 2

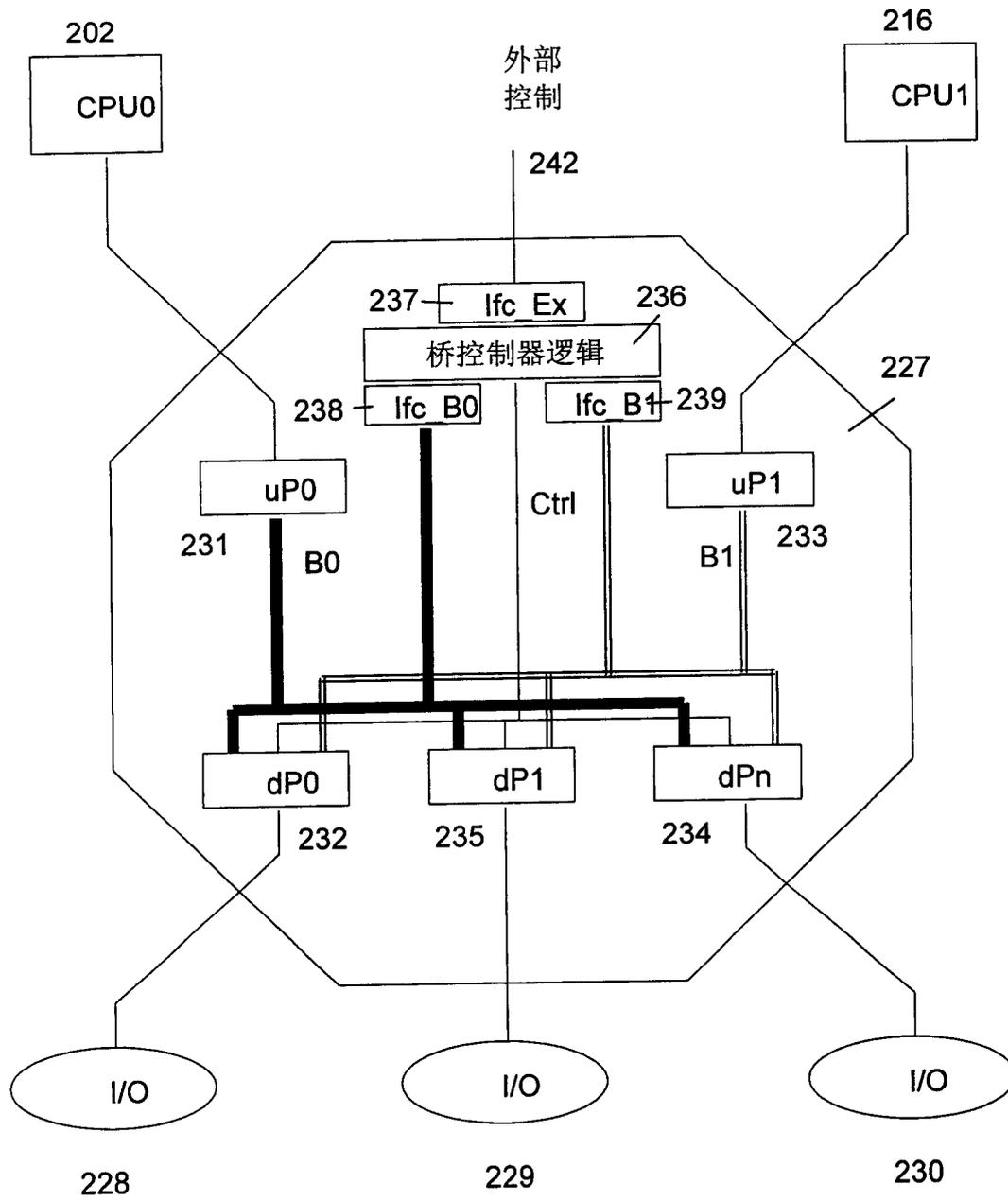


图 3

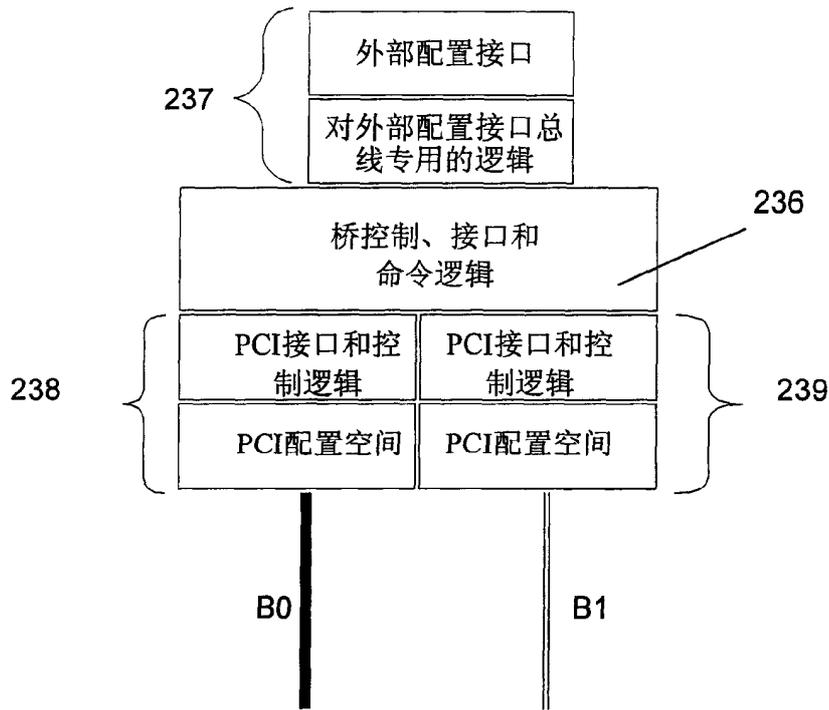


图 4

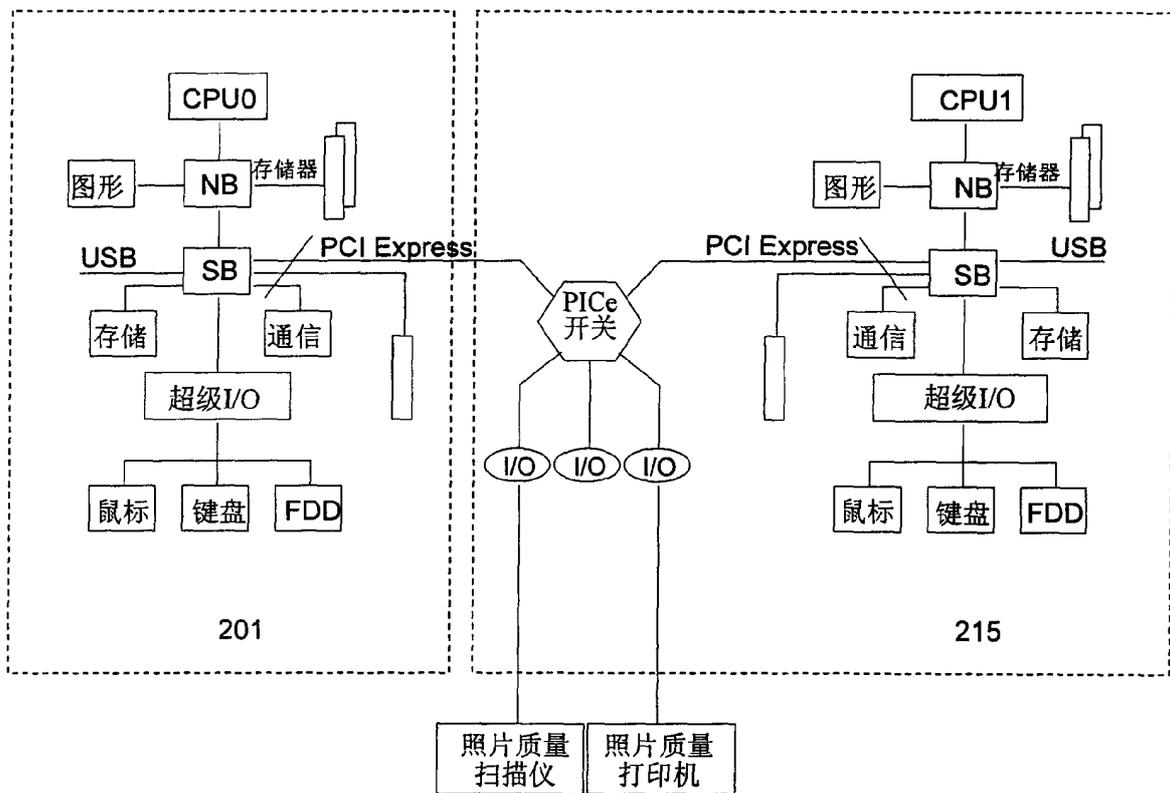


图 5

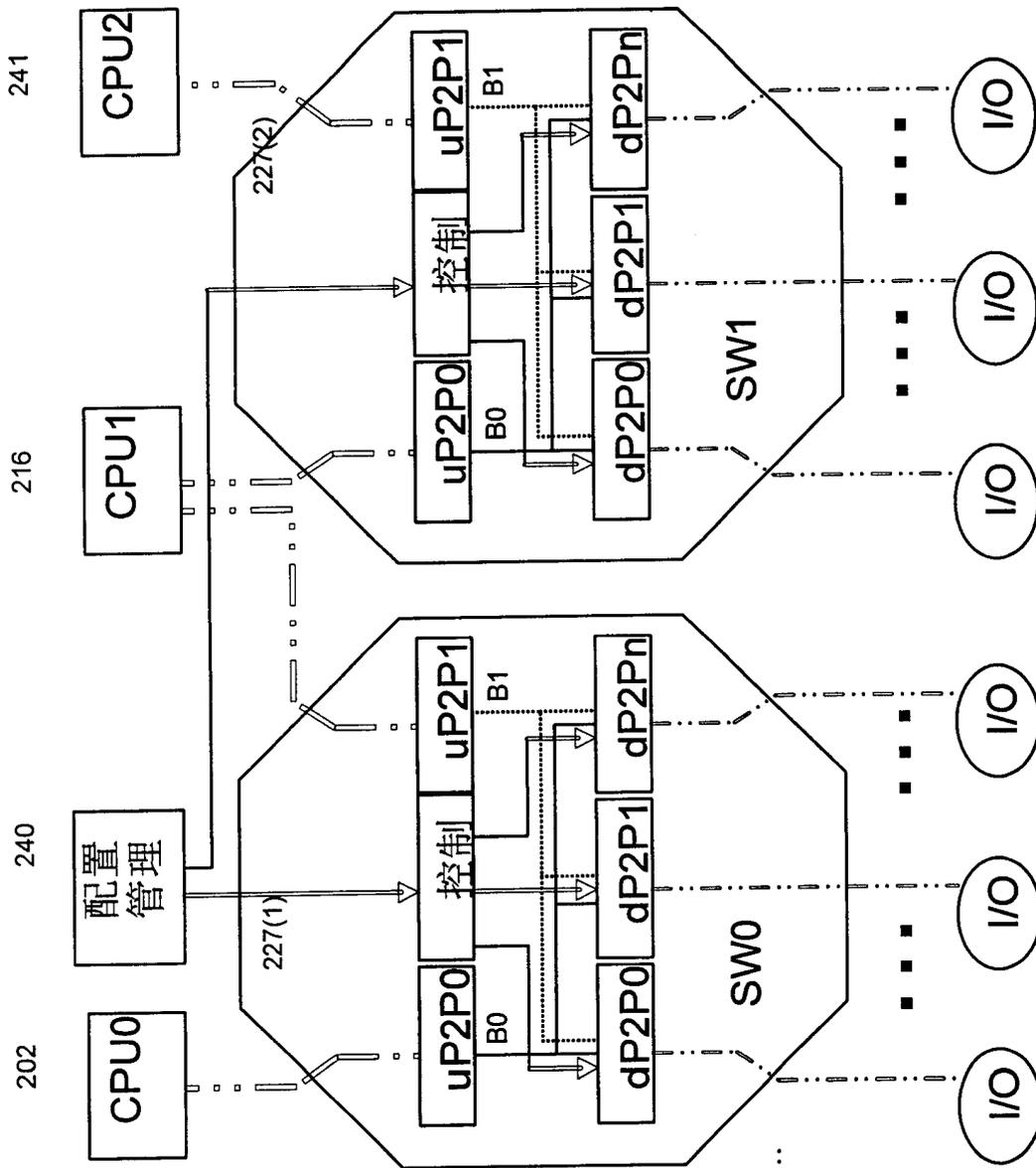


图 6

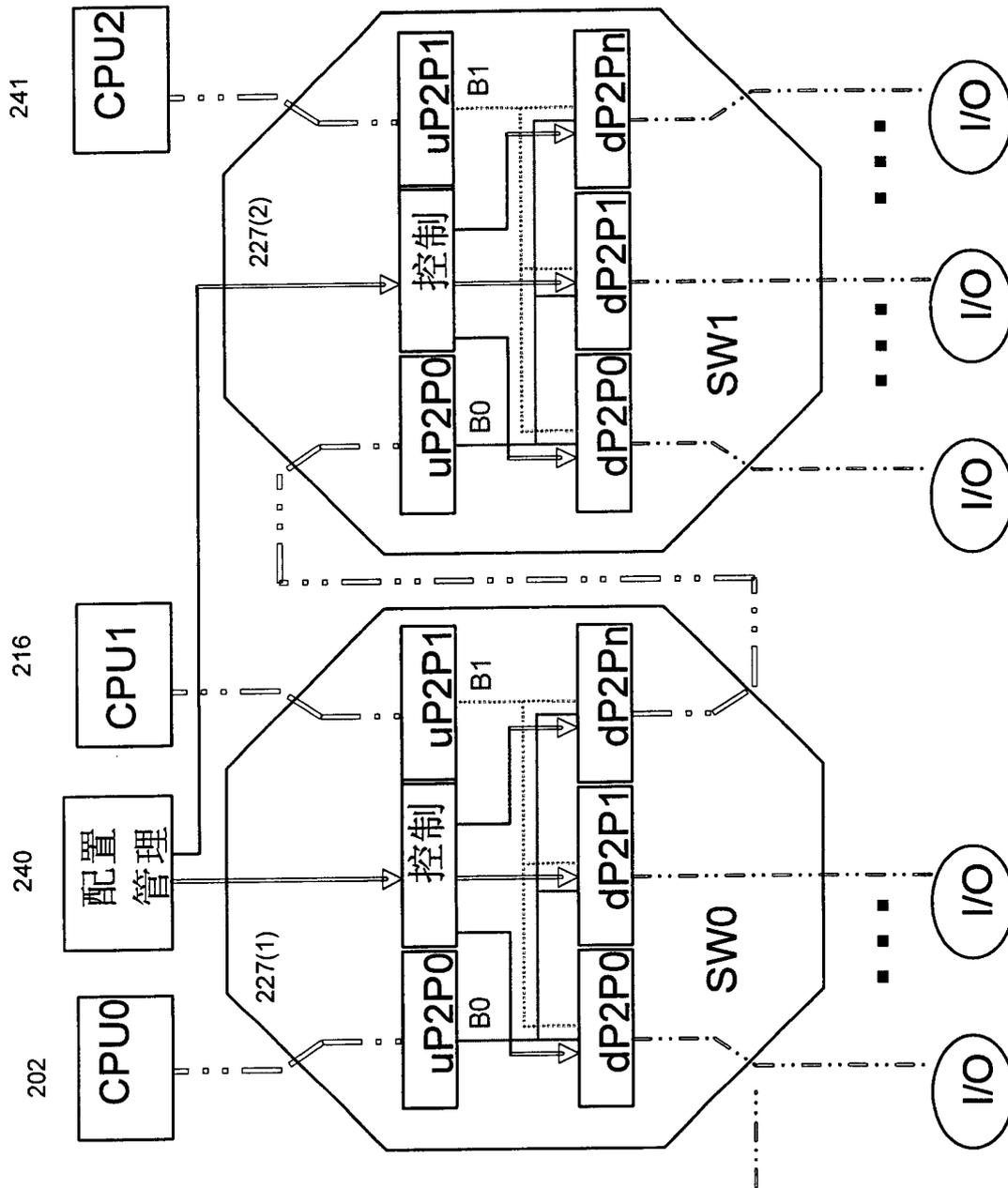


图 7

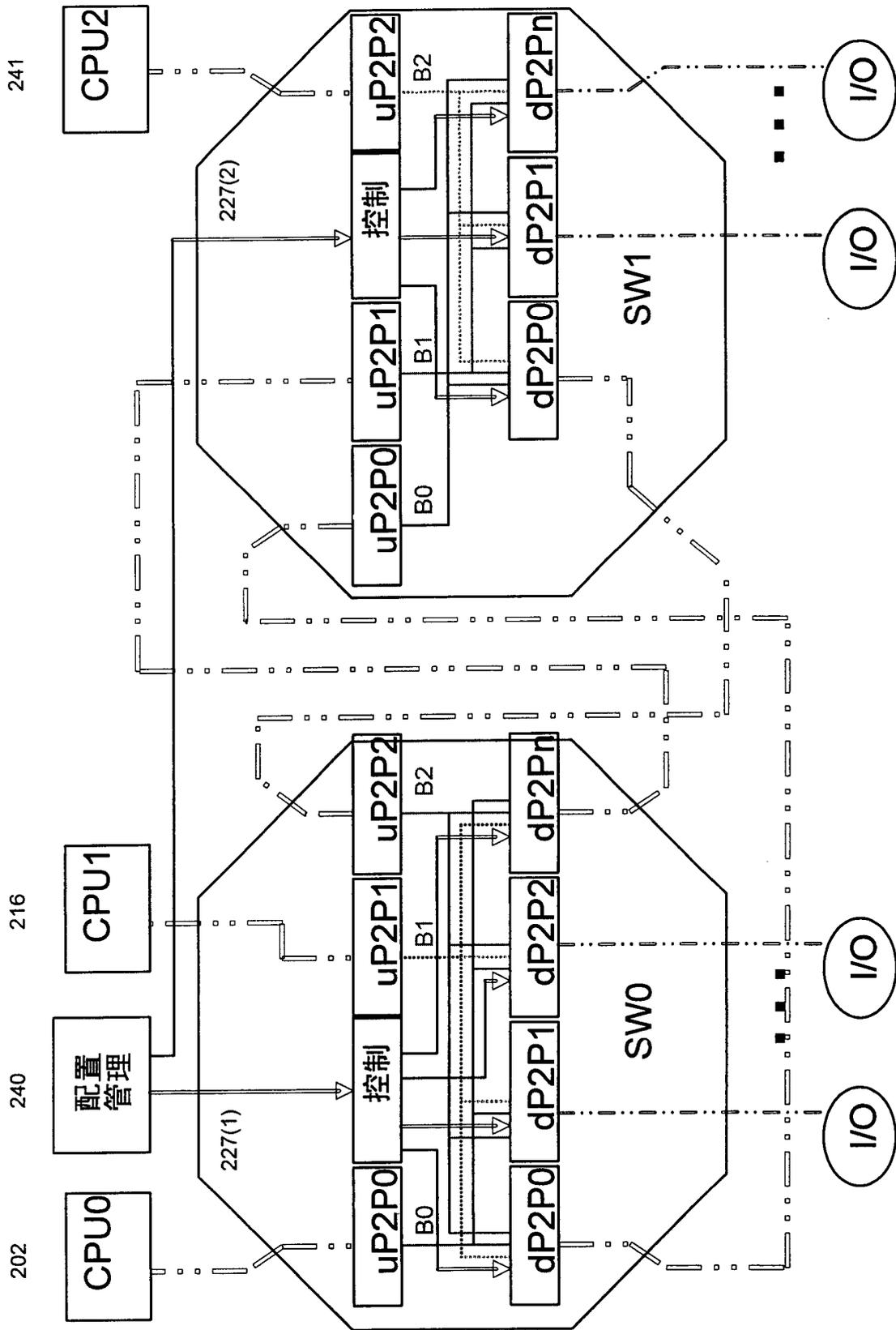


图 8