

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 9/445 (2006.01)



[12] 发明专利申请公布说明书

[21] 申请号 200480039566.X

[43] 公开日 2007 年 1 月 24 日

[11] 公开号 CN 1902584A

[22] 申请日 2004. 12. 1

[21] 申请号 200480039566.X

[30] 优先权

[32] 2004. 1. 7 [33] US [31] 10/752,632

[86] 国际申请 PCT/EP2004/053205 2004. 12. 1

[87] 国际公布 WO2005/069132 英 2005. 7. 28

[85] 进入国家阶段日期 2006. 6. 30

[71] 申请人 国际商业机器公司

地址 美国纽约

[72] 发明人 许玉成(音译)

约汉·诺伯特·迈考利

宋成春(音译)

威廉·格里沃德·舍曼

[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 康建忠

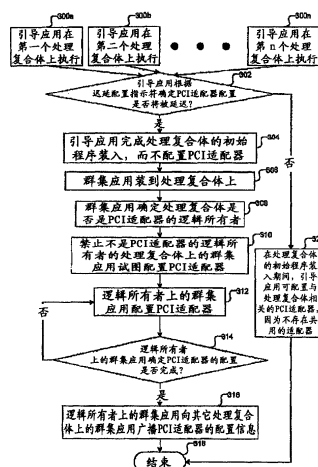
权利要求书 4 页 说明书 8 页 附图 6 页

[54] 发明名称

配置共用资源

[57] 摘要

提供一种方法、系统和制造产品，其中在一些实施例中，确定与资源相关的配置指示符是否指示资源的延迟配置，其中资源由多个处理复合体通过总线共用，其中如果指示了资源的延迟配置，那么在多个处理复合体的初始程序装入期间，阻止配置资源。响应确定与资源相关的配置指示符指示资源的延迟配置，资源仅由共用资源的多个处理复合体中的一个处理复合体配置。



1、一种方法，包括：确定与资源(102)相关的配置指示符(200)是否指示资源的延迟配置，其中所述资源由多个处理复合体(100a,100b,100n)通过总线(104)共用，其中如果指示了所述资源的延迟配置，那么在所述多个处理复合体的初始程序装入期间，阻止配置所述资源；和响应确定与所述资源相关的配置指示符指示资源的延迟配置，仅由共用所述资源的所述多个处理复合体中的一个处理复合体配置所述资源。

2、按照权利要求1所述的方法，其中在处理复合体的初始程序装入期间，执行与资源相关的配置指示符是否指示资源的延迟配置确定，其中所述方法还包括：响应确定与所述资源相关的配置指示符指示资源的延迟配置，完成共用所述资源的所述多个处理复合体的初始程序装入，其中在完成共用所述资源的所述多个处理复合体的初始程序装入之后，仅由所述多个处理复合体中的一个处理复合体配置所述资源。

3、按照权利要求1所述的方法，其中所述多个处理复合体中的所述仅有的一个处理复合体是所述资源的逻辑所有者，其中所述方法还包括：所述逻辑所有者向所述多个处理复合体中的其它处理复合体广播被配置资源的配置信息。

4、按照权利要求1所述的方法，还包括：确定所述多个处理复合体中配置所述资源的仅有的一个处理复合体的故障，其中所述多个处理复合体中的所述仅有的一个处理复合体是所述资源的逻辑所有者；和响应确定所述故障，从所述多个处理复合体中确定所述资源的新的逻辑所有者，其中所述新的逻辑所有者负责资源的后续配置。

5、按照权利要求1所述的方法，其中所述多个处理复合体中的所述仅有的一个处理复合体是资源的初始逻辑所有者，其中所述方法还包括：响应初始逻辑所有者的故障，从所述多个处理复合体中确定所述资源的新的逻辑所有者；确定初始逻辑所有者已从故障状态恢复；响应确定初始逻辑所有者已从故障状态恢复，把所述资源的逻辑所有权从新的逻辑所有者转移给初始逻辑所有者。

6、按照权利要求 1 所述的方法，其中在所述多个处理复合体的初始程序装入期间，执行与资源相关的配置指示符是否指示资源的延迟配置的确定，其中所述方法还包括：响应确定与资源相关的配置指示符指示资源的延迟配置，延迟资源的配置，直到对所述多个处理复合体完成了初始程序装入为止。

7、按照权利要求 1 所述的方法，其中所述多个处理复合体构成多集群系统，其中所述多个处理复合体由主机通过为所述多个处理复合体共用的配置资源访问。

8、按照权利要求 1 所述的方法，其中由在所述多个处理复合体中的第一处理复合体中实现的引导应用执行与资源相关的配置指示符是否指示资源的延迟配置的确定，其中仅由共用资源的多个处理复合体中的一个处理复合体配置资源的操作由在所述多个处理复合体中的所述仅有的一个处理复合体中实现的群集应用执行。

9、按照权利要求 1 所述的方法，其中配置所述资源由跨越所述多个处理复合体的一个群集应用协调。

10、按照权利要求 1 所述的方法，其中共用资源是 PCI 适配器，总线是 PCI 总线，并且配置指示符在 PCI 适配器中实现。

11、一种系统，包括：多个处理复合体；与所述多个处理复合体耦接的总线；由所述多个处理复合体通过所述总线共用的资源；与所述资源相关的配置指示符；确定与资源相关的配置指示符是否指示资源的延迟配置的装置，其中如果指示了所述资源的延迟配置，那么在所述多个处理复合体的初始程序装入期间，阻止配置所述资源；和响应确定配置指示符指示资源的延迟配置，由所述多个处理复合体中的仅有的一个处理复合体配置所述资源的装置。

12、按照权利要求 11 所述的系统，其中确定与资源相关的配置指示符是否指示资源的延迟配置的装置在处理复合体的初始程序装入期间执行所述确定，其中所述系统还包括：响应确定与所述资源相关的配置指示符指示资源的延迟配置，完成共用所述资源的所述多个处理复合体的初始程序装入的装置，其中在完成共用所述资源的所述多个处理复合体

的初始程序装入之后，仅由所述多个处理复合体中的一个处理复合体配置所述资源。

13、按照权利要求 11 所述的系统，资源的逻辑所有者，其中所述多个处理复合体中的所述仅有的一个处理复合体是所述资源的逻辑所有者；和由所述逻辑所有者向所述多个处理复合体中的其它处理复合体广播被配置资源的配置信息的装置。

14、按照权利要求 11 所述的系统，还包括：确定所述多个处理复合体中配置所述资源的仅有的一个处理复合体的故障的装置，其中所述多个处理复合体中的所述仅有的一个处理复合体是所述资源的逻辑所有者；和响应确定所述故障，从所述多个处理复合体中确定所述资源的新的逻辑所有者，其中所述新的逻辑所有者负责资源的后续配置。

15、按照权利要求 11 所述的系统，其中所述多个处理复合体中的所述仅有的一个处理复合体是资源的初始逻辑所有者，其中所述系统还包括：响应初始逻辑所有者的故障，从所述多个处理复合体中确定所述资源的新的逻辑所有者的装置；确定初始逻辑所有者已从故障状态恢复的装置；和响应确定初始逻辑所有者已从故障状态恢复，把所述资源的逻辑所有权从新的逻辑所有者转移给初始逻辑所有者的装置。

16、按照权利要求 11 所述的系统，其中在所述多个处理复合体的初始程序装入期间，执行与资源相关的配置指示符是否指示资源的延迟配置的确定，其中所述系统还包括：响应确定与资源相关的配置指示符指示资源的延迟配置，延迟资源的配置直到对所述多个处理复合体完成了初始程序装入为止的装置。

17、按照权利要求 11 所述的系统，其中所述多个处理复合体构成多集群系统，其中所述多个处理复合体由主机通过为所述多个处理复合体共用的配置资源访问。

18、按照权利要求 11 所述的系统，还包括：在所述多个处理复合体中的仅有的一个处理复合体中实现的群集应用；和在所述多个处理复合体中的第一处理复合体中实现的引导应用，其中由所述引导应用确定与资源相关的配置指示符是否指示资源的延迟配置，其中由共用资源的多

个处理复合体中的一个处理复合体配置资源的操作由所述群集应用完成。

19、按照权利要求 11 所述的系统，还包括跨越所述多个处理复合体的一个群集应用，其中配置所述资源由所述群集应用协调。

20、按照权利要求 11 所述的系统，其中共用资源是 PCI 适配器，总线是 PCI 总线，并且配置指示符在 PCI 适配器中实现。

21、一种保存在计算机可读存储介质上的计算机程序产品，当在计算机系统上运行时，所述计算机程序产品指令计算机系统执行按照任何前述方法权利要求所述的方法。

配置共用资源

技术领域

本发明的公开涉及配置共用资源的方法、系统和制造产品。

背景技术

多集群系统可把多个计算节点耦合在一起。所述多个耦合的计算节点可共同作为单一计算系统工作。多集群系统可被用于并行处理、负载均衡、故障容错等，并实现高可用性系统或者冗余信息技术系统。例如，某些多集群系统可把相同的数据保存在多个计算节点中，一个计算节点可以是一个计算单元、存储单元等。当多集群系统的一个计算节点不可用时，多集群系统的另外的计算节点可被用于代替不可用的计算节点。

诸如外设组件互连(PCI)总线之类的资源可被用于互连设备与处理器和主存储器的本地总线。在某些多集群系统中，多个计算节点通过 PCI 总线可共用 PCI 适配器。主机可通过共用 PCI 适配器与多集群系统的计算节点通信。

如果 PCI 适配器在多集群系统的多个计算节点间被共用，那么即使一个计算节点发生故障，仍然允许主机继续与多集群系统的其它计算节点通信。这样的主机能够访问共用的 PCI 适配器，和访问与未发生故障的计算节点相关的数据。

如果一个以上的计算节点被允许配置共用资源，那么共用资源可能处于出错状态，并且可能不能在计算节点间被共用。

发明内容

本发明提供如独立权利要求中所述的方法、系统和计算机程序产品。

附图说明

现在参见附图，其中相同的附图标记表示对应的部分：

图 1 根据本发明的实施例，图解说明第一计算环境的方框图；

图 2 根据本发明的实施例，图解说明与共用的 PCI 适配器相关的数据结构的方框图；

图 3 根据本发明的实施例，图解说明配置共用 PCI 适配器的逻辑；

图 4 根据本发明的实施例，图解说明转移共用 PCI 适配器的逻辑所有权的逻辑；

图 5 根据本发明的实施例，图解说明第二计算环境的方框图；

图 6 图解说明其中实现本发明的某些方面的计算机体系结构的方框图。

具体实施方式

下面参考构成说明书的一部分，并且图解说明几个实施例的附图进行说明。应当理解可以使用其它实施例，并且在不脱离当前实施例的范围的情况下，可做出结构变化和操作变化。

图 1 根据本发明的一些实施例，图解说明第一计算环境的方框图。

在一些实施例中，多个处理复合体 (complex) 100a...100n 通过 PCI 总线 104 与 PCI 适配器 102 耦接。处理复合体，比如处理复合体 100a 可包括一个或多个计算节点，所述计算节点可包括单处理器或多处理器系统。在一些实施例中，处理复合体 100a...100n 可包括个人计算机，工作站，服务器，大型机，手持式计算机，掌上计算机，电话设备，网络设备，刀片计算机，存储控制器等。

在一些实施例中，PCI 适配器 102 可由 PCI-X 适配器，或者任何 PCI 或 PCI-X 总线兼容装置代替。此外，在备选实施例中，PCI 总线 104 可由 PCI-X 总线或者其它一些总线代替。

所述多个处理复合体 100a...100n 包括多个引导应用 (application) 106a...106n 和多个群集应用 (application) 108a...108n。例如，在一些实施例中，处理复合体 100a 可包括引导应用 106a 和群集应用 108a，处理复合体 100b 可包括引导应用 106b 和群集应用 108b，处理复合体 100n

可包括引导应用 106n 和群集应用 108n。引导应用 106a...106n 和群集应用 108a...108n 可用软件、固件或硬件，或者它们的任意组合来实现。

当被执行时，引导应用 106a...106n 可执行对应处理复合体 100a...100n 的初始程序装入。例如，引导应用 106a 可执行处理复合体 100a 的初始程序装入，引导应用 106b 可执行处理复合体 100b 的初始程序装入，引导应用 106n 可执行处理复合体 100n 的初始程序装入。在处理复合体的初始程序装入期间，处理复合体的操作系统(未示出)、设备驱动器(未示出)等可被装入，使得在完成初始程序装入之后，处理复合体准备好执行其它应用。

当被执行时，群集应用 108a...108n 允许处理节点 100a...100n 共用 PCI 适配器 102。在一些实施例中，只有一个群集应用可配置 PCI 适配器 102，和向其它群集应用广播 PCI 适配器 102 的配置信息。例如，在一些实施例中，群集应用 108a 可配置 PCI 适配器 102，和向可在处理复合体 100a...100n 内执行的 10 群集应用 108b...108n 广播 PCI 适配器 102 的配置信息。虽然示出了多个群集应用 108a...108n，不过在一些实施例中，所述多个群集应用 108a...108n 可以是与处理复合体 100a...100n 相关的分布式群集应用的一部分。

图 2 根据本发明的一些实施例，图解说明与 PCI 适配器 102 相关的数据结构的方框图。在一些实施例中，迟延配置指示符 200 可以与 PCI 适配器 102 相关联。在一些实施例中，迟延配置指示符 200 可以表示单比特数据，所述单比特数据指示 PCI 适配器 102 的配置是否应被延迟，直到处理复合体 100a...100n 的初始程序装入完成为止。

例如，在一些实施例中，如果迟延配置指示符 200 中的单比特数据被设置为 1，那么在处理复合体 100a...100n 的初始程序装入期间，引导应用 106a...106n 不配置 PCI 适配器 102。在一些实施例中，群集应用 108a...108n 中只有一个群集应用，例如群集应用 108a 可配置 PCI 适配器。只有在完成处理复合体 100a...100n 的初始程序装入之后，群集应用 108a...108n 才可开始执行。

图 3 根据本发明的实施例，图解说明配置在处理复合体 100a...100n

中实现的共用 PCI 适配器的逻辑。

控制始于块 300a...300n, 引导应用 106a...106n 在对应的处理复合体 100a...100n 中执行。例如, 引导应用 106a 在处理复合体 100a 中执行(在块 300a), 引导应用 106b 在处理复合体 100b 中执行(在块 300b), 引导应用 106n 在处理复合体 100n 中执行(在块 300n)。处理复合体 100a...100n 中引导应用 106a...106n 的执行可以是并行的。通过引导应用 106a...106n 的执行, 在处理复合体 100a...100n 中开始初始程序装入。

在一些实施例中, 每个引导应用 106a...106n 可根据迟延配置指示符 200 确定(在块 302)PCI 适配器 102 的配置是否将被延迟, 直到处理复合体 100a...100n 的初始程序装入完成为止。如果是, 那么引导应用 106a...106n(在块 304)执行处理复合体 100a...100n 的初始程序装入, 而不配置 PCI 适配器 102。

群集应用 108a...108n 装入(在块 306)对应的处理复合体 100a...100n 中。群集应用 108a...108n 可共同确定(在块 308)为 PCI 适配器 102 的逻辑所有者的一个处理复合体, 作为所述逻辑所有者的处理复合体被包括在所述多个处理复合体 100a...100n 中。例如, 在一些实施例中, 处理复合体 100a 可被确定为 PCI 适配器 102 的逻辑所有者。作为 PCI 适配器 102 的逻辑所有者的处理复合体承担配置 PCI 适配器 102 和向其它处理复合体广播 PCI 适配器 102 的配置信息的职责。

群集应用 108a...108n 禁止(在块 310)不是 PCI 适配器 102 的逻辑所有者的那些处理复合体试图配置 PCI 适配器 102。逻辑所有者中的群集应用配置(在块 312)PCI 适配器 102。例如, 在一些实施例中, 如果处理复合体 100a 是逻辑所有者, 那么群集应用 108a 可配置 PCI 适配器 102。

逻辑所有者中的群集应用确定(在块 314)PCI 适配器 102 的配置是否完成。如果是, 那么逻辑所有者中的群集应用向其它处理复合体的群集应用广播(在块 316)PCI 适配器 102 的配置信息。例如, 如果处理复合体 100a 被确定为 PCI 适配器 102 的逻辑所有者, 那么群集应用 108a 把 PCI 适配器的配置信息分发给群集应用 108b...108n。响应 PCI 适配器 102 的配置信息的广播的完成, 该进程结束(在块 318)。

如果逻辑所有者中的群集应用确定(在块 314)PCI 适配器 102 的配置未完成,那么逻辑所有者中的群集应用继续配置(在块 312) PCI 适配器 102。

如果引导应用 106a...106n 根据迟延配置指示符 200 确定(在块 302) PCI 适配器 102 的配置将不被延迟,那么在处理复合体 100a...100n 的初始程序装入期间,引导应用 106a...106n 可配置(在块 320)与处理复合体 100a...100n 相关的一个或多个 PCI 适配器,因为在处理复合体 100a...100n 中可能不存在任何共用 PCI 适配器。在配置之后,该进程结束(在块 318)。

因此,图 3 的逻辑图解说明了一些实施例,其中共用 PCI 适配器 102 的配置被延迟,直到处理复合体 100a...100n 的初始程序装入完成为止。响应处理复合体 100a...100n 的初始程序装入的完成,所述多个处理复合体 100a...100n 中只有一个可配置共用 PCI 适配器 102,和向其它处理复合体广播共用 PCI 适配器 102 的配置信息。

图 4 根据本发明的实施例,图解说明转移共用 PCI 适配器 102 的逻辑所有权的逻辑。在一些实施例中,转移共用 PCI 适配器的逻辑所有权的逻辑可在群集应用 108a...108n 中实现,所述群集应用 108a...108n 在处理复合体 100a...100n 中执行。

控制开始于块 400,在块 400,作为 PCI 适配器 102 的逻辑所有者的处理复合体发生故障。所述故障可能是由于因软件、硬件或固件错误,逻辑所有者不正确工作的结果。也可能由于其它原因,发生逻辑所有者的故障。

不是逻辑所有者的处理复合体中的群集应用确定(在块 402) PCI 适配器 102 的新的逻辑所有者。例如,在一些实施例中,如果发生故障的逻辑所有者是处理复合体 100a,那么群集应用 108b...108n 从处理复合体 100b...100n 中确定 PCI 适配器 102 的新的逻辑所有者。

新的逻辑所有者中的群集应用承担(在块 404)配置或重新配置 PCI 适配器 102 和向其它处理复合体中的群集应用广播 PCI 适配器 102 的配置信息的职责。例如,如果处理复合体 100a 已发生故障,新的逻辑所有者

是 100b, 那么新的逻辑所有者 100b 可承担配置或重新配置 PCI 适配器 102 和向处理复合体 100c...100n 中的群集应用 108c...108n 广播 PCI 适配器 102 的配置信息的职责, 处理复合体 100a、100b、100c...100n 共用 PCI 适配器 102。

在一些实施例中, 新的逻辑所有者中的群集应用确定(在块 406)发生故障的处理复合体是否已变得可使用, 并且重新加入处理复合体中, 这里所述处理复合体形成一个多集群环境。如果是, 那么在一些实施例中, 新的逻辑所有者中的群集应用可把共用 PCI 适配器 102 的逻辑所有权转移(在块 408)给以前的逻辑所有者。例如, 新的逻辑所有者 100b 可选择地把共用 PCI 适配器 102 的逻辑所有权转回给初始的逻辑所有者 100a, 如果初始的逻辑所有者 100a 不再是故障的处理复合体。如果新的逻辑所有者中的群集应用确定(在块 406)发生故障的处理复合体还未变得可用, 那么新的逻辑所有者继续(在块 404)承担配置共用 PCI 适配器 102 和广播配置信息的职责。

因此, 图 4 图解说明了一些实施例, 其中在共用 PCI 适配器 102 的逻辑所有者发生故障的情况下, 其它处理复合体确定新的逻辑所有者, 所述新的逻辑所有者承担配置 PCI 适配器 102 和向其它起作用的处理复合体广播 PCI 适配器 102 的配置信息的职责。在初始的逻辑所有者从故障状态恢复的情况下, 在一些实施例中, 新的逻辑所有者可把 PCI 适配器的逻辑所有权转回给初始的逻辑所有者。

图 5 根据描述的本发明的一些实施例, 图解说明第二计算环境 500 的方框图。在第二计算环境 500 中, 多个处理复合体 100a...100n 可构成一个多集群系统 502, 其中所述多个处理复合体 100a...100n 中的处理复合体是多集群系统 502 的节点。

多集群系统 502 中的处理复合体 100a...100n 可通过总线 506 与共用资源 504 耦合。在一些实施例中, 共用资源 504 可包括 PCI 适配器 102, 总线 506 可以是 PCI 总线。在一些实施例中, 除 PCI 适配器 102 外的其它资源可构成共用资源 504。共用资源 504 可包括指示共用资源 504 的配置是否应被延迟直到已完成多集群系统 502 的处理复合体 100a...100n 的

初始程序装入为止的迟延配置指示符 508。在完成多集群系统 502 的处理复合体 100a...100n 的初始程序装入之后，处理复合体 100a...100n 中只有一个可配置共用资源 504 和向其它处理复合体广播共用资源的配置信息。

与多集群系统 502 耦接的主机 510 可通过共用资源 504 访问处理复合体 100a...100n 中的数据。即使一个或多个处理复合体 100a...100n 发生故障，主机 510 通过共用资源 504 仍可访问其它处理复合体。

指示共用 PCI 适配器的配置是否应被延迟直到已完成处理复合体的初始程序装入为止的属性与共用 PCI 适配器相关联。在多个处理复合体中实现的群集应用可协调所述多个处理复合体，使得所述多个处理复合体中只有一个处理复合体配置共用 PCI 适配器，和向其它处理复合体广播共用 PCI 适配器的配置信息。如果所述多个处理复合体中的一个以上的处理复合体将配置共用 PCI 适配器，那么共用 PCI 适配器可能处于错误状态，并且不能在多个处理复合体间共用。该实施例通过只由多个处理复合体中的一个处理复合体配置 PCI 适配器，使 PCI 适配器可在多个处理适配器间被共用。

另外的实现细节

通过利用标准的编程和/或工程技术来产生软件、固件、硬件或它们的任意组合，上面描述的技术可被实现成方法、设备或制造产品。这里使用的术语“制造产品”指的是在硬件逻辑电路(例如集成电路芯片，可编程门阵列(PGA)，专用集成电路(ASIC)等)或者计算机可读介质(例如磁存储介质，比如硬盘驱动器，软盘，磁带)，光存储器(例如 CD-ROM，光盘等)，易失性和非易失性存储器件(例如 EEPROM，ROM，PROM，RAM，DRAM，SRAM，固件，可编程逻辑等)中实现的代码或逻辑。计算机可读介质中的代码由处理器访问和执行。由其实现实施例的代码还可通过传输介质访问，或者通过网络从文件服务器访问。这种情况下，其中实现所述代码的制造产品可包括传输介质，比如网络传输线路，无线传输介质，通过空间传播的信号，无线电波，红外信号等。当然，本领域的技术人员会认识到在不脱离实施例的范围的情况下，可对这种配

置进行许多修改，并且制造产品中包含本领域中已知的任何信息承载介质。

图 6 图解说明其中可实现本发明的一些方面的计算机体系结构的方框图。任何处理复合体 100a...100n 或者主机 510 可实现计算机体系结构 600，计算机体系结构 600 具有处理器 602，存储器 604(例如易失性存储器)和存储装置 606(例如非易失性存储装置，磁盘驱动器，光盘驱动器，磁带驱动器等)。存储装置 606 可包括内部存储装置，附加的存储装置或者网络可访问存储装置。存储装置 606 中的程序可按照本领域中已知的方式被装入存储器 604 中并由处理器 602 执行。该体系结构还可包括使得能够与网络通信的网卡 608。该体系结构还可包括至少一个输入装置 610，例如键盘，触摸屏，指示笔，语音激活的输入装置等，和至少一个输出装置 612，例如显示器，扬声器，打印机等。

图 3 和 4 的逻辑说明了按照特定顺序发生的具体操作。此外，所述操作可并行进行也可顺序进行。在另外的实施例中，某些逻辑操作可按照不同的顺序执行、修改或删除，并且仍然实现本发明的实施例。此外，可向上述逻辑中增加步骤，增加的步骤仍然与实施例相符。另外步骤可由单一进程或分布式进程执行。

出于举例说明的目的，用独立模块的形式说明了许多软件和硬件组件。这样的组件可被结合成数目较少的组件，或者被分成数目更多的组件。另外，描述成由特定组件执行的操作可由其它组件执行。

因此，只是出于举例说明的目的给出了实施例的上述描述。实施例的上述描述不是穷尽的，也不意图把本发明局限于公开的具体形式。

图1

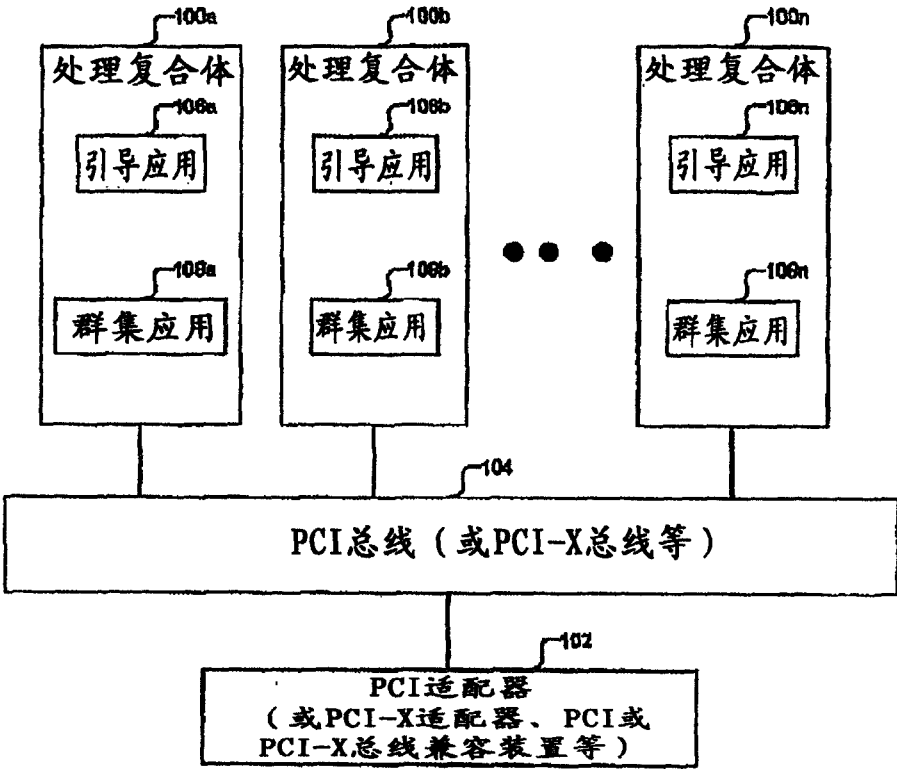


图 2

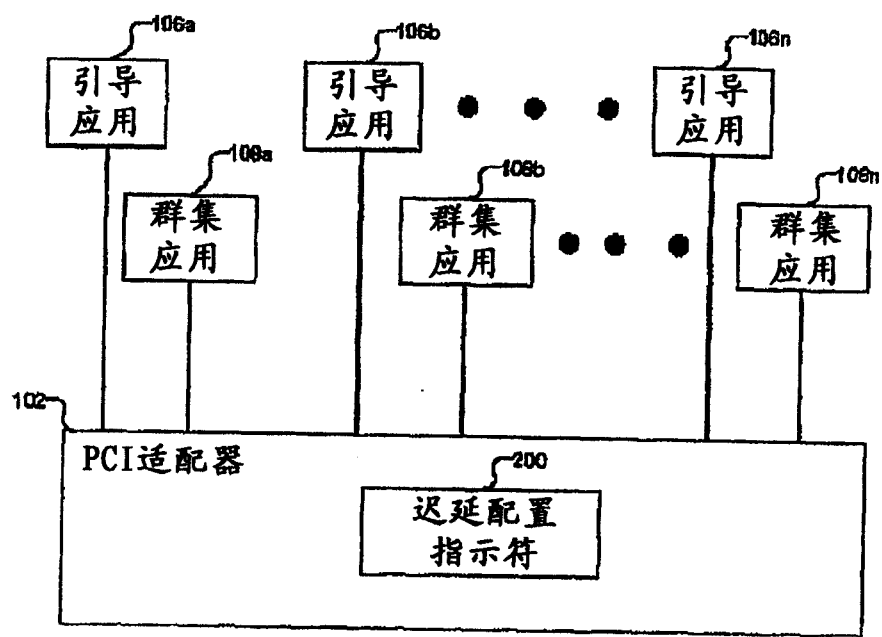


图3

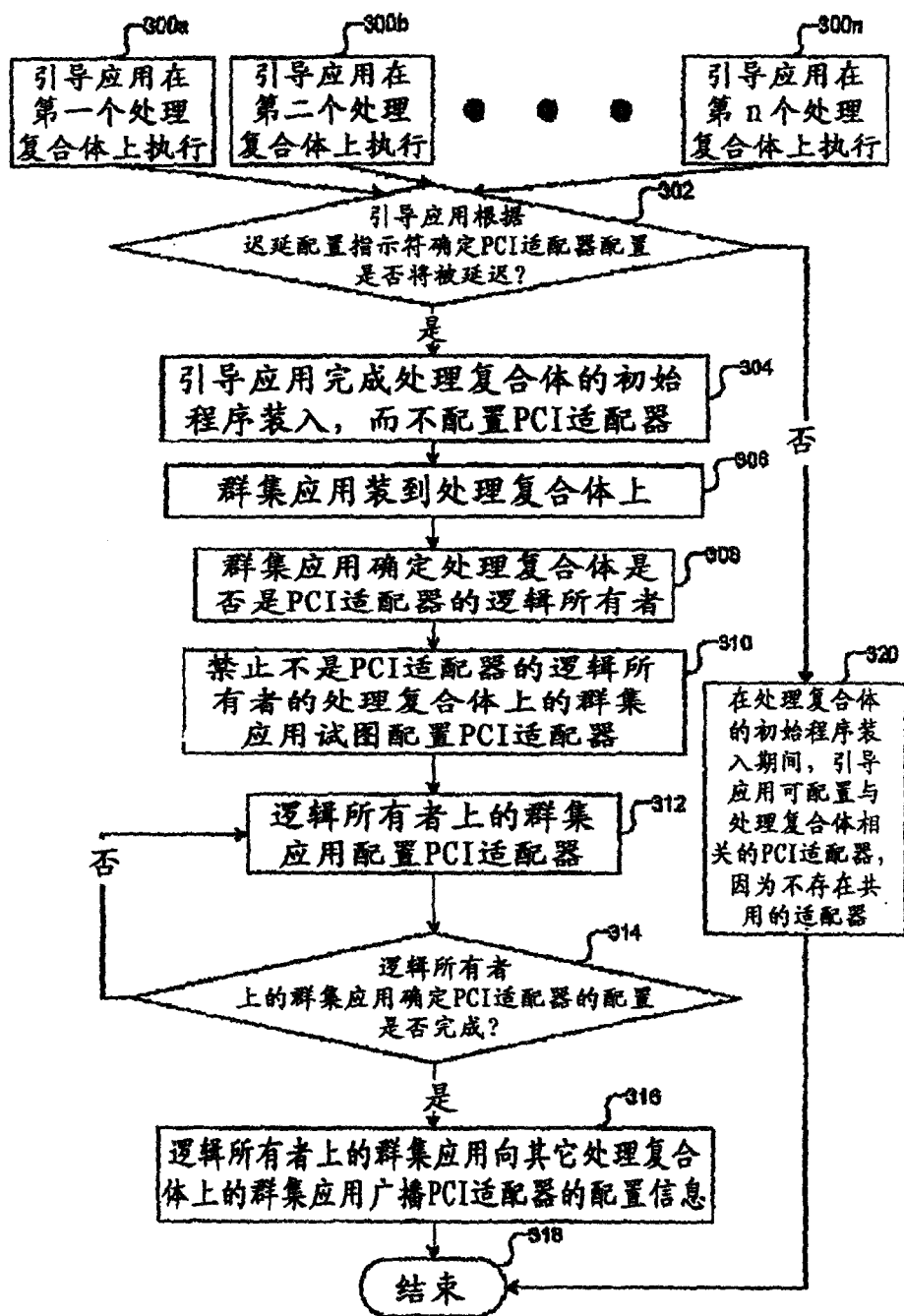


图 4

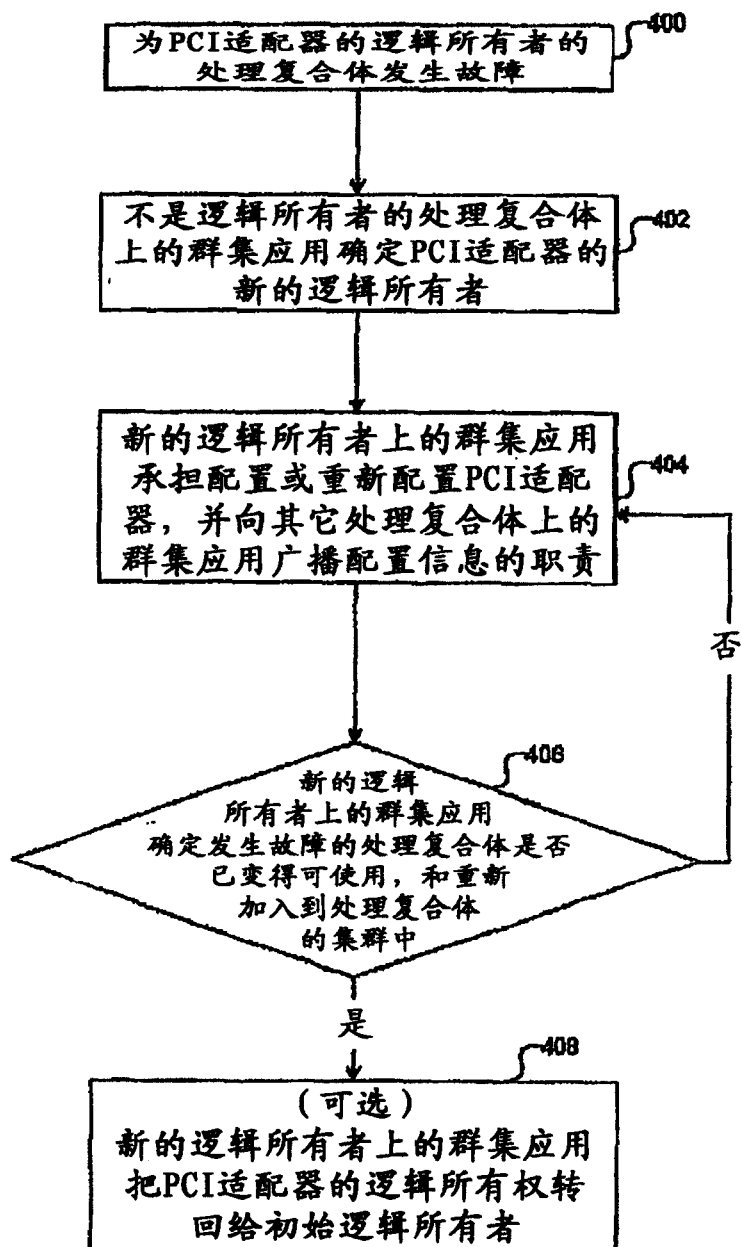


图 5

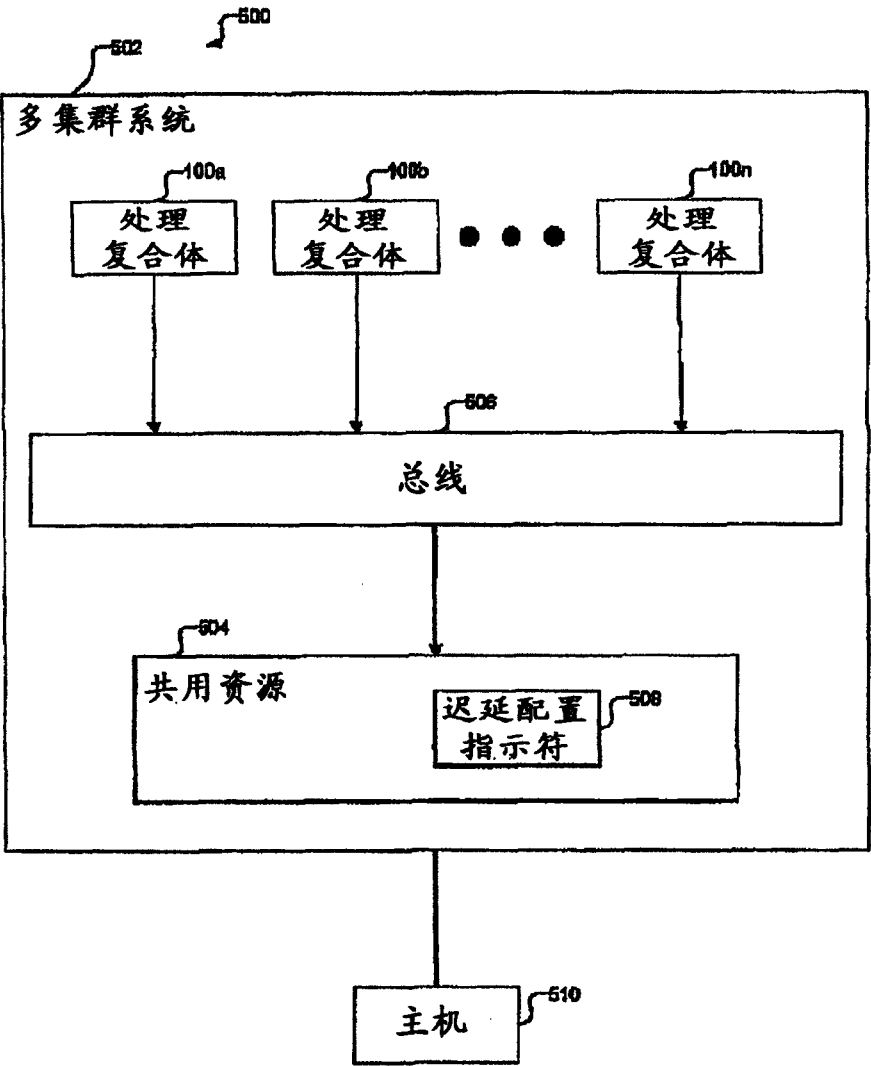


图 6

