[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/30 (2006.01)
G06F 9/46 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200680031546.7

[43] 公开日 2008 年 8 月 27 日

[11] 公开号 CN 101253500A

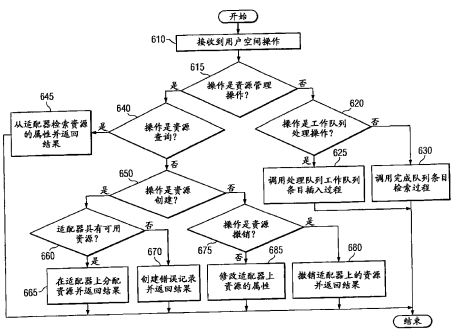
[22] 申请日 2006.8.11
[21] 申请号 200680031546.7
[30] 优先权
[32] 2005.8.31 [33] US [31] 11/216,832
[86] 国际申请 PCT/EP2006/065249 2006.8.11
[87] 国际公布 WO2007/025850 英 2007.3.8
[85] 进入国家阶段日期 2008.2.28
[71] 申请人 国际商业机器公司
地址 美国纽约
[72] 发明人 W·T·博伊德 J·L·赫非尔德
A·梅纳三世 R·雷西奥
M·维加
[74] 专利代理机构 北京市中咨律师事务所
代理人 于 静 李 峥
权利要求书 7 页 说明书 30 页 附图 10 页

[54] 发明名称

用于管理 I/O 的方法

[57] 摘要

提供了一种系统、方法和计算机程序产品，其使得用户空间中间件或应用能够在没有来自本地操作系统(OS)的运行时参与的情况下，将基于文件名的存储请求直接传递至物理 I/O 适配器。 所提供的机制使用可以包括文件名保护表(FNPT)和文件扩展保护表(FEPT)在内的文件保护表(FPT)数据结构来控制用户空间和用户空间外的输入/输出(I/O)操作。 所述 FNPT 具有由所述 OS 的文件系统所管理的每个文件的条目以及指向所述 FEPT 的区段的指针。 所述 FEPT 中的每个条目均可以包括保护域以及其它的保护表上下文信息，可以根据这些信息来检查 I/O 请求，以确定提交所述 I/O 请求的应用实例是否可以访问所述 I/O 请求中所标识的文件。



1. 一种用于管理 I/O 的方法，其包括以下步骤:

从与应用实例关联的处理队列接收处理队列条目，其中所述处理队列条目引用文件;

使用文件保护表数据结构检验关联于所述处理队列条目的文件与所述应用实例相关联; 以及

如果所述处理队列条目所引用的文件与所述应用实例关联，则处理所述处理队列条目，其中在没有主机系统的系统映像干预的情况下，在输入/输出（I/O）适配器中直接从所述应用实例接收所述处理队列条目。

2. 根据权利要求 1 的方法，其中在耦合于运行所述应用实例的主机系统的 I/O 适配器中实现所述方法。

3. 根据权利要求 1 或 2 的方法，其中所述处理队列条目包括引用了文件名保护表中的条目的文件名关键字（FN_Key）值，并且其中所述文件名保护表具有由操作系统的文件系统管理的每个文件的条目。

4. 根据权利要求 3 的方法，其中所述处理队列条目包括引用了文件扩展保护表中的条目的文件扩展关键字（FE_Key）值，并且其中所述文件扩展保护表具有分配给由所述操作系统的文件系统管理的文件的每组线性块地址的条目。

5. 根据权利要求 4 的方法，其中所述文件保护表数据结构包括所述文件名保护表的 I/O 适配器常驻高速缓存部分以及所述文件扩展保护表的 I/O 适配器常驻高速缓存部分。

6. 根据权利要求 4 或 5 的方法，其中文件保护数据结构包括所述文件名保护表和文件扩展保护表，并且其中所述文件名保护表和文件扩展保护表驻留在所述主机系统上。

7. 根据权利要求 4 至 6 中任何一项的方法，其中所述检验步骤进一步包括以下步骤:

处理所述处理队列条目中的 FN_Key 值，以便标识与所述 FN_Key 相

对应的文件名保护表条目；

处理所述文件名保护表条目，以便标识与所述文件名保护表条目相对应的文件扩展保护表的区段；

处理所述处理队列条目中的 FE_Key 值，以便标识与所述 FE_Key 相对应的文件扩展保护表条目；以及

确定是否分配了由所述文件扩展保护表条目所标识的存储设备的一个或多个部分来由所述应用实例进行访问。

8. 根据权利要求 7 的方法，其中所述确定步骤进一步包括以下步骤：

将所述文件扩展保护表条目中所引用的第一保护域和与所述处理队列条目关联的处理队列上下文中所含的第二保护域进行比较；以及

如果所述第一保护域与所述第二保护域相匹配，则确定分配了由所标识的文件扩展保护表条目引用的存储设备的一个或多个部分来由所述应用实例进行访问。

9. 根据权利要求 8 的方法，其中如果所述处理队列条目所引用的文件与所述应用实例关联，则所述处理所述处理队列条目的步骤进一步包括以下步骤：

基于所述文件扩展保护表条目在存储块地址表中进行查找操作，以便标识与所述文件扩展保护表条目相对应的至少一个存储块地址表条目；以及

在由包括在所述文件扩展保护表条目中的存储块地址所引用的存储设备中的存储位置上进行 I/O 操作。

10. 根据权利要求 4 至 9 中任何一项的方法，其进一步包括以下步骤：

基于所述处理队列条目中所提供的 FN_Key 和 FE_Key，标识与所述处理队列条目的目标文件相关联的存储设备部分的获准访问类型；以及

基于与所述处理队列条目的目标文件相关联的存储设备部分的获准访问类型，在所述处理队列条目上进行验证检查，其中仅当成功完成在所述处理队列条目上的验证检查时才处理所述处理队列条目。

11. 根据权利要求 10 的方法，其中所述标识步骤进一步包括以下步

骤:

从文件名保护表（FNPT）数据结构中检索与所述 FN_Key 相对应的文件名保护表条目;

标识与所述文件名保护表条目相对应的文件扩展保护表（FEPT）数据结构的区段;

基于所标识的 FEPT 的区段以及所述 FE_Key，从所述 FEPT 数据结构中检索 FEPT 条目；以及

标识所检索的 FEPT 条目中的验证信息，用于在进行对所述获准访问类型的标识中使用以及用于进行所述验证检查。

12. 根据权利要求 11 的方法，其中所述 FEPT 条目存储了一个或多个访问控制值，所述一个或多个访问控制值标识了在与所述 FEPT 条目关联的存储设备部分上获准的访问类型。

13. 根据权利要求 12 的方法，其中所述一个或多个访问控制值包括有效标识访问控制值和获准操作访问控制值。

14. 根据权利要求 13 的方法，其中如果所述有效标识访问控制值具有第一值，则所述 FEPT 条目有效，并且其中如果所述获准操作访问控制值具有所述第一值，则准许在所述存储设备的关联部分上进行读和写操作。

15. 根据权利要求 13 或 14 的方法，其中在所述处理队列条目上进行验证检查的步骤进一步包括以下步骤:

确定所述 FEPT 条目的有效标识访问控制值的值是否指示所述 FEPT 条目有效。

16. 根据权利要求 15 的方法，其中如果所述 FEPT 条目的有效标识访问控制值指示所述 FEPT 条目有效，并且所述处理队列条目正请求的访问类型是读操作，那么成功完成在所述处理队列条目上的验证检查。

17. 根据权利要求 15 的方法，其中如果所述 FEPT 条目的有效标识访问控制值指示所述 FEPT 条目有效，所述获准操作访问控制值指示准许写操作，并且所述处理队列条目正请求的访问类型是写操作，那么成功完成在所述处理队列条目上的验证检查。

18. 根据权利要求 11 至 17 中任何一项的方法，其中在所述处理队列条目上进行验证检查的步骤进一步包括以下步骤:

确定与所述处理队列条目的目标文件关联的存储设备部分是否处在与所述 FEPT 条目关联的存储设备部分的范围内；以及

如果与所述处理队列条目的目标文件关联的存储设备部分处在与所述 FEPT 条目关联的存储设备部分的范围之外，则拒绝对于与所述处理队列条目的目标文件关联的存储设备部分的访问。

19. 一种用于管理 I/O 的装置，其包括:

处理器；以及

耦合于所述处理器的存储器，其中所述处理器:

从与应用实例关联的处理队列接收处理队列条目，其中所述处理队列条目引用文件;

使用存储在所述存储器中的文件保护表数据结构检验关联于所述处理队列条目的文件与所述应用实例相关联；以及

如果所述处理队列条目所引用的文件与所述应用实例关联，则处理所述处理队列条目，其中在没有主机系统的系统映像干预的情况下，在输入/输出（I/O）适配器中直接从所述应用实例接收所述处理队列条目。

20. 根据权利要求 19 的装置，其中数据处理设备是耦合于运行所述应用实例的主机系统的 I/O 适配器，并且其中，在所述 I/O 适配器中进行所述处理器的接收、检验和处理操作。

21. 根据权利要求 19 或 20 的装置，其中所述处理队列条目包括引用了文件名保护表中的条目的文件名关键字（FN_Key）值，并且其中所述文件名保护表具有由操作系统的文件系统管理的每个文件的条目。

22. 根据权利要求 21 的装置，其中所述处理队列条目包括引用了文件扩展保护表中的条目的文件扩展关键字（FE_Key）值，并且其中所述文件扩展保护表具有分配给由所述操作系统的文件系统管理的文件的每组线性块地址的条目。

23. 根据权利要求 22 的装置，其中所述文件保护表数据结构包括所

述文件名保护表的 I/O 适配器常驻高速缓存部分以及所述文件扩展保护表的 I/O 适配器常驻高速缓存部分。

24. 根据权利要求 22 或 23 的装置，其中文件保护数据结构包括所述文件名保护表和文件扩展保护表，并且其中所述文件名保护表和文件扩展保护表驻留在所述主机系统上。

25. 根据权利要求 22 至 24 中任何一项的装置，其中所述处理器通过以下操作检验所述处理队列条目所引用的文件与所述应用实例关联:

处理所述处理队列条目中的 FN_Key 值，以便标识与所述 FN_Key 相对应的文件名保护表条目;

处理所述文件名保护表条目，以便标识与所述文件名保护表条目相对应的文件扩展保护表的区段;

处理所述处理队列条目中的 FE_Key 值，以便标识与所述 FE_Key 相对应的文件扩展保护表条目；以及

确定是否分配了由所述文件扩展保护表条目所标识的存储设备的一个或多个部分来由所述应用实例进行访问。

26. 根据权利要求 25 的装置，其中所述处理器通过以下操作确定是否分配了由所标识的文件扩展保护表条目引用的存储设备的一个或多个部分来由所述应用实例进行访问:

将所述文件扩展保护表条目中所引用的第一保护域和与所述处理队列条目关联的处理队列上下文中所含的第二保护域进行比较；以及

如果所述第一保护域与所述第二保护域相匹配，则确定分配了由所标识的文件扩展保护表条目引用的存储设备的一个或多个部分来由所述应用实例进行访问。

27. 根据权利要求 26 的装置，其中如果所述处理队列条目所引用的文件与所述应用实例关联，则所述处理器通过以下操作处理所述处理队列条目:

基于所述文件扩展保护表条目在存储块地址表中进行查找操作，以便标识与所述文件扩展保护表条目相对应的至少一个存储块地址表条目；以

及

发布命令，以便在由包括在所述文件扩展保护表条目中的存储块地址所引用的存储设备中的存储位置上进行 I/O 操作。

28. 根据权利要求 22 至 27 中任何一项的装置，其中所述处理器进一步：

基于所述处理队列条目中所提供的 FN_Key 和 FE_Key，标识与所述处理队列条目的目标文件相关联的存储设备部分的获准访问类型；以及

基于与所述处理队列条目的目标文件相关联的存储设备部分的获准访问类型，在所述处理队列条目上进行验证检查，其中仅当成功完成在所述处理队列条目上的验证检查时才处理所述处理队列条目。

29. 根据权利要求 28 的装置，其中所述处理器通过以下操作进行标识：

从文件名保护表（FNPT）数据结构检索与所述 FN_Key 相对应的文件名保护表条目；

标识与所述文件名保护表条目相对应的文件扩展保护表（FEPT）数据结构的区段；

基于所标识的 FEPT 的区段以及所述 FE_Key，从所述 FEPT 数据结构中检索 FEPT 条目；以及

标识所检索的 FEPT 条目中的验证信息，用于在进行对所述获准访问类型的标识中使用以及用于进行所述验证检查。

30. 根据权利要求 29 的装置，其中所述 FEPT 条目存储了一个或多个访问控制值，所述一个或多个访问控制值标识了在与所述 FEPT 条目关联的存储设备部分上获准的访问类型。

31. 根据权利要求 30 的装置，其中所述一个或多个访问控制值包括有效标识访问控制值和获准操作访问控制值。

32. 根据权利要求 31 的装置，其中如果所述有效标识访问控制值具有第一值，则所述 FEPT 条目有效，并且其中如果所述获准操作访问控制值具有所述第一值，则准许在所述存储设备的关联部分上进行读和写操作。

33. 根据权利要求 31 或 32 的装置，其中所述处理器通过以下操作在所述处理队列条目上进行验证检查：

确定所述 FEPT 条目的有效标识访问控制值的值是否指示所述 FEPT 条目有效。

34. 根据权利要求 33 的装置，其中如果所述 FEPT 条目的有效标识访问控制值指示所述 FEPT 条目有效，并且所述处理队列条目正请求的访问类型是读操作，那么成功完成在所述处理队列条目上的验证检查。

35. 根据权利要求 33 的装置，其中如果所述 FEPT 条目的有效标识访问控制值指示所述 FEPT 条目有效，所述获准操作访问控制值指示准许写操作，并且所述处理队列条目正请求的访问类型是写操作，那么成功完成在所述处理队列条目上的验证检查。

36. 根据权利要求 29 至 35 中任何一项的装置，其中所述处理器通过以下操作在所述处理队列条目上进行验证检查：

确定与所述处理队列条目的目标文件关联的存储设备部分是否处在与所述 FEPT 条目关联的存储设备部分的范围内；以及

如果与所述处理队列条目的目标文件关联的存储设备部分处在与所述 FEPT 条目关联的存储设备部分的范围之外，则拒绝对于与所述处理队列条目的目标文件关联的存储设备部分的访问。

37. 一种包括程序代码装置的计算机程序，当所述程序在计算机上运行时，所述程序代码装置适于实现权利要求 1 至 18 中任何一项的方法的所有步骤。

用于管理I/O的方法

技术领域

本发明一般涉及主计算机与输入/输出（I/O）适配器之间的通信协议。更具体地，本发明针对的是一种系统和方法，其用于在没有来自本地操作系统（OS）（或在虚拟系统中，本地管理体（hypervisor））的运行时参与的情况下，使得用户空间中间件或应用能够将基于文件名的存储请求直接传递至物理I/O适配器。

背景技术

根据现有的技术情况，操作系统不允许诸如数据库的用户空间中间件或应用直接访问通过操作系统的本地文件系统的文件模式I/O接口所标识的永久性存储器。因此，用户空间中间件必须在每次进行I/O操作时调用操作系统（OS）调用并引发多次任务切换。当中间件或应用将存储请求传送给OS时导致第一任务切换。在OS完成处理中间件或应用存储请求并将存储请求传递至存储适配器之后，当OS将控制传递回给用户空间中间件或应用时，发生第二任务切换。

当存储适配器完成关联的I/O存储操作并中断正在由应用进行的处理以便OS可以处理存储适配器的完成时，发生第三任务切换。当OS结束处理存储适配器的完成并将控制返回给向OS传送存储请求的中间件或应用时，发生最后的任务切换。除了这些任务切换之外，存储适配器通常还具有单个请求队列来处理来自操作系统的工作。

上述四次任务切换可被视为浪费的处理器周期，因为对于正在切换的线程的所有工作均会停止，直到任务切换完成。在某些服务器中，用户空间中间件或应用程序所进行的存储操作数可能相当大。现代的高端服务器每秒可以有数百万次这些操作，这导致每秒数百万次的任务切换。

发明内容

鉴于上述内容，获得一种这样的方法、系统和具有计算机可读指令的计算机程序产品会是有利的，即其用于处理输入/输出（I/O）存储请求，其中，对此类任务切换进行了最小化。此外，获得一种改进的方法、系统和计算机指令会是有利的，即其在没有来自本地操作系统（OS）（或在虚拟系统中，本地管理体）的运行时参与的情况下使得用户空间中间件或应用能够将基于文件名的 I/O 存储请求直接传递至物理 I/O 适配器。将该机制应用于 InfiniBand、TCP/ IP 卸载引擎、启用 RDMA（远程直接存储器访问）的 NIC（网络接口控制器）、iSCSI 适配器、iSER（用于 RDMA 的 iSCSI 扩展）适配器、并行 SCSI 适配器、光纤通道适配器、串行附加 SCSI 适配器、ATA 适配器、串行 ATA 适配器以及任何其它类型的存储适配器也会是有利的。

进一步地，获得一种改进的方法、系统和计算机指令会是有利的，即其使得保护机制能够确保从应用实例直接发送至物理 I/O 适配器的基于文件名的存储请求仅被完成到先前已经为了随该应用实例的用户空间外 I/O 而分配的存储设备部分。此外，获得一种这样的方法、系统和计算机指令会是有利的，即其使得能够创建、修改、查询和删除用于促进应用实例与物理 I/O 适配器之间基于文件名的直接 I/O 操作的数据结构条目。另外，获得一种这样的方法、系统和计算机指令会是有利的，即其用于处理用户空间操作以便进行存储设备资源管理和直接 I/O 操作数据结构管理。最后，获得一种这样的方法、系统和计算机指令会是有利的，即其使用运行在主机系统上的操作系统的文件系统来实现以上目的。

本发明提供了一种方法、计算机程序产品和数据处理系统，其使得用户空间中间件或应用能够在没有来自本地操作系统（OS）（或在虚拟系统中，本地管理体）的运行时参与的情况下，使用运行在主机系统上的操作系统的文件系统将基于文件名的存储请求直接传送至物理 I/O 适配器。本发明中所描述的机制应用于 InfiniBand 主机通道适配器、TCP / IP 卸载引

擎、启用 RDMA（远程直接存储器访问）的 NIC（网络接口控制器）、iSCSI 适配器、iSER（用于 RDMA 的 iSCSI 扩展）适配器、并行 SCSI 适配器、光纤通道适配器、串行附加 SCSI 适配器、ATA 适配器、串行 ATA 适配器以及任何其它类型的存储适配器。

具体而言，本发明针对的是一种用于提供和使用文件保护表（FPT，file protection table）数据结构来控制用户空间以及用户空间外的输入/输出（I/O）操作的机制。在本发明的一个方面中，所述 FPT 包括文件名保护表（FNPT，file name protection table），其具有由操作系统的文件系统管理的每个文件的条目。所述 FNPT 中的条目包括指向与文件名相对应的文件扩展保护表（FEPT，file extension protection table）的区段的指针。所述 FEPT 中的条目可以包括关键字实例（key instance）和保护域，以及其它的保护表上下文信息，可以根据这些信息来检查 I/O 请求以确定提交所述 I/O 请求的应用实例是否可以访问与所述 I/O 请求中所标识的文件名相对应的存储设备部分。以这样的方式，只有已经分配给所述应用实例的那些存储设备部分才可以被所述应用实例访问。此外，只有为其分配了所述存储设备部分的应用实例才可以访问所述存储设备部分。

所述 FPT 可以进一步包括 LBA 表，在所述 LBA 表中的是标识了与所述文件扩展保护表（FEPT）中的条目相关联的逻辑块地址的 LBA 表条目。所述 LBA 表可以用于将基于文件名的 I/O 请求中所引用的 LBA 映射到物理存储设备的 LBA。本发明进一步提供了一种机制，其用于处理用户空间操作，以便管理所述文件名保护表、文件扩展保护表和 LBA 表中的条目的创建、修改、查询和删除。这样的机制与物理 I/O 适配器的存储管理接口进行连接，以便对与应用实例关联的文件、文件扩展以及 LBA 进行分配、修改、查询和解除分配。

另外，本发明提供了用于处理用户空间操作以便生成工作队列条目来将基于文件名的 I/O 操作直接传递至物理 I/O 适配器的机制。此外，本发明提供了这样的机制，即其用于在工作队列条目已经由物理 I/O 适配器进行处理时，从所述物理 I/O 适配器中检索完成队列条目以便通知应用实例

完成处理。

如下文所阐述的，在本发明的一个示例性实施例中，提供了一种在其中处理队列条目是从与应用实例关联的处理队列接收的方法、计算机程序产品、装置和系统，其中所述处理队列条目引用文件，使用文件保护表数据结构检验关联于所述处理队列条目的文件与所述应用实例关联，并且如果所述处理队列条目所引用的文件与所述应用实例相关联，则处理所述处理队列条目。可以在没有主机系统的系统映像干预的情况下在输入/输出（I/O）适配器中直接从所述应用实例接收所述处理队列条目。可以在耦合于运行所述应用实例的主机系统的 I/O 适配器中实现所述方法，并且可以在耦合于运行所述应用实例的主机系统的 I/O 适配器上执行所述计算机程序产品。

所述处理队列条目可以包括引用了文件名保护表中的条目的文件名关键字（FN_Key）值。所述文件名保护表可以具有由操作系统或系统映像的文件系统管理的每个文件的条目。

所述处理队列条目可以包括引用了文件扩展保护表中的条目的文件扩展关键字（FE_Key）值，并且其中所述文件扩展保护表具有分配给由操作系统或系统映像的文件系统管理的文件的每组线性块地址的条目。所述文件保护表数据结构可以包括所述文件名保护表的 I/O 适配器常驻高速缓存部分以及所述文件扩展保护表的 I/O 适配器常驻高速缓存部分。所述文件保护数据结构可以包括所述文件名保护表以及所述文件扩展保护表。所述文件名保护表和文件扩展保护表可以驻留在所述主机系统上。

通过以下操作可以检验所述处理队列条目所引用的文件是与所述应用实例关联的：处理所述处理队列条目中的 FN_Key 值，以便标识与所述 FN_Key 相对应的文件名保护表条目；处理所述文件名保护表条目，以便标识与所述文件名保护表条目相对应的文件扩展保护表的区段；处理所述处理队列条目中的 FE_Key 值，以便标识与所述 FE_Key 相对应的文件扩展保护表条目；以及确定是否分配了由所述文件扩展保护表条目所标识的存储设备部分来由所述应用实例进行访问。

可以通过将所述文件扩展保护表条目中所引用的第一保护域和与所述处理队列条目关联的处理队列上下文中所含的第二保护域进行比较，确定是否分配了由标识的文件扩展保护表条目所引用的存储设备的一个或多个部分来由所述应用实例进行访问。基于该比较，如果所述第一保护域与所述第二保护域相匹配，则可以确定分配了由所标识的文件扩展保护表条目引用的存储设备的一个或多个部分来由所述应用实例进行访问。

如果所述处理队列条目所引用的文件与所述应用实例相关联，则可以通过基于所述文件扩展保护表条目，在存储块地址表中进行查找操作，以便标识与所述文件扩展保护表条目相对应的至少一个存储块地址表条目，来实现对所述处理队列条目的处理。然后可以在由包括在所述文件扩展保护表条目中的存储块地址所引用的存储设备中的存储位置上进行 I/O 操作。

本发明的机制可以进一步包括：基于所述处理队列条目中所提供的 FN_Key 和 FE_Key，标识与所述处理队列条目的目标文件相关联的存储设备部分的获准访问类型。此外，基于与所述处理队列条目的目标文件相关联的存储设备部分的获准访问类型，可以在所述处理队列条目上进行验证检查，其中仅当成功完成在所述处理队列条目上的验证检查时才处理所述处理队列条目。在标识与所述处理队列条目的目标文件相关联的存储设备部分的获准访问类型中，可以从文件名保护表（FNPT）数据结构中检索与所述 FN_Key 相对应的文件名保护表条目，并且可以标识与所述文件名保护表条目相对应的文件扩展保护表（FEPT）数据结构的区段。基于所述 FEPT 的标识区段和 FE_Key，可以从所述 FEPT 数据结构中检索 FEPT 条目，并且可以标识所检索的 FEPT 条目中的验证信息，用于在进行获准访问类型的标识中使用以及用于进行所述验证检查。

所述 FEPT 条目可以存储一个或多个访问控制值，其标识了在与所述 FEPT 条目关联的存储设备部分上获准的访问类型。所述一个或多个访问控制值可以包括有效标识访问控制值和获准操作访问控制值。如果所述有效标识访问控制值具有第一值，则所述 FEPT 条目有效，并且其中如果所

述获准操作访问控制值具有所述第一值，则准许在所述存储设备的关联部分上进行读和写操作。

举例来说，通过确定所述 FEPT 条目的有效标识访问控制值的值是否指示所述 FEPT 条目有效，本发明的机制可以在所述处理队列条目上进行验证检查。如果所述 FEPT 条目的有效标识访问控制值指示所述 FEPT 条目有效，并且所述处理队列条目正请求的访问类型是读操作，那么成功完成在所述处理队列条目上的验证检查。如果所述 FEPT 条目的有效标识访问控制值指示所述 FEPT 条目有效，所述获准操作访问控制值指示准许写操作，并且所述处理队列条目正请求的访问类型是写操作，那么成功完成在所述处理队列条目上的验证检查。

此外，通过确定与所述处理队列条目的目标文件相关联的存储设备部分是否处于与所述 FEPT 条目相关联的存储设备部分的范围内，本发明可以在所述处理队列条目上进行验证检查。如果与所述处理队列条目的目标文件相关联的存储设备部分处在与所述 FEPT 条目相关联的存储设备部分的范围之外，则可以拒绝对与所述处理队列条目的目标文件相关联的存储设备部分的访问。

举例来说，根据本发明的装置可以包括处理器和耦合于所述处理器的存储设备（例如存储器）。举例来说，根据本发明的系统可以包括处理器以及耦合于所述处理器的 I/O 适配器。

将在以下对本发明的示例性实施例的详细描述中对本发明的这些和其它特征和优点进行描述，或者鉴于以下对本发明的示例性实施例的详细描述，本发明的这些和其它特征和优点对于本领域的普通技术人员将变得显而易见。

附图说明

在所附权利要求中阐述了被认为是本发明特色的新颖特征。然而，当结合附图阅读时，通过参照以下对说明性实施例的详细描述，可以最好地理解本发明本身以及优选的使用模式、进一步的目的及其优点，在附图中：

图 1 是依照本发明的示例性实施例的主处理器节点的功能框图；

图 2 是依照本发明的示例性实施例说明了用于启用用户空间外基于文件名的存储 I/O 访问的主处理器节点的主要操作元件的示图；

图 3 是依照本发明的示例性实施例说明了用于转换和保护基于文件名的存储的示例性控制结构的示图；

图 4 是依照本发明的示例性实施例说明了用于从用户空间中间件或应用实例将存储请求传递至存储适配器的示例性控制结构的示图；

图 5 是依照本发明的示例性实施例说明了用于保证允许用户空间中间件或应用实例所提交的基于文件名的存储 I/O 请求引用基于文件名的存储 I/O 请求中所引用的文件的示例性控制结构的示图；

图 6 是依照本发明的示例性实施例概括了用于处理用户空间操作的调用的示例性操作的流程图；

图 7 是概括了当所调用的用户空间操作是要求进行生成和处理或工作队列元素的工作队列操作时本发明的一个示例性实施例的示例性操作的流程图；

图 8 是概括了当进行验证检查以确定工作队列条目是否有效以及是否可以由物理 I/O 适配器处理时本发明的一个示例性实施例的示例性操作的流程图；

图 9 是概括了当所调用的用户空间操作是完成队列检索过程操作时本发明的一个示例性实施例的示例性操作的流程图；

图 10 是依照本发明的示例性实施例概括了当创建文件保护表条目时本发明的一个示例性实施例的示例性操作的流程图；

图 11 是概括了当处理作为资源修改操作的用户空间操作时本发明的一个示例性实施例的示例性操作的流程图；

图 12 是概括了当处理查询用户空间操作时本发明的一个示例性实施例的示例性操作的流程图；以及

图 13 是概括了当处理撤销（destroy）或删除用户空间操作时本发明的一个示例性实施例的示例性操作的流程图。

具体实施方式

本发明应用于使用诸如 PCI 系列 I/O 适配器、虚拟 I/O 适配器、端点设备、虚拟端点设备等的 I/O 适配器来直接依附于存储器或通过网络依附于存储器的任何通用或专用主机。网络可以包括端节点、交换机、路由器和互连这些组件的链路。网络链路可以是光纤通道、以太网、InfiniBand、高级交换互连、其它标准存储网络互连，或者使用专有或标准协议的专有链路。虽然下文的描述和说明将参照网络和主节点的特定布置，但是应当理解，下面的示例性实施例只是示例性的，并且可以在不背离本发明的范围的情况下，对具体描述和说明的布置进行修改。

重要的是要注意，本发明可以采取全硬件实施例、全软件实施例或者既含有硬件元素又含有软件元素的实施例的形式。在示例性实施例中，以软件实现本发明，其包括但不限于固件、常驻软件、微码等。

此外，本发明可以采取可访问于计算机可用或计算机可读介质的计算机程序产品的形式，该计算机可用或计算机可读介质提供由计算机或任何指令执行系统使用的或者与计算机或任何指令执行系统结合使用的程序代码。对于该描述来说，计算机可用或计算机可读介质可以是能够容纳、存储、通信、传播或传送由指令执行系统、装置或设备使用的或者与指令执行系统、装置或设备结合使用的程序的任何装置。

介质可以是电子、磁性、光学、电磁、红外或半导体系统（或装置或设备）或者传播介质。计算机可读介质的例子包括半导体或固态存储器、磁带、可装卸计算机磁盘、随机访问存储器（RAM）、只读存储器（ROM）、硬磁盘和光盘。光盘的当前的例子包括只读光盘存储器（CD-ROM）、读/写光盘（CD-R/W）和 DVD。

适于存储和/或执行程序代码的数据处理系统可以包括通过系统总线直接地或间接地耦合于存储元件的至少一个处理器。存储元件可以包括在程序代码的实际执行期间所采用的局部存储器、大容量存储器，以及为了减少在执行期间必须从大容量存储器检索代码的次数而提供对至少一些程

序代码的临时存储的高速缓冲存储器。

输入/输出或 I/O 设备（包括但不限于键盘、显示器、指点设备等）可以直接地或者通过插入 I/O 控制器耦合于系统。网络适配器耦合于系统，从而使得数据处理系统能够适于通过介入专用或公用网络耦合于其它的数据处理系统或远程打印机或存储设备。调制解调器、电缆调制解调器和以太网卡正是几种当前可用类型的网络适配器。

现参照附图，并且特别参照图 1，其依照本发明的一个示例性实施例描绘了主节点的功能框图。在该例中，主节点 102 包括通过链路 101 互连的两个处理器 I/O 层次 100 和 103。为了便于描述主节点 102 的元件，仅完全描绘了处理器 I/O 层次 100，且处理器 I/O 层次 103 具有类似的（尽管未示出）如下文讨论的元件布置。

如所示出的，处理器 I/O 层次 100 包括处理器芯片 107，处理器芯片 107 包括一个或多个处理器及其关联的高速缓存。处理器芯片 107 通过链路 108 连接至存储器 112。处理器芯片上的链路之一，例如链路 120，连接至 PCI 系列 I/O 桥接器 128。PCI 系列 I/O 桥接器 128 具有一个或多个 PCI 系列（PCI、PCI_X、PCI_Express 或任何将来推出的 PCI）链路，该链路用于通过诸如链路 132、136 和 140 的 PCI 链路连接其它的 PCI 系列 I/O 桥接器或 PCI 系列 I/O 适配器（例如 PCI 系列适配器 1 145 和 PCI 系列适配器 2 144）。诸如 PCI 系列适配器 1 145 的 PCI 系列适配器可以用于通过诸如到网络 164 的链路 156 这样的网络链路连接至依附网络的存储器 152，链路 156 连接至交换机或路由器 160，而交换机或路由器 160 又通过链路 158 连接至依附网络的存储器 152。诸如 PCI 系列适配器 2 144 的 PCI 系列适配器还可以用于通过链路 148 连接直接依附的存储设备 162。

重要的是要注意，诸如 PCI 系列适配器 1 145 或 PCI 系列适配器 2 144 这样的 PCI 系列适配器可以与主节点 102 上的其它组件集成。例如，PCI 系列适配器 1 145 或 PCI 系列适配器 2 144 可以与 PCI 系列 I/O 桥接器 128 集成。其它例子是诸如 PCI 系列适配器 1 145 或 PCI 系列适配器 2 144 这样的 PCI 系列适配器可以与处理器芯片 107 集成。

虽然将关于 PCI 系列适配器来描述本发明的示例性实施例，但是应当理解本发明并不限于此类适配器。相反，物理 I/O 适配器可以是包括 PCI 系列适配器、虚拟 I/O 适配器、端点设备、虚拟端点设备、虚拟 I/O 适配器端点设备等在内的任何类型的 I/O 适配器。举例来说，在题为“Data Processing System，Method and Computer Program Product for Creation and Initialization of a Virtual Adapter on a Physical Adapter that Supports Virtual Adapter Level Virtualization，”（2005 年 2 月 25 日提出申请，在此通过引用的方式将其纳入本说明书）的共同受让和共同未决的美国专利申请 11/065, 829 中描述了可以随本发明一起使用的虚拟 I/O 适配器的一个例子。在不背离本发明的范围的情况下，可以使用其它类型的 I/O 适配器。

现参照图 2，其描绘了与本发明的一个示例性实施例相关的系统组件的功能框图。在所描绘的例子中，物理 I/O 适配器 200 是诸如图 1 中的 PCI 系列适配器 1 145 或 PCI 系列适配器 2 144 这样的 PCI 适配器的例子。

在该例中，图 2 中所示的物理 I/O 适配器 200 包括诸如处理队列集 236 这样的处理队列（PQs）的一个集合，及其关联的处理队列上下文，例如 PQ 上下文 204。举例来说，处理队列（PQs）可以包括工作队列（例如发送队列和/或接收队列）以及完成队列。工作队列用于将基于文件名的 I/O 存储请求直接提交给物理 I/O 适配器。为了直接访问存储设备的部分，使用本发明的机制，将基于文件名的 I/O 存储请求中的文件名转换成线性块地址（Linear Block Address）。线性块地址（LBA）是从存储设备的逻辑起点的块（即，存储设备的固定大小部分）的索引。完成队列用于将工作队列条目的完成传达回给提交了基于文件名的 I/O 存储请求的应用实例。

物理 I/O 适配器 200 还具有诸如 FPT 上下文 208 这样的文件保护表（FPT）上下文，其用于容纳诸如 FPT 232 或 FPT 252 这样的主机常驻文件保护表的上下文。FPT 上下文 208 还可以用于容纳 FPT 232 或 252 本身或来自主机常驻 FPT 232 或 FPT 252 的条目的高速缓存。

FPT 232 和 252 驻留在诸如 OS 1 220 或 OS 2 240 这样的操作系统（OS）

中。OS（例如 OS 1 220 或 OS 2 240）可以驻留在管理体 216 之上，管理体 216 是管理物理硬件资源的分区和虚拟化以及控制 OS 执行的软件、固件或二者的混合。OS 可以托管一个或多个中间件或应用实例。在图 2 中，OS 1 220 正托管两个中间件或应用实例 App 1 224 和 App 2 228。类似地，OS 2 240 正托管应用 App 1 224 和 App 2 228。OS 在诸如处理器 212 的处理器上运行。

中间件或应用实例（例如 App 1 224）使用诸如处理队列集 236 这样的处理队列集来将基于文件名的 I/O 存储请求传递至物理 I/O 适配器。当物理 I/O 适配器 200 处理基于文件名的 I/O 存储请求时，物理 I/O 适配器 200 使用在基于文件名的 I/O 存储请求中所传递的关键字来在 FPT 上下文 208 中查找条目。如果 FPT 上下文 208 与用于处理队列的 PQ 上下文 204 关联于相同的保护域，那么处理基于文件名的 I/O 存储请求。否则，错误地完成基于文件名的 I/O 存储请求。

接下来转至图 3，其描绘了文件保护表（FPT）的例子。图 3 中示出了三个表：文件名保护表 302、文件扩展保护表 312 和线性块地址（LBA）表 322，其一起可以构成文件保护表数据结构。文件名保护表 302 含有由操作系统或系统映像 300 的文件系统管理的每个文件的条目。文件名保护表 302 中的条目指向与文件名保护表条目所表示的文件相对应的文件扩展保护表 312 的区段。

文件扩展保护表 312 含有每个文件扩展的条目。这些条目中的每个条目均描述了访问控制、文件名、指向线性块地址（LBA）表 322 的指针（其含有与对应的文件扩展保护表条目相关联的 LBA 的范围），以及稍后将在本说明书中涵盖的其它字段。在所描绘的例子中，文件扩展保护表 312 含有每个逻辑卷（LV）的条目，并且因而文件扩展保护表 312 是 LV 文件扩展保护表 312。

可以将文件扩展保护表 312 分段成一组文件扩展保护表段，例如文件扩展保护表段 1 314。可以使用包括 B 树、由非叶节点中的指针和叶节点中的指针组成的树、简单链接的列表等在内的若干数据结构来互连区段。

在所描绘的例子中，文件扩展保护表段 1 314 使用简单链接的列表，其中该表中的第一条目是指向含有文件扩展保护表条目的下一表的指针。

文件扩展保护表条目 N 320 描绘了诸如文件扩展保护表段 1 314 这样的文件扩展保护表段中的示例条目。文件扩展保护表段 1 314 中的每个条目均含有用于定义该条目的一组字段。文件扩展保护表条目 N 320 含有以下字段：访问控制、保护域、关键字实例、文件名、逻辑卷号、SCSI 标识符号、SCSI 逻辑单元号、LBA 表大小、扇区大小、长度、LBA 表指针。

在一个示例性实施例中，适配器的 FE_Key 映射逻辑 386（举例来说，其可以是 I/O 适配器的处理器中的逻辑，或者可以是单独的专用逻辑单元）对文件扩展保护表条目（例如文件扩展保护表条目 N 320）中的字段进行所有的检查。未通过 FE_Key 映射逻辑 386 的任何检查均导致操作错误地完成。在出现错误的情况下，操作系统（OS）可以拆卸在操作中进行传递的中间件或应用实例，或者采取不那么激烈的手段，例如返回带有错误完成的操作。

访问控制字段描述了文件扩展保护表（FEPT）条目是否有效以及可以对 FEPT 条目进行何种类型的操作。可以对该条目进行的可能的操作是：读、写，以及读/写。如果通过中间件或应用实例传递的基于文件名的 I/O 存储请求访问有效的 FEPT 条目，那么操作通过有效/无效检查。如果通过中间件或应用实例传递的基于文件名的存储 I/O 请求尝试进行读访问操作并且 FEPT 条目设置了有效比特，那么操作通过该检查。如果通过中间件或应用实例传递的基于文件名的存储 I/O 请求尝试进行写访问操作并且 FEPT 条目设置了读/写比特，那么操作通过该检查。

保护域字段用于将 FEPT 条目与处理队列（PQ）上下文进行关联。也就是说，如果中间件或应用实例用于在基于文件名的存储 I/O 请求中传递的 PQ 上下文在其保护域字段中含有与 FEPT 条目的保护域字段相同的值，那么对二者进行关联并且操作通过该检查。如果 PQ 上下文和 FEPT 条目中的这些保护域之间存在不匹配，那么操作无法通过该检查。

关键字实例用于将来自中间件或应用实例的基于文件名的 I/O 存储请

求中所传递的文件扩展关键字与存储在 FEPT 条目中的文件扩展关键字进行比较。如果二者相匹配，则操作通过该检查。如果关键字实例与在基于文件名的存储 I/O 请求中所传递的存储关键字不匹配，那么操作不通过该检查。

文件扩展关键字或“FE_Key”具有两个字段－第一字段是进入 FEPT 的索引，例如偏移，并且第二字段是将要与第一字段所指向的 FEPT 条目中的关键字实例进行比较的关键字实例。当中间件或应用实例提交基于文件名的 I/O 存储请求时，适配器使用文件名和第一字段来从 FEPT 获取条目。举例来说，这可以通过使用文件名或文件名关键字标识文件名保护表 302 中指向 FEPT 312 的区段的起始地址的条目来完成。然后可以使用文件扩展关键字的第一字段中的索引或偏移来标识 FEPT 312 中的特定条目。然后，适配器将 FEPT 条目内的关键字实例与通过中间件或应用实例传递的第二字段进行比较。

文件名字段是任选的，并且如果包括了文件名字段，则可以将其用于标识与 FEPT 312 条目关联的文件名和/或文件名关键字。文件名字段可以用于对基于文件名的存储 I/O 请求中所传递的文件名或文件名关键字进行检查。如果二者匹配，那么操作通过该检查；否则如果二者不匹配，则操作无法通过该检查。

逻辑卷号是任选的，并且如果包括了逻辑卷号，则可以用其来对在中间件或应用实例的基于文件名的存储 I/O 请求中所传递的 LV 号与存储在 LV 文件扩展保护表条目中的 LV 号进行比较。如果二者匹配，则操作通过该检查。如果逻辑卷号与通过基于文件名的存储 I/O 请求中所传递的 LV 号不匹配，那么操作无法通过该检查。

SCSI 标识符号（ID）和 SCSI 逻辑单元号（LUN）用于将条目分别与特定的 SCSI 设备和该设备内的特定 LUN 相关联。

LBA 表大小用于定义与 FEPT 条目关联的每个 LBA 表段（例如 LBA 表段 1 324）可含的条目的最大数目。扇区大小用于定义与 FEPT 条目关联的盘上的每个扇区的大小。长度字段用于定义与 FEPT 条目关联的盘 LBA

集（the set of disk LBAs）的总长度。

FEPT 条目 320 的 LBA 表指针指向 LBA 表 322 中的一个或多个对应的 LBA 表条目。因而，利用 LBA 表指针字段，可以标识与 FEPT 312 中的 FEPT 条目关联的线性块地址，以便提供对与处理队列（在适配器 316 中从该处理队列接收基于文件名的 I/O 请求）相关联的物理存储设备上的存储位置的线性块地址的访问。

还可以将 LBA 表 322 分段成一组 LBA 表段，例如 LBA 表段 1 324。可以使用包括 B 树、由非叶节点中的指针和叶节点中的指针组成的树、简单链接的列表等在内的若干数据结构来互连区段。在所描绘的例子中，LBA 表段 1 324 使用简单链接的列表，其中，该表中的第一条目是指向含有 LBA 表条目的下一表的指针。

诸如 LBA 表段 1 324 这样的 LBA 表段中的每个条目均描述了与该条目关联的盘线性块地址（LBA）的范围。对于该描述来说，条目可以使用起始 LBA 和长度、起始 LBA 和结束 LBA 等。

诸如适配器 316 的物理 I/O 适配器可以选择存储整个文件保护表、部分文件保护表，或者不存储文件保护表。所示出的适配器 316 具有持有一个区段的文件名保护表高速缓存和文件扩展保护表高速缓存，例如高速缓存的文件名保护表段 1 390 和文件扩展保护表段 1 392。

类似地，适配器 316 可以选择存储整个 LBA 表、部分 LBA 表，或者不存储 LBA 表。在所描绘的例子中，所示出的适配器 316 具有持有一个区段的 LBA 表高速缓存，例如高速缓存的 LBA 表段 1 398。

接下来参照图 4，其依照本发明的示例性实施例示出了用于为用户空间中间件或应用实例将基于文件名的 I/O 存储请求传递至物理 I/O 适配器的示例性控制结构的示例图。为了进行说明，举例来说，所示出的系统映像（其可以是诸如 Windows XPTM、AIXTM、LinuxTM 等的操作系统，或者诸如基于文件名的 I/O 存储服务器或文件模式 I/O 存储服务器的专用软件映像）具有使用存储或网络适配器从存储设备调用存储操作的应用。为了进行以下描述，可以交换使用术语“系统映像”和“操作系统”来指

代系统映像，即系统存储器的当前内容，其可以包括操作系统和任何运行的应用实例。

诸如系统映像 1 412 的系统映像具有与存储适配器 420 关联的设备驱动器，例如适配器驱动器 440。适配器驱动器 440 可以含有处理队列（PQ）表后备存储 444，其含有适配器的 PQ 表中的条目的副本，例如系统映像 1 的处理队列表段 1 400。

当应用实例 X 432 进行基于文件名的 I/O 访问时，应用实例通过使用处理队列（PQ）门铃 436 来通知关联的适配器 420。举例来说，PQ 1 门铃 436 通知适配器 420: 在用于实现应用实例 X 432 与适配器 420 之间通信的处理队列集的发送队列 428 中存在存储工作请求。

来自 PQ 1 门铃 436 的数据提供了需要由加法器 422 添加到适配器 420 中的挂起工作请求的当前数目的工作请求数。也就是说，由中间件或应用实例生成的基于文件名的 I/O 请求发送可以包括存储在发送队列中作为工作队列条目的多个实际工作请求。PQ 1 门铃 436 标识出作为基于文件名的 I/O 请求的一部分的工作请求的数目。

提供工作请求数作为 PQ 计数字段，该 PQ 计数字段存储在与系统映像关联的相关处理队列表条目 PQ N 中，例如来自系统映像 1 的 PQ 段 1 的高速缓存的 PQ 条目 N 424。一旦完成存储工作请求，就将用于通知应用已经完成工作请求的消息添加到完成队列 450。

如图 4 中所示，来自系统映像 1 的 PQ 段 1 的高速缓存的 PQ 条目 N 424 包括 PQ 上下文信息，其包括 PQ 头地址、PQ 起始地址、PQ 结束地址、PQ 计数，以及附加的 PQ 上下文信息。PQ 起始地址字段存储应用的处理队列 428 中的第一工作队列条目的系统存储地址。PQ 结束地址字段存储与处理队列 428 最后的工作队列条目关联的最后的系统存储地址。PQ 头地址字段存储适配器打算处理的下一处理队列条目的系统存储地址。适配器在处理循环处理队列中的处理队列条目时更改 PQ 头地址。PQ 计数字段存储已由应用实例 432 递送但尚未由适配器处理的处理队列条目的数目。

接下来参照图 5，其依照本发明的一个示例性实施例，提供了对用于

保证用户空间中间件或应用实例所提交的基于文件名的 I/O 存储请求被授权引用该基于文件名的 I/O 存储请求中所引用的存储设备区域的示例性控制结构的描述。图 5 集中于通过确保只有与存储设备上的那些存储块相关联的应用实例才是仅有的可以访问那些存储块的应用实例来保护与应用实例关联的存储块。

如图 5 中所示，系统映像 1 500 托管应用实例 X 532。该应用实例 X 532 使用以上参照图 4 所描述的机制来实现基于文件名的 I/O 存储请求。该机制使用处理队列 528 将基于文件名的 I/O 存储请求作为工作队列条目（WQEs）（例如 WQE 536）提交给所期望的物理 I/O 适配器，例如适配器 516。将基于文件名的 I/O 存储工作请求放入发送队列 528，发送队列 528 是作为与应用实例 X 532 和适配器 516 相关联的处理队列集的一部分的工作队列。适配器 516 上的处理队列上下文 517（例如在来自系统映像（SI）1 的 PQ 段 1 的高速缓存的 PQ 条目 N 524 中的处理队列上下文）含有保护域字段 518。

当应用 X 532 提交诸如基于文件名的 I/O 存储请求 536 这样的基于文件名的 I/O 存储请求时，部分请求将含有 FN_Key 538 和 FE_Key 539。取决于本发明的特定实现，FN_Key 538 被系统映像 500 用作进入文件名保护表（FNPT）510 的索引或者被适配器 516 用作进入适配器 516 的 FNPT 高速缓存 530 中的高速缓存的文件名保护表段 535 的索引。举例来说，FN_Key 538 可以是进入 FNPT 510 或高速缓存的 FNPT 段 535 的偏移，其中 FN_Key 538 准许标识 FNPT 510 或高速缓存的 FNPT 段 535 中与文件名 I/O 存储工作请求的目标文件的文件名相对应的特定条目。

FE_Key 539 由系统映像 500 用于访问文件扩展保护表（FEPT）502 的区段中由对应于 FN_Key 538 的 FNPT 条目所引用的特定条目。可选地，在优选实施例中，FE_Key 539 可以由适配器 516 用来访问 FEPT 的高速缓存段 545 中与 FN_Key 538 所标识的高速缓存的 FNPT 条目相对应的特定条目。

只有在适配器的文件名保护表高速缓存 530 和文件扩展保护表段高速

缓存 540 中分别存在所需区段的情况下，才进行对高速缓存的文件名保护表段 535 和高速缓存的文件扩展保护表段 545 的访问。举例来说，如果适配器的高速缓存 530 和 540 内不存在所需区段，则可能需要将所需要的文件名和/或文件扩展保护表段从系统映像 500 加载到适配器的高速缓存 530 和 540 中。可选地，FE_Key 检查逻辑 519 可以在系统映像 500 中直接访问文件名和/或文件扩展保护表段，例如，文件名保护表段 511 和/或文件扩展保护表段 1 504。

当应用实例或中间件请求分配操作系统的文件系统中的文件时，生成 FN_Key 和 FE_Key。也就是说，操作系统将分配适当的存储设备块来存储文件并且将为该文件在 FNPT 和 FEPT 中生成条目。作为生成这些条目的一部分，操作系统会将 FN_Key 和 FE_Key 分派给表中的条目并且将这些关键字报告回给发出请求的应用实例、中间件等。另外，当将新的文件扩展添加到现有文件时（例如，当文件的大小增加到超过已分配的存储设备部分时），可以进一步生成和分派文件扩展保护表条目以及由此的 FE_Key。应用实例、中间件等然后可以在提交文件名 I/O 请求时将这些关键字用作与应用实例、中间件等关联的处理队列中的工作队列条目。

如以上所提及的，FN_Key 和 FE_Key 用于在文件名保护表 510 和文件扩展保护表 502，或者在高速缓存的文件名保护表段 535 和高速缓存的文件扩展保护表 545 中查找分别与 FN_Key 和 FE_Key 关联的条目。举例来说，FN_Key 可以具有用于与存储在文件名保护表 510/530 的条目中的 FN_Key 实例相比较的值。类似地，FE_Key 可以具有用于与 FEPT 段 504/540 中的文件扩展保护表条目的关键字实例字段相比较以便标识与 FE_Key 匹配的条目的值。可选地，举例来说，FN_Key 和 FE_Key 可以是在表中用于从段起始地址偏移至表中的特定条目的偏移。

在本发明的优选实施例中，适配器 516 中的 FE_Key 检查逻辑 519 用于基于如上所述的 FN_Key 和 FE_Key 来进行对 FNPT 和 FEPT 中的条目的查找。此后，FE_Key 检查逻辑 519 进行保护域检查以检验来自适配器 516 中的 PQ 上下文 524 的保护域与由基于文件名的 I/O 存储请求 536 中的

FN_Key 和 FE_Key 所指向的保护表条目 N 520 中的保护域相匹配。未通过 FE_Key 检查逻辑 519 的任何检查均导致操作错误地完成。在这样的情况下，操作系统（例如系统映像 1 500）可以拆卸在操作中进行传递的中间件或应用实例（例如应用实例 X 532），或者可以采取不那么激烈的手段，例如返回带有错误完成的操作。

假设通过了先前在上面讨论的所有检查，那么由适配器 516 处理基于文件名的 I/O 存储请求，以便向/从由高速缓存的 LBA 表段 550（或者可选地，与系统映像 500 关联的 LBA 表段 570）中与 FEPT 段中的文件扩展保护表条目相对应的条目所引用的物理存储设备 560（例如硬盘）的线性块地址读、写或读/写数据。

本发明使用 FNPT、FEPT 和 LBA 表来管理“用户空间”和“用户空间外”的基于文件名的 I/O 操作。用户空间是用于运行用户应用的系统存储器部分。在“用户空间”中进行的基于文件名的 I/O 操作包括与创建、修改、查询和删除 FNPT、FEPT 和 LBA 表条目有关的操作、应用对工作队列请求的提交和处理、系统映像所进行的其它 I/O 操作，等等。就本发明而言，在“用户空间外”进行的基于文件名的 I/O 操作包括在 I/O 适配器 516 中进行以便促进验证和执行对诸如物理存储设备 560 的物理存储设备的 I/O 请求的操作。

在应用实例与物理 I/O 适配器之间基于文件名的直接 I/O 操作期间，上述数据结构和机制用于控制诸如应用 X 532 的应用对存储设备 560 的各部分的访问。以下描述提供了与依照先前的上述机制来分配资源、创建工作队列条目和处理完成队列条目的方法有关的细节。

图 6 是依照本发明的示例性实施例概括了用于处理用户空间操作的调用的示例性操作的流程图。在本发明的示例性实施例中，图 6 中所概括的操作是由系统映像或操作系统响应于用户空间操作的调用而进行的。虽然示例性实施例使得这些操作在系统映像或操作系统中进行，但是本发明并不限于此。相反，举例来说，可以在用户空间应用、管理体等中进行该操作。

应当理解，可以通过计算机程序指令实现图 6 中流程图说明以及此后描述的后续附图中的流程图说明的块的组合。可以将这些计算机程序指令提供给处理器或其它可编程数据处理装置来产生机器，从而使得在处理器或其它可编程数据处理装置上执行的指令创建用于实现流程图块中所指定的功能的装置。还可以将这些计算机程序指令存储在可以指导处理器或其它可编程数据处理装置以特定方式运行的计算机可读存储器或存储介质中，从而使得存储在计算机可读存储器或存储介质中的指令产生包括实现流程图块中所指定的功能的指令装置在内的制品。

相应地，流程图说明的块支持用于实现指定功能的装置的组合、用于实现指定功能的步骤以及用于实现指定功能的程序指令装置的组合。还应该理解，流程图说明中的每个块以及流程图说明中块的组合可以由实现指定功能或步骤的基于硬件的专用计算机系统或者由专用硬件和计算机指令的组合来实现。

如图 6 中所示，操作开始于调用用户空间操作（步骤 610）。例如，可以借助于用户管理接口、自动脚本/工作流等来实现调用。可以通过应用实例、系统映像等进行调用。可以实现此类调用的用户管理接口的一个例子是高级交互执行（AIX）操作系统中的原始模式（raw mode）I/O。其它操作系统可以具有类似的接口。调用该用户管理接口用于像创建卷、撤销卷之类的管理操作以及诸如读或写的功能操作。

对关于正调用的用户空间操作是否是资源管理操作进行确定（步骤 615）。操作系统在此限制对基础硬件的访问，从而使得应用不能访问与其它应用关联的资源。因而，资源管理操作是必须由操作系统实现的操作，因为没有其它备选方案用于将应用的访问限于其拥有的资源。这样的操作的例子包括创建卷、查询卷、撤销卷。非资源管理操作是这样的操作，即在该操作下，通过本发明的机制，物理适配器可以将应用的访问限于其拥有的资源。非资源管理操作的例子是读和写操作。

如果操作不是资源管理操作，那么该操作便是处理队列操作。因此，对关于操作是否用于工作队列处理（例如，与发送队列中的条目关联的处

理）进行确定（步骤 620）。如果是的话，则调用工作队列条目插入过程来创建工作队列条目（步骤 625）。如先前所讨论的以及此后在图 7 中概括的，该工作队列条目插入过程用于将工作请求提交给 I/O 适配器。

如果操作并非用于工作队列处理，那么调用完成队列条目检索过程（步骤 630）。如此后较为详细描述的，完成队列条目检索过程用于为已由物理 I/O 适配器完成的工作请求而从物理 I/O 适配器中检索完成队列条目。

如果用户空间操作是资源管理操作（步骤 615），那么对关于该操作是否是资源查询操作进行确定（步骤 640）。如果该操作是资源查询操作，那么系统映像/操作系统从物理 I/O 适配器检索资源属性并将结果返回给调用用户空间操作的元件，例如，系统映像或应用实例（步骤 645）。如此后较为详细讨论的，举例来说，该操作用于从 LBA 表条目和文件扩展保护表条目获取属性信息。

如果操作不是资源查询操作，那么对关于该操作是否是资源创建操作进行确定（步骤 650）。如果该操作是资源创建操作，则对关于物理 I/O 适配器是否具有可用于分配给调用用户空间操作的元件的资源进行确定（步骤 660）。举例来说，如以上所讨论的，适配器保护表中的每个文件扩展保护表条目均含有 LBA 表大小、扇区大小和长度。这些参数可以限制可供适配器用于分配的资源数。因而，物理 I/O 适配器可以确定没有足够的资源可用于分配给调用用户空间操作的元件。

如果存在可供分配的足够资源，那么在物理 I/O 适配器上分配这些资源，并且物理 I/O 适配器将该分配的结果返回给进行调用的元件（步骤 665）。如果没有可供分配的足够资源，那么可以生成错误记录并将其返回给调用用户空间操作的元件（步骤 670）。

如果操作不是资源创建操作（步骤 650），那么对关于该操作是否是资源撤销操作（文中也称为“删除”或“解除分配”操作）进行确定（步骤 675）。如果该操作是资源撤销操作，那么撤销物理 I/O 适配器上的资源并将操作结果返回给调用用户空间操作的元件（步骤 680）。如果操作不是资源撤销操作，那么该操作是资源修改操作并在物理 I/O 适配器上修

改指定资源的属性（步骤685）。该操作然后终止。

图7是概括了当所调用的用户空间操作是要求进行生成和处理或者工作队列元素的工作队列操作时本发明的示例性操作的流程图。举例来说，图7中所示出的操作对应于图6中的步骤625。

如图7所示，当应用实例将一个或多个工作队列条目添加到与应用实例和适配器关联的处理队列集的工作队列（例如发送队列）时，操作开始（步骤710）。如以上所讨论的，该工作队列条目包括FN_Key、FE_Key、保护域、将要进行的I/O操作的标识符，以及视情况的逻辑卷号和/或SCSI LUN。

将处理队列门铃消息从应用实例发送至物理 I/O 适配器以通知物理I/O适配器最新递送的工作请求（步骤715）。在本发明的一个示例性实施例中，发送处理队列门铃消息涉及对与工作队列关联的门铃地址进行编程的 I/O 写入。如以上所讨论的，门铃消息用于将附加工作请求添加到物理I/O适配器的高速缓存的处理队列条目中的处理队列计数。

此后，物理 I/O 适配器根据文件保护表条目（即文件名保护表条目和文件扩展保护表条目）中存储的数据，对工作队列条目中所存储的信息进行验证检查（步骤720）。如以上所讨论的，这些检查可以包括基于FN_Key来查找文件名保护表中的条目以便由此标识文件扩展保护表的区段，并且然后基于FE_Key来查找所标识的区段内的文件扩展保护表条目。该检查还可以进一步包括，例如，检查所标识的文件扩展保护表条目中的保护域、逻辑卷号、SCSI标识号、SCSI逻辑单元号等与工作队列条目中的类似值之间的匹配。此后将较为详细地描述这些检查。

对关于是否成功完成了所有检查进行确定（步骤 725）。如果成功完成了所有检查，则物理I/O适配器使用线性块地址（LBA）表将所标识的文件扩展保护表条目中引用的文件转换成LBA（例如，借助于LBA表指针），并且进行 LBA 包容性（containment）检查（步骤 730）。因为应用实例与存储设备在不同的空间中操作，所以通过应用实例生成的基于文件名的I/O存储请求所引用的地址可能不同于存储设备的实际物理地址。

LBA表条目为分配给特定文件的存储设备提供有关实际物理LBA的信息，如从对应的文件扩展保护表条目确定的。因而，可以借助于文件名保护表和文件扩展保护表，实现在基于文件名的I/O存储请求（以及由此的工作队列条目）中所引用的文件与LBA表中所引用的LBA之间的映射，以便确定基于文件名的I/O操作将要导向的实际物理LBA。

举例来说，文件扩展保护表条目中的LBA表指针可以用于访问LBA表中与文件扩展保护表条目相对应的一个或多个条目。从与文件扩展保护表条目相对应的LBA表条目，可以标识与文件扩展保护表条目相对应的盘线性块地址（LBA）的范围。然后可以使用这些LBA将工作队列条目中所引用的文件映射到物理存储设备的LBA。

返回到图7，对关于是否成功完成了LBA包容性检查进行确定（步骤735）。这些LBA包容性检查是这样的检查，即其确定与基于文件名的I/O操作（以及由此的工作队列条目）中所引用的文件相对应的映射LBA是否属于如在对应的LBA表条目中所标识的分配给应用实例的LBA。举例来说，如果应用实例尝试访问未分配给该应用实例的存储设备部分，那么至少一个LBA包容性检查会失败。如果未成功完成验证检查或包容性检查中的任何一个，则生成错误结果（步骤740）。

如果成功完成了验证和包容性检查，则物理I/O适配器将工作队列条目标记为有效（步骤750）并且实现与工作队列条目关联的所有功能，例如读、写、读/写（步骤755）。此后，或者在步骤740中生成错误结果之后，物理I/O适配器创建与工作队列条目关联的完成队列条目，并且进行直接存储器访问（DMA）操作以便将完成队列条目发送给应用实例（步骤760）。

然后对关于是否请求了完成队列事件进行确定（步骤765）。如果是的话，则物理I/O适配器生成完成队列事件（步骤770）并且终止操作。也就是说，在递送到处理队列的发送和接收队列的工作请求完成之后，将完成消息放入完成队列并且如果应用对其进行请求，则可以生成事件。

在图7中重要的是要注意，在步骤710和715之后，系统映像或操作

系统不参与工作队列条目的处理。相反，物理 I/O 适配器实现所有必需的操作来进行有效性和包容性检查，实现与工作队列条目关联的功能，生成完成队列条目，以及将完成队列条目发送至主机。因而，通过本发明，可以避免 I/O 操作期间在已知系统中经历的多次任务切换，如以上在本发明的背景技术中所描述的，因为在 I/O 操作已经由操作系统或系统映像提交之后，在该 I/O 操作的实际检查和处理期间不必涉及操作系统或系统映像。仅再次利用操作系统或系统映像来检索与所处理的工作队列条目关联的完成队列条目，以及将该完成队列条目传递给应用。

图 8 中说明了用于确定工作队列条目是否有效以及是否可由物理 I/O 适配器处理而进行的示例性验证检查。举例来说，图 8 中所概括的验证检查操作可以对应于图 7 中的步骤 720 和 725。

如图 8 中所示，操作开始于从工作队列（例如发送队列）为基于文件名的 I/O 操作检索下一工作队列条目（步骤 810）。然后根据高速缓存的或系统映像常驻文件名保护表条目和文件扩展保护表条目检查该工作队列条目，以便确定是否可以进行对应的基于文件名的 I/O 操作。首先，使用工作队列条目中的 FN_Key 来查找对应于 FN_Key 的文件名保护表条目（步骤 812）。文件名保护表条目包括指向与文件名保护表条目相对应的文件扩展保护表的区段的起始地址的指针（步骤 814）。然后使用工作队列条目中的 FE_Key 来标识在所标识的文件扩展保护表的区段中的条目（步骤 816）。然后检索所标识的文件扩展保护表条目的字段的数据，用于认证应用实例对于与工作队列条目中由 FN_Key 标识的文件相对应的存储设备部分的访问（步骤 820）。

取决于本发明的特定实现，以上对于文件名保护表和文件扩展保护表中的条目的标识可以以多种不同的方式来实现。在一个例子中，FN_Key 和 FE_Key 是表中从所标识的起始地址的偏移。在其它例子中，工作队列条目中的 FN_Key 和 FE_Key 具有这样的值，即将该值与文件名保护表和文件扩展保护表的条目中的关键字实例相比较，从而标识具有匹配值的条目。在不背离本发明的范围的情况下，可以使用用于标识每个表中的特定

条目的其它机制。

在从所标识的文件扩展保护表条目中检索数据之后，对关于是否已经通过以上查找操作找到有效的文件扩展保护表条目进行确定（步骤 830）。如果没有，则生成并返回错误结果（步骤 840）。如以上所提及的，这可以通过查看文件扩展保护表的访问控制中的有效/无效比特来确定是否已将该比特设置成有效值来实现。此外，如果文件扩展保护表条目无效，则错误结果可以是，例如，拆卸在生成工作队列条目的工作请求中进行传递的中间件或应用实例，或者可以采取不那么激烈的手段，例如返回带有错误完成的操作。

如果已经找到了有效的文件扩展保护表条目，那么检查关联的文件扩展保护表条目是否支持将要结合工作队列条目而进行的 I/O 操作（步骤 850）。举例来说，将适配器保护表条目的访问控制与工作队列条目中的 I/O 操作标识符进行比较，以确定文件扩展保护表条目是否指示可否进行 I/O 操作。

如果基于文件扩展保护表条目中的访问控制的设置不能进行 I/O 操作，那么该操作生成并返回错误结果（步骤 840）。如果按照文件扩展保护表条目的指示可以进行 I/O 操作，那么对关于工作队列条目的保护域是否对应于文件扩展保护表条目的保护域进行确定（步骤 860）。如果保护域不匹配，那么该操作生成并返回错误结果（步骤 840）。

如果保护域匹配，那么可以对文件扩展保护表条目中的附加信息进行附加检查，并且可以对关于这些检查是否成功进行确定（步骤 870）。如以上所提及的，这些附加检查可以包括，例如，检查文件扩展保护表的文件名字段以确定文件名与通过工作队列条目传递的文件名是否匹配（如果工作队列条目中存在文件名的话）。类似地，如果工作队列条目具有关联的 LV 号标识符和/或 SCSI LUN 标识符，那么可以对此信息进行附加检查。与先前的检查一样，如果这些检查的结果是工作队列条目与适配器保护表条目不匹配，那么生成并返回错误结果（步骤 840）。应当理解，步骤 870 是任选的并且不是本发明的所有实施例中都出现此步骤。

如果通过了所有检查，则将工作队列条目预先标记为可由物理 I/O 适配器处理的有效工作队列条目（步骤 880）。此有效性的预先标记仅意味着工作队列条目已通过第一组有效性检查。如上所述，工作队列条目在由物理 I/O 适配器处理之前还必须通过包容性检查。在步骤 880 之后，关于有效性检查而言操作结束，但是，如图 7 中所示，操作继续整个操作中的步骤 730 或 740。

应当理解，虽然图 8 说明了为了处理基于文件名的 I/O 操作所进行的一系列检查，但是本发明并不限于所描绘的特定检查系列。相反，图 8 中所概括的操作仅是示例性的并且可以在不背离本发明的范围的情况下做出许多修改。举例来说，可以根据需要修改实现各种有效性检查的顺序，以便在不同操作顺序的情况下实现不同系列的有效性检查。此外，除了图 8 中所示出的有效性检查之外，或者代替图 8 中所示出的有效性检查，还可以将其它有效性检查用于本发明的示例性实施例。

图 9 是概括了当所调用的用户空间操作是完成队列检索过程操作时本发明的示例性操作的流程图。例如，图 9 中所示出的操作对应于图 6 中的步骤 630。

如图 9 中所示，操作开始于轮询完成队列以确定是否存在准备要处理的任何完成队列条目（步骤 910）。对关于是否任何完成队列条目准备要被处理进行确定（步骤 920）。如果否，则将空结果返回给用户空间应用（步骤 930）。如果存在准备要被处理的完成队列条目，则将下一完成队列条目返回给用户空间应用（步骤 940）并且操作终止。

应当注意，图 6 至图 9 中所描述的以上操作适用于非虚拟和虚拟系统这二者中基于文件名的直接 I/O 操作。在虚拟系统中，仅有的附加可能是由操作系统或系统映像调用管理体或其它虚拟化机制以便在资源创建、修改、查询或删除期间帮助维护相连范围的虚拟 LBA。

如以上所讨论的，关于图 6 中所概括的操作，本发明的机制涉及确定所调用的用户空间操作是否针对创建、查询、修改或删除对于应用与适配器之间基于文件名的直接 I/O 的资源分配。基于这些确定，操作系统或系

统映像可以调用用于创建、修改、查询或删除资源分配的各种操作。现将参照图 10 至图 13 并根据本发明的文件名保护表、文件扩展保护表和线性块地址表来描述这些操作中的每一个。应当理解，可以为虚拟和非虚拟系统实现图 10 至图 13 中所示出的操作。因而，举例来说，可以基于逻辑卷、SCSI 标识符或 SCSI 逻辑单元号来进行操作以便创建、修改、查询和删除或撤销文件名、文件扩展和 LBA 条目。

图 10 是依照本发明的示例性实施例概括了当在 LBA 表中创建 LBA 条目时本发明的示例性操作的流程图。举例来说，图 10 中所概括的操作对应于图 6 中的步骤 665。

如图 10 中所示，操作开始于接收请求创建一个或多个文件保护表条目的用户空间操作，即分配与特定文件关联的一组 LBA 并且通过该组 LBA 使得应用实例和/或系统映像的直接 I/O 访问成为可能（步骤 1010）。响应于接收到创建用户空间操作，操作系统或系统映像使用物理 I/O 适配器的存储管理接口来请求物理 I/O 适配器创建一个或多个文件保护表条目（步骤 1020）。可以用多种不同的方式来实现存储管理接口。举例来说，存储管理接口可以是在其中可以将资源管理操作从系统映像传递至适配器的队列。

然后对关于 I/O 适配器是否具有足够的资源来完成请求进行确定（1030）。例如，I/O 适配器可以检查文件保护表以确定条目是否可用，并且如果否，则确定是否可以创建另一文件保护表段。如果这些确定中的任一确定是肯定的，即文件保护表可以接纳分配，那么步骤 1030 中的确定是：I/O 适配器具有足够的资源；否则确定是：I/O 适配器没有足够的资源可用于分配。

如果有足够的资源可用于将所请求的文件和对应的 LBA 存储空间分配给应用实例，那么创建适当的文件名保护表、文件扩展保护表和 LBA 条目（步骤 1040）。LBA 条目标识映射到应用实例所请求的文件的物理存储设备 LBA。文件扩展保护表条目为分配给文件的 LBA 标识访问控制、域保护、文件名等。举例来说，可以从请求文件分配的应用实例和用于提交

文件分配请求的应用实例的处理队列（例如保护域）获取该信息。

物理 I/O 适配器然后将创建用户空间操作的结果返回给应用实例（步骤 1050）。该结果可以包括，例如，为文件名保护表生成的 FN_Key 和 FE_Key 以及为文件创建的文件扩展保护表条目。另外，物理 I/O 适配器还可以通知应用实例可以由应用实例用来对物理 I/O 适配器进行基于文件名的直接 I/O 的 LBA。

如果没有足够的资源来分配所请求的文件，那么物理 I/O 适配器不创建文件保护表条目（步骤 1060）。物理 I/O 适配器然后将所得到的错误作为创建用户空间操作的结果返回给应用实例（步骤 1050）。操作然后终止。

图 11 是概括了当处理作为资源修改操作的用户空间操作时本发明的示例性操作的流程图。举例来说，图 11 中所概括的操作可以对应于图 6 的步骤 685。

如图 11 中所示，操作开始于从应用实例、系统映像等接收请求修改一个或多个文件保护表条目的用户空间操作（步骤 1110）。系统映像然后使用物理 I/O 适配器的存储管理接口来请求物理适配器修改与应用实例或系统映像所标识的文件名关联的一个或多个文件保护表条目（步骤 1120）。对关于物理 I/O 适配器是否具有足够的资源来完成修改请求进行确定（步骤 1130）。

文件名保护表条目具有固定的字段集，并且因而，在已经创建文件名保护表条目之后，资源不足的情况不会用于该文件名保护表条目。文件扩展保护表在附加文件扩展被创建（即向特定文件分配 LBA）时向其添加条目，并且因而受到 LBA 表段大小的限制。可以向 LBA 表段添加附加条目，并且如前所述，存在 LBA 表段可能用尽资源的情况。如果物理 I/O 适配器没有足够的资源可用于完成修改请求，则物理 I/O 适配器会将错误消息返回给应用实例，指示无法完成修改（步骤 1140）。

如果存在足够的可用资源，则对关于正在被修改的文件保护表条目上是否存在任何有效 I/O 事务进行确定（步骤 1150）。如果正在被修改的文件保护表条目上存在有效 I/O 事务，则物理 I/O 适配器启动计时器并且等

待达到静点（quiescent point）（步骤 1160）。静点是指在该点处，被修改的文件保护表条目上没有有效 I/O 事务。该检查以及等待静点是必要的，以便不会对文件保护表条目做出修改，该修改将导致系统损坏，因为有效 I/O 事务操作在先前的文件保护表条目属性下。

然后对关于是否在计时器超时之前达到静点进行确定（步骤 1170）。如果否，则将错误消息返回给应用实例，指示无法完成修改（步骤 1140）。如果在计时器超时之前达到静点，则物理 I/O 适配器修改文件保护表条目的属性（步骤 1180），并且将经修改资源的属性返回给应用实例（步骤 1190）。操作然后终止。

图 12 是概括了当处理查询用户空间操作时本发明的示例性操作的流程图。举例来说，图 12 中所概括的操作可以对应于图 6 的步骤 645。

如图 12 中所示，操作开始于从应用实例、系统映像等接收请求查询文件保护表条目的属性的用户空间操作（步骤 1210）。响应于接收到该用户空间操作，系统映像使用适配器的存储管理接口来请求物理 I/O 适配器查询一个或多个文件保护表条目（步骤 1220）。物理 I/O 适配器然后将文件保护表条目的属性返回给应用实例（步骤 1230）。

图 13 是概括了当处理撤销或删除用户空间操作时本发明的示例性操作的流程图。举例来说，图 13 中所示出的操作对应于图 6 的步骤 680。例如，如果操作系统或系统映像允许由中间件或应用实例减少逻辑卷，则可以撤销或删除文件保护表条目。举例来说，该减少然后可以导致撤销或删除 LBA 表条目、文件扩展保护表，以及甚至是文件名保护表条目。

如图 13 中所示，操作开始于接收撤销或删除用户空间操作（步骤 1310）。响应于接收到撤销或删除用户空间操作，系统映像使用物理 I/O 适配器的存储管理接口来请求物理 I/O 适配器撤销或删除一个或多个文件保护表条目（步骤 1320）。对关于 I/O 事务在正被删除或撤销的文件保护表条目上是否有效进行确定（步骤 1330）。

如果 I/O 事务在文件保护表条目上有效，则物理 I/O 适配器启动计时器并且等待达到静点（步骤 1340）。然后对关于是否在计时器超时之前达

到静点进行确定（步骤 1350）。如果否，则物理 I/O 适配器创建错误结果并将错误结果返回给应用实例（步骤 1360）。如果在计时器超时之前达到静点，或者如果文件保护表条目上没有有效 I/O 事务，则物理 I/O 适配器撤销或删除现有的文件保护表条目（步骤 1370），并且将结果返回给应用实例（步骤 1380）。当通过操作系统或系统映像撤销或删除文件保护表条目时，从文件保护表段中移除条目，并且释放盘中的 LBA 且使之可供其它应用使用。

应当注意，上述流程图中所概括的操作参照了在一个或多个文件保护表条目上进行的操作。当进行这样的操作时，可能还要求改变到其它的文件保护表条目。例如，在创建和修改操作期间，中间件或应用实例可以通过在 LBA 表中创建附加条目来增加与特定文件关联的 LBA 数。这又要求文件扩展保护表中的附加条目指向新的 LBA 表条目。类似地，在删除或撤销操作期间，操作系统或系统映像撤销一个或多个 LBA 表条目或区段，并且然后将关联的文件扩展保护表条目的访问控制字段设置成无效。

因而，利用本发明，检查所调用的用户空间操作以查看该操作是资源查询操作、资源创建操作、资源撤销操作、资源修改操作、工作队列操作还是完成队列操作。基于该确定，实现用于查询、创建、撤销和修改资源分配、工作队列条目以及完成队列条目的对应操作。因而，如果应用要求资源以便进行基于文件名的直接 I/O 操作，需要修改资源分配以便进行此类直接 I/O 操作，或者需要撤销资源分配，则本发明提供了实现这些目的的机制。另外，应用可以提交用于处理的工作队列条目，并且处理完成队列条目以获取与已由物理 I/O 适配器完成处理的工作队列条目有关的信息。以这样的方式，通过本发明的机制可以管理基于文件名的直接 I/O 操作。

此外，如上述示例性实施例所说明的，本发明提供了用于处理基于文件名的 I/O 操作的多个数据结构和机制。这些数据结构和机制提供了使用文件保护表访问控制来处理队列到线性块地址转换。该机制确保只有与文件（以及因而对应的存储设备部分）关联的应用才可以实际访问存储设备

部分。包括关键字检查和保护域检查在内的多个验证检查用于维护此安全级别。这些检查确保应用实例正在访问有效的适配器保护表条目，并且应用有权访问与有效的文件扩展保护表条目关联的存储设备部分。

应当注意，虽然本发明的示例性实施例的以上机制利用操作系统或系统映像来进行有关创建和管理文件保护表条目的许多操作，但是这些操作通常并不随适配器所处理的每个工作请求而实现。也就是说，操作系统或系统映像仅参与建立文件保护表条目和注册具有关联的文件/LBA 的应用实例/中间件。不需要操作系统或系统映像来处理中间件或应用实例所提交的每个实际工作请求，因为应用和适配器可以使用文件保护表和上述机制来处理工作请求。因此，本发明消除了如以上在本发明的背景技术中所解释的现有技术机制所需的上下文切换及其关联的开销。

重要的是要注意到，虽然已经在全功能数据处理系统的环境中描述了本发明，但是本领域的普通技术人员应该理解，能够以指令的计算机可读介质的形式和各种形式来分布本发明的诸过程，并且本发明可等同地应用而与实际用于实现分布的信号承载介质的特定类型无关。计算机可读介质的例子包括可记录型介质，例如软盘、硬盘驱动、RAM、CD-ROM、DVD-ROM，以及传输型介质，例如数字和模拟通信链路、使用例如像射频和光波传输之类的传输形式的有线或无线通信链路。计算机可读介质可以采取编码格式的形式，可以对其解码以便在特定的数据处理系统中实际使用。

已经出于说明和描述的目的给出了本发明的描述，且并不旨在以所公开的形式穷举或限制本发明。对本领域的普通技术人员来说，很多修改和变形将是显而易见的。选择和描述实施例是为了最好地解释本发明的原理、实际应用，以及使本领域的普通技术人员能够针对适于预期的特定用途的各种实施例以及各种修改来理解本发明。

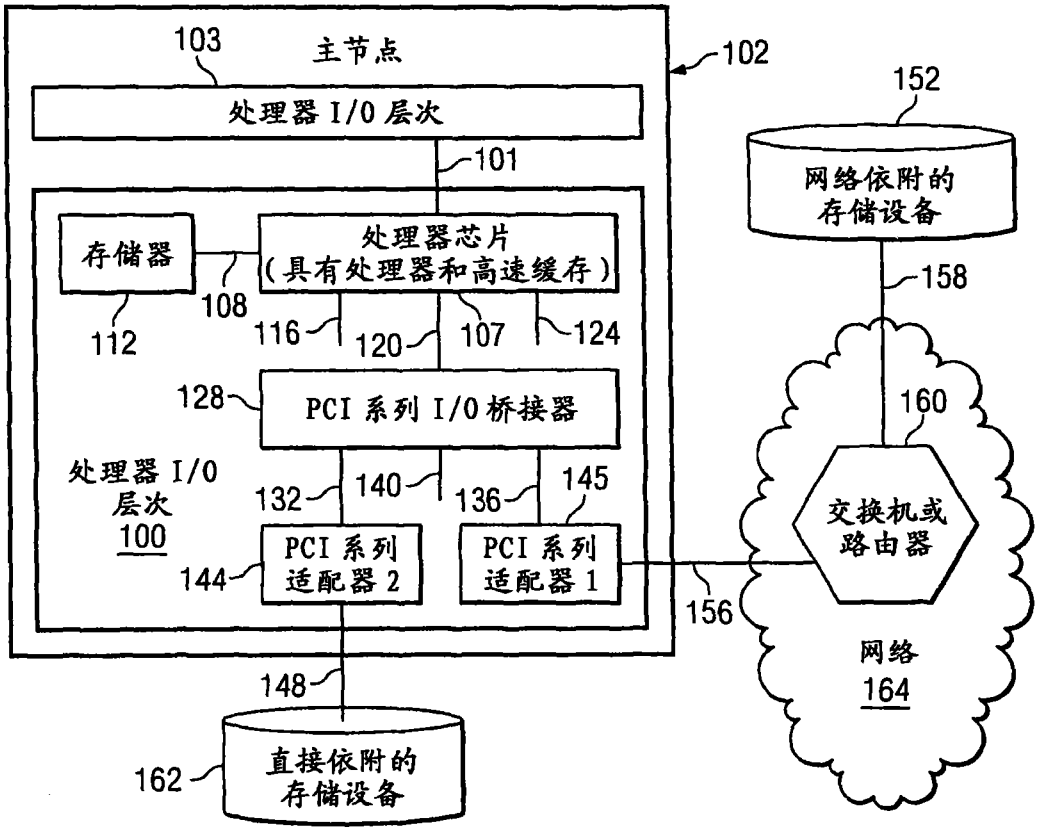

103
主节点
102
处理器 I/O 层次
101
152
网络依附的
存储设备
存储器
处理器芯片
(具有处理器和高速缓存)
158
108
112
116
120
107
124
128
PCI 系列 I/O 桥接器
160
处理器 I/O
层次
100
132
140
136
145
交换机或
路由器
PCI 系列
适配器 2
PCI 系列
适配器 1
144
156
网络
164
148
直接依附的
存储设备
162

图 1

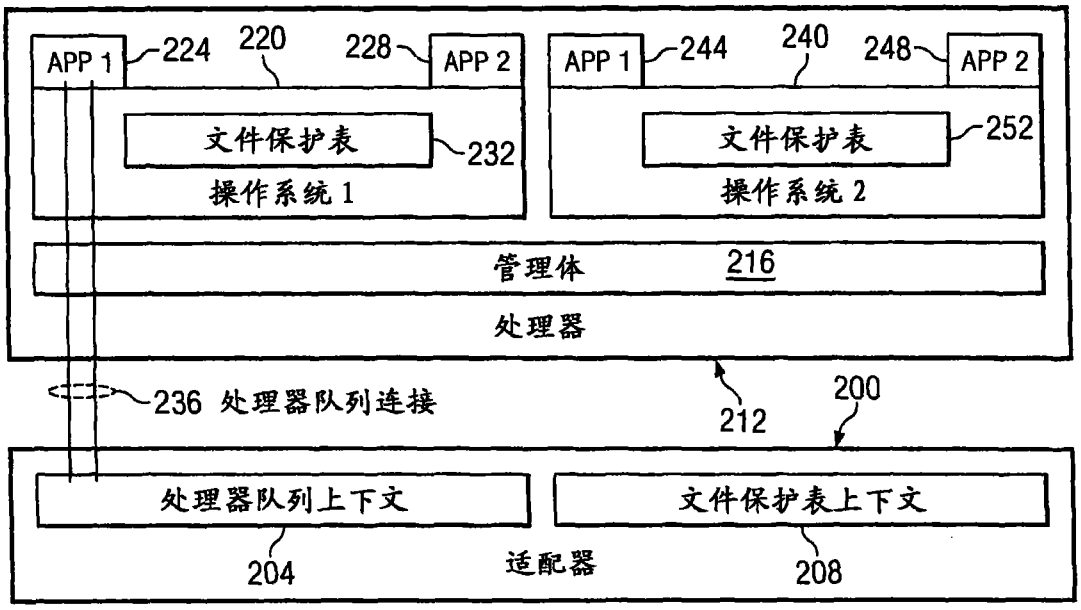

APP 1
224
220
228
APP 2
APP 1
244
240
248
APP 2
文件保护表
232
文件保护表
252
操作系统 1
操作系统 2
管理体
216
处理器
236 处理器队列连接
212
200
处理器队列上下文
文件保护表上下文
204
适配器
208

图 2

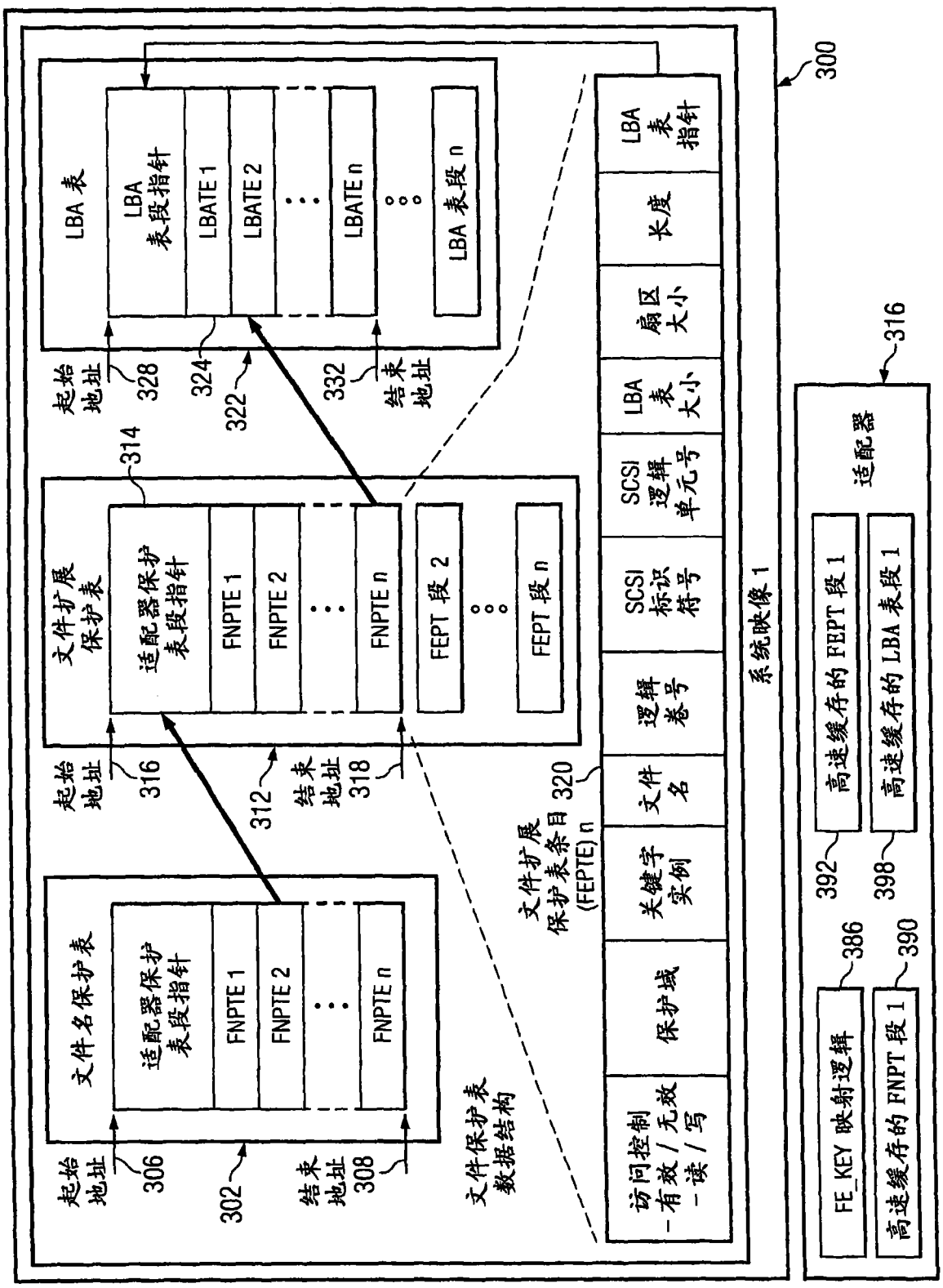

300
系统映像 1
LBA 表
LBA 表段指针
LBATE 1
LBATE 2
LBATE n
LBA 表段 n
起始地址
328
324
322
332
结束地址
文件扩展保护表
适配器保护表段指针
FNPTE 1
FNPTE 2
FNPTE n
FEPT 段 2
FEPT 段 n
314
起始地址
316
312
结束地址
318
文件扩展保护表条目 (FEPTE) n
320
访问控制 - 有效 / 无效 - 读 / 写
保护域
关键字实例
文件名
逻辑卷号
SCSI 标识符号
SCSI 逻辑单元号
LBA 表大小
扇区大小
长度
LBA 表指针
文件名保护表
适配器保护表段指针
FNPTE 1
FNPTE 2
FNPTE n
起始地址
306
302
结束地址
308
文件保护表数据结构
适配器
316
高速缓存的 FEPT 段 1
392
高速缓存的 LBA 表段 1
398
FE_KEY 映射逻辑
386
高速缓存的 FNPT 段 1
390

图 3

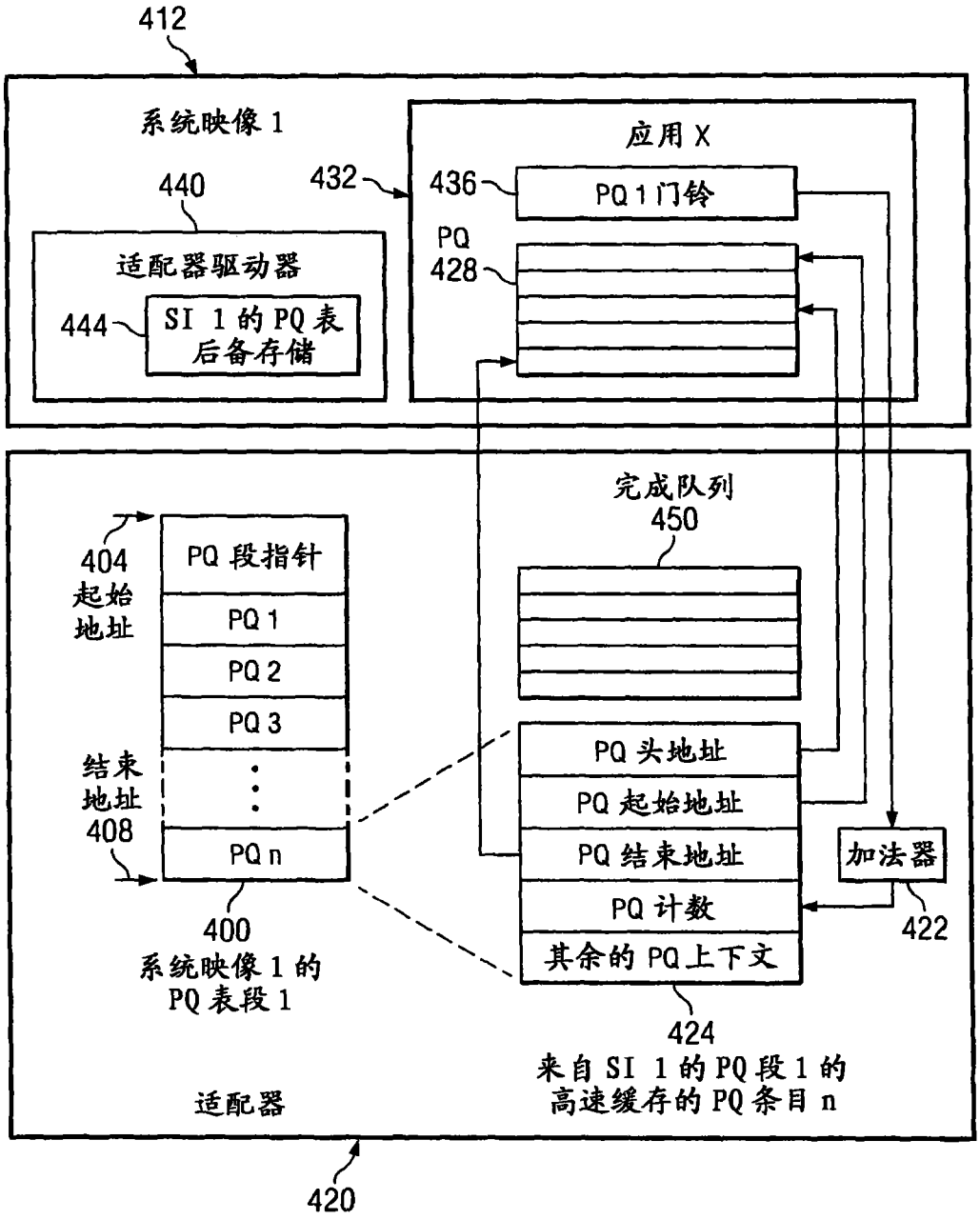

412
系统映像 1
440
适配器驱动器
444
SI 1 的 PQ 表
后备存储
应用 X
432
436
PQ 1 门铃
PQ
428
完成队列
450
404
起始
地址
PQ 段指针
PQ 1
PQ 2
PQ 3
结束
地址
408
PQ n
400
系统映像 1 的
PQ 表段 1
PQ 头地址
PQ 起始地址
PQ 结束地址
PQ 计数
其余的 PQ 上下文
加法器
422
424
来自 SI 1 的 PQ 段 1 的
高速缓存的 PQ 条目 n
适配器
420

图 4

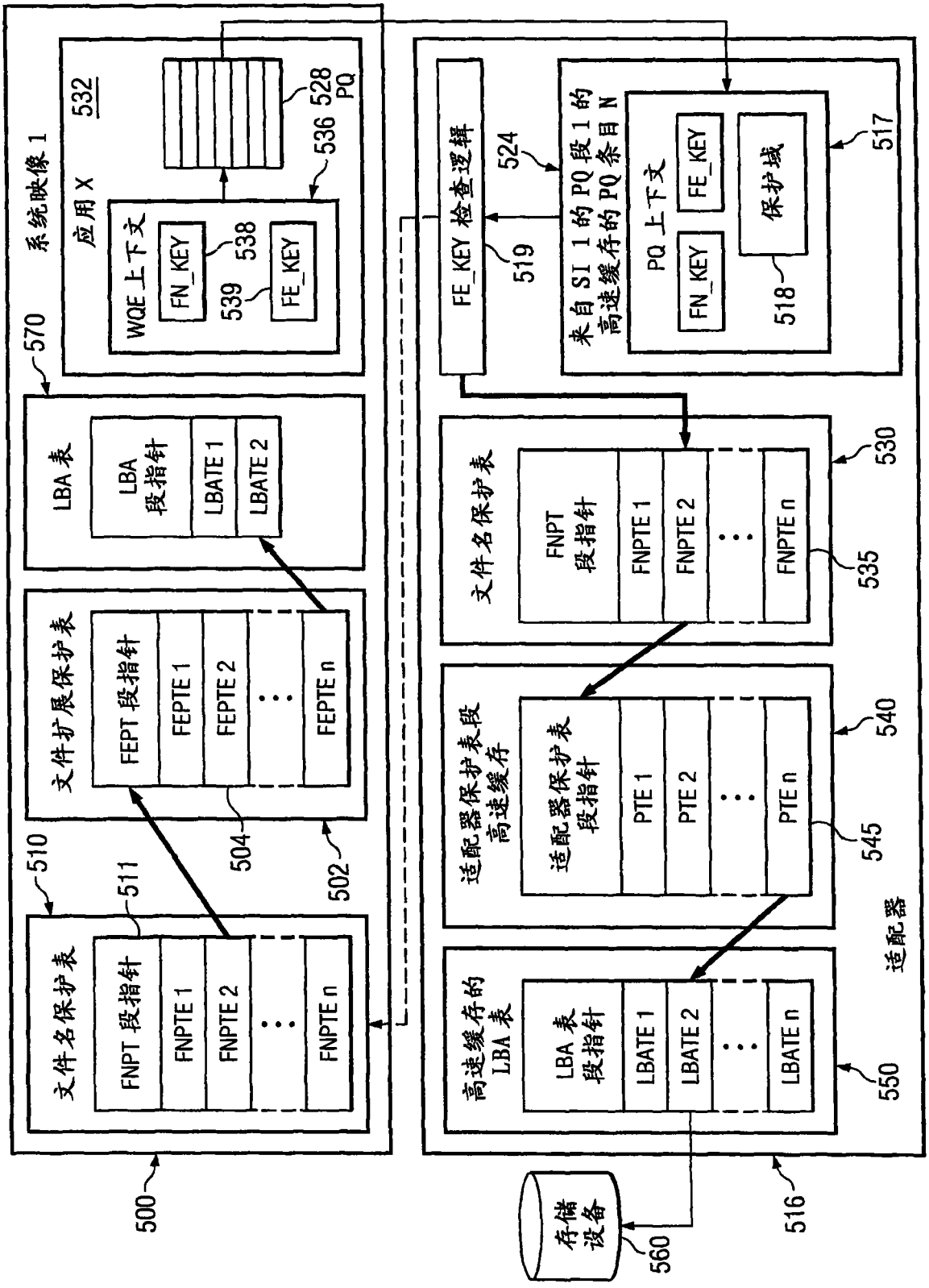

系统映像 1
570
应用 X
532
WQE 上下文
FN_KEY
539
538
FE_KEY
536
528
PQ
LBA 表
LBA 段指针
LBATE 1
LBATE 2
文件扩展保护表
FEPT 段指针
FEPTE 1
FEPTE 2
FEPTE n
504
502
510
511
文件名保护表
FNPT 段指针
FNPTE 1
FNPTE 2
FNPTE n
500
FE_KEY 检查逻辑
524
519
来自 SI 1 的 PQ 段 1 的高速缓存的 PQ 条目 N
517
PQ 上下文
FN_KEY
FE_KEY
保护域
518
文件名保护表
FNPT 段指针
FNPTE 1
FNPTE 2
FNPTE n
530
535
适配器保护表段高速缓存
适配器保护表段指针
PTE 1
PTE 2
PTE n
540
545
高速缓存的 LBA 表
LBA 表段指针
LBATE 1
LBATE 2
LBATE n
550
适配器
516
存储设备
560

图 5

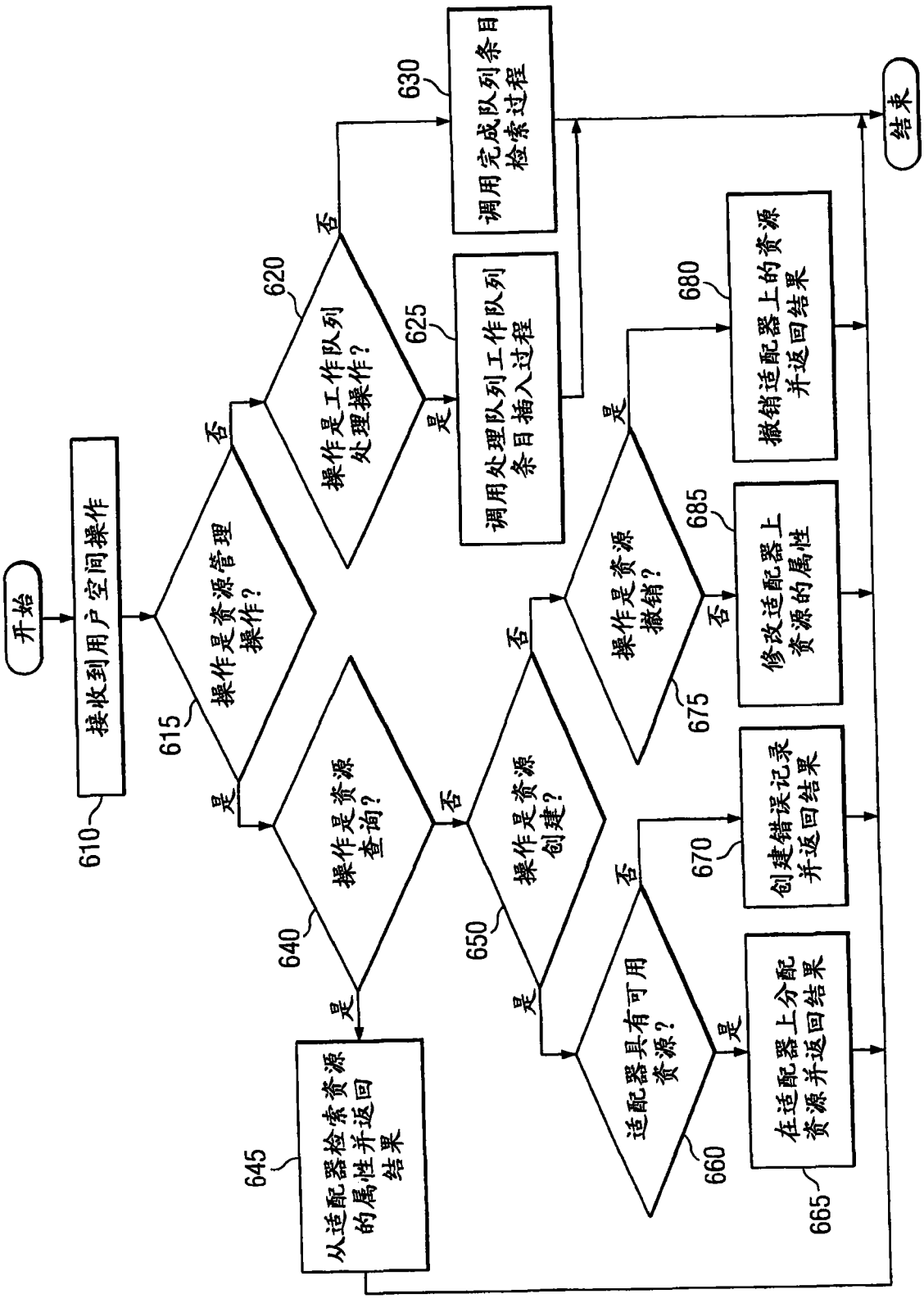

开始
接收到用户空间操作
610
操作是资源管理操作?
615
是
否
操作是工作队列处理操作?
620
是
否
调用处理队列工作队列条目插入过程
625
调用完成队列条目检索过程
630
操作是资源查询?
640
是
否
从适配器检索资源的属性并返回结果
645
操作是资源创建?
650
是
否
适配器具有可用资源?
660
是
否
在适配器上分配资源并返回结果
665
创建错误记录并返回结果
670
操作是资源撤销?
675
是
否
修改适配器上资源的属性
685
撤销适配器上的资源并返回结果
680
结束

图 6

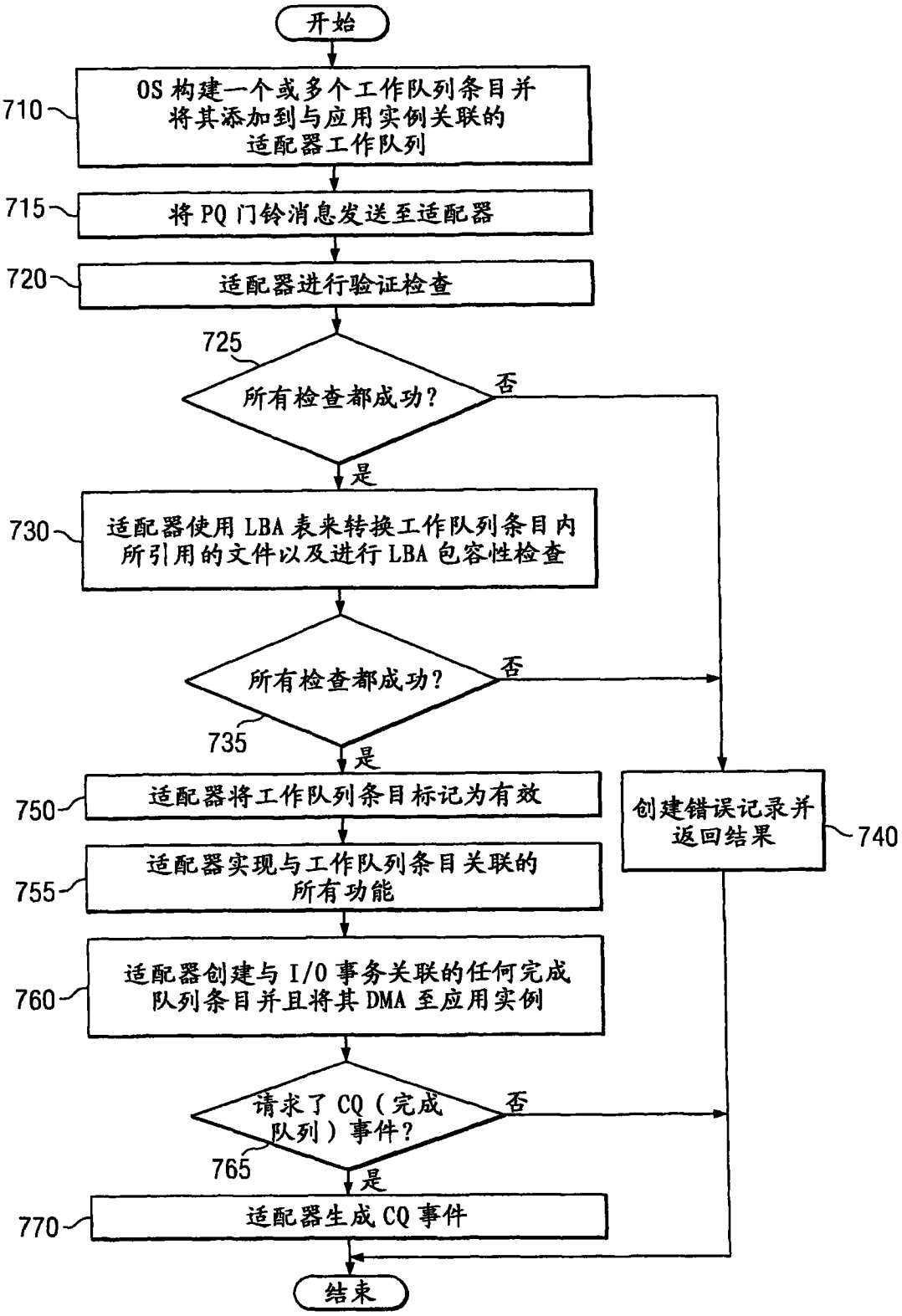

开始
OS 构建一个或多个工作队列条目并将其添加到与应用实例关联的适配器工作队列
710
将 PQ 门铃消息发送至适配器
715
适配器进行验证检查
720
725
所有检查都成功？
否
是
适配器使用 LBA 表来转换工作队列条目内所引用的文件以及进行 LBA 包容性检查
730
所有检查都成功？
否
735
是
适配器将工作队列条目标记为有效
750
创建错误记录并返回结果
740
适配器实现与工作队列条目关联的所有功能
755
适配器创建与 I/O 事务关联的任何完成队列条目并且将其 DMA 至应用实例
760
请求了 CQ（完成队列）事件？
否
765
是
适配器生成 CQ 事件
770
结束

图 7

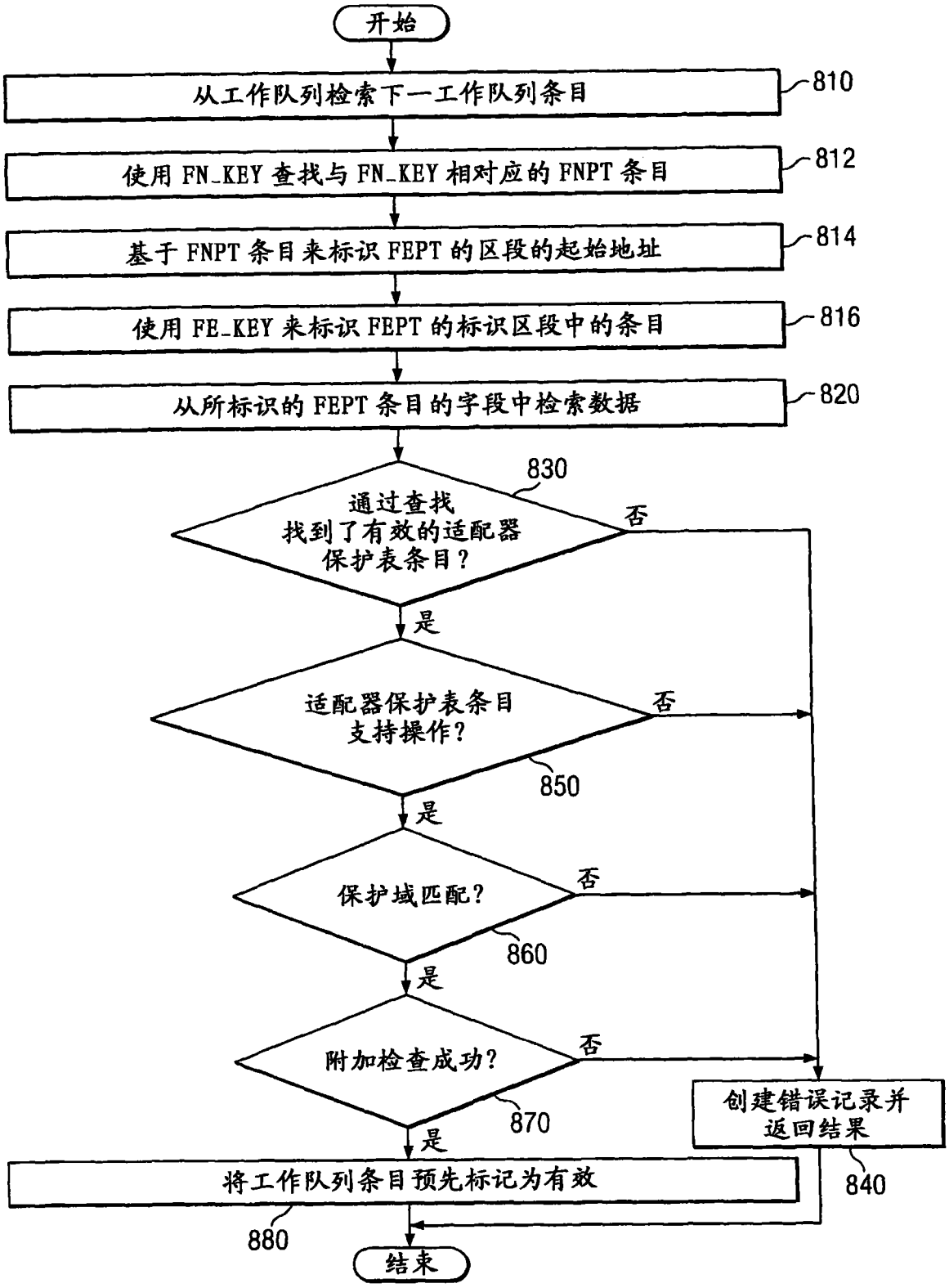

开始
从工作队列检索下一工作队列条目
810
使用 FN_KEY 查找与 FN_KEY 相对应的 FNPT 条目
812
基于 FNPT 条目来标识 FEPT 的区段的起始地址
814
使用 FE_KEY 来标识 FEPT 的标识区段中的条目
816
从所标识的 FEPT 条目的字段中检索数据
820
830
通过查找
找到了有效的适配器
保护表条目？
否
是
适配器保护表条目
支持操作？
否
850
是
保护域匹配？
否
860
是
附加检查成功？
否
870
是
创建错误记录并
返回结果
840
将工作队列条目预先标记为有效
880
结束

图 8

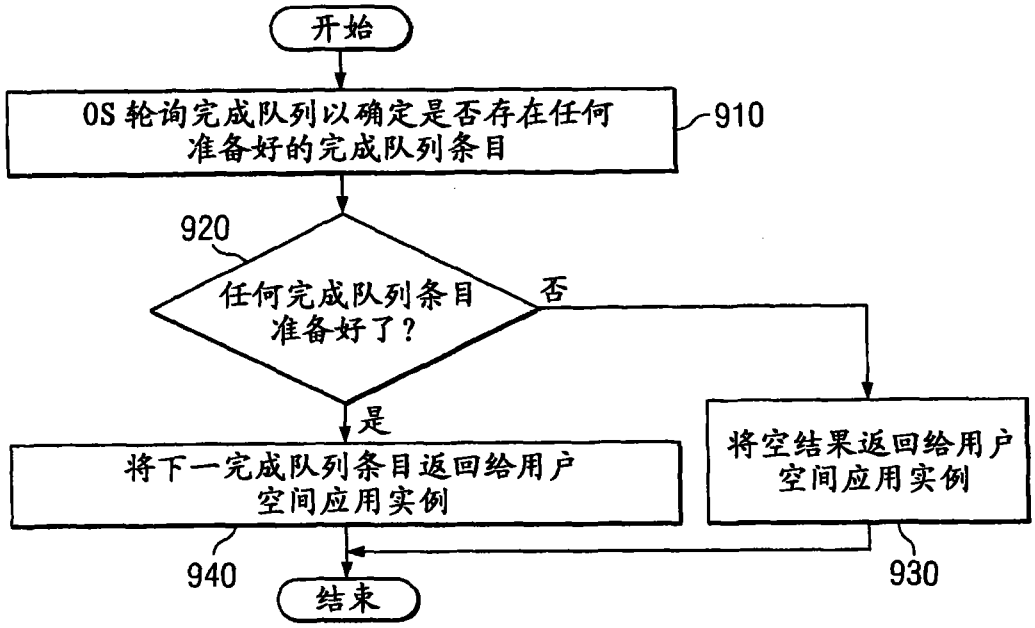

开始
OS 轮询完成队列以确定是否存在任何准备好的完成队列条目
910
920
任何完成队列条目准备好了?
否
是
将下一完成队列条目返回给用户空间应用实例
将空结果返回给用户空间应用实例
940
930
结束

图 9

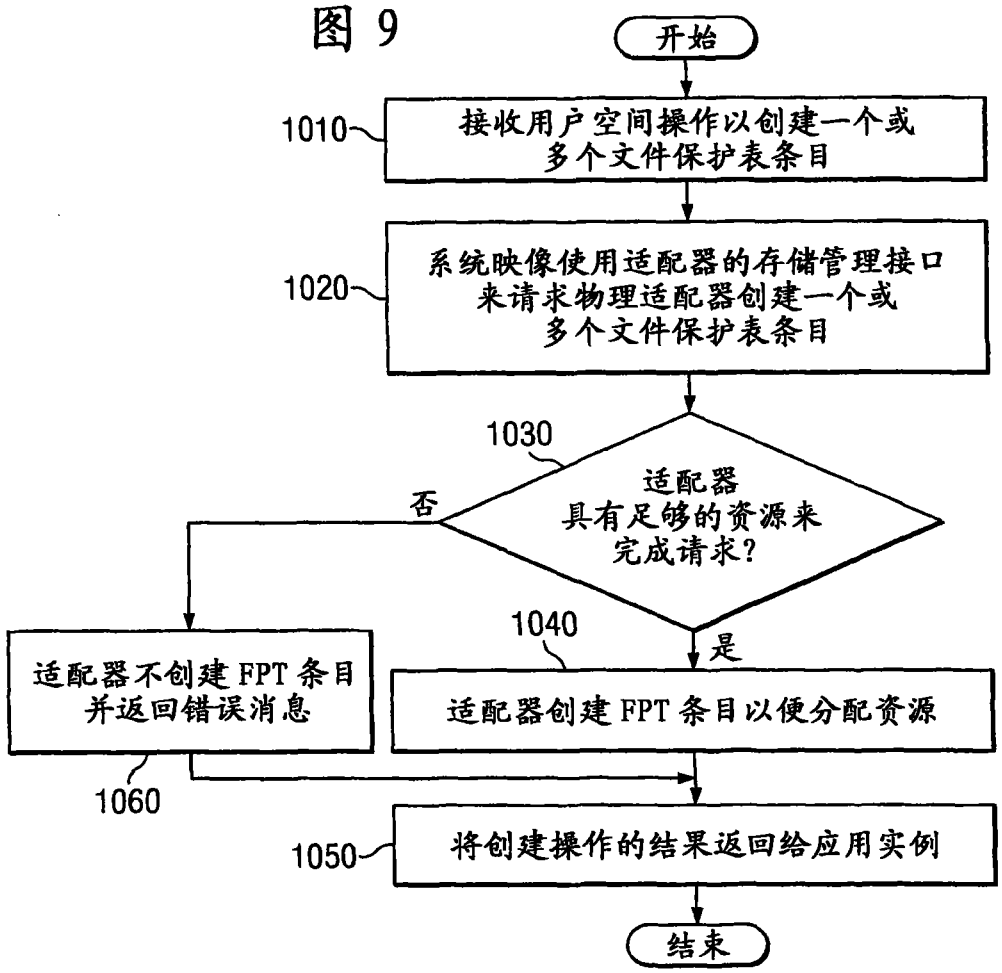

开始
1010
接收用户空间操作以创建一个或多个文件保护表条目
1020
系统映像使用适配器的存储管理接口来请求物理适配器创建一个或多个文件保护表条目
1030
适配器具有足够的资源来完成请求?
否
是
1040
适配器不创建 FPT 条目并返回错误消息
适配器创建 FPT 条目以便分配资源
1060
1050
将创建操作的结果返回给应用实例
结束

图 10

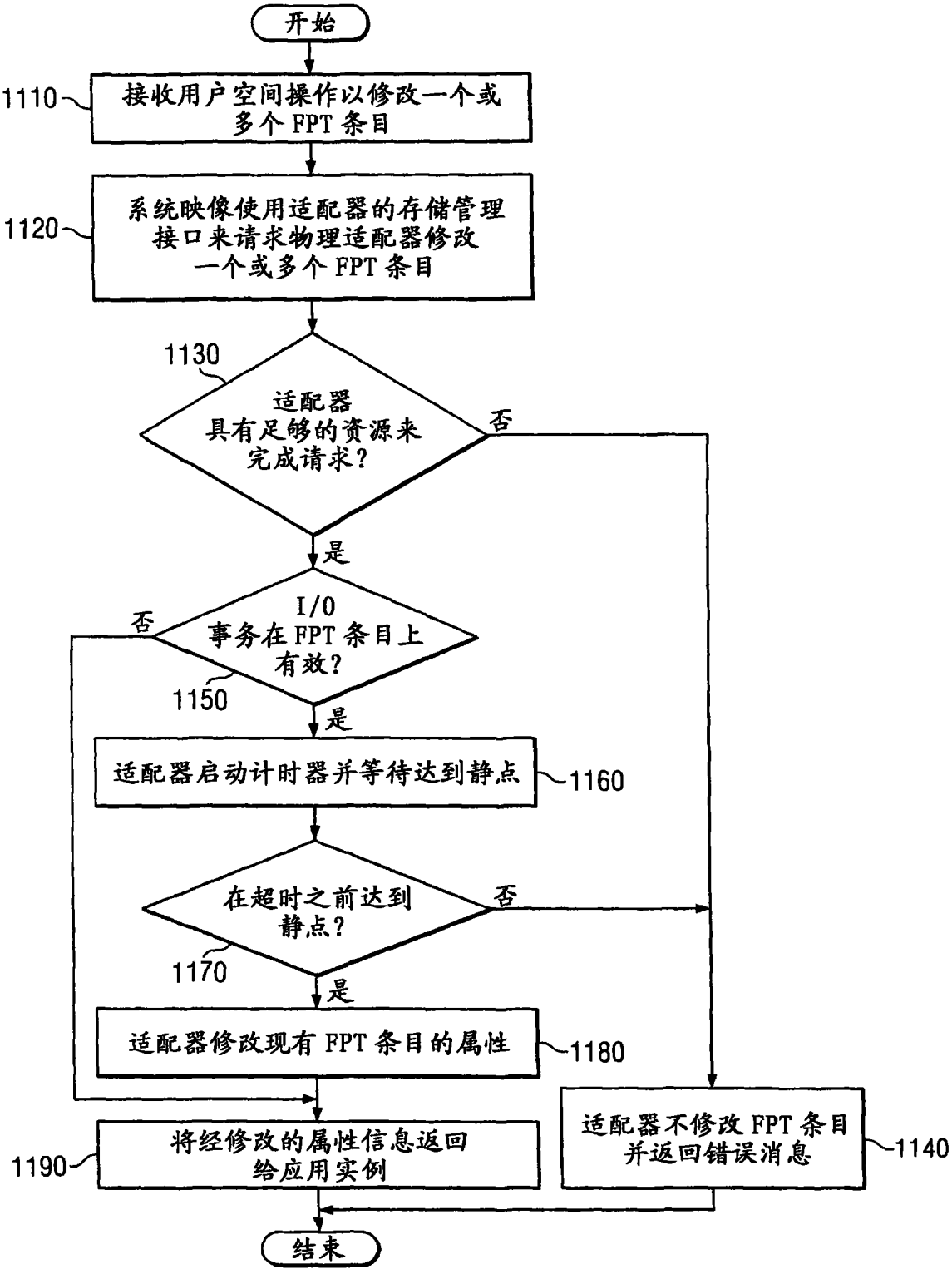

开始
1110
接收用户空间操作以修改一个或多个 FPT 条目
1120
系统映像使用适配器的存储管理接口来请求物理适配器修改一个或多个 FPT 条目
1130
适配器具有足够的资源来完成请求?
否
是
I/O 事务在 FPT 条目上有效?
否
1150
是
适配器启动计时器并等待达到静点
1160
在超时之前达到静点?
否
1170
是
适配器修改现有 FPT 条目的属性
1180
适配器不修改 FPT 条目并返回错误消息
1140
1190
将经修改的属性信息返回给应用实例
结束

图 11

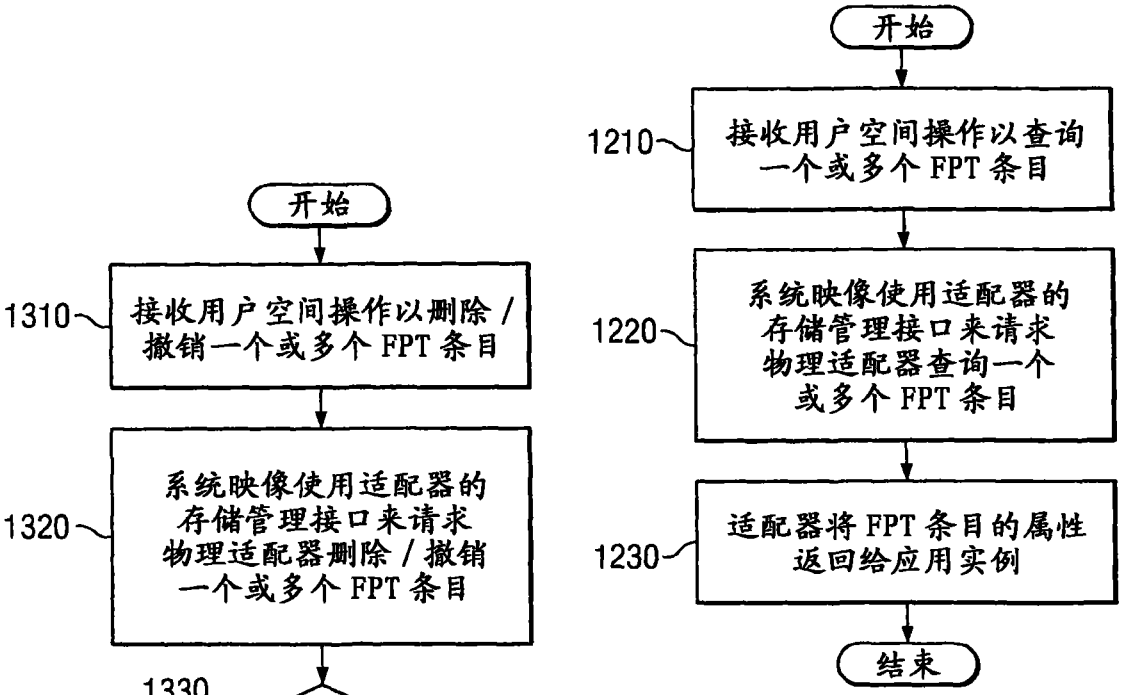

开始
1210
接收用户空间操作以查询一个或多个 FPT 条目
1220
系统映像使用适配器的存储管理接口来请求物理适配器查询一个或多个 FPT 条目
1230
适配器将 FPT 条目的属性返回给应用实例
结束

图 12

开始
1310
接收用户空间操作以删除/撤销一个或多个 FPT 条目
1320
系统映像使用适配器的存储管理接口来请求物理适配器删除/撤销一个或多个 FPT 条目
1330
I/O 事务在 FPT 条目上有效?
否
是
1340
适配器启动计时器并等待达到静点
在超时之前达到静点?
否
1350
1370
是
适配器删除/撤销现有的 FPT 条目
适配器不删除/撤销 FPT 条目并返回错误消息
1360
1380
将结果返回给应用实例
结束

图 13