

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.

G06F 3/06 (2006.01)

G06F 11/00 (2006.01)



[12] 发明专利说明书

专利号 ZL 200510083312.4

[45] 授权公告日 2007 年 7 月 11 日

[11] 授权公告号 CN 1326026C

[22] 申请日 2005.7.12

[21] 申请号 200510083312.4

[30] 优先权

[32] 2004. 8. 5 [33] US [31] 10/912,973

[73] 专利权人 国际商业机器公司

地址 美国纽约

[72] 发明人 约翰·查尔斯·埃利奥特

罗伯特·阿卡拉·库博

卡尔·埃文·琼斯

格雷格·史蒂文·卢卡斯

[56] 参考文献

CN1146048A 1997.3.26

US5938771A 1999.8.17

US5928367A 1999.7.27

US5758054A 1998.5.26

US2004/0078666A1 2004.4.22

US6112277A 2000.8.29

审查员 郑宗玉

[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 吴丽丽

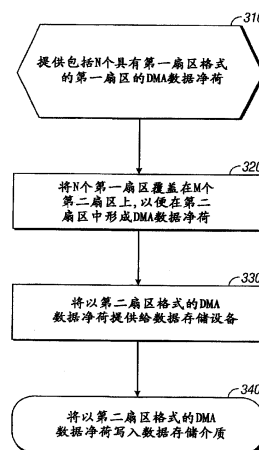
权利要求书 5 页 说明书 17 页 附图 7 页

[54] 发明名称

将数据净荷从第一扇区格式转换成第二扇区
格式的系统和方法

[57] 摘要

该方法形成多个 DMA 数据净荷，其中每个 DMA 数据净荷包括多个第一扇区，并且该方法将一个或多个那些 DMA 数据净荷发送到扇区格式转换设备。然后，该方法将第 i 个 DMA 数据净荷覆盖在部分或全部多个第二扇区上以便形成第 i 个转换后的 DMA 数据净荷，在数据队列中排队第 i 个转换后的 DMA 数据净荷，将第 i 个转换后的 DMA 数据净荷传送到数据存储设备，以及将第 i 个转换后的 DMA 数据净荷写入信息存储介质。



1. 一种将数据净荷从第一扇区格式转换成第二扇区格式的方法，包括步骤：

提供信息存储和检索系统，该系统包括：由以第一扇区格式写入的多个第一扇区组成的信息；设备控制器；数据存储设备，其中，所述数据存储设备包括：由每个都包括第二扇区格式的多个第二扇区组成的信息存储介质；包括数据队列，缓冲器和非易失性存储器的扇区格式转换设备；互连所述扇区格式转换设备和所述设备控制器的通信链路；其中，所述扇区格式转换设备与所述数据存储设备互连；

设置电源故障告警时间间隔；

形成多个 DMA 数据净荷，其中每个 DMA 数据净荷包括多个第一扇区，其中，所述多个 DMA 数据净荷的每一个都包括开始逻辑块地址和结束逻辑块地址；

将一个或多个所述 DMA 数据净荷发送到所述扇区格式转换设备；

由所述扇区格式转换设备将第 i 个 DMA 数据净荷覆盖在部分或全部多个第二扇区上，以便形成第 i 个转换后的 DMA 数据净荷；

在所述数据队列中排队所述第 i 个转换后的 DMA 数据净荷；

将所述第 i 个转换后的 DMA 数据净荷发送到所述数据存储设备；

将所述第 i 个转换后的 DMA 数据净荷写入所述信息存储介质。

2. 如权利要求 1 所述的方法，进一步包括步骤：

将所述一个或多个 DMA 数据净荷存储在所述缓冲器中；以及从所述缓冲器检索第 i 个 DMA 数据净荷。

3. 如权利要求 1 所述的方法，进一步包括步骤：

提供多个飞行中指示符，其中，每个飞行中指示符包括第一值或第二值，以及其中，第 i 个飞行中指示符与第 i 个转换后的 DMA 数据净荷有关；

当将所述第 i 个转换后的 DMA 数据净荷发送到所述数据存储设

备时, 将第 i 个飞行中指示符设置成第二值;

在将所述第 i 个转换后的 DMA 数据净荷写入所述信息存储介质后, 由所述数据存储设备生成第 i 个写入完成信号;

在接收到所述第 i 个写入完成信号时, 将第 i 个飞行中指示符复位到所述第一值。

4. 如权利要求 3 所述的方法, 进一步包括提供包括微码的扇区格式转换设备的步骤, 其中所述微码包含多个标志, 其中所述多个标志包括所述多个飞行中指示符。

5. 如权利要求 3 所述的方法, 进一步包括提供包括微码的扇区格式转换设备的步骤, 其中所述微码包含多个位, 以及其中所述多个位包括所述多个飞行中指示符。

6. 如权利要求 3 所述的方法, 其中所述复位步骤进一步包括通过所述扇区格式转换设备复位所述第 i 个飞行中指示符。

7. 如权利要求 3 所述的方法, 进一步包括步骤:

确定是否已经接收到电源故障告警;

如果接收到电源故障告警, 则执行下列步骤:

开始所述电源故障告警时间间隔;

在所述电源故障告警时间间隔期间, 将一个或多个转换后的 DMA 数据净荷从所述队列提供给所述数据存储设备, 其中, 所述队列包括 P 个排队的转换后的 DMA 数据净荷;

确定所述电源故障告警时间间隔是否到期;

如果所述电源故障告警时间间隔到期, 则执行下列步骤:

确定是否已经将第 j 个转换后的 DMA 数据净荷写入所述信息存储介质, 其中, j 大于或等于 1 并且小于或等于 P ;

如果在所述电源故障告警时间间隔到期时, 还未将第 j 个转换后的 DMA 数据净荷写入所述信息存储介质, 则执行下列步骤:

将第 j 个转换后的 DMA 数据净荷保存在非易失性存储器中;

失去到所述信息存储和检索系统的市电;

恢复到所述信息存储和检索系统的市电;

确定在市电失去时 Q 个转换后的 DMA 数据净荷仍然在飞行中；
从所述非易失性存储器检索在失去市电前还没有写入所述信息存储介质的每个转换后的 DMA 数据净荷；以及

将在失去市电前还没有写入所述信息存储介质的每个转换后的 DMA 数据净荷提供给所述数据存储设备。

8. 如权利要求 7 所述的方法，进一步包括步骤：

完成向所述数据存储设备发送已经转换后的 DMA 数据净荷，其中，在该发送期间接收到所述电源故障告警；

设置用于该发送的已经转换后的 DMA 数据净荷的飞行中指示符。

9. 如权利要求 7 所述的方法，进一步包括步骤：

在恢复到所述信息存储和检索系统的市电时，确定是否将第 k 个飞行中指示符设置成所述第二值；

如果将第 k 个飞行中指示符设置成所述第二值，则从所述非易失性存储器检索第 k 个转换后的 DMA 数据净荷。

10. 如权利要求 1 所述的方法，进一步包括从由 520 字节扇区、524 字节扇区和 528 字节扇区组成的组选择所述第一扇区格式的步骤。

11. 如权利要求 1 所述的方法，进一步包括将所述第二扇区格式设置成包括 512 字节扇区的步骤。

12. 如权利要求 1 所述的方法，进一步包括设置所述电源故障告警时间间隔等于 3 毫秒的步骤。

13. 如权利要求 1 所述的方法，进一步包括提供包括 RAID 控制器的设备控制器的步骤。

14. 如权利要求 1 所述的方法，进一步包括提供 S-ATA 数据存储设备的步骤。

15. 如权利要求 11 所述的方法，进一步包括在接收到所述电源故障告警后的 4 毫秒或更长时间内向所述信息存储和检索系统提供市电的步骤。

16. 一种用于将数据净荷从第一扇区格式转换成第二扇区格式的

系统，包括：由以第一扇区格式写入的多个第一扇区组成的信息；设备控制器；数据存储设备，其中，所述数据存储设备包括：由每个都包括第二扇区格式的多个第二扇区组成的信息存储介质；包括数据队列，缓冲器和非易失性存储器的扇区格式转换设备；互连所述扇区格式转换设备和所述设备控制器的通信链路，其中，所述扇区格式转换设备与所述数据存储设备互连，所述系统进一步包括：

用于形成多个 DMA 数据净荷的装置，其中每个 DMA 数据净荷包括多个第一扇区，其中，所述多个 DMA 数据净荷的每一个都包括开始逻辑块地址和结束逻辑块地址；

用于将所述多个 DMA 数据净荷的一个或多个发送到所述扇区格式转换设备的装置；

用于由所述扇区格式转换设备将第 i 个 DMA 数据净荷覆盖在部分或全部多个第二扇区上，以便形成第 i 个转换后的 DMA 数据净荷的装置；

用于在所述数据队列中排队所述第 i 个转换后的 DMA 数据净荷的装置；

用于将所述第 i 个转换后的 DMA 数据净荷发送到所述数据存储设备的装置；

用于将所述第 i 个转换后的 DMA 数据净荷写入所述信息存储介质的装置。

17. 如权利要求 16 所述的系统，进一步包括：

用于将所述一个或多个 DMA 数据净荷存储在所述缓冲器中的装置；以及

用于从所述缓冲器检索第 i 个 DMA 数据净荷的装置。

18. 如权利要求 16 所述的系统，进一步：

用于将多个飞行中指示符的第 i 个飞行中指示符与第 i 个转换后的 DMA 数据净荷相关联的装置，其中，所述多个飞行中指示符的每个包括第一值或第二值；

用于当将所述第 i 个转换后的 DMA 数据净荷发送到所述数据存

储设备时，将第 i 个飞行中指示符设置成第二值的装置；

用于在将所述第 i 个转换后的 DMA 数据净荷写入所述信息存储介质后，由所述数据存储设备生成第 i 个写入完成信号的装置；

用于在接收到所述第 i 个写入完成信号时，将第 i 个飞行中指示符复位到所述第一值的装置。

19. 如权利要求 16 所述的系统，进一步包括：

用于从由 520 字节扇区、524 字节扇区和 528 字节扇区组成的组选择所述第一扇区格式的装置。

20. 如权利要求 16 所述的系统，进一步包括：

用于将所述第二扇区格式设置成包括 512 字节扇区的装置。

21. 如权利要求 16 所述的系统，进一步包括：

用于设置所述电源故障告警时间间隔等于 3 毫秒的装置。

22. 如权利要求 16 所述的系统，其中，所述设备控制器包括 RAID 控制器。

23. 如权利要求 16 所述的系统，其中，所述数据存储设备包括 S-ATA 设备。

将数据净荷从第一扇区格式转换成 第二扇区格式的系统和方法

技术领域

本发明涉及将数据净荷从第一扇区格式转换成第二扇区格式同时在电源损失期间保持数据完整性的装置和方法。

背景技术

使用数据存储和检索系统来存储由一个或多个主机系统提供的信息。这些数据存储和检索系统接收将信息写入一个或多个辅助存储设备的请求，以及从那些一个或多个辅助存储设备检索信息的请求。在接收写入请求后，系统将从主机检索的信息存储在数据超高速缓存中。在某些实现中，还将那一信息的拷贝存储在非易失性存储设备中。在接收读取请求时，系统从一个或多个辅助存储设备检索信息并将那一信息移动到数据超高速缓存。因此，该系统连续地向和从存储设备以及向和从数据超高速缓存移动信息。

然而，如果以第一扇区格式写入位于数据超高速缓存中的数据，以及如果以第二扇区格式写入位于一个或多个存储设备中的数据，那么在数据超高速缓存和一个或多个存储设备间，需要扇区格式转换设备。申请人已经发现所需要的是将数据净荷从第一扇区格式转换成第二扇区格式同时在电源损失事件期间保持数据完整性的扇区格式转换设备，以及使用那个设备的方法。申请人的装置和方法将数据净荷从第一扇区格式转换成第二扇区格式，同时在电源损失事件期间保持数据完整性。在某些实施例中，申请人的装置和方法符合 SFF-8045 规范下的电源故障告警协议。

发明内容

申请人的发明包括将数据净荷从第一扇区格式转换成第二扇区格式同时在电源损失期间保持数据完整性的装置和方法。该方法提供信息存储和检索系统，它包括：由以第一扇区格式写入的多个第一扇区组成的信息；设备控制器；数据存储设备，其中，数据存储设备包括由多个第二扇区组成的信息存储介质，每个第二扇区包括第二扇区格式；包括数据队列、缓冲器和非易失性存储器的扇区格式转换设备；互连扇区格式转换设备和设备控制器的通信链路；其中，扇区格式转换设备与数据存储设备互连。

该方法进一步设置电源故障告警时间间隔；以及形成多个 DMA 数据净荷，每个 DMA 数据净荷包括多个第一扇区，其中，多个 DMA 数据净荷的每一个都包括开始逻辑块地址和结束逻辑块地址；以及将 DMA 数据净荷的一个或多个发送到扇区格式转换设备。

然后，该方法将第 i 个 DMA 数据净荷覆盖在部分或全部多个第二扇区上以便形成第 i 个转换后的 DMA 数据净荷，在数据队列中排队第 i 个转换后的 DMA 数据净荷，将第 i 个转换后的 DMA 数据净荷传送到数据存储设备，以及将第 i 个转换后的 DMA 数据净荷写入信息存储介质。

附图说明

通过阅读下述结合附图的详细描述，将更好地理解本发明，在附图中类似的标记用来表示类似的元件，以及其中：

图 1 是示出了申请人的数据存储和检索系统的一个实施例的框图；

图 2 是示出了将多个数据存储设备互连到控制器的多个扇区格式转换设备的框图；

图 3 是概括申请人的方法的第一实施例的流程图；

图 4 是概述申请人的方法的第二实施例的流程图；

图 5 是表示将包括以第一扇区格式写入的多个扇区的 DMA 数据净荷映射到具有第二扇区格式的多个扇区的框图；

图 6 是概述申请人的电源告警故障算法的步骤的流程图；以及图 7 是描述图 6 的电源告警故障算法的一部分的框图。

具体实施方式

参考附图，在下述说明书的优选实施例中描述了本发明，在附图中类似的标记表示相同或类似的元件。本发明将描述为实现在包括两个集群、多个主机适配器、多个设备适配器和数据超高速缓存的信息存储和检索系统中。然而，对申请人的在电源损失期间保持数据完整性的方法的下述描述，不意味着将申请人的发明限定于数据处理应用，因为总的来说本发明能应用于扇区格式转换。

现在参考图 1，信息存储和检索系统 100 能经由通信链路 395，与主机 390 进行通信。图 1 所示的实施例显示了单个主机。在其他实施例中，申请人的信息存储和检索系统能与多个主机进行通信。

主机 390 包括计算机系统，诸如大型机、个人计算机、工作站及其组合，计算机系统包括操作系统，诸如 Windows、AIX、Unix、MVS、LINUX 等（Windows 是微软公司的注册商标，AIX 是 IBM 公司的注册商标以及 MVS 是 IBM 公司的商标，以及 UNIX 是在美国和通过 Open Group 排他许可的其他国家的注册商标）。在某些实施例中，主机 390 进一步包括存储管理程序。主机 390 中的存储管理程序可以包括现有技术中已知的存储管理型程序（诸如在 IBM MVS 操作系统中实现的 IBM DFSMS）的功能，该程序管理向数据存储和检索系统传送数据。

在某些实施例中，申请人的信息存储和检索系统 100 包括位于四个主机架（bay）101、106、111 和 116 中的多个主机适配器 102-105、107-110、112-115 和 117-120。在其他实施例中，申请人的信息存储和检索系统包括少于 16 个的主机适配器。与申请人的系统的任何实施例中设置的主机适配器的数量无关，那些主机适配器的每一个都包括同样有权访问中央处理/超高速缓存元件 130 和 140 的共享资源。每个主机适配器可以包括一个或多个光纤信道端口、一个或多个 FICON 端

口、一个或多个 ESCON 端口或一个或多个 SCSI 端口。每个主机适配器通过互连总线 121 连接到两个集群，以便每一集群都能处理来自任何主机适配器的 I/O。

处理器部分 130 包括处理器 132 和超高速缓存 134。在某些实施例中，处理器部分 130 进一步包括存储器 133。在某些实施例中，存储器设备 133 包括随机存取存储器。在某些实施例中，存储器设备 133 包括非易失性存储器。

处理器部分 140 包括处理器 142 和超高速缓存 144。在某些实施例中，处理器部分 140 进一步包括存储器 143。在某些实施例中，存储器设备 143 包括随机存取存储器。在某些实施例中，存储器设备 143 包括非易失性存储器。

I/O 部分 160 包括多个设备适配器，诸如设备适配器 165、166、167 和 168。I/O 部分 170 进一步包括多个设备适配器，诸如设备适配器 175、176、177 和 178。

在申请人的系统的某些实施例中，在位于申请人的信息存储和检索系统中的单个卡上，将一个或多个主机适配器、处理器部分 130 和一个或多个设备适配器封装在一起。类似地，在某些实施例中，一个或多个主机适配器、处理器部分 160 和一个或多个设备适配器位于申请人的信息存储和检索系统中设置的另一个卡上。在这些实施例中，申请人的系统 100 包括与多个数据存储设备互连的两个卡。

在图 1 所示的实施例中，将十六个数据存储设备组织成两个阵列，即阵列“A”和阵列“B”。图 1 所示的实施例显示了两个存储设备阵列。在其他实施例中，申请人的信息存储和检索系统包括两个以上的存储设备阵列。每个存储阵列对于主机来说看起来象是一个或多个逻辑设备。

在某些实施例中，一个或多个数据存储设备包括多个硬盘驱动单元。在图 1 所示的实施例中，盘阵列“A”包括盘驱动器 181、182、183、191、192、193 和 194。盘阵列“B”包括盘驱动器 185、186、187、188、196、197 和 198。在某些实施例中，阵列“A”和“B”利用 RAID 协议。

在某些实施例中，阵列“A”和“B”包括有时称为 JBOD 阵列（即“正好一串盘（Just a Bunch of Disks）”）的阵列，其中，阵列不是根据 RAID 配置的。如本领域的技术人员将意识到的那样，RAID（独立盘冗余阵列）阵列（rank）包括以盘驱动器阵列形状构造的独立的盘驱动器以便获得超出单个大驱动器的性能、容量和/或可靠性。

图 1 所示的实施例示出了两个存储设备阵列。在其他实施例中，申请人的系统包括单个存储设备阵列。在其他实施例中，申请人的系统包括两个以上的存储设备阵列。

现在，参考图 2，与用在申请人的信息存储和检索系统中的数据存储介质、例如多个硬盘无关，包括一个或多个那些存储介质的每个数据存储设备通过通信链路 215 互连到设备控制器 210。在某些实施例中，控制器 210 包括处理器 212 和存储器 214。在某些实施例中，存储器 214 包括非易失性存储器。

在某些实施例中，设备控制器 210 包括设备适配器，诸如例如设备适配器 165（图 1）。在某些实施例中，设备控制器 210 包括 RAID 控制器。在某些实施例中，通信链路 215 包括光纤信道判优环路（Fibre Channel Arbitrated Loop）。在图 2 所示的实施例中，通信链路 215 包括链路段 225、245、265 和 285。

图 2 所示的环路结构包括申请人的系统的一个实施例。在其他实施例中，链路 215 包括开关光纤或开关和环路拓扑结构的组合。在其他实施例中，链路 215 包括开关的双 FC-AL 环路，其中设备控制器 210 连接到两个 FC-AL 环路。每个环路包括一个或多个光纤信道开关。转换设备 220、240 等点对点地连接到每个环路上的开关。在其他实施例中，每个环路包含执行在每个环路上面向（surface）多个点对点设备的切换功能的单个转换设备。

在一些实施例中，转换设备包括执行从第一扇区格式到第二扇区格式（“扇区大小转换”）的扇区大小转换的元件，其中，通信协议对第一扇区格式和第二扇区格式是相同的。在其他实施例中，转换设备包括执行“扇区大小转换”和从第一通信协议到第二通信协议的转换

（“协议转换”）的元件。在一些实施例中，第一通信协议是光纤信道协议，第二通信协议是串行 ATA 或 ATA。在其他实施例中，第一通信协议是串行附加 SCSI，第二通信协议是串行 ATA 或 ATA。

在所有实施例中，申请人的扇区格式转换设备能够进行扇区大小转换。在某些实施例中，申请人的扇区格式转换设备还能够进行协议转换。在此提到的扇区格式转换设备包括能够进行扇区格式转换的设备以及能够进行扇区格式转换和协议转换的设备。

每个存储设备通过分别将存储设备 230、250、270 和 290 互连到通信环路 215 的格式转换设备，诸如设备 220、240、260 和 280，互连到通信链路 215。扇区格式转换设备 220 通过通信链路 232 将存储设备 230 互连到环路 215。

通信链路 225 互连扇区格式转换设备 220 和扇区格式转换设备 240。扇区格式转换设备 240 通过通信链路 252，互连存储设备 250 与环路 215。通信链路 245 互连扇区格式转换设备 240 和扇区格式转换设备 260。扇区格式转换设备 260 通过通信链路 272，互连存储设备 270 和环路 215。

通信链路 265 互连扇区格式转换设备 260 和扇区格式转换设备 280。扇区格式转换设备 280 通过通信链路 292，互连存储设备 290 和环路 215。通信链路 285 互连扇区格式转换设备 280 和通信环路 215。单独地从本地驱动线路、诸如 ATA 或 S-ATA 选择通信链路 232、252、272 和 292。

每个扇区格式转换设备包括数据缓冲器，诸如数据缓冲器 222、242、262 和 282。每个扇区格式转换设备进一步包括处理器，诸如处理器 224、244、264 和 284。每个扇区格式转换设备进一步包括永久存储器，诸如永久存储器 226、246、266 和 286。就永久存储器来说，申请人是指非易失性存储器，即经受得住市电（utility power）损失的存储器。在某些实施例中，分别单独地从包括一个或多个 EEPROMs（电可擦可编程只读存储器）、一个或多个闪存、电池备用 RAM、硬盘驱动器、它们的组合等的组选择非易失性存储器设备 226、246、

266 和 286。在某些实施例中，每个扇区格式转换设备进一步包括数据队列，诸如数据队列 228、248、268 和 288。

在某些实施例中，将从一个或多个主机接收的数据初始地写入位于申请人的信息存储和检索系统、诸如系统 100（图 1）中的数据超高速缓存、诸如例如数据超高速缓存 134（图 1）。在某些实施例中，使用第一扇区格式，将那一数据写入数据超高速缓存。在某些实施例中，主机使用不同于盘的固有扇区大小（native sector size）的扇区大小，使用链路 395 上的块存储协议，将数据写入存储子系统。在其他实施例中，存储子系统将其自己的数据增加到由主机所写的的数据上以便生成其自己的扇区大小。对于主机正使用基于分块的协议、诸如 SCSI 的情形，通过预先挂起和/或将数据追加到扇区，或通过将由 zSeries 系统使用的诸如 ECKD 数据格式的基于记录的数据分成固定块扇区，来完成上述操作。

在某些实施例中，第一扇区格式包括每扇区 520 个字节数据。在某些实施例中，第一扇区格式包括每扇区 524 个字节数据。在某些实施例中，第一扇区格式包括每扇区 528 字节数据。在申请人的方法的某些实施例中，在将数据提供到一个或多个附加存储设备前，将数据从上述第一扇区格式转换成第二扇区格式。在某些实施例中，第二扇区格式包括每扇区 512 个字节数据。

图 3 概述申请人的将数据从第一扇区格式转换成第二扇区格式的方法的步骤。在步骤 310 中，申请人的方法提供包括（N）个第一扇区的直接存储器存取（“DMA”）数据净荷。就“第一扇区”来说，申请人是指包括以第一扇区格式写入的数据的扇区。在某些实施例中，第一扇区格式包括 520 个字节扇区。在其他实施例中，第一扇区格式包括 524 个字节扇区。在其他实施例中，第一扇区格式包括 528 个字节扇区。

在某些实施例中，通过一个或多个主机、诸如例如主机 390（图 1），将包括（N）个第一扇区的 DMA 数据净荷提供给位于信息存储和检索系统、诸如例如系统 100（图 1）中的扇区格式转换设备、诸如

设备 220 (图 2)。在某些实施例中, 将包括 (N) 个第一扇区的 DMA 数据净荷从位于信息存储和检索系统、诸如例如系统 100 (图 1) 中的数据超高速缓存、诸如例如数据超高速缓存 134 (图 1, 2) 提供给位于相同信息存储和检索系统中的扇区格式转换设备、诸如设备 220 (图 2)。

在步骤 320, 申请人的方法将包括 (N) 个第一扇区的 DMA 数据净荷转换成包括 (M) 个第二扇区的转换后的 DMA 数据净荷。关于“第二扇区”, 申请人是指包括以第二扇区格式写入的数据的扇区。在某些实施例中, 第二扇区格式包括 512 个字节扇区。在某些实施例中, 通过扇区格式转换设备、诸如设备 220 (图 2) 执行步骤 320。如本领域的技术人员将意识到的那样, 如果第一扇区格式包括每扇区 (X) 个字节, 以及如果第二扇区格式包括每扇区 (Y) 个字节, 其中 (X) 大于 (Y), 那么 (N) 小于 (M)。另外, 如果 (X) 小于 (Y), 那么 (N) 大于 (M)。

在某些实施例中, 在步骤 320, 申请人的方法将步骤 310 中提供的 DMA 数据净荷从第一扇区格式转换成包括以第二扇区格式写入的数据的转换后的 DMA 数据净荷。在某些实施例中, 步骤 320 使用覆盖操作, 形成所转换后的 DMA 数据净荷。

现在参考图 3 和 5, 在步骤 310 中, 申请人的方法提供父 DMA 数据净荷, 即 DMA 数据净荷 510, 它包括写入到总共 A+1 个连续第一扇区的数据。提供连续的第一扇区进一步提供顺序相关数据以便正确地构造父 DMA 数据净荷。包括父 DMA 净荷的数据可以驻留在主机中的数据超高速缓存上, 或者可以在先前的操作中生成。

在步骤 320, 申请人的方法将在步骤 310 中接收的父 DMA 数据净荷覆盖在多个连续第二扇区上以便形成转换后的 DMA 数据净荷 540。在步骤 320, 在将 DMA 数据净荷 510 覆盖在多个连续第二扇区上以便形成转换后的 DMA 数据净荷 540 之前, 申请人的方法仅预取和读取开始的第二扇区、即扇区 550 以及末尾的第二扇区、即扇区 560。在某些实施例中, 通过扇区格式转换设备, 诸如设备 220 (图 2), 执

行步骤 320。

提供下述例子 I 来进一步示例说明本领域的技术人员如何构成和使用本发明以及如何识别目前提供的优选实施例。然而，该例子不视为对仅由附加权利要求限定的本发明的范围的限制。

例子 I

在该例子 I 中，DMA 净荷 510 包括约 8.4 千字节的数据。另外在该例子中，第一扇区格式包括每扇区约 524 字节。在该例子 I 中，DMA 净荷 510 包括约 16 个第一扇区，即图 5 中的“A”等于 15。DMA 数据净荷 510 从以第一扇区格式写入的第 (i) 个第一扇区、即扇区 520 开始，并且用以第一扇区格式写入的第 (i)+15 个第一扇区、即扇区 530 结束。

在该例子 I 中，第二扇区格式包括 512 字节扇区。另外，在该例子中，转换后的 DMA 数据净荷 540 包括约 8.4 千字节数据。因此，在该例子 I 中，转换后的 DMA 数据净荷 540 包括约 16.4 个连续第二扇区。因此，在该例子 I 中，转换后的 DMA 数据净荷 540 从以第二扇区格式写入的第 (j) 个扇区、即扇区 550 开始，并且扩展到第 17 个连续第二扇区、即第二扇区 560 中。

使用该例子，并再次参考图 5 所示的实施例，开始转换后的 DMA 数据净荷 540 的第 (j) 个第二扇区、即扇区 550 包括两个数据部分，即数据部分 552 和数据部分 554。数据部分 552 包括“旧”数据，即先前写入到扇区 550 的数据。数据部分 554 包括写入到扇区 550、作为 DMA 数据净荷 540 的开始部分的“新”数据。

类似地，再次参考图 5 所示的实施例，结束转换后的 DMA 净荷 540 的第 (j)+16 个第二扇区、即扇区 560 包括两个数据部分，即数据部分 562 和数据部分 564。数据部分 562 包括写入到扇区 560 的“新”数据，以及数据部分 564 包括“旧”数据，即先前写入到扇区 550 作为 DMA 数据净荷 540 的结束部分的数据。

象利用任何数据存储和检索系统那样，要求维持所存储的数据与操作的主机的视图一致。必须将操作的排序维持成以由主机将数据写

入数据存储和检索系统的顺序，将数据施加到存储设备。因此，在形成 DMA 数据净荷中使用的“旧”数据必须考虑通过先前形成的排队操作何种数据将被写入存储设备，以及使用适当的数据（来自先前排队操作的数据或从存储设备取出的数据）来从主机的角度维持数据的一致性。

再次参考图 3，在步骤 330 中，申请人的方法将在步骤 320 中形成的转换后的 DMA 数据净荷提供给目标存储设备，诸如存储设备 230（图 2）。在步骤 340 中，目标存储设备以第二扇区格式将所转换后的 DMA 数据净荷写入位于数据存储设备中的存储介质。

在某些实施例中，通过扇区格式转换设备、诸如设备 220，接收步骤 310 的 DMA 数据净荷，并将其存储在位于那一扇区格式转换设备内的缓冲器、诸如缓冲器 222 中。在某些实施例中，在步骤 330 中，基本上与步骤 320 的扇区转换同步地将所转换后的 DMA 数据净荷提供到存储设备。在其他实施例中，排队所转换的 DMA 数据净荷，以便稍后传送到存储设备。

在某些实施例中，使用图 4 所示的步骤，实现图 3 的方法。在步骤 410，申请人的方法提供 DMA 数据净荷，诸如例如 DMA 数据净荷 510（图 5），它包括多个连续第一扇区。

在某些实施例中，在步骤 410 中，控制器，诸如控制器 210（图 2）经由通信链路 205 从数据超高速缓存 134（图 1，2）检索第 i 至第 $i+A$ 个第一扇区（包括第 $i+A$ 个第一扇区）。如本领域的技术人员将意识到的，多个第一扇区的每一个都具有在那个数据超高速缓存上的开始和结束逻辑块地址。另外，在步骤 410，控制器、诸如控制器 210（图 2）将多个第一扇区提供给扇区格式转换设备。在某些实施例中，接收扇区格式转换设备将 DMA 数据净荷存储在缓冲器、诸如例如缓冲器 222（图 2）中。

在步骤 420，申请人的方法将步骤 410 的 DMA 数据净荷覆盖在多个连续第二扇区上，如以上结合步骤 320（图 3）所述的那样，以便形成第 k 个转换后的 DMA 数据净荷。在某些实施例中，通过扇区格

式转换设备、诸如设备 220 (图 2) 执行步骤 420。

在步骤 430, 申请人的方法排队步骤 420 的第 k 个转换后的 DMA 数据净荷。在某些实施例中, 步骤 430 包括在位于扇区格式转换设备、诸如设备 220 中的队列、诸如队列 228 中排队第 k 个转换后的 DMA 数据净荷。在某些实施例中, 通过扇区格式转换设备、诸如设备 220 执行步骤 430。

在步骤 440, 申请人的方法将在步骤 420 中形成的第 k 个转换后的 DMA 数据净荷从队列、诸如队列 228 提供到目标数据存储设备、诸如数据存储设备 230。在某些实施例中, 通过扇区格式转换设备、诸如设备 220 执行步骤 440。

在步骤 450, 申请人的方法提供和设置第 k 个飞行中 (Inflight) 指示符来表示第 k 个转换后的 DMA 数据净荷处于正在飞行中 (In Flight)。为本申请的目的, “正在飞行中”的 DMA 数据净荷是提供给还没有接收到写入完成信号的目标数据存储设备的转换后的 DMA 数据净荷。

在某些实施例中, 申请人的方法为提供给数据存储设备的每个转换后的 DMA 数据净荷创建离散的飞行中指示符。在某些实施例中, 申请人的飞行中指示符包括能被打开或关闭的标志。在某些实施例中, 这种标志设置在位于扇区格式转换设备中的设备微码中。在某些实施例中, 申请人的飞行中指示符包括能设置成“0”或“1”的位。在某些实施例中, 这一位能设置在位于扇区格式转换设备中的设备微码中。另外, 在步骤 450, 申请人的方法设置第 k 个飞行中指示符, 即与第 k 个转换后的 DMA 数据净荷有关的飞行中指示符, 以表示第 k 个转换后的 DMA 数据净荷为“正在飞行中”。在某些实施例中, 通过扇区格式转换设备、诸如设备 220 执行步骤 450。

在某些实施例中, 基本上同步地执行步骤 440 和 450。在某些实施例中, 步骤 440 先于步骤 450。而在其他实施例中, 步骤 450 先于步骤 440。

在步骤 460 中, 目标数据存储设备、诸如数据存储设备 230 接收

第 k 个转换后的 DMA 数据净荷，以及将该第 k 个转换后的 DMA 数据净荷写入位于其中的数据存储介质、诸如数据存储介质 236 中。

在步骤 470，目标数据存储设备报告第 k 个转换后的 DMA 数据净荷已经被写入介质中。该报告能包括本领域的技术人员已知的任何信令方法。例如，目标数据存储驱动器可以通过通信接口连接到发布（issuing）扇区格式转换设备。在某些实施例中，目标数据存储设备向发布扇区格式转换设备提供写入完成信号。另外，在某些实施例中，发布扇区格式转换设备轮询目标数据存储设备以断定第 k 个转换后的 DMA 数据净荷的状态。

在步骤 480，申请人的扇区格式转换设备从目标数据存储设备接收已经将第 k 个转换后的 DMA 数据净荷写入介质的报告。在某些实施例中，步骤 470 和 480 基本上同步地进行。

在步骤 490，申请人的方法复位第 k 个飞行中指示符来表示已经将第 k 个 DMA 数据净荷写入介质。在某些实施例中，步骤 490 包括删除第 k 个飞行中指示符。在其他实施例中，步骤 490 包括将第 k 个飞行中位从“1”设置成“0”。在其他实施例中，步骤 490 包括关闭第 k 个飞行中指示符标志。

图 4 的申请人方法同步地执行多个步骤。例如，在步骤 420，第一扇区格式的 DMA 数据净荷正被转换成具有第二扇区格式的 DMA 数据净荷，同时，在步骤 430，排队先前转换后的 DMA 数据净荷，同时，在步骤 440，正在将所转换后的 DMA 数据净荷提供给一个或多个数据存储设备，同时，在步骤 460，通过一个或多个数据存储设备，正在将排队的 DMA 数据净荷写入一个或多个存储介质，同时，在步骤 470，一个或多个数据存储设备正提供写入完成信号。

申请人的将 DMA 数据净荷从第一扇区格式转换成包括第二扇区的转换后的 DMA 数据净荷的方法符合 SFF 规范 SFF-8045 中的名为“Power Failure Warning”的章节 6.4.8.2（在下文中称为“PFW 规范，在此引入以供参考）。本领域的技术人员将意识到，SFF 委员会是以即时（prompt）的方式为解决盘工业需要而形成的特定组。当在 1990

成立时，最初目标限制于定义用于盘驱动器的事实上的机械外壳以便它们能装配到膝上型计算机和其他小型产品中。在 1992 年 11 月，SFF 委员会目标拓宽到包含需要即时工业行为的其他区域。SFF 规范在范围上是窄的，以便降低开发时间。在其他需要中，PFW 规范批准当写超高速缓存时，数据存储设备必须停止在块边界上向非易失性存储器写入数据。

在某些实施例中，一个或多个申请人的数据存储设备，诸如例如数据存储设备 230，包括有时称为“S-ATA”的设备。如本领域的技术人员将意识到的，S-ATA 设备采用串行 ATA 盘接口技术。

S-ATA 基于串行信令技术。串行 ATA 是点对点连接，并且允许多个端口聚集在单个控制器、诸如例如控制器 210（图 2）中。SATA 存储设备不一定支持 PFW 规范。即使当使用一个或多个 S-ATA 存储设备时，申请人的方法也是符合 PFW 规范的。

在由申请人的装置接收 PFW 信号同时形成转换后的 DMA 数据净荷和/或处理那些转换后的 DMA 数据净荷的情况下，申请人的方法变成 PFW 算法。使用该 PFW 算法，在接收 PFW 信号后，申请人的方法启动 PFW 时间间隔，完成已经转换后的 DMA 数据净荷的传输，同时尽可能地在 PFW 时间间隔期间继续将先前转换后的 DMA 数据净荷发送到数据存储设备。在 PFW 时间间隔到期时，申请人的方法保存用于在 PFW 时间间隔到期时仍然在飞行中的每个 DMA 数据净荷的 DMA 净荷元数据。

在申请人的装置产生和/或接收 PFW 信号同时执行图 4 的方法的情况下，申请人的方法变成申请人的 PFW 算法。申请人的 PFW 算法在图 6 中进行了概述。现在，参考图 4 和 6，在检测 PFW 信号时，申请人的方法转到步骤 610。步骤 620 包括接收那一 PFW 信号。

图 4 示出了申请人的从步骤 420 转到步骤 610 的方法。如上所述，申请人的方法可以同步地执行步骤 410、420、430、440、450、460、470、480 和/或 490。如果正执行步骤 410、420、430、440、450、460、470、480 和/或 490 的任何一个或多个以及接收 PFW 信号，则申请人

的方法转到步骤 610。

在步骤 610，申请人的方法设置电源故障告警时间间隔。PFW 规范尤其要求在断言 PFW 信号后，将全部电源提供给一个或多个数据存储设备达至少 4 毫秒。在申请人的方法的某些实施例中，PFW 时间间隔被设置成小于 4 毫秒。在某些实施例中，申请人的方法将 PFW 时间间隔设置成约 3 毫秒。在该 PFW 时间间隔期间，申请人的方法完成已经转换后的 DMA 净荷的正在进行中的传输同时向一个或多个数据存储设备提供尽可能多的转换后的 DMA 数据净荷。

在图 6 所示的实施例中，在步骤 620 中接收 PFW 信号前，示出了步骤 610。在某些实施例中，可以在接收 PFW 信号前的任何时间执行步骤 610。在某些实施例中，将 PFW 时间间隔被编码在位于申请人的扇区格式转换设备中的固件中。在某些实施例中，在系统启动期间设置 PFW 时间间隔。在某些实施例中，PFW 时间间隔可由系统所有者/操作者进行调整。

在某些实施例中，基本上与在步骤 620 中接收 PFW 信号同步地执行步骤 610。在这些实施例中，在接收 PFW 信号时，申请人的方法立即确定排队的转换后的 DMA 数据净荷的数量以及确定系统性能，诸如设备适配器带宽。基于这些确定，申请人的方法设置适当的 PFW 时间间隔。在这些同步实施例中，通过扇区格式转换设备、诸如设备 220（图 2）执行步骤 610。

在步骤 620 中接收 PFW 信号后，申请人的方法转到步骤 625 和步骤 630。在步骤 630，申请人的方法开始在步骤 610 中设置的 PFW 时间间隔。在某些实施例中，由扇区格式转换设备、诸如设备 220（图 2）执行步骤 630。申请人的方法从步骤 630 转到步骤 640。

在步骤 625，申请人的方法完成已经转换后的 DMA 数据净荷的传输。在某些实施例中，由扇区格式转换设备、诸如设备 220（图 2）执行步骤 625。申请人的方法从步骤 625 转到步骤 635，其中，该方法设置用于步骤 625 的 DMA 数据净荷的飞行中指示符。在某些实施例中，由扇区格式转换设备、诸如设备 220（图 2）执行步骤 635。申请

人的方法从步骤 635 转到步骤 640。

在步骤 640，申请人的方法继续将转换后的 DMA 数据净荷从数据队列、诸如队列 228（图 2）提供到一个或多个目标数据存储设备、诸如数据存储设备 230（图 2）。在某些实施例中，步骤 640 进一步包括当在步骤 620 中接收 PFW 信号时确定排队的转换后的 DMA 数据净荷的数量（P）。在某些实施例中，由扇区格式转换设备、诸如设备 220（图 2）执行步骤 640。

提供下述例子以便向本领域的技术人员进一步示例说明如何构成和使用本发明以及如何识别目前的优选实施例。然而，该例子不视为对仅由附加权利要求限定的本发明的范围的限制。

例子 II

现在参考图 7，包括第 k、第 k+1 和第 k+2 个转换后的 DMA 数据净荷的转换后的 DMA 数据净荷 710、720 和 730 正在飞行中，并且在飞行队列中仍然是未完成的。在时间 T_0 传送第 k 个转换后的 DMA 数据净荷。在时间 T_1 传送第 k+1 个转换后的 DMA 数据净荷。在时间 T_2 传送第 k+2 个转换后的 DMA 数据净荷。

在时间 T_3 ，申请人的方法开始向存储设备传送第 k+3 个转换后的 DMA 数据净荷，即转换的净荷 740。当传送所转换后的 DMA 数据净荷 740 时，在完成传输所转换的净荷 740 前，在时间 T_4 ，接收 PFW 信号。

使用图 6 的申请人的电源故障告警算法，在步骤 625，申请人的方法完成转换后的 DMA 数据净荷 740 的传输，以及在步骤 635，申请人的方法使转换后的 DMA 数据净荷 740 处于正在飞行中的状态。

在申请人的方法正执行步骤 625 和 635 时，在步骤 640 中，申请人的方法同时正在跟踪到一个或多个目标数据存储设备的正在飞行的转换后的 DMA 数据净荷 710、720 和 730 的状态。

再次参考图 6，在步骤 650，申请人的方法确定 PFW 时间间隔是否到期。在某些实施例中，由扇区格式转换设备、诸如设备 220（图 2）执行步骤 650。如果申请人的方法在步骤 650 确定 PFW 时间间隔尚未

到期，那么该方法从步骤 650 转到步骤 640 并继续。

或者，如果申请人的方法在步骤 650 确定 PFW 时间间隔已经到期，那么该方法从步骤 650 转到步骤 660，其中该方法保存在 PFW 时间间隔到期时仍然在飞行中的每个转换后的 DMA 净荷。

在某些实施例中，由扇区格式转换设备、诸如设备 220（图 2）执行步骤 660。在某些实施例中，将在步骤 660 存储的飞行中的转换后的 DMA 数据净荷写入位于扇区格式转换设备、诸如设备 220 中的非易失性存储器、诸如存储器 226。在某些实施例中，将在步骤 660 中存储的飞行的转换后的 DMA 数据净荷写入位于设备控制器、诸如设备控制器 210 中的非易失性存储器、诸如存储器 214。

在步骤 670，使市电恢复到申请人的信息存储和检索系统。关于“市电”，申请人是指由申请人的信息存储和检索系统外的商业和/或本发电机生产工厂连续提供的电源。

在步骤 680，申请人的方法检索在步骤 660 中保存的飞行中的转换后的 DMA 数据净荷。在某些实施例中，由与数据超高速缓存、诸如数据超高速缓存 134 互连以及一个或多个扇区格式转换设备、诸如设备 220 互连的控制器、诸如控制器 210 执行步骤 680。在其他实施例中，由一个或多个扇区格式转换设备执行步骤 680。

申请人的方法从步骤 680 转到步骤 690，其中，该方法将在步骤 680 中检索的转换后的 DMA 数据净荷提供给数据存储设备。申请人的方法从步骤 690 转到步骤 430，并且如上所述那样继续。

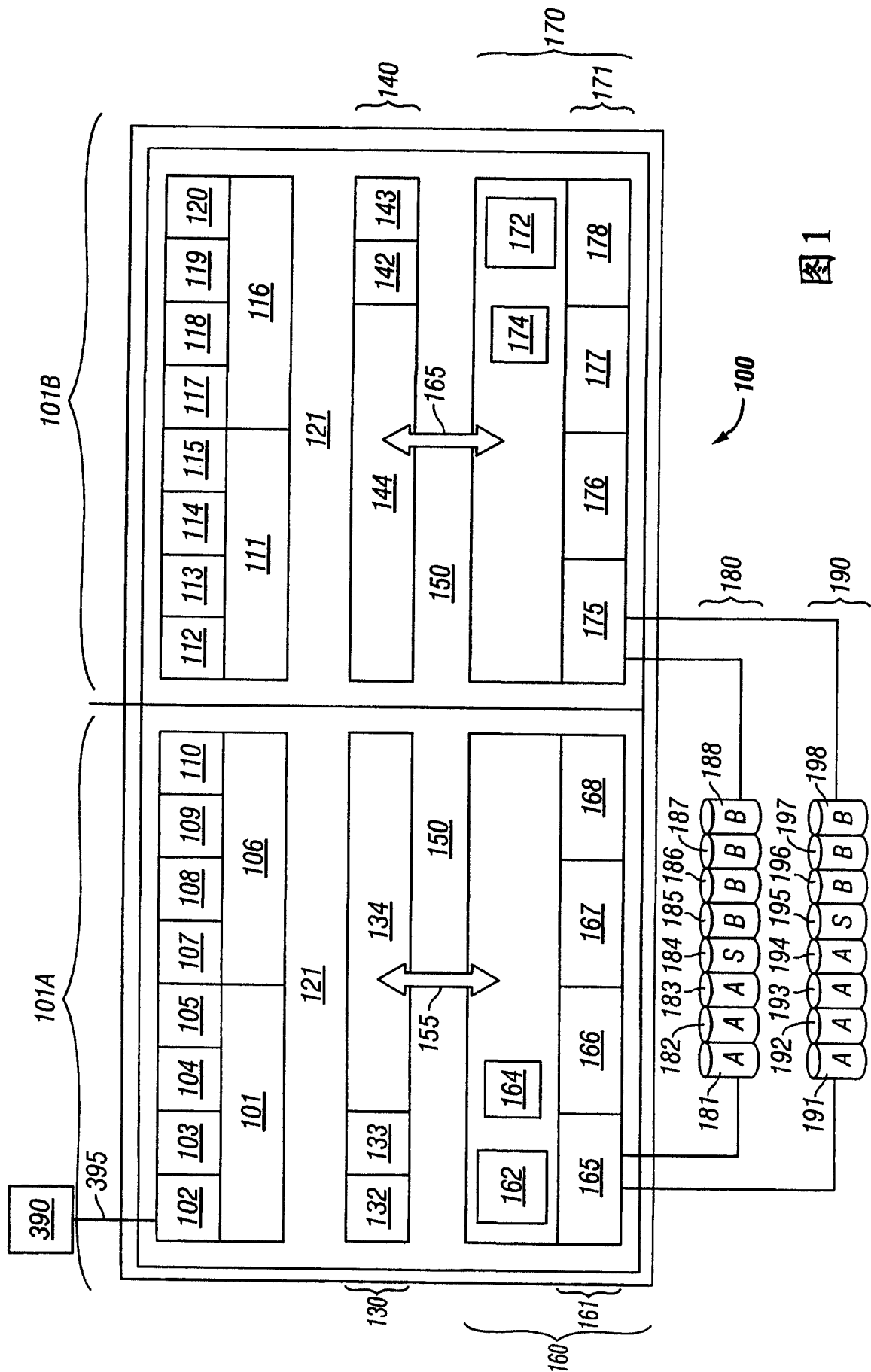
在图 3、4 和/或 6 中所述的申请人的方法的实施例可以单独地实现。此外，在某些实施例中，可以组合、删除或重排序在图 3、4 和/或 6 所述的各个步骤。

在某些实施例中，申请人的发明包括驻留在存储器、诸如例如存储器 214（图 2）、存储器 226（图 2）、存储器 246（图 2）、存储器 266（图 2）和/或存储器 286（图 2）中的指令，其中，分别由处理器 212（图 2）、224（图 2）、244（图 2）、264（图 2）和 284（图 2）执行那些指令，以便执行在图 3 中所述的步骤 310、320、330 和 340、

在图 4 中所述的步骤 410、420、430、440、450、460、470、480 和/或 490、和/或在图 6 中所述的步骤 610、620、630、640、650、660、670、680 和/或 690。

在其他实施例中，申请人的发明包括驻留在任何其他计算机程序产品中的指令，其中由系统 100 外部或内部的计算机执行那些指令，以便执行在图 3 中所述的步骤 310、320、330 和 340、在图 4 中所述的步骤 410、420、430、440、450、460、470、480 和/或 490、和/或在图 6 中所述的步骤 610、620、630、640、650、660、670、680 和/或 690。在任一情况下，可以将指令编码在信息存储介质、包括例如磁性信息存储介质、光学信息存储介质、电子信息存储介质等中。关于“电子存储介质”，申请人是指例如诸如 PROM、EPROM、EEPROM、闪速 PROM、小型闪存、智能介质等的设备。

尽管已经详细地示例说明了本发明的优选实施例，但是显然在不背离如在下述权利要求中阐述的本发明的范围的情况下，本领域的技术人员会想到对那些实施例的改进和修改。



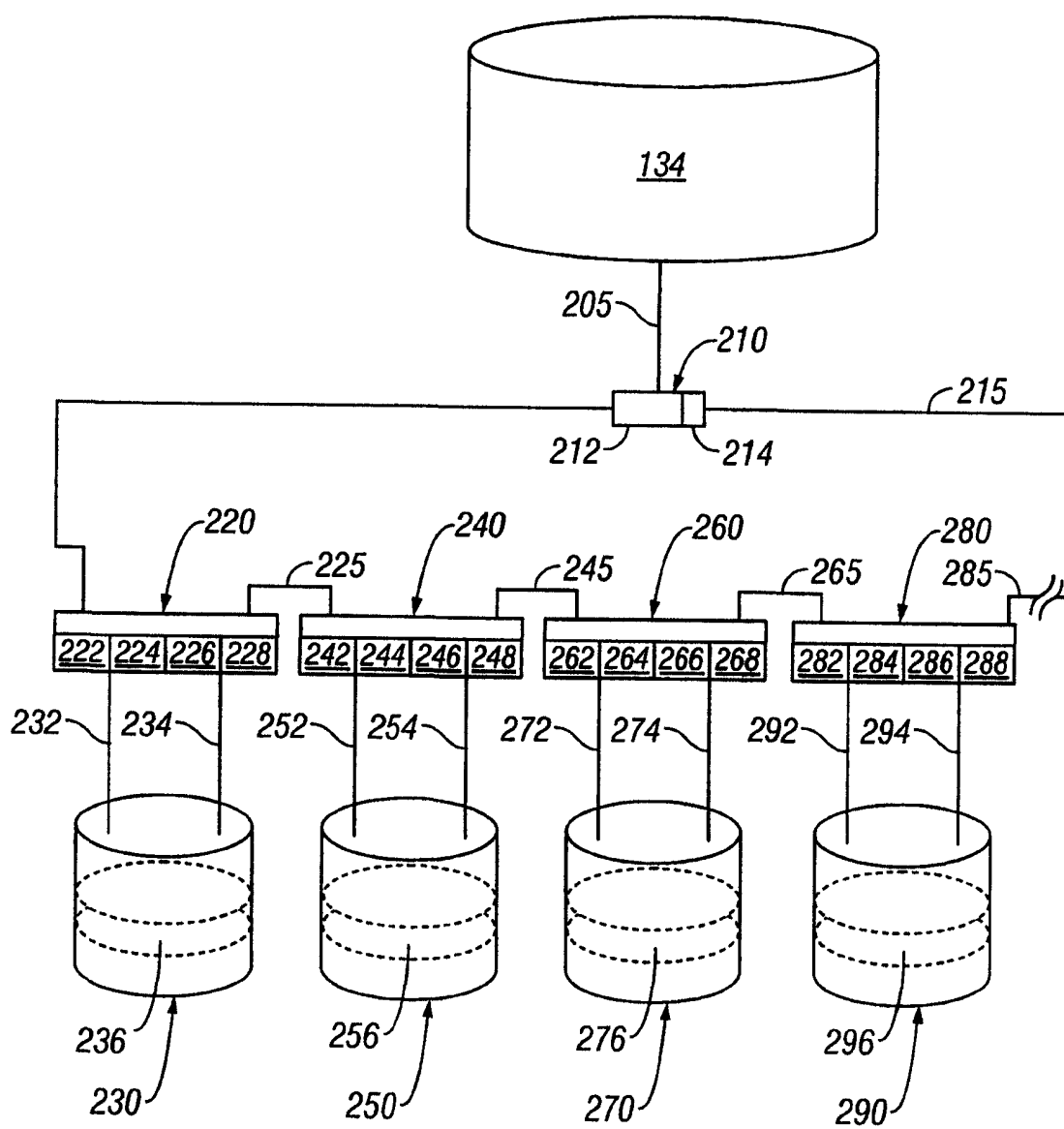


图 2

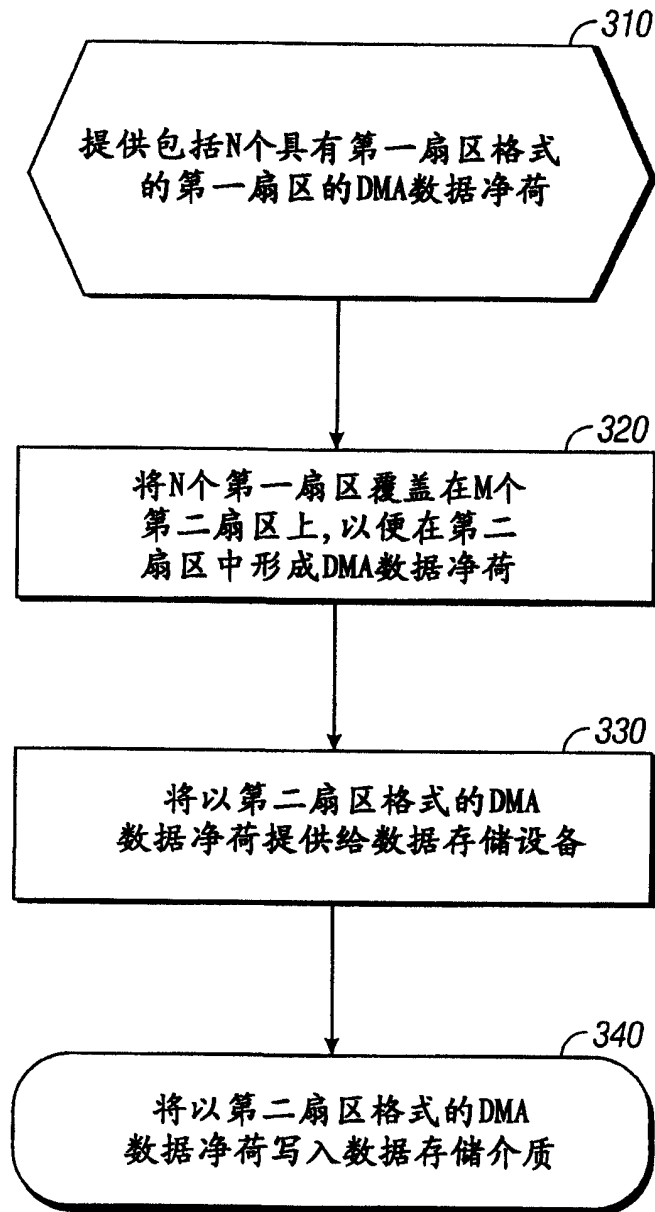


图3

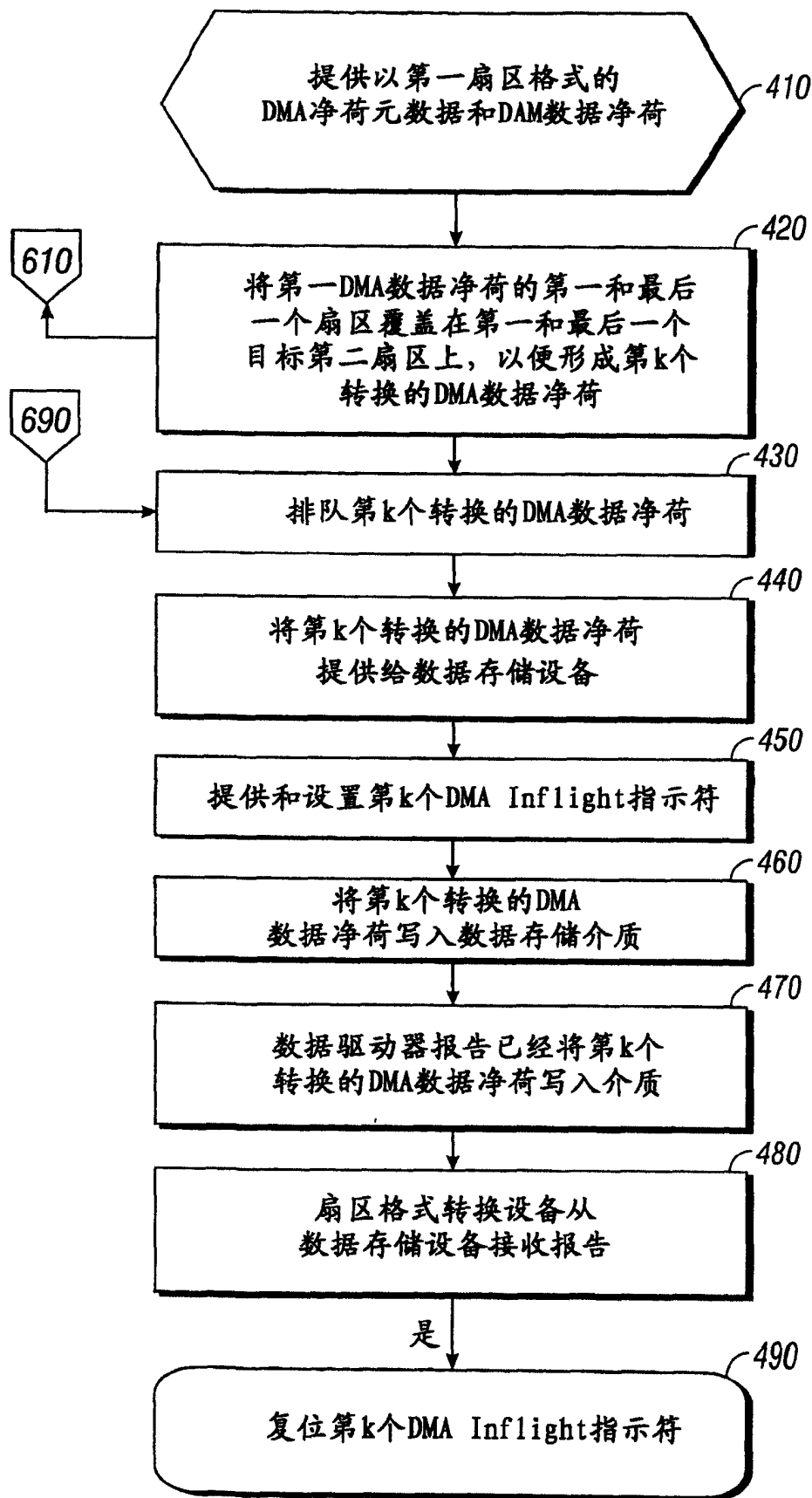


图4

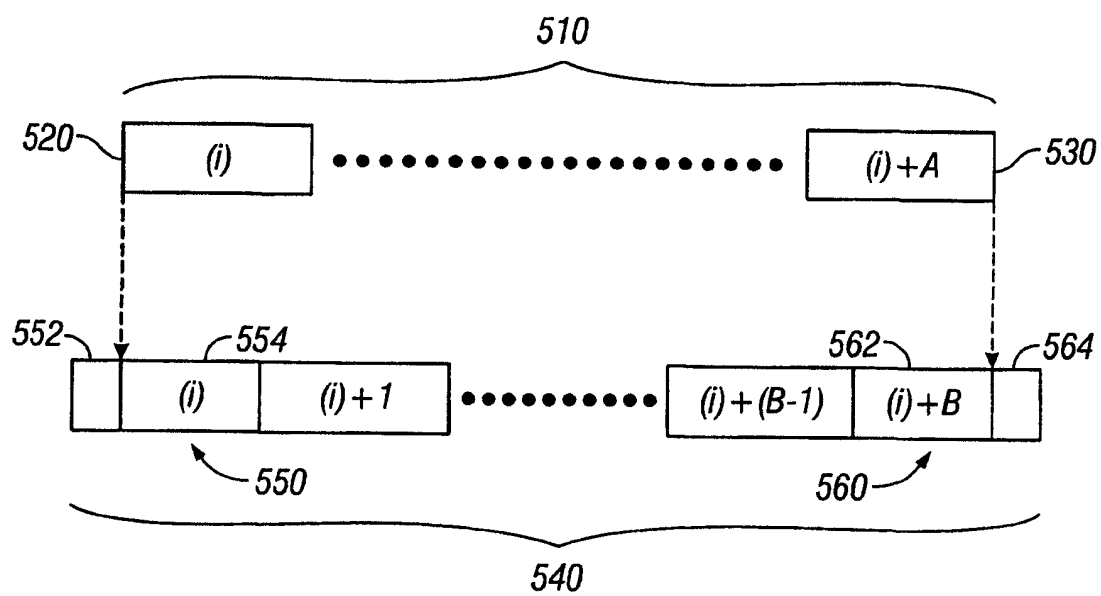


图 5

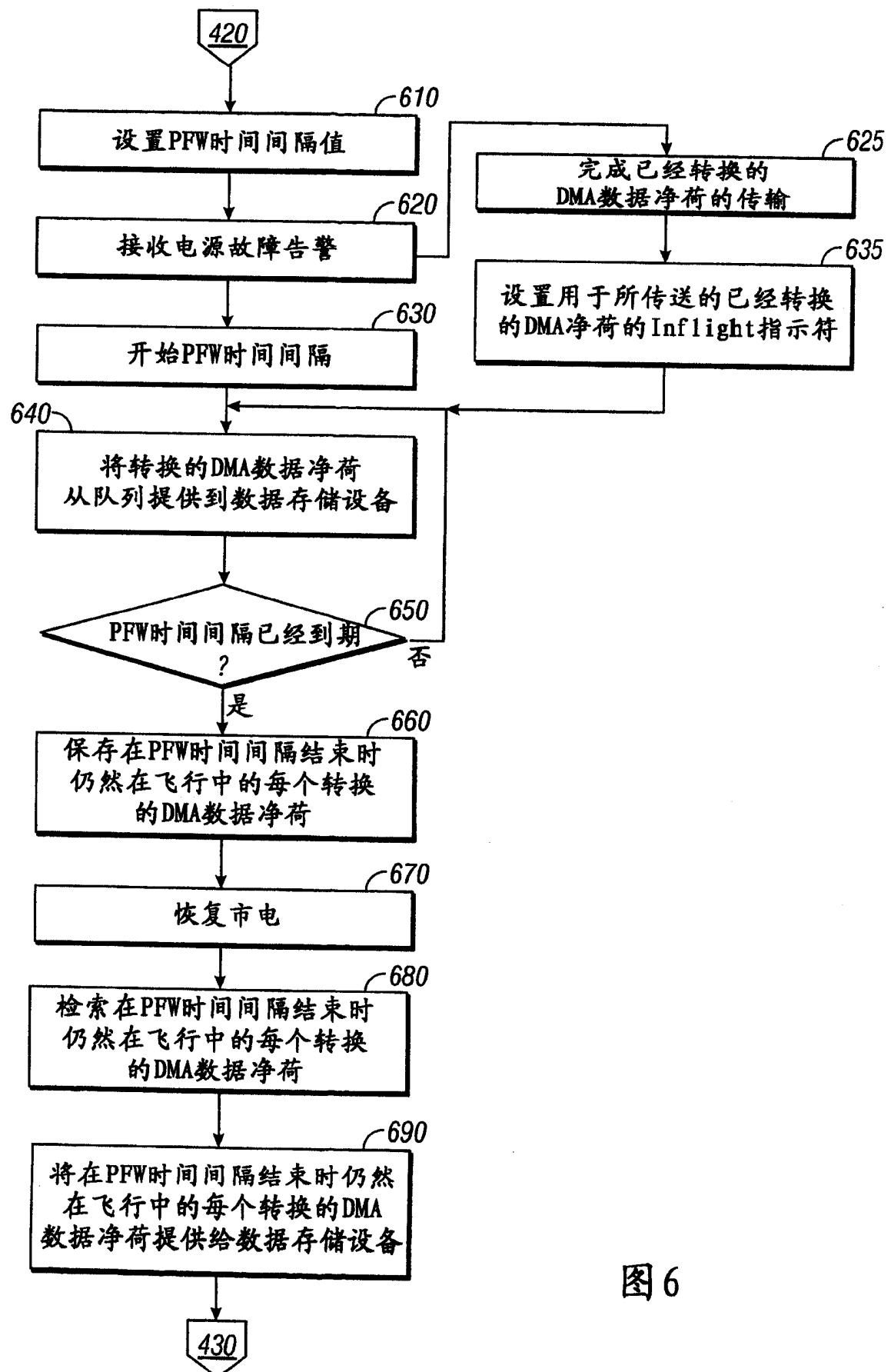


图6

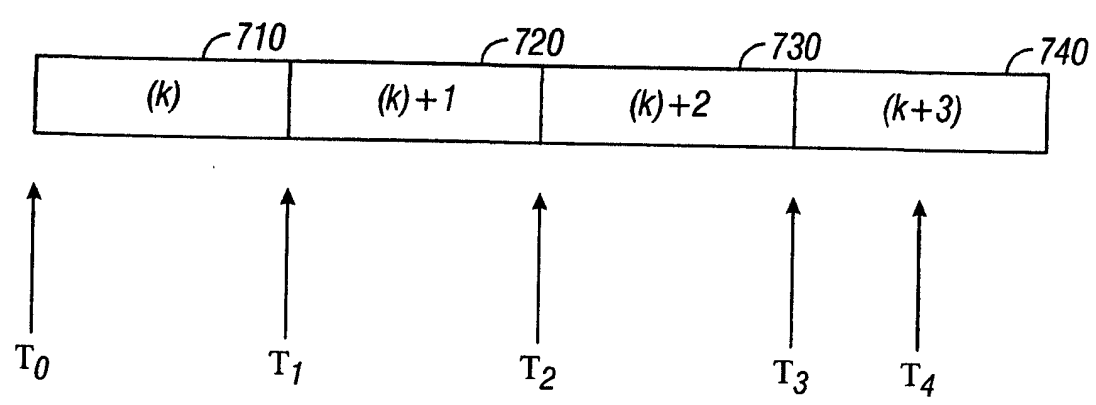


图 7