

(12) 发明专利

(10) 授权公告号 CN 101689131 B

(45) 授权公告日 2013.03.20

(21) 申请号 200780051010.6

(74) 专利代理机构 北京安信方达知识产权代理

(22) 申请日 2007.12.06

有限公司 11262

(30) 优先权数据

代理人 韩龙 阎斌斌

60/873,111 2006.12.06 US

(51) Int. Cl.

60/974,470 2007.09.22 US

G06F 11/10 (2006.01)

(85) PCT申请进入国家阶段日

(56) 对比文件

2009.08.06

US 2004/0250019 A1, 2004.12.09,

(86) PCT申请的申请数据

W0 98/28685 A1, 1998.07.02,

PCT/US2007/086705 2007.12.06

审查员 齐慧峰

(87) PCT申请的公布数据

WO2008/127458 EN 2008.10.23

(73) 专利权人 弗森 - 艾奥公司

权利要求书 7 页 说明书 61 页 附图 22 页

地址 美国犹他州

(72) 发明人 大卫·弗林 乔纳森·撒切尔

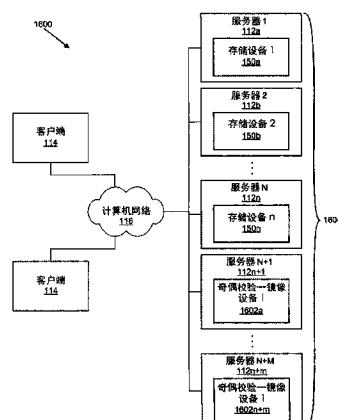
迈克尔·扎佩

(54) 发明名称

用于共享的、前端、分布式 RAID 的装置、系统
和方法

(57) 摘要

本发明公开了一种共享的前端分布式 RAID。多存储请求接收器模块 (2302) 接收来自至少两个客户端 (114) 的至少两个存储请求，以将文件或对象数据存储在存储设备集 (1604) 的一个或多个存储设备 (150) 中。所述存储请求是并发的存储请求并且具有至少一部分共有数据。存储设备集 (1604) 包括形成 RAID 群组的自主存储设备 (150)。每个存储设备 (150) 能够独立地通过网络 (116) 接收来自客户端 (114) 的存储请求。条带化模块 (2304) 计算条带模式并将每个条带的 N 个数
据段写入 N 个存储设备 (150)。奇偶校验 - 镜像模块 (2306) 将 N 个数据段的集写入奇偶校验 - 镜像存储设备 (1602)。定序器模块 (2308) 确保第一存储请求执行完成后才执行第二存储请求。



B

CN 101689131 B

1. 一种用于在前端、分布式 RAID 存储系统中通过一个或多个客户端管理可靠地存储共享数据的装置，所述装置包括：

多存储请求接收器模块，该多存储请求接收器模块接收来自至少两个客户端的至少两个存储请求，以将数据存储在存储设备集的一个或多个存储设备中，所述数据包括文件的数据或对象的数据，所述存储请求具有至少一部分共有数据，并且所述存储请求是并发的，这是由于所述存储请求的到达使得一个存储请求还没有完成时，至少两个存储请求中的另一个存储请求就到达，所述存储设备集包括形成 RAID 群组的自主存储设备，每个存储设备能够独立地通过网络接收来自客户端的存储请求；

条带化模块，该条带化模块为所述数据计算条带模式，所述条带模式包括一个或多个条带，每个条带包括 N 个数据段的集，该条带化模块还将条带的所述 N 个数据段写入所述存储设备集中的 N 个存储设备，其中，所述 N 个数据段中的每一个被分别写入所述存储设备集中的不同的存储设备并被分配给所述条带；

奇偶校验 - 镜像模块，该奇偶校验 - 镜像模块将所述条带的 N 个数据段的集写入所述存储设备集中的一个或多个奇偶校验 - 镜像存储设备，所述奇偶校验 - 镜像存储设备是除所述 N 个存储设备之外的设备；及

定序器模块，该定序器模块通过在执行第二存储请求之前接收来自存储设备中的每一个的应答，确保来自第一客户端的第一存储请求完成之后才执行来自第二客户端的所述第二存储请求，并发的所述至少两个存储请求包括所述第一存储请求和所述第二存储请求。

2. 如权利要求 1 所述的装置，其中，所述至少两个存储请求中的第一存储请求包括更新所述数据的一个或多个数据段，所述至少两个存储请求的第二存储请求包括下列之一：更新所述数据的一个或多个数据段和读取所述更新的数据的一个或多个数据段中的一个或多个。

3. 如权利要求 1 所述的装置，其中，所述应答确认存储请求已完成。

4. 如权利要求 1 所述的装置，其中，所述定序器模块通过下列方法之一选择在所述第二存储请求之前执行的第一存储请求：根据到达的顺序定序并发存储请求、根据时间戳定序并发存储请求及利用选择标准定序并发存储请求。

5. 如权利要求 1 所述的装置，进一步包括主控制器，所述主控制器包括所述定序器模块并通过所述多存储请求接收器模块、所述条带化模块和所述奇偶校验 - 镜像模块控制服务并发的所述至少两个存储请求。

6. 如权利要求 5 所述的装置，其中，所述主控制器运行在下列设备中的一个设备之中：

所述存储设备集中的存储设备；

客户端；及

第三方 RAID 管理设备。

7. 如权利要求 5 所述的装置，其中，所述主控制器包括主控制器群组中的一个，每个主控制器能够服务来自两个或多个客户端的存储请求，所述存储请求指令所述存储设备集的存储设备。

8. 如权利要求 7 所述的装置，进一步包括主验证模块，该主验证模块在执行接收到的存储请求之前确认服务所述接收到的存储请求的主控制器正在控制先于一个或多个并发存储请求的执行的所述存储请求的执行，其他主控制器接收所述一个或多个并发存储请

求,所述主控制器服务的存储请求与所述其他主控制器接收到的所述一个或多个并发存储请求至少有一部分数据相同。

9. 如权利要求 7 所述的装置,进一步包括主确定模块,该主确定模块发送主确定请求并接收响应所述主确定请求的主确定响应,所述主确定请求包括确定所述主控制器群组中的哪一个主控制器被分配给执行存储请求的查询,所述主确定响应使得所述主确定模块能够识别所述主控制群组中被分配为执行所述存储请求的主控制器,其中,所述主确定模块在发送所述存储请求之前从所述主确定响应中确定被标识为分配给执行所述存储请求的主控制器,所述存储请求包括为其他存储请求所共有的数据部分。

10. 如权利要求 9 所述的装置,其中,所述主确定模块运行在客户端内。

11. 如权利要求 9 所述的装置,进一步包括主错误模块,该主错误模块为响应下列状态之一而返回错误指示:

由主控制器控制的多存储请求接收器模块接收了不由所述主控制器控制的存储请求;及

所述主确定模块在所述存储请求执行完成时确定所述主控制器不再是确定的主控制器。

12. 如权利要求 5 所述的装置,进一步包括主错误模块,该主错误模块为响应下述状态而返回错误指示:由主控制器控制的多存储请求接收器模块接收了不由所述主控制器控制的存储请求。

13. 如权利要求 5 所述的装置,其中,所述主控制器控制传送给一个或多个次级主控制器的存储请求,所述次级主控制器控制用于存储在所述存储设备集的存储设备上的数据的存储请求。

14. 如权利要求 13 所述的装置,其中,所述主控制器起次级主控制器的作用,所述次级主控制器用于指令存储在所述存储设备集的存储设备上的数据的存储请求。

15. 如权利要求 5 所述的装置,其中,所述主控制器控制传送给一个或多个次级主控制器的存储请求,每个所述次级主控制器控制用于存储在对所述次级主控制器来说唯一的存储设备集的存储设备上的数据的存储请求。

16. 如权利要求 5 所述的装置,其中,所述主控制器包括奇偶校验 - 镜像存储设备中的一个。

17. 如权利要求 1 所述的装置,其中,下述操作中的至少一种发生在所述存储设备集的存储设备、客户端和第三方 RAID 管理设备中的一个设备上:接收存储数据的请求、计算条带模式和将 N 个数据段写入所述 N 个存储设备、将 N 个数据段的集写入奇偶校验 - 镜像存储设备。

18. 如权利要求 1 所述的装置,进一步包括奇偶校验生成模块,该奇偶校验生成模块为所述条带计算奇偶校验数据段并将所述奇偶校验数据段存储在奇偶校验 - 镜像存储设备上,由所述奇偶校验 - 镜像存储设备上的 N 个数据段的集计算奇偶校验条带。

19. 如权利要求 1 所述的装置,进一步包括奇偶校验更替模块,该奇偶校验更替模块为每个条带变更被分配为一个或多个用于所述条带的所述奇偶校验 - 镜像存储设备的所述存储设备集中的存储设备。

20. 一种在前端、分布式 RAID 存储系统中通过一个或多个客户端管理可靠地存储共享

数据的系统,所述系统包括 :

形成 RAID 群组的自主存储设备的存储设备集,所述存储设备独立地通过网络接收来自客户端的存储请求,其中,所述存储设备集中的一个或多个所述自主存储设备被指定为用于条带的奇偶校验 - 镜像存储设备;

多存储请求接收器模块,该多存储请求接收器模块接收来自至少两个客户端的至少两个存储请求,以将数据存储在所述存储设备集的一个或多个存储设备中,所述数据包括文件的数据或对象的数据,所述存储请求具有至少一部分共有数据,并且所述存储请求是并发的,这是由于所述存储请求的到达使得一个存储请求还没有完成时,两个存储请求中的另一个存储请求就到达;

条带化模块,该条带化模块为所述数据计算条带模式,所述条带模式包括一个或多个条带,每个条带包括 N 个数据段的集,该条带化模块还将条带的所述 N 个数据段写入所述存储设备集中的 N 个存储设备,其中,所述 N 个数据段中的每一个被分别写入所述存储设备集中的不同的存储设备并被分配给所述条带;

奇偶校验 - 镜像模块,该奇偶校验 - 镜像模块将所述条带的 N 个数据段的集写入所述存储设备集中的一个或多个奇偶校验 - 镜像存储设备,所述奇偶校验 - 镜像存储设备是除所述 N 个存储设备之外的设备;

定序器模块,该定序器模块通过在执行第二存储请求之前接收来自存储设备中的每一个的应答,确保来自第一客户端的第一存储请求完成之后才执行来自第二客户端的所述第二存储请求,并发的所述至少两个存储请求包括所述第一存储请求和所述第二存储请求。

21. 如权利要求 20 所述的系统,进一步包括一个或多个服务器,所述服务器包括所述存储设备集的存储设备。

22. 如权利要求 21 所述的系统,进一步包括所述一个或多个服务器内的至少两个客户端。

23. 一种用于在前端、分布式 RAID 存储系统中通过一个或多个客户端管理共享数据存储的方法,所述方法包括 :

接收来自至少两个客户端的至少两个存储请求,以将数据存储在存储设备集的一个或多个存储设备中,所述数据包括文件的数据或对象的数据,所述存储请求具有至少一部分共有数据,并且所述存储请求是并发的,这是由于所述存储请求的到达使得一个存储请求还没有完成时,两个存储请求中的另一个存储请求就到达,所述存储设备集包括形成 RAID 群组的自主存储设备,每个存储设备能够独立地通过网络接收来自客户端的存储请求;

为所述数据计算条带模式,所述条带模式包括一个或多个条带,每个条带包括 N 个数据段的集;

将条带的所述 N 个数据段写入所述存储设备集中的 N 个存储设备,其中,所述 N 个数据段中的每一个被分别写入所述存储设备集中的不同的存储设备并被分配给所述条带;

将所述条带的 N 个数据段的集写入所述存储设备集中的一个或多个奇偶校验 - 镜像存储设备,所述奇偶校验 - 镜像存储设备是除所述 N 个存储设备之外的设备;

通过在执行第二存储请求之前接收来自存储设备中的每一个的应答,确保来自第一客户端的第一存储请求完成之后才执行来自第二客户端的所述第二存储请求,并发的所述至少两个存储请求包括所述第一存储请求和所述第二存储请求。

24. 如权利要求 23 所述的方法,进一步包括在执行接收到的存储请求之前确认服务所述接收到的存储请求的主控制器正在控制先于一个或多个并发存储请求的执行的所述存储请求的执行,其他主控制器接收所述一个或多个并发存储请求,所述主控制器服务的存储请求与所述其他主控制器接收到的所述一个或多个并发存储请求至少有一部分数据相同。

25. 如权利要求 23 所述的方法,进一步包括发送主确定请求并接收响应所述主确定请求的主确定响应,所述主确定请求包括确定主控制器群组中的哪一个主控制器被分配为执行存储请求的查询,所述主确定响应使得所述主确定模块能够识别所述主控制群组中被分配给执行所述存储请求的主控制器,其中,主确定模块在发送所述存储请求之前从所述主确定响应中确定被标识为分配给执行所述存储请求的主控制器,所述存储请求包括为其他存储请求所共有的数据部分。

26. 如权利要求 25 所述的方法,进一步包括为响应下列状态之一而返回错误指示:

通过未被确定为用于存储请求所引用的数据的主控制器接收所述存储请求;及在所述存储请求执行完成时确定用于所述存储请求引用的数据的所述主控制器不再是确定的主控制器。

27. 一种在前端、分布式 RAID 存储系统中通过一个或多个客户端管理可靠地存储共享数据的方法,该方法包括:

接收来自至少两个客户端的至少两个并发存储请求,以将数据存储在存储设备集的一个或多个存储设备中,至少一个并发存储请求包括存储数据的请求,并发的存储请求具有相同的至少一部分数据,存储设备集包括形成 RAID 群组的自主存储设备,每个存储设备独立地从客户端直接接收存储请求;

为数据计算条带模式,所述条带模式包括一个或多个条带,每个条带包括 N 个数据段的集;

将条带的 N 个数据段写入存储设备集中的 N 个存储设备,其中,N 个数据段中的每一个被分别写入不同的存储设备;及

通过在执行第二存储请求之前接收来自存储设备中的每一个的应答,确保第一存储请求完成之后才执行至少两个并发存储请求中的所述第二存储请求。

28. 如权利要求 27 的方法,还包括执行接收到的存储请求之前确认服务接收到的存储请求的主控制器正在控制先于一个或多个并发存储请求的执行的存储请求的执行,其他主控制器接收一个或多个并发存储请求,所述主控制器服务的存储请求与所述其他主控制器接收到的一个或多个并发存储请求至少有一部分数据相同。

29. 如权利要求 27 的方法,还包括发送主确定请求并接收响应主确定请求的主确定响应,所述主确定请求包括确定主控制器群组中的哪一个主控制器被分配为执行存储请求的查询,主确定响应使得主确定模块能够识别被分配为执行所述存储请求的主控制群组中的主控制器,其中,主确定模块在发送所述存储请求之前从所述主确定响应中确定被标识为被分配为执行所述存储请求的主控制器,所述存储请求包括为其他存储请求所共有的数据部分。

30. 如权利要求 29 的方法,还包括为响应下列状态之一而返回错误指示:

通过主控制器接收存储请求,该主控制器未被确定为用于存储请求所引用的数据的主

控制器；及

在所述存储请求执行完成时，确定用于所述存储请求引用的数据的主控制器不再是确定的主控制器。

31. 一种在前端、分布式 RAID 存储系统中通过一个或多个客户端管理可靠地存储共享数据的装置，该装置包括：

多存储请求接收器模块，用于接收来自至少两个客户端的至少两个并发存储请求，以将数据存储在存储设备集的一个或多个存储设备中，至少一个并发存储请求包括存储数据的请求，并发的存储请求具有相同的至少一部分数据，存储设备集包括形成 RAID 群组的自主存储设备，每个存储设备独立地从客户端直接接收存储请求，而无须干预存储控制器；

条带化模块，用于为数据计算条带模式，所述条带模式包括一个或多个条带，每个条带包括 N 个数据段的集，并且所述条带化模块将条带的 N 个数据段写入存储设备集中的 N 个存储设备，其中，N 个数据段中的每一个被分别写入不同的存储设备；及

定序器模块，用于通过在执行第二存储请求之前接收来自存储设备中的每一个的应答，确保第一存储请求完成之后才执行至少两个并发存储请求中的所述第二存储请求。

32. 如权利要求 31 的装置，其中至少两个存储请求中的第一存储请求包括更新数据的一个或多个数据段，并且至少两个存储请求中的第二存储请求包括下述各项之一：更新数据的一个或多个数据段；读取所述更新的数据中的一个或多个数据段中的一个或多个。

33. 如权利要求 31 的装置，其中所述应答告知存储请求的完成。

34. 如权利要求 31 的装置，其中定序器模块通过下述方法之一在第二请求之前选择用于执行的第一存储请求：

通过到达顺序定序并发存储请求、通过时间戳定序并发存储请求和利用选择标准定序并发存储请求。

35. 如权利要求 31 的装置，其中该装置包括控制服务至少两个并发存储请求的主控制器。

36. 如权利要求 35 的装置，其中主控制器在下述设备之一中运行：

存储设备集的存储设备；

客户端；和

第三方 RAID 管理设备。

37. 如权利要求 35 的装置，其中主控制器包括主控制器群组中的一个，每个主控制器能够服务来自两个或多个客户端的存储请求，存储请求指令存储设备集的存储设备。

38. 如权利要求 37 的装置，还包括主验证模块，该主验证模块在执行接收到的存储请求之前确认服务接收到的存储请求的主控制器正在控制先于一个或多个并发存储请求的执行的存储请求的执行，其他主控制器接收一个或多个并发存储请求，所述主控制器服务的存储请求与其他主控制器接收到的一个或多个并发存储请求至少有一部分数据相同。

39. 如权利要求 37 的装置，还包括主确定模块，该主确定模块发送主确定请求并接收响应所述主确定请求的主确定响应，所述主确定请求包括确定所述主控制器群组中的哪一个主控制器被分配给执行存储请求的查询，所述主确定响应使得所述主确定模块能够识别被分配为执行所述存储请求的所述主控制器群组中的主控制器，其中，所述主确定模块在发送所述存储请求之前从所述主确定响应中确定被标识为分配给执行所述存储请求的主控

制器,所述存储请求包括为其他存储请求所共有的数据部分。

40. 如权利要求 37 的装置,其中主确定模块运行在客户端内。

41. 如权利要求 39 的装置,还包括主错误模块,该主错误模块为响应下列状态之一而返回错误指示:

由主控制器控制的多存储请求接收器模块接收了不由所述主控制器控制的存储请求;及

主确定模块在存储请求执行完成时确定主控制器不再是确定的主控制器。

42. 如权利要求 35 所述的装置,进一步包括主错误模块,该主错误模块为响应下述状态而返回错误指示:由主控制器控制的多存储请求接收器模块接收了不由所述主控制器控制的存储请求。

43. 如权利要求 35 所述的装置,其中,所述主控制器控制传送给一个或多个次级主控制器的存储请求,次级主控制器控制用于存储在所述存储设备集的存储设备上的数据的存储请求。

44. 如权利要求 43 的装置,其中,主控制器起用于指令存储在所述存储设备集的存储设备上的数据的存储请求的次级主控制器的作用。

45. 如权利要求 35 的装置,其中,主控制器控制传送给一个或多个次级主控制器的存储请求,每个次级主控制器控制用于存储在对次级主控制器来说唯一的存储设备集的存储设备上的数据的存储请求。

46. 如权利要求 31 的装置,还包括奇偶校验 - 镜像模块,该奇偶校验 - 镜像模块将条带的 N 个数据段的集写入存储设备集中的一个或多个奇偶校验 - 镜像存储设备,奇偶校验 - 镜像存储设备是除所述 N 个存储设备之外的设备。

47. 如权利要求 46 的装置,其中,接收存储数据的请求、计算条带模式和将 N 个数据段写入 N 个存储设备、以及将 N 个数据段的休写入奇偶校验 - 镜像存储设备中的一项发生在下述设备之一上:

存储设备集的存储设备;

客户端;和

第三方 RAID 管理设备。

48. 如权利要求 46 的装置,还包括偶校验生成模块,该奇偶校验生成模块为所述条带计算奇偶校验数据段并将所述奇偶校验数据段存储在奇偶校验 - 镜像存储设备上,由所述奇偶校验 - 镜像存储设备上的 N 个数据段的集计算奇偶校验条带。

49. 如权利要求 46 的装置,还包括奇偶校验更替模块,该奇偶校验更替模块为每个条带变更被分配为一个或多个用于所述条带的所述奇偶校验 - 镜像存储设备的所述存储设备集中的存储设备。

50. 一种在前端、分布式 RAID 存储系统中通过一个或多个客户端管理可靠地存储共享数据的系统,所述系统包括:

形成 RAID 群组的自主存储设备的存储设备集,存储设备独立地从客户端直接接收存储请求;

多存储请求接收器模块,用于接收来自至少两个客户端的至少两个并发存储请求,用于存储存储设备集的一个或多个存储设备中的数据,至少一个并发存储请求包括存储数据

的请求，并发的存储请求具有相同的至少一部分数据；

条带化模块，用于为数据计算条带模式，所述条带模式包括一个或多个条带，每个条带包括 N 个数据段的集，并且所述条带化模块将条带的 N 个数据段写入存储设备集中的 N 个存储设备，其中，N 个数据段中的每一个被分别写入不同的存储设备；及

定序器模块，用于通过在执行第二存储请求之前接收来自存储设备中的每一个的应答，确保第一存储请求完成之后才执行至少两个并发存储请求中的所述第二存储请求。

51. 如权利要求 50 的系统，还包括一个或多个服务器，服务器包括存储设备集的存储设备。

52. 如权利要求 51 的系统，还包括运行在一个或多个服务器内的所述至少两个客户端。

用于共享的、前端、分布式 RAID 的装置、系统和方法

[0001] 相关申请的交叉引用

[0002] 本申请要求下述申请的优先权 :David Flynn 等人于 2006 年 12 月 6 日提交的题为“Elemental Blade System”的美国临时专利申请 (申请号为 :60/873, 111) ; David Flynn 等人于 2007 年 9 月 22 日提交的题为“Apparatus, System, and Method for Object-Oriented Solid-State Storage”的美国临时专利申请 (申请号为 :60/974, 470)。上述申请通过引用并入本文中。

[0003] 发明背景

技术领域

[0004] 本发明涉及数据存储器, 具体地, 涉及利用共享的、前端的、分布式的独立驱动器冗余阵列 (“RAID”) 的数据存储器。

背景技术

[0005] 传统的 RAID 利用磁盘或其他存储设备的阵列, 其中, 所述存储设备中的每一个的至少一部分被用于 RAID 并形成 RAID 群组。RAID 控制器管理传送到所述 RAID 群组的存储语法。对于冗余系统来说, RAID 控制器具有备用 RAID 控制器, 如果主 RAID 控制器出现故障或不可用, 该备用 RAID 控制器准备好接管主 RAID 控制器。来自试图访问存储在 RAID 中相同数据的多个客户端的存储请求按到达的顺序被顺序地执行。

[0006] 前端、分布式的 RAID 系统包括自主存储设备, 所述自主存储设备中的每一个都包括起分布式 RAID 控制器作用的存储控制器, 并且所述存储设备均能被配置在多个、重叠的、服务多个客户端的 RAID 群组中。必要的时候, 两个客户端可试图访问相同的数据。如果一个存储请求先到达并执行, 通常就不会出现数据的不一致。另一方面, 如果用于同一数据的两个或更多个存储请求同时到达或几乎同时到达, 数据可能会被损坏。

[0007] 例如, 如果数据存储在 RAID 群组中的四个存储设备中, 其中, 所述存储设备中的一个被指定为奇偶校验 - 镜像存储设备, 第一客户端将存储请求发送给作为 RAID 控制器的第一存储控制器, 第二客户端将第二存储请求发送给作为第二 RAID 控制器的第二存储设备, 并且这两个存储请求访问相同的数据, 所述第一存储设备可开始在所述第一存储设备上执行所述存储请求, 然后, 在 RAID 群组中的另一个存储设备上执行所述存储请求。同时, 第二存储设备上的所述第二 RAID 控制器可开始在另一个存储设备上执行所述第二存储请求, 然后, 在 RAID 群组中余下的存储设备中执行所述第二存储请求。这种执行上的不匹配可能是由于下述原因 : 存储设备之间的物理距离、执行时间不一致, 等等。这种方法可能会损坏数据。

发明内容

[0008] 需要一种用于处理访问相同数据的并发存储请求的共享的、前端、分布式 RAID 的系统、装置和方法。有利地是, 这种系统、装置和方法可控制访问数据, 使得执行完一个存储

请求后再执行第二存储请求。

[0009] 本发明是针对现有技术的现况开发出来的，具体地，是针对现有技术中通过现有 RAID 并未完全解决的问题和需要。因此，本发明已经被开发出来以提供克服现有技术中的上述多数或全部缺陷的用于共享的、前端、分布式 RAID 的装置、系统及方法。

[0010] 所述装置具有多个模块，包括多存储请求接收器模块、条带化模块、奇偶校验 - 镜像模块和定序器模块。所述多存储请求接收器模块接收来自至少两个客户端的至少两个存储请求，以将数据存储在存储设备集的一个或多个存储设备中。所述数据包括文件的数据或对象的数据。所述存储请求具有至少一部分共有数据，并且所述存储请求是并发的，这是由于所述存储请求的到达使得一个存储请求还没有完成时，两个存储请求中的另一个存储请求就到达。所述存储设备集包括形成 RAID 群组的自主存储设备，其中，每个存储设备能够独立地通过网络接收来自客户端的存储请求。

[0011] 所述条带化模块为所述数据计算条带模式。所述条带模式包括一个或多个条带。每个条带包括 N 个数据段的集。所述条带化模块还将条带的所述 N 个数据段写入所述存储设备集中的 N 个存储设备，其中，所述 N 个数据段中的每一个被写入所述存储设备集中的不同的存储设备并被分配给所述条带。所述奇偶校验 - 镜像模块将所述条带的 N 个数据段的集写入所述存储设备集中的一一个或多个奇偶校验 - 镜像存储设备。所述奇偶校验 - 镜像存储设备是除所述 N 个存储设备之外的设备。所述定序器模块确保来自第一客户端的第一存储请求完成之后才执行来自第二客户端的第二存储请求，其中，并发的所述至少两个存储请求包括所述第一存储请求和所述第二存储请求。

[0012] 在一种实施方式中，所述至少两个存储请求的第一存储请求包括更新所述数据的一个或多个数据段，所述至少两个存储请求的第二存储请求包括下列之一：更新所述数据的一个或多个数据段和读取所述数据的相同的一个或多个数据段中的一个或多个。在另一种实施方式中，确保第一存储请求完成之后才执行第二存储请求进一步包括：接收来自存储设备集的存储设备中的每一个的应答，所述存储设备在执行所述第二存储请求之前一起接收第一存储请求和存储请求。所述应答确认存储请求已完成。在另一种实施方式中，所述定序器模块通过下列方法之一选择在所述第二存储请求之前执行的第一存储请求：根据到达的顺序定序并发存储请求、根据时间戳定序并发存储请求及利用选择标准定序并发存储请求。

[0013] 在一种实施方式中，所述装置包括主控制器，其中，所述主控制器包括所述定序器模块并通过所述多存储请求接收器模块、所述条带化模块和所述奇偶校验 - 镜像模块控制服务并发的至少两个存储请求。在另一种实施方式中，所述主控制器运行在所述存储设备集中的存储设备、客户端和第三方 RAID 管理设备内。在另一种实施方式中，所述主控制器包括主控制器群组中的一个，每个主控制器能够服务来自一个或多个客户端的存储请求。所述存储请求指令所述存储设备集的存储设备。

[0014] 在一种实施方式中，所述装置包括主验证模块，该主验证模块在执行接收到的存储请求之前确认服务所述接收到的存储请求的主控制器正在控制先于一个或多个并发存储请求的执行的所述存储请求的执行。其他主控制器接收所述一个或多个并发存储请求，服务请求与所述其他主控制器接收到的所述一个或多个并发存储请求至少有一部分数据相同。

[0015] 在另一种实施方式中，所述装置包括主确定模块，该主确定模块发送主确定请求并接收响应所述主确定请求的主确定响应。所述主确定请求包括确定所述主控制器群组中的哪一个主控制器被分配给执行存储请求的查询。所述主确定响应使得所述主确定模块能够识别所述主控制群组中被分配为执行所述存储请求的主控制器。所述主确定模块在发送所述存储请求之前从所述主确定响应中确定被标识为分配给执行所述存储请求的主控制器。所述存储请求包括为其他存储请求所共有的数据部分。在另一种实施方式中，所述主确定模块在客户端内执行。

[0016] 在一种实施方式中，所述装置包括主错误模块，该主错误模块为响应下列状态之一而返回错误指示：由主控制器控制的多存储请求接收器模块接收了不由所述主控制器控制的存储请求；及所述主确定模块在所述存储请求执行完成时确定所述主控制器不再是确定的主控制器。在另一种实施方式中，所述主错误模块为响应下述状态而返回错误指示：由主控制器控制的多存储请求接收器模块接收了不由所述主控制器控制的存储请求。

[0017] 在一种实施方式中，所述主控制器控制传送给一个或多个次级主控制器的存储请求，所述次级主控制器控制用于存储在所述存储设备集的存储设备上的数据的存储请求。在另一种实施方式中，所述主控制器起次级主控制器的作用，所述次级主控制器用于指令存储在所述存储设备集的存储设备上的数据的存储请求。在一种实施方式中，所述主控制器控制传送给一个或多个次级主控制器的存储请求，其中，每个所述次级主控制器控制用于存储在对所述次级主控制器来说唯一的存储设备集的存储设备上的数据的存储请求。在另一种实施方式中，所述主控制器包括奇偶校验-镜像存储设备中的一个。

[0018] 在一种实施方式中，下述操作中的至少一种发生在所述存储设备集的存储设备、客户端和第三方 RAID 管理设备中的一个设备上：接收存储数据的请求、计算条带模式和将 N 个数据段写入所述 N 个存储设备、将 N 个数据段的集写入奇偶校验-镜像存储设备。在另一种实施方式中，所述装置包括奇偶校验生成模块，该奇偶校验生成模块为所述条带计算奇偶校验数据段并将所述奇偶校验数据段存储在奇偶校验-镜像存储设备上，其中，由所述奇偶校验-镜像存储设备上的 N 个数据段的集计算奇偶校验条带。在另一种实施方式中，所述装置包括奇偶校验更替模块，该奇偶校验更替模块为每个条带变更被分配为一个或多个用于所述条带的所述奇偶校验-镜像存储设备的所述存储设备集中的存储设备。

[0019] 还提出了本发明的一种用共享的、前端、分布式 RAID 的系统。所述系统大体上包括上文中关于所述装置描述的模块和实施方式。具体地，在一种实施方式中，所述系统包括形成 RAID 群组的自主存储设备的存储设备集。所述存储设备独立地通过网络接收来自客户端的存储请求。所述存储设备集中的一个或多个所述自主存储设备被指定为用于条带的奇偶校验-镜像存储设备。

[0020] 所述系统包括多存储请求接收器模块，该多存储请求接收器模块接收来自至少两个客户端的至少两个存储请求，以将数据存储在所述存储设备集的一个或多个存储设备中。所述数据包括文件的数据或对象的数据。所述存储请求具有至少一部分共有数据，并且所述存储请求是并发的，这是由于所述存储请求的到达使得一个存储请求还没有完成时，两个存储请求中的另一个存储请求就到达。所述系统包括条带化模块，该条带化模块为所述数据计算条带模式。所述条带模式包括一个或多个条带，每个条带包括 N 个数据段的集。该条带化模块还将条带的所述 N 个数据段写入所述存储设备集中的 N 个存储设备。所述 N

个数据段中的每一个被写入所述存储设备集中的不同的存储设备并被分配给所述条带。

[0021] 所述系统包括奇偶校验 - 镜像模块, 该奇偶校验 - 镜像模块将所述条带的 N 个数据段的集写入所述存储设备集中的一个或多个奇偶校验 - 镜像存储设备, 所述奇偶校验 - 镜像存储设备是除所述 N 个存储设备之外的设备。所述系统包括定序器模块, 该定序器模块确保来自第一客户端的第一存储请求完成之后才执行来自第二客户端的第二存储请求。并发的所述至少两个存储请求包括所述第一存储请求和所述第二存储请求。在一种实施方式中, 所述系统包括一个或多个服务器, 所述服务器包括所述存储设备集的存储设备。在另一种实施方式中, 所述系统包括所述一个或多个服务器内的一个或多个客户端。

[0022] 还提出了本发明的一种用于共享的、前端、分布式 RAID 的方法。公开的实施方式中的该方法大体上包括实现上述与所述装置和系统的运行有关的功能的必要步骤。在一种实施方式中, 所述方法包括接收来自至少两个客户端的至少两个存储请求, 以将数据存储在存储设备集的一个或多个存储设备中。所述数据包括文件的数据或对象的数据。所述存储请求具有至少一部分共有数据, 并且所述存储请求是并发的, 这是由于所述存储请求的到达使得一个存储请求还没有完成时, 两个存储请求中的另一个存储请求就到达。所述存储设备集包括形成 RAID 群组的自主存储设备, 每个存储设备能够独立地通过网络接收来自客户端的存储请求。

[0023] 所述方法包括为所述数据计算条带模式, 其中, 所述条带模式包括一个或多个条带, 每个条带包括 N 个数据段的集。所述方法包括将条带的所述 N 个数据段写入所述存储设备集中的 N 个存储设备。所述 N 个数据段中的每一个被写入所述存储设备集中的不同的存储设备并被分配给所述条带。所述方法包括将所述条带的 N 个数据段的集写入所述存储设备集中的一个或多个奇偶校验 - 镜像存储设备, 所述奇偶校验 - 镜像存储设备是除所述 N 个存储设备之外的设备。所述方法包括确保来自第一客户端的第一存储请求完成之后才执行来自第二客户端的第二存储请求, 并发的所述至少两个存储请求包括所述第一存储请求和所述第二存储请求。

[0024] 在一种实施方式中, 所述方法包括在执行接收到的存储请求之前确认服务所述接收到的存储请求的主控制器正在控制先于一个或多个并发存储请求的执行的所述存储请求的执行。其他主控制器接收所述一个或多个并发存储请求, 服务请求与所述其他主控制器接收到的所述一个或多个并发存储请求至少有一部分数据相同。

[0025] 在另一种实施方式中, 所述方法包括发送主确定请求并接收响应所述主确定请求的主确定响应。所述主确定请求包括确定所述主控制器群组中的哪一个主控制器被分配为执行存储请求的查询。所述主确定响应使得所述主确定模块能够识别所述主控制器群组中被分配给执行所述存储请求的主控制器。主确定模块在发送所述存储请求之前从所述主确定响应中确定被标识为分配给执行所述存储请求的主控制器。所述存储请求包括为所述其他请求所共有的数据部分。

[0026] 在另一种实施方式中, 所述方法包括为响应下列状态之一而返回错误指示: 通过未被确定为用于存储请求所引用的数据的主控制器的主控制器接收所述存储请求; 及在所述存储请求执行完成时确定用于所述存储请求引用的数据的所述主控制器不再是确定的主控制器。

[0027] 本说明书全文所提到的特征、优点或者类似措辞并不意味着可在本发明包含在本

发明的任一单独的实施方式中的情况下实现所有的特征和优点。当然，涉及特征和优点的措辞被理解为意味着：与实施方式一起描述的特定的特征、优点或者特点包括在本发明的至少一种实施方式中。因此，在本说明书全文中，关于特征、优点和类似措辞的讨论可（但未必）涉及同一实施方式。

[0028] 此外，描述的本发明的特征、优点和特点可采用任何合适的方式与一个或多个实施方式结合。相关领域的技术人员可意识到本发明可在不具备特定实施方式的一个或多个具体特征或优点的情况下被实施。在其他例子中，可意识到附加特征和优点出现在某些实施方式中，而不是在本发明的所有实施方式中都出现。

[0029] 通过下面的说明和附加的权利要求，本发明的这些特征和优点将变得更加充分的显而易见，或者可以通过按下文所阐述的实施本发明的方法而获悉。

附图说明

[0030] 为了使本发明的优点更加容易理解，会参考附图中示出的特定实施方式给出上面简要描述的本发明的更具体的说明。在理解到这些附图仅描述了本发明的一般实施方式并且并不因此认为本发明限于此范围的情况下，将通过使用附图并结合更多的具体特征和细节描述和解释本发明，附图中：

[0031] 图 1A 是示意性框图，示出了根据本发明的用于固态存储设备内的数据管理的系统的一种实施方式；

[0032] 图 1B 是示意性框图，示出了根据本发明的用于存储设备内的对象管理的系统的一种实施方式；

[0033] 图 1C 是示意性框图，示出了根据本发明的用于服务器内的存储区域网络的系统的一种实施方式；

[0034] 图 2A 是示意性框图，示出了根据本发明的用于在固态存储设备内的对象管理的装置的一种实施方式；

[0035] 图 2B 是示意性框图，示出了根据本发明的位于固态存储设备内的固态存储设备控制器的一种实施方式；

[0036] 图 3 是示出了根据本发明的位于固态存储设备内的固态存储设备控制器的一种实施方式的示意性框图，该固态存储设备控制器具有写入数据管道和读取数据管道；

[0037] 图 4A 是示意性框图，示出了根据本发明的位于固态存储控制器内的内存库交错控制器的一种实施方式；

[0038] 图 4B 是示意性框图，示出了根据本发明的位于固态存储设备内的内存库交错控制器的一种替代实施方式；

[0039] 图 5A 是示意性流程图，示出了根据本发明的在固态存储设备内采用数据管道管理数据的方法的一种实施方式；

[0040] 图 5B 是示意性流程图，示出了根据本发明的用于服务器内的 SAN 的方法的一种实施方式。

[0041] 图 6 是示意性流程图，示出了根据本发明的在固态存储设备内采用数据管道管理数据的方法的另一种实施方式；

[0042] 图 7 是示意性流程图，示出了根据本发明的在固态存储设备内利用内存库交错管

理数据的方法的一种实施方式；

[0043] 图 8 是示意性框图,示出了根据本发明的在固态存储设备内的收集垃圾的装置的一种实施方式；

[0044] 图 9 是示意性流程图,示出了根据本发明的在固态存储设备内的收集垃圾的方法的一种实施方式

[0045] 图 10 是示意性框图,示出了根据本发明的用于渐进式 RAID 和前端分布式 RAID 的系统的一种实施方式；

[0046] 图 11 是示意性框图,示出了根据本发明的用于渐进式 RAID 的装置的一种实施方式；

[0047] 图 12 是示意性框图,示出了根据本发明的利用渐进式 RAID 更新数据段的装置的一种实施方式；

[0048] 图 13 是示意性流程图,示出了根据本发明的利用渐进式 RAID 管理数据的方法的一种实施方式；

[0049] 图 14 是示意性流程图,示出了根据本发明的利用渐进式 RAID 更新数据段的方法的一种实施方式；

[0050] 图 15 是示意性框图,示出了根据本发明的用于前端分布式 RAID 的装置的一种实施方式；

[0051] 图 16 是示意性流程图,示出了根据本发明的用于前端分布式 RAID 的方法的一种实施方式；

[0052] 图 17 是示意性框图,示出了根据本发明的用于共享的、前端、分布式 RAID 的装置的一种实施方式；

[0053] 图 18 是示意性流程图,示出了根据本发明的用于共享的、前端、分布式 RAID 的方法的一种实施方式。

具体实施方式

[0054] 为了更显著地强调功能性单元运行的独立性,在本说明书中描述的许多功能性单元已被标示为模块。例如,模块可作为硬件电路来实施,所述硬件电路包括自定义 VLSI 电路、门阵列或成品半导体(例如逻辑芯片、晶体管或其他分立元件)。模块也可在可编程硬件设备(如现场可编程门阵列、可编程阵列逻辑、可编程逻辑设备或类似设备)内实施。

[0055] 模块还可在由不同类型的处理器运行的软件中实施。例如,可执行代码的识别模块可以包括一个或多个计算机指令物理块或逻辑块,该计算机指令被作为对象、程序或函数来组织。然而,识别模块的可执行文件不必在物理上位于一起,但是可包括存储在不同位置的不同命令,当这些命令在逻辑上连接在一起时,所述命令包括所述模块并实现所述模块的指定目标。

[0056] 当然,可执行代码的模块可以为一个或许多指令,并且甚至可以分布在若干不同的代码段中、分布在不同的程序中并可分布在多个存储设备中。类似地,可以在此在模块内识别并示出运算数据,并且可以以任何合适的形式体现所述运算数据并在任意合适类型的数据结构中组织所述运算数据。所述运算数据可作为单数据集收集,或者可以分布在不同的位置(包括不同的存储设备),并且可在系统或网络中至少部分地仅作为电信号存在。当

模块或模块的部分在软件中实施时，软件部分被存储在一个或多个计算机可读媒体上。

[0057] 本说明书全文所提到的“一种实施方式”、“实施方式”或类似的措辞意味着与实施方式一起描述的特定的特征、结构或特点包括在本发明的至少一种实施方式中。因此，在本说明书全文中，短语“在一种实施方式中”、“在实施方式中”及类似措辞的出现可（但未必）涉及同一实施方式。

[0058] 提及信号承载媒介可采取任何能够生成信号、导致信号生成或者导致在数字处理设备上执行机器可读命令程序的形式。信号承载媒介可通过下述设备体现：传输线、光盘、数字视频光盘、磁带、伯努利驱动器、磁盘、穿孔卡、闪存、集成电路或其他数字处理装置存储设备。

[0059] 此外，描述的本发明的特征、结构或特点可以以任何合适的方式合并在一种或多种实施方式中。在下文的说明中，提供了大量的具体细节以全面理解本发明的实施方式，所述具体细节比如编程、软件模块、用户选择、网络事务、数据库查询、数据库结构、硬件模块、硬件电路、硬件芯片等等的实例。然而，相关技术领域的技术人员可认识到：本发明可在不具备一个或多个具体实施方式的具体细节的情况下被实施，或者本发明可结合其他方法、组件、材料等实施。在其他例子中，并没有显示或描述公知的结构、材料或操作以使本发明变得清晰。

[0060] 此处包括的示意性流程图大体上是作为逻辑流程图来列举的。就这点而言，描述的顺序和标记的步骤是本方法的一种实施方式的指示性说明。可设想其他在功能上、逻辑上或效果上与图示方法的一个或多个步骤（或其中部分）相同的步骤和方法。此外，使用的格式和符号被用于解释方法的逻辑步骤并被理解为不限制本方法的范围。尽管在流程图中可使用不同的箭头类型和线条类型，但这些箭头类型和线条类型被理解为不限制相应方法的范围。的确，一些箭头或其他连接器可用于仅表示方法的逻辑流程。例如，箭头可表示描述的方法的列举的步骤之间的未指明期间的等待或监测时期。此外，特定方法的步骤的顺序可或可不严格依照所示的对应步骤的顺序。

[0061] 固态存储系统

[0062] 图 1A 是示意性框图，示出了根据本发明的用于固态存储设备内的数据管理的系统 100 的一种实施方式。系统 100 包括固态存储设备 102、固态存储控制器 104、写入数据管道 106、读取数据管道 108、固态存储器 110、计算机 112、客户端 114 和计算机网络 116，这些装置描述如下。

[0063] 系统 100 包括至少一个固态存储设备 102。在另一种实施方式中，系统 100 包括两个或更多个固态存储设备 102，每个固态存储设备 102 可包括非易失性的、固态的存储器 110，所述非易失性的、固态的存储器例如纳米随机存取存储器（“纳米 RAM”或者“NRAM”）、磁电阻式 RAM（“MRAM”）、动态 RAM（“DRAM”）、相变 RAM（“PRAM”）闪存等等。结合图 2 和图 3 更详细地描述了固态存储设备 102。固态存储设备 102 被描述成位于通过计算机网络 116 与客户端 114 相连的计算机 112 内。在一种实施方式中，固态存储设备 102 位于计算机 112 内部并且采用系统总线连接，所述系统总线例如快速外围组件互连（“PCI-e”）总线、串行高级技术附件（“串行 ATA”）总线或类似总线。在另一种实施方式吧，固态存储设备 102 位于计算机 112 外部，并且通过通用串行总线（“USB”）、电气与电子工程师协会（“IEEE”）1394 总线（“火线”）或类似总线连接。在其他实施方式中，固态存储设备 102

采用下述方式与计算机 112 相连接：外围组件互连（“PCI”）express 总线、外部电或光总线扩展或者总线网络解决方案，所述总线网络解决方案例如无限带宽或快速 PCI 高级交换（“PCIe-AS”）或类似技术。

[0064] 在不同的实施方式中，固态存储设备 102 可以是双列直插式内存模块（“DIMM”）、子卡或微型模块的形式。在另一种实施方式中，固态存储设备 102 是位于机架式刀片内的元件。在另一种实施方式中，固态存储设备 102 包含在直接集成到高级集成装置（如主板、笔记本电脑、图形处理器）的封装内。在另一种实施方式中，包括固态存储设备 102 的单独元件直接集成到高级集成装置上而不经过中间封装。

[0065] 固态存储设备 102 包括一个或多个固态存储控制器 104，每个固态存储控制器 104 可包括写入数据管道 106 和读取数据管道 108，而且，每个固态存储控制器 104 还包括固态存储器 110，这将在下文中结合图 2 和图 3 详细说明。

[0066] 系统 100 包括一台或多台连接到固态存储设备 102 的计算机 112。计算机 112 可以是主机、服务器、存储区域网络（“SAN”）的存储控制器、工作站、个人计算机、笔记本电脑、手持式计算机、超级计算机、计算机集群、网络交换机、路由器或设备、数据库或存储设备、数据采集或数据采集系统、诊断系统、测试系统、机器人、便携式电子设备、无线设备或类似设备。在另一种实施方式中，计算机 112 可以是客户端，并且固态存储设备 102 自主运行以应答发送自计算机 112 的数据请求。在这种实施方式中，计算机 112 和固态存储设备 102 可采用下列方式连接：计算机网络、系统总线或其他适于在计算机 112 和自主固态存储设备 102 之间连接的通信手段。

[0067] 在一种实施方式中，系统 100 包括一个或多个客户端 114，所述一个或多个客户端 114 通过一个或多个计算机网络 116 连接到一台或多台计算机 112。客户端 114 可以是主机、服务器、SAN 的存储控制器、工作站、个人计算机、笔记本电脑、手持式计算机、超级计算机、计算机集群、网络交换机、路由器或设备、数据库或存储设备、数据采集或数据采集系统、诊断系统、测试系统、机器人、便携式电子设备、无线设备或类似设备。计算机网络 116 可包括因特网、广域网（“WAN”）、城域网（“MAN”）、局域网（“LAN”）、令牌环网、无线网络、光纤通道网络、SAN、网络附属存储（“NAS”）、ESCON 或类似网络、或者是网络的任意组合。计算机网络 116 还可包括来自 IEEE802 系列网络技术中的网络，如以太网、令牌环网、WiFi、WiMax 及类似网络。

[0068] 计算机网络 116 可包括服务器、交换机、路由器、电缆、无线电和其他用于促进计算机 112 和客户端 114 的网络连接的设备。在一种实施方式中，系统 100 包括通过计算机网络 116 进行对等通信的多台计算机 112。在另一种实施方式中，系统 100 包括通过计算机网络 116 进行对等通信的多个固态存储设备 102。本领域技术人员可认识到其他计算机网络 116 可包括一个或多个计算机网络 116 以及相关设备，所述相关设备具有一个或多个客户端 114、其他计算机或与一台或多台计算机 112 相连的一个或多个固态存储设备 102 之间的单个或冗余连接，所述其他计算机具有一个或多个固态存储设备 102。在一种实施方式中，系统 100 包括两个或更多个通过计算机网络 116 连接到客户端 114 的固态存储设备 102，而不包括计算机 112。

[0069] 存储控制器管理的对象

[0070] 图 1B 是示意性框图，示出了根据本发明的用于存储设备内的对象管理的系统 101

的一种实施方式。系统 101 包括一个或多个存储设备 150 (每一个存储设备 150 都具有存储控制器 152 和一个或多个数据存储设备 154) 和一个或多个请求设备 155。存储设备 150 联网在一起并与一个或多个请求设备 155 连接。请求设备 155 将对象请求发给存储设备 150a。对象请求可以是创建对象的请求、向对象写入数据的请求、从对象读取数据的请求、删除对象的请求、检查对象的请求、复制对象的请求及类似请求。本领域技术人员会认识到其他对象请求。

[0071] 在一种实施方式中,存储控制器 152 和数据存储设备 154 是分离的设备。在另一种实施方式中,存储控制器 152 和数据存储设备 154 集成到一个存储设备 150 上。在另一种实施方式中,数据存储设备 154 为固态存储器 110,而存储控制器 152 为固态存储设备控制器 202。在其他实施方式中,数据存储设备 154 可以为硬盘驱动器、光驱动器、磁带存储器或类似存储设备。在另一种实施方式中,存储设备 150 可包括两个或更多个不同类型的数据存储设备 154。

[0072] 在一种实施方式中,数据存储设备 154 为固态存储器 110,并且被布置为固态存储元件 216、218、220 的阵列。在另一种实施方式中,固态存储器 110 被布置在两个或更多个内存库 (bank) 214a-n 内。下文结合图 2B 更详细地描述了固态存储器 110。

[0073] 存储设备 150a-n 可联网在一起并且可作为分布式存储设备运行。与请求设备 155 连接的存储设备 150a 控制发送到所述分布式存储设备的对象请求。在一种实施方式中,存储设备 150 和关联的存储控制器 152 管理对象并对请求设备 155 来说表现为分布式对象文件系统。在这种情况下,一类分布式对象文件系统的实例是并行对象文件系统。在另一种实施方式中,存储设备 150 和关联的存储控制器 152 管理对象并对请求设备 155 来说表现为分布式对象文件服务器。在这种情况下,一类分布式对象文件服务器的实例是并行对象文件服务器。在这些和其他实施方式中,请求设备 155 可只管理对象或者与存储设备 150 结合而参与管理对象,这通常并不将存储设备 150 的功能限制在为其他客户端 114 充分管理对象的范围内。在退化情况下,每个分布式存储设备、分布式对象文件系统和分布式对象文件服务器能作为单个设备独立运行。联网的存储设备 150a-n 可作为分布式存储设备、分布式对象文件系统、分布式对象文件服务器和它们的任意组合运行,所述组合具有一个或多个为一个或多个请求设备 155 配置的这些功能。例如,存储设备 150 可配置为:为第一请求设备 155a 作为分布式存储设备运行,而请求设备 155b 作为分布式存储设备和分布式对象文件系统为运行。当系统 101 包括一个存储设备 150a 时,存储设备 150a 的存储控制器 152a 管理对象并对请求设备 155 来说表现为对象文件系统或对象文件服务器。

[0074] 在一种实施方式中,其中,存储设备 150 作为分布式存储设备联网在一起,存储设备 150 充当由一个或多个分布式存储控制器 152 管理的独立驱动器冗余阵列 (“RAID”)。例如,写入对象数据段的请求导致所述数据段根据 RAID 级别在数据存储设备 154a-n 中被条带化为具有奇偶校验条带的条带。这种布置的一个好处是这种对象管理系统可在单独的存储设备 150 (无论是存储控制器 152、数据存储设备 154 或存储设备 150 的其他组件) 出现故障时继续使用。

[0075] 当冗余网络用于互连存储设备 150 和请求设备 155 时,所述对象管理系统可在出现网络故障的情况下(只要网络中的一个仍在运行)继续使用。具有一个存储设备 150a 的系统 101 还可包括多个数据存储设备 154a,而存储设备 150a 的存储控制器 152a 可作为

RAID 控制器运行并在存储设备 150a 的数据存储设备 154a 间分割数据段, 存储设备 150a 的存储控制器 152a 可包括根据 RAID 级别的奇偶校验条带。

[0076] 在一种实施方式中, 其中, 一个或多个存储设备 150a-n 是具有固态存储设备控制器 202 和固态存储器 110 的固态存储设备 102, 固态存储设备 102 可配置为 DIMM 配置、子卡、微型模块等, 并保留在计算机 112 内。计算机 112 可以是服务器或具有固态存储设备 102 的类似设备, 固态存储设备 102 联网在一起并作为分布式 RAID 控制器运行。有利地是, 存储设备 102 可采用 PCI-e、PCIe-AS、无限带宽或其他高性能总线、交换总线、网络总线或网络连接, 并且可提供极致密型、高性能的 RAID 存储系统, 在该系统中, 单独的或分布式固态存储控制器 202 自主地在固态存储器 110a-n 间条带化数据段。

[0077] 在一种实施方式中, 请求设备 155 用于与存储设备 150 通信的同一网络可被对等存储设备 150a 使用, 以与对等存储设备 150b-n 通信以实现 RAID 功能。在另一种实施方式中, 可为了 RAID 的目的而在存储设备 150 间使用单独的网络。在另一种实施方式中, 请求设备 155 可通过向存储设备 150 发送冗余请求而参与 RAID 进程。例如, 请求设备 155 可向第一存储设备 150a 发送第一对象写入请求, 而向第二存储设备 150b 发送具有相同数据段的第二对象写入请求以实现简单的镜像。

[0078] 当具有在存储设备 102 内进行对象处理的能力时, 只有存储控制器 152 具有采用一个 RAID 级别存储一个数据段或对象的能力, 而采用不同的 RAID 级别或不采用 RAID 条带化来存储另一数据段或对象。这些多个 RAID 群组可与存储设备 150 内的多个分区相关联。可同时在各种 RAID 群组间支持 RAID0、RAID1、RAID5、RAID6 和复合 RAID 类型 10、50、60, 所述 RAID 群组包括数据存储设备 154a-n。本领域技术人员可认识到还可同时支持的其他 RAID 类型和配置。

[0079] 而且, 由于存储控制器 152 像 RAID 控制器一样自主运行, 所述 RAID 控制器能够执行渐进 RAID 并能够将在数据存储设备 154 间条带化的具有一个 RAID 级别的对象或对象的某些部分转换为另一 RAID 级别, 转换时请求设备 155 不受影响、不参与或者甚至不探测 RAID 级别的变化。在优选实施方式中, 促进 RAID 配置从一个级别变为另一级别可在对象或甚至在包基上自主实现, 并且可由运行在存储设备 150 或存储控制器 152 中的一个上的分布式 RAID 控制模块初始化。通常, RAID 渐进是从高性能和低效率的存储配置 (如 RAID1) 转换为低性能和高存储效率的存储配置 (如 RAID5), 其中, 转换是基于读取频率被动态地初始化。但是, 可以发现, 从 RAID5 到 RAID1 的渐进也是可能的。可配置其他用于初始化 RAID 渐进的进程, 或者可由客户端或外部代理 (如存储系统管理服务器请求) 请求该进程。本领域技术人员可认识到具有存储控制器 152 的存储设备 102 的其他特征和优点, 该存储控制器 152 自主管理对象。

[0080] 具有服务器内 SAN 的固态存储设备

[0081] 图 1C 是示意性框图, 示出了根据本发明的用于服务器内存储区域网络 (“SAN”) 的系统 103 的一种实施方式。系统 103 包括计算机 112, 计算机 112 通常被配置为服务器 (“服务器 112”)。每个服务器 112 包括一个或多存储设备 150, 其中, 服务器 112 和存储设备 150 分别连接到共享网络接口 155。每个存储设备 150 包括存储控制器 152 和相应的数据存储设备 154。系统 103 包括客户端 114、114a、114b, 客户端 114、114a、114b 位于服务器 112 的内部或者外部。客户端 114、114a、114b 可通过一个或多个计算机网络 116 与每个服

务器 112 和每个存储设备 150 通信,所述一个或多个计算机网络 116 大体上与上述的计算机网络类似。

[0082] 存储设备 150 包括 DAS 模块 158、NAS 模块 160、存储通信模块 162、服务器内 SAN 模块 164、通用接口模块 166、网络代理模块 170、虚拟总线模块 172、前端 RAID 模块 174 及后端 RAID 模块 176,这些模块将在下文中描述。模块 158-176 图示为位于存储设备 150 内,模块 158-176 中的每一个的全部或部分可位于存储设备 150、服务器 112、存储控制器 152 内或位于其他位置。

[0083] 服务器 112(如与服务器内 SAN 结合使用的)是具有服务器功能的计算机。服务器 112 至少包括一项服务器功能(如文件服务器功能),而且还可包括其他服务器功能。服务器 112 可以是服务器群的一部分并可服务其他客户端 114。在其他实施方式中,服务器 112 还可以是个人计算机、工作站或其他包括存储设备 150 的计算机。服务器 112 可像访问直接附加存储(“DAS”)、SAN 附加存储或者网络附加存储(“NAS”)那样访问服务器 112 内的一个或多个存储设备 150。参与服务器内 SAN 或 NAS 的存储控制器 152 可位于服务器 112 的内部或外部。

[0084] 在一种实施方式中,服务器内 SAN 装置包括 DAS 模块 158,该 DAS 模块 158 将由服务器 112 的存储控制器 152 控制的至少一个数据存储设备 154 的至少一部分配置为附属于服务器 112 的 DAS 设备,以服务从至少一个客户端 114 传送到服务器 112 的存储请求。在一种实施方式中,第一数据存储设备 154a 被配置为第一服务器 112a 的 DAS,而且,第一数据存储设备 154a 还被配置为第一服务器 112a 的服务器内 SAN 存储设备。在另一种实施方式中,第一数据存储设备 154a 被分割,以使得一个分区为 DAS 而另一个分区为服务器内 SAN。在另一种实施方式中,第一数据存储设备 154a 内的存储空间的至少一部分被配置为第一服务器 112a 的 DAS,而第一服务器 112a 的存储空间的同一部分被配置为第一服务器 112a 的服务器内 SAN。

[0085] 在另一种实施方式中,服务器内 SAN 装置包括 NAS 模块 160,该 NAS 模块 160 将存储控制器 152 配置为用于至少一个客户端 114 的 NAS 设备并服务来自客户端 114 的文件请求。存储控制器 152 还可被配置为用于第一服务器 112a 的服务器内 SAN 设备。存储设备 150 可通过共享网络接口 155 直接连接到计算机网络 116,共享网络接口 155 独立于存储设备 150 位于其内的服务器 112。

[0086] 在一种基本的形式中,用于服务器内 SAN 的装置包括第一服务器 112a 内的第一存储控制器 152a,其中,第一存储控制器 152a 控制至少一个存储设备 154a。第一服务器 112a 包括由第一服务器 112a 和第一存储控制器 152a 共享的网络接口 156,所述服务器内 SAN 装置包括存储通信模块 162,该存储通信模块 162 促进第一存储控制器 152a 和位于第一服务器 112a 外部的至少一个设备之间的通信,以使得第一存储控制器 152a 和外部设备之间的所述通信独立于第一服务器 112a。存储通信模块 162 可允许第一存储控制器 152a 独立地访问网络接口 156a 以进行外部通信。在一种实施方式中,存储通信模块 162 访问网络接口 156a 中的交换机以管理第一存储控制器 152a 和外部设备之间的网络流量。

[0087] 服务器内 SAN 装置还包括服务器内 SAN 模块 164,该服务器内 SAN 模块 164 利用网络协议和总线协议中的一个或两个服务存储请求。服务器内 SAN 模块 164 服务独立于第一服务器 112a 的所述存储请求,并且所述服务请求接收自内部或外部客户端 114a、114。

[0088] 在一种实施方式中,位于第一服务器 112a 外部的设备是第二存储控制器 152b。第二存储控制器 152b 控制至少一个数据存储设备 154b。服务器内 SAN 模块 164 利用第一和第二存储控制器 152a、152b 之间、通过网络接口 156a 的通信服务所述存储请求,第一和第二存储控制器 152a、152b 独立于第一服务器 112a。第二存储控制器 152b 可位于第二服务器 112b 内或位于一些其他设备内。

[0089] 在另一种实施方式中,第一服务器 112a 外部的设备是客户端 114,并且所述存储请求源于外部客户端 114,其中,第一存储控制器 152a 被配置为 SAN 的至少一部分,并且服务器内 SAN 模块 164 通过独立于第一服务器 112a 的网络接口 156a 服务所述存储请求。外部客户端 114 可位于第二服务器 112b 内或可位于第二服务器 112b 的外部。在一种实施方式中,即使当第一服务器 112a 不可用时,服务器内 SAN 模块 164 也能够服务来自外部客户端 114 的存储请求。

[0090] 在另一种实施方式中,生成所述存储请求的客户端 114a 位于第一服务器 112a 的内部,其中,第一存储控制器 152a 被配置为 SAN 的至少一部分,并且服务器内 SAN 模块 164 通过一个或多个网络接口 156a 和系统总线服务所述存储请求。

[0091] 传统的 SAN 配置允许像按直接附加存储 (“DAS”) 访问服务器 112 内的存储设备一样访问远离服务器 112 的存储设备,以使得远离服务器 112 的所述存储设备表现为块存储设备。通常,按 SAN 连接的存储设备需要 SAN 协议,所述 SAN 协议例如光纤通道、互联网小型计算机系统接口 (“iSCSI”)、HyperSCSI、光纤连通性 (“FICON”)、通过以太网的高级技术附件 (“ATA”) 等。服务器内 SAN 包括服务器 112 内的存储控制器 152,同时仍然允许存储控制器 152a 和远程存储控制器 152b 或外部客户端之间利用网络协议和 / 或总线协议的网络连接。

[0092] 通常,SAN 协议是网络协议的形式,并且,出现了更多的网络协议,例如可允许存储控制器 152a 和关联的数据存储设备 154a 被配置为 SAN 并与外部外部客户端 114 或第二存储控制器 152b 通信的无限带宽。在另一种实例中,第一存储控制器 152a 可利用以太网与外部客户端 114 或第二存储控制器 152b 通信。

[0093] 存储控制器 152 可通过总线与内部存储控制器 152 或客户端 114a 通信。例如,存储控制器 152 可通过使用 PCI-e 的总线通信,所述 PCI-e 可支持 PCI 快速输入 / 输出虚拟化 (“PCIe-IOV”)。其他新出现的总线协议允许系统总线扩展超出计算机或服务器 112 并可允许存储控制器 152a 被配置为 SAN。一种这样的总线协议是 PCIe-AS。本发明并不仅限于 SAN 协议,还可利用新出现的网络和总线协议服务存储请求。外部设备(无论是客户端 114 的形式还是外部存储控制器 152b 的形式)可通过扩展系统总线或计算机网络 116 通信。正如此处所使用的,存储请求包括写入数据、读取数据、擦除数据、查询数据的请求等等,并且所述存储请求可包括对象数据、元数据、管理请求以及块数据请求。

[0094] 传统的服务器 112 通常具有控制访问服务器 112 内的设备的根联合体。通常,服务器 112 的所述根联合体具有网络接口 156,从而使得服务器 112 控制任何通过网络接口 156 的通信。然而,在服务器内 SAN 装置的优选实施方式中,存储控制器 152 能够独立地访问网络接口 156,从而使得客户端 114 可直接地与形成 SAN 的第一服务器内 112a 内的一个或多个存储控制器 152a 通信,或者使得一个或多个第一存储控制器 152a 可与第二存储控制器 152b 或其他远程存储控制器 152 联网在一起以形成 SAN。在这种优选实施方式中,远

离第一服务器 112a 的设备可通过单独的、共享的网络地址访问第一服务器 112a 或第一存储控制器 152a。在一种实施方式中，服务器内 SAN 装置包括通用接口模块 166，该通用接口模块 166 配置网络接口 156、存储控制器 152 和服务器 112，以使得可使用共享网络地址访问服务器 112 和存储控制器 152。

[0095] 在其他实施方式中，服务器 112 包括两个或更多个网络接口 156。例如，服务器 112 可通过一个网络接口 156 通信，而存储设备 150 可通过另一个接口通信。在另一个实例中，服务器 112 包括多个存储设备 150，每个存储设备 150 具有网络接口 156。本领域技术人员会认识到具有一个或多个存储设备 150 和一个或多个网络接口 156 的服务器 112 的其他配置，其中，一个或多个存储设备 150 访问独立于服务器 112 的网络接口 156。本领域技术人员还可认识到扩展这些不同的配置的方法以支持网络冗余并提高可用性。

[0096] 有利地是，服务器内 SAN 装置大大降低了传统 SAN 的复杂性和花费。例如，典型的 SAN 需要具有外部存储控制器 152 和关联的数据存储设备 154 的服务器 112。这占用了机架上的额外空间并且需要电缆、交换机等。配置传统的 SAN 所需的电缆、交换机和其他的开销占用了空间、降低了带宽并且昂贵。服务器内 SAN 装置允许存储控制器 152 和关联的存储器 154 适合服务器 112 的形体尺寸，并因此减少了需要的空间和费用。服务器内 SAN 还允许通过内部和外部高速数据总线使用相对高速的通信的连接。

[0097] 在一种实施方式中，存储设备 150 为固态存储设备 102，存储控制器 152 为固态存储控制器 104，而数据存储设备 154 为固态存储器 110。由于此处所述的固态存储设备 102 的速度，这种实施方式是有利的。此外，固态存储设备 102 可被配置为位于 DIMM 中，所述 DIMM 可方便地装配在服务器 112 内并仅需要少量的空间。

[0098] 服务器 112 中的一个或多个内部客户端 114a 还可通过服务器 112 的网络接口 156 连接到计算机网络 116，并且服务器 112 通常控制所述客户端的连接。这种方法具有一些好处。客户端 114a 可直接地本地访问或远程访问存储设备 150，并且客户端 114a 可初始化客户端 114a 的存储器和存储设备 150 之间的本地或远程直接存储器存取（“DMA”，“RDMA”）数据的传送。

[0099] 在另一种实施方式中，当利用本地连接的存储设备 150 作为 DAS 设备、网络连接的存储设备 150、网络连接的固态存储设备 102（这些设备作为服务器内 SAN、外部 SAN 和混合 SAN 的一部分）时，位于服务器 112 内部或外部的客户端 114、114a 可通过一个或多个网络 116 对客户端 114 起文件服务器的作用。存储设备 150 可同时参与 DAS、服务器内 SAN、SAN、NAS 等（及其中的任意的组合）。此外，每个存储设备 150 可以以下方式被分割：第一分区使存储设备 150 作为 DAS 可用，第二分区使存储设备 150 作为服务器内 SAN 内的元件可用，第三分区使存储设备 150 作为 NAS 可用，第四分区使存储设备 150 作为 SAN 的元件可用，等等。类似地，存储设备 150 可被分割为符合安全性和存取控制要求。本领域技术人员会认识到可以构建和支持下述设备或系统的任意组合和排列：存储设备、虚拟存储设备、存储网络、虚拟存储网络、专用存储器、共享存储器、平行文件系统、平行对象文件系统、块存储设备、对象存储设备、存储装置、网络装置及类似设备。

[0100] 此外，通过将存储设备 150 直接地连接到计算机网络 116，存储设备 150 彼此之间能够互相通信并能够起服务器内 SAN 的作用。通过计算机网络 116 连接的服务器 112 内的客户端 114a 和客户端 114 可像访问 SAN 那样访问存储设备 150。通过将存储设备 150 移

到服务器 112 内并使其具备将存储设备 150 配置为 SAN 的能力,服务器 112/ 存储设备 150 的结合使得在常规 SAN 中不需要专用的存储控制器、光纤通道网络和其他设备。服务器内 SAN 系统 103 具有能够使存储设备 150 与客户端 114 和计算机 112 共享共用的资源（如电源、制冷、管理和物理空间）的优点。例如,存储设备 150 可插在服务器 112 的空的插槽中并提供 SAN 或 NAS 的所有工作性能、可靠性和可用性。本领域技术人员会认识到服务器内 SAN 系统 103 的其他特征和优点。

[0101] 在另一种配置中,多个服务器内 SAN 存储设备 150a 被布置在单独的服务器 112a 基础架构内。在一种实施方式中,服务器 112a 由一个或多个利用 PCI 快速 IOV 互连的内部刀片服务器客户端 114a 组成,而没有外部网络 156、外部客户端 114、114b 或外部存储设备 150b。

[0102] 此外,服务器内 SAN 存储设备 150 可通过一个或多个计算机网络 116 与对等存储设备 150 通信,所述对等存储设备 150 位于计算机 112(图 1 中的每一台计算机)内,或者不通过计算机 112 而直接连接到计算机网络 116 以形成同时具有 SAN 和服务器内 SAN 的全部功能的混合 SAN。这种灵活性具有以下优点:简化了扩展性和在不同的可能的固态存储网络实施方案之间的移植。本领域技术人员会认识到放置和互连固态控制器 104 的其他组合、配置、实施方案和布局。

[0103] 当网络接口 156a 仅能被运行在服务器 112a 内的一个代理控制时,运行在所述代理中的链路建立模块 168 能够通过连接到外部存储设备 150b 和客户端 114、114b 的网络接口 156a 建立内部客户端 114a 和存储设备 150a/ 第一存储控制器 152a 之间的通信通路。在优选的实施方式中,一旦建立了所述通信通路,单独的内部存储设备 150a 和内部客户端 114a 能够建立和管理它们自己的命令队列,并通过网络接口 156a 和独立于控制网络接口 156a 的网络代理或代理的 RDMA 将命令和数据同时双向地直接传送给外部存储设备 150b 和客户端 114、114b。在一种实施方式中,链路建立模块 168 在初始化过程(如硬件的启动或初始化)期间建立通信链路。

[0104] 在另一种实施方式中,网络代理模块 170 指令至少一部分用于通过第一服务器 112 服务存储请求的命令,而至少与所述存储请求关联的数据(也可能是其他命令)在第一存储控制器 152a 和独立于第一服务器的外部存储设备 150b 之间通信。在另一种实施方式中,网络代理模块 170 代表内部存储设备 150a 和客户端 114a 转发命令或数据。

[0105] 在一种实施方式中,第一服务器 11a 包括位于第一服务器 112a 内的一个或多个服务器,并包括虚拟总线模块 172,该虚拟总线模块 172 允许第一服务器 112a 内的所述一个或多个服务器通过分享的虚拟总线独立地访问一个或多个存储控制器 152a。所述虚拟总线可利用高级总线协议(如 PCIe-IOV)建立。支持 IOV 的网络接口 156a 可允许所述一个或多个服务器和所述一个或多个存储控制器 152a 独立地控制一个或多个网络接口 156a。

[0106] 在不同的实施方式中,服务器内 SAN 装置允许两个或更多个存储设备 150 被配置在 RAID 中。在一种实施方式中,服务器内 SAN 装置包括将两个或更多个存储控制器配置为 RAID 的前端 RAID 模块 174。当来自客户端 114、114a 的存储请求包括存储数据的请求时,前端 RAID 模块 174 通过将所述数据以符合特定应用的 RAID 级的形式写入所述 RAID 服务所述存储请求。第二存储控制器 152 可位于第一服务器 112a 的内部或者外部。前端 RAID 模块 174 允许将存储控制器 152 配置成 RAID,从而使得存储控制器对发送所述存储请求的

客户端 114、114a 可见。这种方法允许被指定为主机的存储控制器 152 或客户端 114、114a 管理条纹和校验信息。

[0107] 在另一种实施方式中,服务器内 SAN 装置包括后端 RAID 模块 176,该后端 RAID 模块 176 将由存储控制器 152 控制的两个或更多个数据存储设备 154 配置为 RAID。当来自所述客户端 114 的存储请求包括存储数据的请求时,后端 RAID 模块 176 通过将所述数据以符合应用的 RAID 级的形式写入所述 RAID 服务所述存储请求,从而使得客户端 114、114a 像访问由第一存储控制器 152 控制的单个数据存储设备 154 那样访问被配置为 RAID 的存储设备 154。这种 RAID 应用允许以如下方式将由存储控制器 152 控制的数据存储设备配置成 RAID :配置成 RAID 的过程对任何访问数据存储设备 154 的客户端 114、114a 来说是透明的。在另一种实施方式中,前端 RAID 和后端 RAID 都具有多级 RAID。本领域技术人员会认识到将存储设备 154 以符合此处所述的固态存储控制器 104 和关联的固态存储器 110 的形式配置为 RAID 的其他方法。

[0108] 用于存储控制器管理的对象的装置

[0109] 图 2A 是示意性框图,示出了根据本发明的用于存储设备内的对象管理的装置 200 的一种实施方式。装置 200 包括存储控制器 152,该存储控制器 152 具有 :对象请求接收器模块 260、解析模块 262、命令执行模块 264、对象索引模块 266、对象请求排队模块 268、具有消息模块 270 的封包器 302、及对象索引重建模块 272,上述模块描述如下。

[0110] 存储控制器 152 大体上与图 1B 中的系统 101 描述的存储控制器 152 类似,并且可以是图 2 描述的固态存储设备控制器 202。装置 200 包括对象请求接收器模块 260,该对象请求接收器模块 260 接收来自一个或多个请求设备 155 的对象请求。例如,对于存储对象数据请求,存储控制器 152 在数据存储设备 154 中以数据包的形式存储数据段,该数据存储设备 154 与存储控制器 152 相连接。所述对象请求通常由存储在或将要被存储在一个或多个对象数据包中的数据段指令存储控制器 152 管理的对象。对象请求可请求存储控制器 152 创建对象,该对象随后会通过可利用本地或远程直接内存读取 (“DMA”、“RDMA”) 转换的稍后的对象请求来填充数据。

[0111] 在一种实施方式中,对象请求为将对象的全部或一部分写入先前创建的对象的写入请求。在一个实例中,所述写入请求用于对象的数据段。可将所述对象的其他数据段写入存储设备 150 或者写入其他存储设备。在另一个实例中,所述写入请求用于整个对象。在另一个实例中,所述对象请求为从由存储控制器 152 管理的数据段中读取数据。在又一种实施方式中,所述对象请求为删除请求,以删除数据段或对象。

[0112] 有利地是,存储控制器 152 能接受不仅仅写新对象或为已存在的对象添加数据的写入请求。例如,由对象请求接收器模块 260 接收的写入请求可包括 :在由存储控制器 152 存储的数据前添加数据的请求、在已存储的数据中插入数据的请求或者替换数据的一段的请求。由存储控制器 152 保持的对象索引提供了这些复杂写操作所需要的灵活性,所述写操作在其他存储控制器内不可用,但是目前仅在服务器和其他计算机文件系统内的存储控制器外可用。

[0113] 装置 200 包括解析模块 262,该解析模块 262 将所述对象请求解析为一条或多条命令。通常,解析模块 262 将所述对象请求解析为一个或多个缓存。例如,所述对象请求中的一条或多条命令可被解析为命令缓存。通常,解析模块 262 准备对象请求,以使得所述对象

请求中的信息可以被存储控制器 152 理解并执行。本领域技术人员会认识到将对象请求解析为一条或多条命令的解析模块 262 的其他功能。

[0114] 装置 200 包括命令执行模块 264，该命令执行模块 264 执行从所述对象请求解析出的命令。在一种实施方式中，命令执行模块 264 执行一条命令。在另一种实施方式中，命令执行模块 264 执行多条命令。通常，命令执行模块 264 解释解析自所述对象请求的命令（如写入命令），然后创建、排列并且执行子命令。例如，解析自对象请求的写入命令可指令存储控制器 152 存储多个数据段。所述对象请求还可包括必要属性（如加密、压缩等）。命令执行模块 264 可命令存储控制器 152 压缩所述数据段、加密所述数据段、创建一个或多个数据包并为每个数据包关联包头、使用媒体加密密钥加密所述数据包、添加错误修正码并将所述数据包存储在指定位置。在指定位置存储所述数据包，并且其他子命令还可被分解为其他更低级别的子命令。本领域技术人员会认识到命令执行模块 264 能执行一条或多条解析自对象请求的命令的其他方法。

[0115] 装置 200 包括对象索引模块 266，该对象索引模块 266 在对象索引中创建对象项，以响应创建对象或存储所述对象数据段的存储控制器 152。通常，存储控制器 152 从所述数据段中创建数据包，并且在存储所述数据段时，所述数据包存储的位置即被指定。同数据段一起接收的或作为对象请求的一部分接收的对象元数据可采用类似方法存储。

[0116] 对象索引模块 266 在存储所述数据包和分配所述数据包的物理地址时创建进入对象索引的对象项。所述对象项包括所述对象的逻辑标识符和一个或多个物理地址之间的映射，所述一个或多个物理地址对应于存储控制器 152 存储一个或多个数据包和任何对象元数据包的位置。在另一种实施方式中，在存储所述对象的数据包之前在所述对象索引中创建项。例如，如果存储控制器 152 较早地确定存储所述数据包的物理地址，则对象索引模块 266 可较早地在所述对象索引中创建项。

[0117] 通常，当对象请求或对象请求组导致对象或数据段被修改时（可能在读修改写操作期间），所述对象索引模块 266 更新所述对象索引中的项以符合修改的对象。在一种实施方式中，所述对象索引创建新对象并在所述对象索引为所述修改的对象创建新项。通常，当仅有对象的一部分被修改时，所述对象包括修改过的数据包和一些保持不变的数据包。在这种情况下，所述新项包括到未变的数据包（与最初写入它们的位置相同）的映射和到写入新位置的修改后的对象的映射。

[0118] 在另一种实施方式中，对象请求接收器模块 260 接收对象请求，该对象请求包括擦除数据块或其他对象元的命令，存储控制器 152 可至少存储一个包（如擦除包，该擦除包具有对象的引用、与对象的关系和擦除的数据块的大小的信息）。此外，这可进一步表明擦除的对象元素被填充为 0。因此，擦除对象请求可用于仿真被擦除的实际的内存或存储器，并且，所述实际的内存或存储器实际上具有合适的内存 / 存储器的一部分，所述合适的内存 / 存储器实际上以 0 存储在所述内存 / 存储器的单元中。

[0119] 有利地是，创建具有项（该项表明了数据段和对象元数据之间的映射）的对象索引允许存储控制器 152 自主的处理和管理对象。这种能力允许在存储设备 150 中十分灵活地存储数据。一旦创建了对象的索引项，存储控制器 152 可有效地处理后继关于所述对象的对象请求。

[0120] 在一种实施方式中，存储控制器 152 包括对象请求排队模块 268，该对象请求排队

模块 268 在解析模块 262 解析之前将一个或多个由对象请求接收器模块 260 接收到的对象排队。对象请求排队模块 268 允许在接收对象请求时和在执行所述对象请时之间的灵活性。

[0121] 在另一种实施方式中,存储控制器 152 包括封包器 302,该封包器 302 根据一个或多个数据段创建一个或多个数据包,其中,数据包的大小适于存储在数据存储设备 154 内。在下文中结合图 3 更详细地描述封包器 302。在一种实施方式中,封包器 302 包括为每个包创建包头的消息模块 270。所述包头包括包标识符和包长度。所述包标识符把所述包与对象(为该对象生成所述包)联系起来。

[0122] 在一种实施方式中,由于包标识符包含足够的信息以确定对象和在对象内的包含在包内的对象元素之间的关系,因此每个包包括自包含的包标识符。然而,更有效的优选实施方式是在容器中存储包。

[0123] 容器是一种数据结构,这种数据结构有助于更有效的存储数据包并帮助建立对象和数据包、元数据包和其他与存储在容器内的对象有关的包之间的关系。注意到存储控制器 152 通常以处理作为对象的一部分接收的对象元数据的类似方式处理数据段。通常,“包”可指包含数据的数据包、包含元数据的元数据包或其他包类型的其他包。对象可存储在一个或多个容器中,并且容器通常包括仅用于一个唯一的对象的包。对象可分布在多个容器之间。容器通常存储在单个逻辑擦除块内(存储部)并且通常不分散在逻辑擦除块间。

[0124] 在一个实例中,容器可分散在两个或更多个逻辑/虚拟页间。通过将容器与对象关联起来的容器标签确定容器。容器可包含 0 个到许多个包并且容器内的这些包通常来自一个对象。包可以有许多对象元素类型(包括对象属性元、对象数据元、对象索引元和类似的元素类型)。可以创建包括不止一个对象元类型的混合包。每个包可包含 0 个到许多个同一类型的元。容器内的每个包通常都包含标识与对象关系的唯一标识符。

[0125] 每个包与一个容器相关联。在优选实施方式中,容器被限于擦除块,以使得在每个擦除块的起始部分或在擦除块的起始部分附近能发现容器包。这有助于将数据丢失限制在具有损坏的包头的擦除块范围内。在这种实施方式中,如果对象索引不可用并且擦除块内的包头损坏,由于可能没有可靠的机制确定后继包的位置,从损坏的包头到擦除块尾的内容可能会丢失。在另一种实施方式中,更可靠的方法是采用限于页的边界的容器。这种实施方式需要更多包头开销。在另一种实施方式中,容器可流经页面和擦除块边界。这种方法需要较少的包头开销,但是,如果包头损坏,则有可能会丢失更多部分的数据。对这些实施方式来说,使用一些类型的 RAID 以进一步保证数据完整性是可以预期的。

[0126] 在一种实施方式中,装置 200 包括对象索引重建模块 272,该对象索引重建模块 272 采用来自存储在数据存储设备 154 中的包头的信息重建所述对象索引中的项。在一种实施方式中,对象索引重建模块 272 通过读取包头(以确定每个包所属的对象)和序列信息(以确定数据或元数据在对象中所属的位置)来重建所述对象索引的项。对象索引重建模块 272 采用每个包的物理地址信息和时间戳或序列信息以创建包的物理地址和对象标识符和数据段序列间的映射。对象索引重建模块 272 使用时间戳或序列信息以再现索引变更的顺序并通常因此重建最近的状态。

[0127] 在另一种实施方式中,对象索引重建模块 272 采用包头信息以及容器包信息放置包以识别包的物理位置、对象标识符和每个包的序列号,从而在所述对象索引中重建项。在

一种实施方式中，在写入数据包时，擦除块被戳记上时间，或者赋给擦除块序列号，并且擦除块的时间戳或序列信息和来自容器头和包头的信息一起使用以重建对象索引。在另一种实施方式中，当擦除块恢复时，时间戳或序列信息被写入该擦除块。

[0128] 当对象索引存储在易失性存储器中时，如果不能重建所述对象索引，错误、失电、或其他导致存储控制器 152 未存储所述对象索引而停工的因素可能会成为问题。对象索引重建模块 272 允许所述对象索引存储在具有易失性存储体优点（如快速存取）的易失性存储体中。对象索引重建模块 272 允许自主地快速重建所述对象索引，而并不需要依靠位于存储设备 150 外的设备。

[0129] 在一种实施方式中，易失性存储体中的所述对象索引周期性地存储在数据存储设备 154 内。在具体的实例中，所述对象索引或“索引元数据”周期性地存储固态存储器 110 中。在另一种实施方式中，所述索引元数据存储在固态存储器 110n（与固态存储器 110a-110n-1 存储包分离）中。独立于数据和对象元数据管理所述索引元数据，所述数据和对象元数据传送自请求设备 155 并且由存储控制器 152/ 固态存储控制器 202 管理。管理和存储与其他来自对象的数据和元数据分离的索引元数据允许有效的数据流，同时存储控制器 152/ 固态存储设备控制器 202 并不会不必要地处理对象元数据。

[0130] 在一种实施方式中，其中，由对象请求接收器模块 260 接收到的对象请求包括写入请求，存储控制器 152 通过本地或远程直接存储器存取（“DMA”、“RDMA”）操作接收来自请求设备 155 的内存的一个或多个对象数据段。在优选实例中，存储控制器 152 在一次或多次 DMA 或 RDMA 操作中从请求设备 155 的内存中读取数据。在另一实例中，请求设备 155 在一次或多次 DMA 或 RDMA 操作中将所述数据段写入存储控制器 152。在另一种实施方式中，其中，所述对象请求包括读请求，存储控制器 152 在一次或多次 DMA 或 RDMA 操作中将对象的一个或多个数据段传送给请求设备 155 的内存。在优选实例中，存储控制器 152 在一次或多次 DMA 或 RDMA 操作中将数据写入请求设备 155 的内存。在另一实例中，请求设备在一次或多次 DMA 或 RDMA 操作中从存储控制器 152 中读取数据。在另一实施方式中，存储控制器 152 在一次或多次 DMA 或 RDMA 操作中从请求设备 155 的内存中读取对象命令请求集。在另一实例中，请求设备 155 在一次或多次 DMA 或 RDMA 操作中将对象命令请求集写入存储控制器 152。

[0131] 在一种实施方式中，存储控制器 152 仿真块存储，并且在请求设备 155 和存储控制器 152 之间通信的对象包括一个或多个数据块。在一种实施方式中，请求设备 155 包括驱动器，以使得存储设备 150 表现为块存储设备。例如请求设备 155 可与请求设备 155 期望数据存储的物理地址一起发送特定大小的一组数据。存储控制器 152 接收所述数据块，并将与所述数据块一起传送的物理块地址或者将物理块地址的转化形式作为对象标识符。然后，存储控制器 152 通过随意地封包所述数据块和存储数据块将所述数据块存储为对象或对象的数据段。然后，对象索引模块 266 利用基于物理块的对象标识符和存储控制器 152 存储所述数据包的实际物理位置在所述对象索引中创建项，所述数据包包括来自所述数据块的数据。

[0132] 在另一种实施方式中，存储控制器 152 通过接收块对象仿真块存储。块对象可包括块结构中的一个或多个数据块。在一种实施方式中，存储控制器 152 像处理任意其他对象一样处理所述块对象。在另一种实施方式中，对象可代表整个块设备、块设备的分区或块

设备的一些其他逻辑子元件或物理子元件,所述块设备包括磁道、扇区、通道及类似设备。值得特别注意的是将块设备 RAID 组重映射到支持不同 RAID 构建(如渐进 RAID)的对象。本领域技术人员会认识到将传统的或未来的块设备映射到对象的其他方法。

[0133] 固态存储设备

[0134] 图 2B 是示出了根据本发明的位于固态存储设备 102 内的固态存储设备控制器 202 的一种实施方式 201 的示意性框图,该固态存储设备控制器 202 包括写入数据管道 106 和读取数据管道 108。固态存储设备控制器 202 可包括若干固态存储控制器 0-N,104a-n,每个固态存储控制器都控制固态存储器 110。在描述的实施方式中,示出了两个固态控制器:固态控制器 0104a 和固态控制器 N104n,并且它们中的每一个都控制固态存储器 110a-n。在描述的实施方式中,固态存储控制器 0 104a 控制数据通道,以使得附属固态存储器 110a 存储数据。固态存储控制器 N 104n 控制与存储的数据关联的索引元数据通道,以使得关联的固态存储器 110n 存储索引元数据。在替代的实施方式中,固态存储设备控制器 202 包括具有单个固态存储器 110a 的单个固态控制器 104a。在另一种实施方式中,存在大量的固态存储控制器 104a-n 和关联的固态存储器 110a-n。在一种实施方式中,一个或多个固态控制器 104a-104n-1(与它们的关联固态存储器 110a-110n-1 连接)控制数据,而至少一个固态存储控制器 104n(与其关联固态存储器 110n 连接)控制索引元数据。

[0135] 在一种实施方式中,至少一个固态控制器 104 是现场可编程门阵列(“FPGA”)并且控制器功能被编入 FPGA。在特定的实施方式中,FPGA 是 Xilinx® 公司的 FPGA。在另一种实施方式中,固态存储控制器 104 包括专门设计为固态存储控制器 104 的组件(如专用集成电路(“ASIC”)或自定义逻辑解决方案)。每个固态存储控制器 104 通常包括写入数据管道 106 和读取数据管道 108,结合图 3 进一步描述了这两个管道。在另一种实施方式中,至少一个固态存储控制器 104 由 FPGA、ASIC 和自定义逻辑组件的组合组成。

[0136] 固态存储器

[0137] 固态存储器 110 是非易失性固态存储元件 216、218、220 的阵列,该阵列布置在内存库 214 中并且通过双向存储输入输出(I/O)总线 210 并行访问。在一种实施方式中,存储 I/O 总线 210 能够在任何一个时刻进行单向通信。例如,当将数据写入固态存储器 110 时,不能从固态存储器 110 中读取数据。在另一种实施方式中,数据可同时双向地流动。然而,双向(如此处针对数据总线使用的)指在同一时间数据仅在一个方向流动的数据通路,但是,当在双向数据总线上流动的数据被阻止时,数据可在所述双向总线上沿相反方向流动。

[0138] 固态存储元件(如 SSS 0.0 216a)通常被配置为芯片(一个或多个小片的封装)或电路板上的小片。正如所描述的那样,固态存储元件(如 216a)独立于或半独立于其他固态存储元件(如 218a)运行,即使这些元件被一起封装在芯片包、芯片包的堆栈或一些其他封包元件内。正如所描述的,一列固态存储元件 216、218、220 被指定为内存库 214。正如所描述的,可以有“n”个内存库 214a-n 并且每个内存库可以有“m”个固态存储元件 216a-m,218a-m,220a-m,从而在固态存储器 110 中成为固态存储元件 216、218、220 的 n*m 阵列。在一种实施方式中,固态存储器 110a 在每个内存库 214(有 8 个内存库 214)中包括 20 个固态存储元件 216、218、220,并且,固态存储器 110n 在每个内存库 214 中(只有一个内存库 214)包括两个固态存储元件 216、218。在一种实施方式中,每个固态存储元件 216、218、220 由单层单元(“SLC”)设备组成。在另一种实施方式中,每个固态存储元件 216、218、220 由

多层单元（“MLC”）设备组成。

[0139] 在一种实施方式中,用于多个内存库的固态存储元件被封包在一起,所述多个内存库共享公用存储 I/O 总线 210a 行 (如 216b、218b、220b)。在一种实施方式中,固态存储元件 216、218、220 的每个芯片可具有一个或多个小片,而一个或多个芯片垂直堆叠且每个小片可被独立存取。在另一种实施方式中,固态存储元件 (如 SSS 0.0 216a) 的每个小片可具有一个或多个虚拟小片,每个芯片可具有一个或多个小片,而一个或多个小片中的一些或全部垂直堆叠且每个虚拟小片可被独立存取。

[0140] 在一种实施方式中,每组有四个堆,每堆有两个小片垂直堆叠,从而形成 8 个存储元件 (如 SSS 0.0-SSS 0.8) 216a-220a,每个存储元件位于分离的内存库 214a-n 内。在另一种实施方式中,20 个存储元件 (如 SSS 0.0-SSS 20.0) 216 形成虚拟内存库 214a,因此八个虚拟内存库中的每一个都具有 20 个存储元件 (如 SSS 0.0-SSS 20.8) 216、218、220。通过存储 I/O 总线 210 将数据发送到固态存储器 110,并发送到存储元件 (SSS 0.0-SSS 0.8) 216a、218a、220a 的特定组的所有存储元件。存储控制总线 212a 用于选择特定的内存库 (如内存库 -0 214a),从而通过连接到所有内存库 214 的存储 I/O 总线 210 接收到的数据仅被写入选定的内存库 214a。

[0141] 在优选实施方式中,存储 I/O 总线 210 由一个或多个独立 I/O 总线 (包括 210a-a-m,210n, a-m 的“II0Ba-m”) 组成,其中,每一行内的固态存储元件共享独立 I/O 总线中的一条,所述独立 I/O 总线中的一条平行访问每个固态存储元件 216、218、220,从而使得同时访问所有的内存库 214。例如,存储 I/O 总线 210 的一个通道可同时访问每个内存库 214a-n 的第一固态存储元件 216a、218a、220a。存储 I/O 总线 210 的第二通道可同时访问每个内存库 214a-n 的第二固态存储元件 216b、218b、220b。固态存储元件 216、218、220 的每一行都被同时访问。在一种实施方式中,其中,固态存储元件 216、218、220 是多层的 (物理堆叠的),固态存储元件 216、218、220 的所有物理层被同时访问。正如此处所使用的,“同时”还包括几乎同时的访问,其中,以略有不同的时间间隔访问设备以避免切换噪声。在这种情况下,同时被用于与连续的或系列的访问相区别,其中,命令和 / 或数据被单独地并相继地发送。

[0142] 通常,采用存储控制总线 212 独立地选择内存库 214a-n。在一种实施方式中,采用芯片选通或芯片选择来选择内存库 214。当芯片选择和芯片使能均可用时,存储控制总线 212 可选择多层固态存储元件 216、218、220 中的一层。在其他实施方式中,存储控制总线 212 使用其他命令来单独地选择多层固态存储元件 216、218、220 中的一层。还可通过控制和地址信息的结合来选择固态存储元件 216、218、220,所述控制和地址信息在存储 I/O 总线 210 和存储控制总线 212 上传输。

[0143] 在一种实施方式中,每个固态存储元件 216、218、220 被分割成擦除块,并且每个擦除块被分割成页。典型的页的容量为 2000 字节 (“2kB”)。在一个实例中,固态存储元件 (如 SSS 0.0) 包括两个寄存器并能编程为两页,从而双寄存器固态存储元件 216、218、220 具有 4kB 的容量。20 个固态存储元件 216、218、220 的内存库 214 就会有 80kB 的页访问容量,同时同一地址流出存储 I/O 总线 210 的通道。

[0144] 在固态存储元件 216、218、220 的内存库 214 中的这一组 80kB 大小的页可称为虚拟页。类似地,内存库 214a 的每个存储元件 216a-m 的擦除块可被分组以形成虚拟块。在优

选实施方式中,当在固态存储元件 216、218、220 中接收到擦除命令时,擦除位于固态存储元件 216、218、220 内的页擦除块。然而,在固态存储元件 216、218、220 内的擦除块、页、平面层或其他逻辑和物理部分的大小和数量预计会随着技术的进步而变化,可以预期的是,与新配置一致的许多实施例是可能的并与本文的一般描述相一致。

[0145] 通常,当将包写入固态存储元件 216、218、220 内的特定位置时,其中,拟将所述包写入特定页内的位置,所述特定页对应于特定内存库的特定元件的特定擦除块的页,在发送所述包之后通过存储 I/O 总线 210 发送物理地址。所述物理地址包含足够的信息,以使得固态存储元件 216、218、220 将所述包导入页内的指定位置。由于存储元件行 (如 SSS 0.0-SSS 0.N 216a、218a、220a) 上的存储元件通过存储 I/O 总线 210a.a 内的合适总线同时被访问,为了到达合适的页并将所述数据包写入在存储元件行 (SSS 0.0-SSS 0.N 216a、218a、220a) 中具有相似地址的页,存储控制总线 212 同时选择内存库 214a(包括具有要将所述数据包写入其内的正确页的固态存储元件 SSS 0.0 216a)。

[0146] 类似地,在存储 I/O 总线 210 上传输的读命令需要同时在存储控制总线 212 上传输的命令,以选择单个的内存库 214a 和内存库 214 内的合适页。在优选实施方式中,读命令读取整个页,并且由于在内存库 214 内存在许多并行的固态存储元件 216、218、220,读命令读取整个虚拟页。然而,所述读命令可分割为子命令,这将在下文中结合内存库交错进行解释。还可以在写操作中访问虚拟页。

[0147] 可通过存储 I/O 总线 210 发出的擦除块擦除命令以擦除擦除块,该擦除块具有特定的擦除块地址以擦除特定的擦除块。通常,可通过存储 I/O 总线 210 的并行通路发送擦除块擦除命令以擦除虚拟擦除块,每个虚拟擦除块具有特定的擦除块地址以擦除特定的擦除块。同时,通过存储控制总线 212 选择特定的内存库 (如内存库 -0 214a) 以防止擦除所有的内存库 (内存库 1-N 214b-n) 中的具有类似地址的擦除块。还可采用存储 I/O 总线 210 和存储控制总线 212 的结合将其他命令发送到特定位置。本领域技术人员会认识到采用双向存储 I/O 总线 210 和存储控制总线 212 选择特定存储单元的其他方法。

[0148] 在一种实施方式中,将包顺序地写入固态存储器 110。例如,包流到存储元件 216 的内存库 214a 的存储写入缓冲器,并且当所述缓冲器饱和时,所述包被编程入指定的虚拟页。然后所述包再次填充所述存储写入缓冲器,并且当所述存储缓冲器再次饱和时,所述包被写入下一虚拟页。所述下一个虚拟页可位于同一个内存库 214a 内或可位于另一个内存库 (如 214b) 内。这个过程 (一个虚拟页接一个虚拟页) 通常一直持续到虚拟块被填满时。在另一种实施方式中,当这个过程 (一个虚拟擦除块接一个虚拟擦除块) 持续时,数据流可继续越过虚拟擦除块边界。

[0149] 在读、修改、写操作中,在读操作中定位并读取与所述对象关联的数据包。已被修改的修改对象的数据段并不写入读取它们的位置。取而代之,修改的数据段再次被转化为数据包并随后被写入正在被写入的虚拟页中的下一个可用位置。各个数据包的所述对象索引项被修改为指向包含已修改的数据段的包。所述对象索引中用于与同一对象 (未被修改) 关联的数据包的项 (或多个项) 会包括指向未被修改的数据包的源位置的指针。因此,如果源对象保持不变 (例如保持所述对象的先前版本不变),所述源对象将在所述对象索引中具有指向所有与最初写入的一样的数据包的指针。新对象将在所述对象索引中具有指向一些源数据包的指针和指向正在被写入的虚拟页中的修改的数据包的指针。

[0150] 在复制操作中,所述对象索引包括用于源对象的项,该源对象映射到若干存储在固态存储器 110 中的包。当复制完拷贝时,创建了新对象并在所述对象索引中创建将所述新对象映射到源包的新项。还将所述新对象写入固态存储器 110,且所述新对象的地址映射到所述对象索引中的新项。新对象包可用于确定在源对象中的包,该包被引用以防在未复制的源对象中发生改变以防对象索引丢失或损坏。

[0151] 有利地是,顺序地写入包有助于更平滑地使用固态存储器 110 并允许固态存储设备控制器 202 监测固态存储器 110 内的存储热点和不同虚拟页的层使用状况。相继地写入包还可有助于建立强大、高效的垃圾收集系统,这将在下文中详细描述。本领域技术人员会认识到顺序地存储数据包的其他好处。

[0152] 固态存储设备控制器

[0153] 在不同的实施方式中,固态存储设备控制器 202 还可包括数据总线 204、局部总线 206、缓冲控制器 208、缓冲器 0-N 222a-n,主控制器 224、直接存储器存取 (“DMA”) 控制器 226、存储器控制器 228、动态存储器阵列 230、静态随机存储器阵列 232、管理控制器 234、管理总线 236、连接系统总线 240 的网桥 238 和杂项逻辑块 242,这些将在下文中描述。在其他实施方式中,系统总线 240 与一个或多个网络接口卡 (“NIC”) 244 相连接,这些网络接口卡中的一些可包括远程 DMA (“RDMA”) 控制器 246、一个或多个中央处理器 (“CPU”) 248、一个或多个外部存储器控制器 250 和关联的外部存储器阵列 252、一个或多个存储控制器 254、对等控制器 256 和专用处理器 258,这将在下文描述。连接到系统总线 240 的组件 244-258 可位于计算内 112 内或者可以为其他设备。

[0154] 通常,固态存储控制器 104 通过存储 I/O 总线 210 与固态存储器 110 进行数据通信。在典型的实施方式中,固态存储器布置在内存库 214 内,且每个内存库 214 包括多个并行访问的存储元件 216、218、220,存储 I/O 总线 210 是多条总线的阵列,每一条总线用于内存库 214 内的存储元件 216、218、220 的每一行。正如此处所使用的,术语“存储 I/O 总线”可指一条存储 I/O 总线 210 或多条独立的数据总线 204 的阵列。在优选实施方式中,访问存储元件的行(如 216、218a、220a)的每条存储 I/O 总线 210 可包括在存储元件 216、218a、220a 的行中访问的存储部(如擦除块)的逻辑 - 物理映射。如果第一存储部失效、部分失效、不可访问或出现一些其他问题时,这种映射允许映射到存储部的物理地址的逻辑地址重映射到不同的存储部。相对于图 4A 和 4B 中的重映射模块 430 进一步解释了重映射。

[0155] 还可通过系统总线 240、网桥 238、局部总线 206、缓冲器 222 并最终通过数据总线 204 将数据从请求设备 155 传送到固态存储控制器 104。数据总线 204 通常连接到一个或多个由缓冲控制器 208 控制的缓冲器 222a-n。缓冲控制器 208 通常控制数据从局部总线 206 传递到缓冲器 222 并通过数据总线 204 传递到管道输入缓冲器 306 和输出缓冲器 330。为了解决时钟域差异、防止数据冲突等等,缓冲控制器 208 通常控制在缓冲器 222 中暂时存储来自请求设备 155 的数据的方式,并控制此后传送给数据总线 204(或相反)的方式。缓冲控制器 208 通常与主控制器 224 结合使用以协调数据流。当数据到达时,所述数据会到达系统总线 240 并通过网桥 238 传递给局部总线 206.

[0156] 通常,数据在主控制器 224 和缓冲控制器 208 的控制下从局部总线 206 传递给一个或多个数据缓冲器 222。然后,所述数据通过固态控制器 104 从缓冲器 222 流向数据总线 204 并到达固态存储器 110(如 NAND 闪存或其他存储媒体)。在优选实施方式中,数据

与与所述数据一起到达的关联的带外元数据（“对象元数据”）采用一个或多个的数据通道被送达，所述数据通道包括一个或多个固态存储控制器 104a-104n-1 和关联的固态存储器 110a-110n-1，而至少一个通道（固态存储控制器 104n、固态存储器 110n）用于带内元数据（如索引信息和其他固态存储设备 102 内部生成的元数据）。

[0157] 局部总线 206 通常为双向总线或总线组，所述双向总线或总线组允许数据和命令在固态存储设备控制器 202 内部的设备间通信，也允许命令和数据在固态存储设备 102 内部的设备和与系统总线 240 连接的设备 244-258 之间通信。网桥 238 有助于在局部总线 206 和系统总线 240 之间的通信。本领域技术人员会认识到其他实施方式，如总线 240、206、204、210 和网桥 238 的环结构或交换式星形配置和功能。

[0158] 系统总线 240 通常是计算机、安装有或连接有固态存储设备 102 的其他设备的总线。在一种实施方式中，系统总线 240 可以为 PCI-e 总线、串行高级技术附件（“串行 ATA”）总线、并行 ATA 或类似总线。在另一种实施方式中，系统总线 240 为外部总线，例如小型计算机系统接口（“SCSI”）、防火墙、光纤通道、USB、PCIe-As 或类似总线。固态存储设备 102 可被封装为适于置于设备内部或被封装为外部连接设备。

[0159] 固态存储设备控制器 202 包括在固态存储设备 102 内控制较高级别功能的主控制器 224。在不同的实施方式中，主控制器 224 通过解释对象请求和其他请求来控制数据流，指导创建索引，该索引将与数据关联的对象标识符映射到关联的数据（或协调的 DMA 请求等）的物理地址。主控制器 224 完全地或部分地控制此处描述的许多功能。

[0160] 在一种实施方式中，主控制器 224 采用嵌入式控制器。在另一种实施方式中，主控制器 224 采用局部存储器，如动态存储器阵列 230（动态随机存取存储器“DRAM”）、静态存储器阵列 232（静态随机存取存储器“SRAM”）等。在一种实施方式中，采用主控制器 224 控制局部存储器。在另一实施方式中，主控制器 224 通过存储器控制器 228 访问局部存储器。在另一种实施方式中，所述主控制器 224 运行 Linux 服务器并可支持各种常用服务器接口，如万维网、超文本标记语言（“HTML”）等。在另一种实施方式中，主控制器 224 采用纳米处理器。可采用可编程或标准逻辑或上述控制器类型的任意组合来构建主控制器 224。本领域技术人员会认识到主控制器 224 的许多实施方式。

[0161] 在一种实施方式中，其中，存储控制器 152/ 固态存储设备控制器 202 管理多个数据存储设备 / 固态存储器 110a-n，主控制器 224 在内部控制器（如固态存储控制器 104a-n）之间分配工作负载。例如，主控制器 224 可分割将要被写入数据存储设备（如固态存储器 110a-n）中的对象，使得每个附属的数据存储设备存储所述对象的一部分。这种特征是允许更快地存储和访问对象的性能增强。在一种实施方式中，主控制器 224 利用 FPGA 实施。在另一种实施方式中，位于主控制器 224 内的固件可通过管理总线 236、通过网络连接到 NIC244 的系统总线 240 或其他连接到系统总线 240 的设备更新。

[0162] 在一种实施方式中，管理对象的主控制器 224 仿真块存储，从而使得计算机 112 或其他连接到存储设备 / 固态存储设备 102 的设备将存储设备 / 固态存储设备 102 视为块存储设备并将数据发送给存储设备 / 固态存储设备 120 中的特定物理地址。然后，主控制器 224 分配块并像存储对象一样存储数据块。然后，主控制器 224 将块和与块一起发送的物理地址映射到由主控制器 224 确定的实际位置。映射存储在对象索引中。通常，对于块仿真来说在计算机 112、客户端 114 或其他希望将存储设备 / 固态存储设备 102 当成块存储设备

来使用的设备中提供有块设备应用程序接口（“API”）。

[0163] 在另一种实施方式中，主控制器 224 与 NIC 控制器 244 和嵌入式 RDMA 控制器 246 协同运行以提供准时的 RDMA 数据和命令集传输。NIC 控制器 244 可隐藏在非透明端口后以使得能够使用自定义的驱动器。同样地，客户端 114 上的驱动器可通过采用标准栈 API 的并与 NIC244 结合运行的 I/O 存储驱动器访问计算机网络 116。

[0164] 在一种实施方式中，主控制器 224 也是独立驱动器冗余阵列（“RAID”）控制器。当数据存储设备 / 固态存储设备 120 与一个或多个其他数据存储设备 / 固态存储设备 120 联网时，主控制器 224 可以是用于单层 RAID、多层 RAID、渐进 RAID 等的 RAID 控制器。主控制器 224 还允许一些对象存储在 RAID 阵列内而其他对象不通过 RAID 存储。在另一种实施方式中，主控制器 224 可以是分布式 RAID 控制器元件。在另一种实施方式中，主控制器 224 可包括许多 RAID、分布式 RAID 和另行描述的其他功能。

[0165] 在一种实施方式中，主控制器 224 与单个或多个网络管理器（如交换机）协同运行以建立路由、平衡带宽使用率、故障转移等。在另一种实施方式中，主控制器 224 与集成专用逻辑器件（通过局部总线 206）和关联的驱动器软件协同运行。在另一种实施方式中，主控制器 224 与附属专用处理器 258 或逻辑器件（通过外部系统总线 240）和关联的驱动器软件协同运行。在另一种实施方式中，主控制器 224 与远程专用逻辑器件（通过计算机网络 116 和关联的驱动器软件）协同运行。在另一种实施方式中，主控制器 224 与局部总线 206 或附属于硬盘驱动器（“HDD”）存储控制器的外部总线协同运行。

[0166] 在一种实施方式中，主控制器 224 与一个或多个存储控制器 254 通信，其中存储设备 / 固态存储设备 120 可表现为通过 SCSI 总线、因特网 SCSI（“iSCSI”）、光纤通道等连接的存储设备。同时，存储设备 / 固态存储设备 120 可自主地管理对象并可表现为对象文件系统或分布式对象文件系统。还可通过对等控制器 256 和 / 或专用处理器 258 访问主控制器 224。

[0167] 在另一种实施方式中，主控制器 224 与自主集成管理控制器协同运行以周期性地验证 FPGA 码和 / 或控制器软件、在运行（复位）时验证 FPGA 码和 / 或在通电（复位）期间验证控制器软件、支持外部复位请求、支持由于检查包而超时的复位请求，并支持电压、电流、功率、温度及其他环境测量和阈值中断设置。在另一种实施方式中，主控制器 224 管理垃圾收集以释放擦除块用于再次使用。在另一种实施方式中，主控制器 224 管理耗损均衡。在另一种实施方式中，主控制器 224 允许数据存储设备 / 固态存储设备 102 被分割成多个虚拟设备并允许基于分区的媒体加密。在又一种实施方式中，主控制器 224 支持具有高级的、多位的 ECC 修正的固态存储控制器 104。本领域技术人员会认识到位于存储控制器 152 内（或更具体地说位于固态存储设备 102 内）的主控制器 224 的其他特征和功能。

[0168] 在一种实施方式中，固态存储设备控制器 202 包括存储器控制器 228，该存储器控制器 228 控制动态随机存储器阵列 230 和 / 或静态随机存储器阵列 232。如上所述，存储器控制器 228 可独立于主控制器 224 使用或与主控制器 224 集成使用。存储器控制器 228 通常控制验证一些存储器类型，如 DRAM（动态随机存储器阵列 230）和 SRAM（静态随机存储器阵列 232）。在其他实例中，存储器控制器 228 还控制其他存储器类型，如电可擦可编程只读存储器（“EEPROM”）等。在其他实施方式中，存储器控制器 228 控制两种或更多种存储器类型且存储器控制器 228 可包括不止一个控制器。通常，存储器控制器 228 在可行情况

下控制尽可能多的 SRAM232，并且通过 DRAM230 补足 SRAM232。

[0169] 在一种实施方式中，所述对象索引存储在存储器 230、232 中并周期性的被卸载到固态存储器 110n 或其他非易失性存储器的通道内。本领域技术人员会认识到存储器控制器 228、动态存储器阵列 230、静态存储器阵列 232 的其他运用和配置。

[0170] 在一种实施方式中，固态存储设备控制器 202 包括 DMA 控制器 226，该 DMA 控制器 226 控制在下列设备之间的 DMA 操作：存储设备 / 固态存储设备 102、一个或多个外部存储器控制器 250、关联的外部存储器阵列 252 和 CPU248。应该注意到，外部存储器控制器 250 和外部存储器阵列 252 之所以被称为外部是因为它们位于存储设备 / 固态存储设备 102 的外部。此外，DMA 控制器 226 还可通过 NIC244 和关联的 RDMA 控制器 246 控制请求设备的 RDMA 操作。DMA 和 RDMA 在下文中有详细说明。

[0171] 在一种实施方式中，固态存储设备控制器 202 包括连接到管理总线 236 的管理控制器 234。通常管理控制器 234 管理存储设备 / 固态存储设备 102 的环境指标和状态。管理控制器 234 可通过管理总线 236 监测设备温度、风扇转速、电力供应设置等。管理控制器 234 可支持电可擦可编程序只读存储器（“EEPROM”）以存储 FPGA 码和控制器软件。通常，管理总线 236 连接到存储设备 / 固态存储设备 102 内的不同组件。管理控制器 234 可通过局部总线 206 进行警报、中断等的通信或可包括单独的到系统总线 240 或其他总线的连接。在一种实施方式中，管理总线 236 为内部集成电路（“I²C”）总线。本领域技术人员会认识到通过管理总线 236 连接到存储设备 / 固态存储设备 102 的组件的管理控制器 234 的其他功能和运用。

[0172] 在一种实施方式中，固态存储设备控制器 202 包括杂项逻辑块 242，该杂项逻辑块 242 可被定制为专用。通常，当固态设备控制器 202 或主控制器 224 被配置为使用 FPGA 或其他可配置控制器时，可基于特定应用、用户需求、存储需求等而包括定制逻辑。

[0173] 数据管道

[0174] 图 3 是示出了根据本发明的位于固态存储设备 102 内的固态存储设备控制器 104 的一种实施方式 300 的示意性框图，该固态存储设备控制器具有写入数据管道 106 和读取数据管道 108。实施方式 300 包括数据总线 204、局部总线 206 和缓冲控制器 208，这些设备大体上类似于相对于图 2 中固态存储设备控制器 202 描述的设备。所述写入数据管道 106 包括封包器 302 和纠错码（“ECC”）发生器 304。在其他实施方式中，所述写入数据管道 106 包括输入缓冲器 306、写入同步缓冲器 308、写入程序模块 310、压缩模块 312、加密模块 314、垃圾收集器旁路 316（部分位于所述读取数据管道 108 内）、媒体加密模块 318 和写入缓冲器 320。读取数据管道 108 包括读取同步缓冲器 328、ECC 纠错模块 322、解包器 324、对齐模块 326 和输出缓冲器 330。在另一种实施方式中，读取数据管道 108 可包括媒体解密模块 332、垃圾收集器旁路 316 的一部分、解密模块 334、解压缩模块 336 和读取程序模块 338。固态存储控制器 104 还可包括控制与状态寄存器 340 和控制队列 342、内存库交错控制器 344、同步缓冲器 346、存储总线控制器 348 及多路转换器（“MUX”）350。固态控制器 104 的组件和关联的写入数据管道 106 和读取数据管道 108 描述如下。在其他实施方式中，可采用同步固态存储器 110 并且可不使用同步缓冲器 308、328。

[0175] 写入数据管道

[0176] 写入数据管道 106 包括封包器 302，该封包器直接地或间接地通过另一写入数据

管道 106 的级接收将要被写入固态存储器的数据或元数据段，并创建一个或多个大小适于固态存储器 110 的包。所述数据或元数据段通常是对象的一部分，但也可包括整个对象。在另一种实施方式中，所述数据段是数据块的一部分，但也可包括整个数据块。通常，对象接收自计算机 112、客户端 114 或其他计算机或设备并被以流向固态存储设备 102 或计算机 112 的数据段的形式传送给固态存储设备 102。数据段也可被称为另一名称（如数据包裹），本文所提及的数据段包括对象或数据块的全部或一部分。

[0177] 每个对象被存为一个或多个包。每个对象可具有一个或多个容器包。每个包包含包头。所述包头可包括包头类型字段。类型字段可包括数据、对象属性、元数据、数据段定界符（多包）、对象结构、对象连接及类似物。所述包头还可包括关于包的大小的信息（如包内的数据的字节数）。所述包的长度可由包类型确定。一个实例可能是利用数据包包头的偏移值来确定对象内数据段的位置。本领域技术人员会认识到其他包含在由封包器 302 添加到数据上的包头内的信息和其他添加到数据包的信息。

[0178] 每个包包括包头，还可能包括来自所述数据和元数据段的数据。每个包的包头包括用于将包与包所属对象联系起来的相关信息。例如，所述包头可包括对象标识符和偏移值，该偏移值表明了用于数据包形成的数据段、对象或数据块。所述包头还可包括存储总线控制器 348 用以存储包的逻辑地址。所述包头还可包括关于包的大小的信息（如包内字节数）。所述包头还可包括序列号，当生成数据段或对象时，该序列号识别数据段相对于对象内的其他包所属的位置。所述包头可包括包头类型字段。类型字段可包括数据、对象属性、元数据、数据段定界符（多包）、对象结构、对象连接及类似物。本领域技术人员会认识到其他包含在由封包器 302 加到数据上的包头内的信息和其他添加到数据包的信息。

[0179] 写入数据管道 106 包括 ECC 发生器 304，该 ECC 发生器为一个或多个接收自封包器 302 的包生成一个或多个纠错码（“ECC”）。ECC 发生器 304 通常采用纠错算法生成 ECC，该 ECC 与包一起存储。与包一起存储的 ECC 通常用于探测和纠正由于传送和存储而引起的错误。在一种实施方式中，包作为长度为 N 的未编码块流入 ECC 发生器 304。计算并添加长度为 S 的并发位，并作为长度为 N+S 的编码块输出。N 和 S 的值依赖于算法的特点，该算法被选择用于实现特定的性能、效率和鲁棒性指标。在优选实施方式中，在 ECC 块和包之间并没有固定关系；包可包括不止一个 ECC 块；ECC 块可包括不止一个包；且第一包可在 ECC 块内的任何位置终止而第二包可始于同一 ECC 块内的第一包终止的位置。在优选实施方式中，ECC 算法不能被动态修改。在优选实施方式中，与数据包一起存储的 ECC 足够稳健以在两个以上的位内纠正错误。

[0180] 有利地是，采用允许不止一位的修正甚至是两位修正的稳健 ECC 算法允许延长固态存储器 110 的使用寿命。例如，如果固态存储器 110 内使用闪存作为存储媒体，闪存在每个擦除周期内可被写入大约 100000 次不出现错误。这种使用期限可通过稳健 ECC 算法延长。固态存储设备 102 板载有 ECC 发生器 304 和相应的 ECC 纠错模块 322，固态存储设备 102 可在其内部纠正错误并具有比采用不甚稳健的 ECC 算法（如单位错误修正）更长的使用寿命。然而，在其他实施方式中，ECC 发生器 304 可采用不甚稳健的算法并可修正单位或双位错误。在另一种实施方式中，固态存储设备 110 可包括不甚可靠的存储器以增加容量，所述不甚可靠的存储器例如多级单元（“MLC”）闪存，所述不甚可靠的存储器在没有稳健 ECC 算法的情况下可以不充分可靠。

[0181] 在一种实施方式中,写入数据管道包括输入缓冲器 306,该输入缓冲器接收将要被写入固态存储器 110 的数据段并存储输入的数据段直到写入数据管道 106 的下一级,例如封包器 302(或其他更复杂写入数据管道 106 的其他级)准备处理下一个数据段。通过使用适当容量的数据缓冲器,输入缓冲器 306 通常允许写入数据管道 106 接收和处理数据段之间存在速率差异。输入缓冲器 306 还允许数据总线 204 将数据传送给写入数据管道 106 的速率大于写入数据管道 106 能支持的速率,从而改进数据总线 204 运行的效率。通常,当写入数据管道 106 不包括输入缓冲器 306 时,缓冲功能在别处(如固态存储设备 102)实现,但所述别处位于写入数据管道 106 外、位于计算机内,例如当使用远程直接存储器读取(“RMAD”)时,如在网络接口卡(“NIC”)内或其他设备上。

[0182] 在另一种实施方式中,写入数据管道 106 还包括写入同步缓冲器 308,该写入同步缓冲器 308 在将包写入固态存储器 110 之前缓冲接收自 ECC 发生器 304 的包。写入同步缓冲器 308 位于本地时钟域和固态存储时钟域之间的边界上,并且提供缓冲以解决时钟域差异。在其他实施方式中,可采用同步固态存储器 110 而移除同步缓冲器 308、328。

[0183] 在一种实施方式中,写入数据管道 106 还包括媒体加密模块 318,该媒体加密模块 318 直接地或间接地从封包器 302 接收一个或多个包,并在将包发送给 ECC 发生器 304 之前利用对固态存储设备 102 唯一的加密密钥加密所述一个或多个包。通常,整个包(包括包头)都被加密。在另一种实施方式中,并不加密包头。在本文中,在一种实施方式中,加密密钥被理解为意味着在外部管理的秘密加密密钥,这种密钥将固态存储器 110 和需要加密保护的设备集成在一起。媒体加密模块 318 和相应的媒体解密模块 332 为存储在固态存储器 110 中数据提供安全等级。例如,当数据利用媒体加密模块加密时,如果固态存储器 110 连接到不同的固态存储控制器 104、固态存储设备 102 或计算机 112,通常,在不使用同一加密密钥(在将数据写入固态存储器 110 期间使用)时,如果不经过合理的努力,则不能读取固态存储器 110 的内容。

[0184] 在典型的实施方式中,固态存储设备 102 不将所述加密密钥存储在非易失性存储器中并且不允许从外部访问所述加密密钥。在初始化期间为固态存储控制器 104 提供加密密钥。固态存储设备 102 可使用并存储非秘密性加密临时值,该非秘密性加密临时值与加密密钥结合使用。不同的临时值可与每个包一起存储。为了加强保护,加密算法可利用唯一临时值在多个包之间分割数据段。所述加密密钥可接收自客户端 114、计算机 112、密钥管理器或其他管理固态存储控制器 104 使用的加密密钥的设备。在另一种实施方式中,固态存储器 110 可具有两个或更多个分区,并且固态存储控制器 104 显得就像有两个或更多个固态存储控制器 104,每一个固态存储控制器 104 在固态存储器 110 内的单个分区上运行。在这种实施方式中,唯一的媒体加密密钥可与每个分区一起使用。

[0185] 在另一种实施方式中,写入数据管道 106 还包括加密模块 314,该加密模块 314 在将数据段发送给封包器 302 之前直接地或间接地加密接收自输入缓冲器 306 的数据或元数据段,利用与数据段一同接收的加密密钥来加密数据段。加密模块 314 与媒体加密模块 318 不同,这是由于:加密模块 314 用以加密数据的加密密钥对存储在固态存储设备 102 内的数据来说不是共同的并在对象基础上可能不同,并且加密密钥可不与数据段一起接收(如下所述)。例如,加密模块 314 用以加密数据段的加密密钥可与数据段一起被接收或可作为将对象写入数据段所属位置的命令的一部分被接收。固态存储设备 102 可在每个与加密密

钥结合使用的对象包中使用并存储非秘密性加密临时值。不同的临时值可与每个包一起存储。为了通过加密算法加强保护,可利用唯一临时值在多个包之间分割数据段。在一种实施方式中,媒体加密模块 318 使用的临时值与加密模块 314 使用的临时值相同。

[0186] 加密密钥可接收自客户端 114、计算机 112、密钥管理器或其他保存用于加密数据段的加密密钥的设备。在一种实施方式中,加密密钥被从固态存储设备 102、计算机 112、客户端 114 或其他外部代理中的一个传送到固态存储控制器 104,所述外部代理能够执行工业标准方法以安全地传送并保护私有密钥和公共密钥。

[0187] 在一种实施方式中,加密模块 314 利用与第一包一起接收的第一加密密钥加密第一包,并利用与第二包一起接收的第二加密密钥加密第二包。在另一种实施方式中,加密模块 314 利用与第一包一起接收的第一加密密钥加密第一包,而将第二数据包传递给下一级(未经加密)。有利地是,包括在固态存储设备 102 的写入数据管道 106 内的加密模块 318 允许对象接对象或段接段的数据加密,而不需要单独的文件系统或其他外部系统来追踪不同的用于存储相应用对象或数据段的加密密钥。每个请求设备 155 或相关密钥管理器独立地管理加密密钥,该加密密钥仅用于加密请求设备 155 发送的对象或数据段。

[0188] 在另一种实施方式中,写入数据管道 106 包括压缩模块 312,该压缩模块 312 在将数据段发送给封包器 302 之前为元数据段压缩数据。压缩模块 312 通常利用本领域技术人员熟知的压缩程序来压缩数据或元数据段以减少段占用的存储空间大小。例如,如果数据段包括一串 512 个 0 位,压缩模块 312 可用表明 512 个 0 位的编码或令牌来替换这 512 个 0 位,其中,所述编码所占的空间比 512 个 0 位所占的空间要小得多。

[0189] 在一种实施方式中,压缩模块 312 利用第一压缩程序压缩第一段,而输送第二段(未经压缩)。在另一种实施方式中,压缩模块 312 利用第一压缩程序压缩第一段并利用第二压缩程序压缩第二段。在固态存储设备 102 内具有这种灵活性是有利的,以便客户端或其他将数据写入固态存储设备 102 内的设备中每一个都可指定压缩程序或以便一个设备指定压缩程序而另一个设备指定无压缩。还可根据每个对象类型或对象类基础的默认设置来选择压缩程序。例如,特定对象的第一对象可以能够废除默认压缩程序设置,同一对象类和对象类型的第二对象可采用默认压缩程序,而同一对象类和对象类型的第三对象可不压缩。

[0190] 在一种实施方式中,写入数据管道 106 包括垃圾收集器旁路 316,该垃圾收集器旁路 316 接收来自读取数据管道的 108(在垃圾收集系统中作为数据旁路的一部分)的数据段。垃圾收集系统通常标记不再有效的包,不再有效的原因通常是由于包被标记为删除或包已被修改且修改过的数据存储在不同的位置。在某一时刻,垃圾收集系统确定存储器的某个区域可被恢复。之所以确定某个区域可被恢复可能是由于:缺乏可用的存储空间、标记为无效的数据百分比达到阈值、有效数据的合并、存储器的该区域错误检出率达到阈值或基于数据分布提高性能等。垃圾收集算法可考虑大量的因素以确定何时存储器的区域将要被恢复。

[0191] 一旦存储器的区域被标记为恢复,该区域内的有效包通常必须被重新存放。垃圾收集器旁路 316 允许将包读入读取数据管道 108,并允许然后将包直接传送给写入数据管道 106 而不会将包路由出固态存储控制器 104。在优选实施方式中,垃圾收集器旁路 316 是运行在固态存储设备 102 内的自主垃圾收集系统的一部分。这允许固态存储设备 102 管理

数据,从而数据系统地传播到整个固态存储器 110 以提升性能、数据可靠性并避免过度使用和不充分使用固态存储器 110 的任何一个位置或区域,并且延长了固态存储器 110 的使用寿命。

[0192] 垃圾收集器旁路 316 协调将数据段插入写入数据管道 106 而其他数据段由客户端 114 或其他设备写入。在描述的实施方式中,垃圾收集器旁路 316 位于写入数据管道 106 内的封包器 302 之前、读取数据管道内的解包器 314 之后,但也可位于写入和读取数据管道 106、108 内的其他位置。可在清洗写入数据管道 106 期间使用垃圾收集器旁路 316,以填充虚拟页的剩余部分,从而提升固态存储器 110 内的存储效率并因此减少垃圾收集的频率。

[0193] 在一种实施方式中,写入数据管道 106 包括写入缓冲器 320,该写入缓冲器 320 为了高效的写操作而缓冲数据。通常,写入缓冲器 320 包括用于包的足够容量,以填充固态存储器 110 内的至少一个虚拟页。这允许写操作将数据的整个页没有中断地发送给固态存储器 110。通过选择写入数据管道 106 的写入缓冲器 320 的容量并将读取数据管道 108 内的缓冲器的容量选为同样大小容量或比固态存储器 110 内存储写入缓冲器的容量大,由于单个写入命令可被设计为将数据的整个虚拟页发送给固态存储器 110,从而以单条命令替代多条命令,写入和读取数据的效率更高。

[0194] 当填充写入缓冲器 320 时,固态存储器 110 可用于其他读操作。这是有利的,原因是:当将数据写入存储写入缓冲器时和注入数据缓冲器的数据失速时,具有更小容量的写入缓冲器的或不具有写入缓冲器的其他固态设备可绑定固态存储器。读操作会被拦截直到整个存储写入缓冲器被填充或被编程。用于不具写入缓冲器或具有小容量的写入缓冲器的系统的另一种方法是清洗未满的存储写入缓冲器以使得能进行读操作。同样地,由于需要多写入 / 编程周期来填充页,因此这种方法的效率低下。

[0195] 对于描述的具有容量比虚拟页容量大的写入缓冲器 320 的实施方式,单个的写入命令(包括大量子命令)的后续命令可以是单个程序命令,以将来自每个固态存储元件 216、218、220 中的存储写入缓冲器的数据页传递给每个固态存储元件 216、218、220 中的指定页。这种技术带来的好处是:减少了部分页编程,众所周知,这降低了数据的可靠性和稳定性并在当缓冲器填充时,为读命令和其他命令释放了目标内存库。

[0196] 在一种实施方式中,写入缓冲器 320 为交替缓冲器,其中,所述交替缓冲器的一侧被填充,然后当所述交替缓冲器的另一侧被填充时,所述交替缓冲器的一侧被指定为在适当的时间传送数据。在另一种实施方式中,写入缓冲器 320 包括先进先出(“FIFO”)寄存器,该 FIFO 寄存器的容量比数据段虚拟页的容量大。本领域技术人员会认识到允许在将数据写入固态存储器 110 之前存储数据虚拟页的其他写入缓冲器 320 配置。

[0197] 在另一种实施方式中,写入缓冲器 320 的容量比虚拟页小,从而少于一页的信息可被写入固态存储器 110 内的存储写入缓冲器。在这种实施方式中,为了防止写入数据管道 106 的失速阻止读操作,采用需要从一个位置移动到另一个位置的垃圾收集系统将数据排队,这个过程是垃圾收集进程的一部分。为了防止写入数据管道 106 中的数据失速,可通过垃圾收集器旁路 316 将所述数据供应给写入缓冲器 320 并然后将所述数据供应给固态存储器 110 中的存储写入缓冲器,从而在编程所述数据之前填充虚拟页的页面。这样,写入数据管道 106 中的数据失速不会使读取自固态存储设备 102 的数据失速。

[0198] 在另一种实施方式中,写入数据管道 106 包括写入程序模块 310,该写入程序模块

310 具有写入数据管道 106 内的一个或多个用户可定义的功能。写入程序模块 310 允许用户自定义写入数据管道 106。用户可基于特定数据请求或应用自定义写入数据管道 106。当固态存储控制器 104 为 FPGA 时, 用户可相对轻松地编程具有自定义命令和功能的写入数据管道 106。用户还可利用写入程序模块 310 以使 ASIC 包括自定义功能, 然而自定义 ASIC 可能比使用 FPGA 时更困难。写入程序模块 310 可包括缓冲器和旁路机制, 以允许第一数据段在写入程序模块 310 中执行, 而第二数据段通过写入数据管道 106 可继续传送。在另一种实施方式中, 写入程序模块 310 可包括能通过软件编程的处理器内核。

[0199] 应注意, 写入程序模块 310 被示为位于输入缓冲器 306 和压缩模块 312 之间, 然而写入程序模块 310 可位于写入数据管道 106 内的任何位置, 并且可分布在不同的级 302–320 之间。此外, 在不同的、已编程的且独立运行的级 302–320 之间可分布有多个写入程序模块 310。此外, 级 302–320 的顺序可以改变。本领域技术人员会认识到基于特定用户需求的级 302–320 的顺序的可行改变。

[0200] 读取数据管道

[0201] 读取数据管道 108 包括 ECC 纠错模块 322, 该 ECC 纠错模块 322 通过使用与请求包中的每个 ECC 块一起存储的 ECC 来确定接收自固态存储器 110 的请求包的 ECC 块中是否存在错误。然后, 如果存在任何错误并且所述错误可使用 ECC 修正, 则 ECC 纠错模块 322 修正请求包中的任何错误。例如, 如果 ECC 能够探测 6 位的错误但只能修正 3 位的错误, 那么 ECC 纠错模块 322 修正具有 3 位错误的请求包 ECC 块。ECC 纠错模块 322 通过把出错的位改变为正确的 1 或 0 状态来修正出错的位, 从而请求数据包与其被写入固态存储器 110 并且为包生成 ECC 时一致。

[0202] 如果 ECC 纠错模块 322 确定请求包包含了比 ECC 能修正的位数多的出错位, 则 ECC 纠错模块 322 不能修正请求包毁坏的 ECC 块的错误并发送中断。在一种实施方式中, ECC 纠错模块 322 发送中断以及指示请求包出错的消息。所述消息可包括指出 ECC 纠错模块 322 不能修正错误或 ECC 纠错模块 322 没有能力修正错误的信息。在另一种实施方式中, ECC 纠错模块 322 与所述中断和 / 或消息一起发送请求包中毁坏的 ECC 块。

[0203] 在优选的实施方式中, 请求包中毁坏的 ECC 块或毁坏的 ECC 块的一部分 (不能被 ECC 纠错模块 322 修正) 由主控制器 224 读取, 并被修正和返回给 ECC 纠错模块 322 以被读取数据管道 108 进一步处理。在一种实施方式中, 请求包中毁坏的 ECC 块或毁坏的 ECC 块的一部分被发送给请求数据的设备。请求设备 155 可修正所述 ECC 块或用另一拷贝替换数据 (如备份或镜像拷贝), 然后可使用请求数据包的替换的数据或将所述替换的数据返回给读取数据管道 108。请求设备 155 可使用出错请求包中的包头信息以识别替换毁坏请求包或替换包所属的对象所需的数据。在另一种优选实施方式中, 固态存储控制器 104 采用一些类型的 RAID 存储数据并能够恢复毁坏的数据。在另一种实施方式中, ECC 纠错模块 322 发送中断和 / 或消息, 并且接收设备停止与请求数据包关联的读操作。本领域技术人员会认识到 ECC 纠错模块 322 确定请求包的一个或多个 ECC 块为毁坏的且 ECC 纠错模块 322 不能修正错误后采取的其他选择和操作。

[0204] 读取数据管道 108 包括解包器 324, 该解包器 324 直接地或间接地接收来自 ECC 修正模块 322 的请求包 ECC 块, 并检查和删除一个或多个包头。解包器 324 可通过检查包头内的包标识符、数据长度、数据位置等验证包头。在一种实施方式中, 所述包头包括散列码,

该散列码可用于验证传递给读取数据管道 108 的包为请求包。解包器 324 还从请求包中删除由封包器 302 添加的包头。解包器 324 可被指定为不对某些包起作用而将这些包未经修改地向前传送。一个实例可以是容器标签，当对象索引重建模块 272 需要包头信息时，该容器标签在重建进程期间被请求。另外的实例包括传送不同类型的包（预定在固态存储设备 102 内使用）。在另一种实施方式中，解包器 324 操作可以依赖于包的类型。

[0205] 读取数据管道 326 包括对齐模块 326，该对齐模块 326 接收来自解包器 324 的数据并删除多余的数据。在一种实施方式中，发送给固态存储器 110 的读命令恢复数据包。请求数据的设备可不需要恢复的数据包内的所有数据，并且对齐模块 326 删除多余的数据。如果恢复页内的所有数据都是请求的数据，对齐模块 326 不删除任何数据。

[0206] 对齐模块 326 在数据段传输到下一级之前以与请求数据段的设备兼容的形式按对象的数据段重新格式化数据。通常，由于数据由读取数据管道 108 处理，数据段或包的大小在不同级间改变。对齐模块 326 使用接收到的数据以将数据格式化为适于发送给请求设备 155 的数据段，该数据段还适于连接在一起以形成响应。例如，来自第一数据包的一部分的数据可与来自第二数据包的一部分的数据结合。如果数据段比由请求设备 155 请求的数据大，对齐模块 326 可丢弃不需要的数据。

[0207] 在一种实施方式中，读取数据管道 108 包括读取同步缓冲器 328，该读取同步缓冲器 328 在读取数据管道 108 处理之前缓冲一个或多个读取自固态存储器 110 的请求包。读取同步缓冲器 328 位于固态存储时钟域和本地总线时钟域之间的边界上并提供缓冲以解决时钟域差异。

[0208] 在另一种实施方式中，读取数据管道 108 包括输出缓冲器 330，该输出缓冲器 330 接收来自对齐模块 326 的请求包并在数据包传送到所述请求设备 155 前存储该包。输出缓冲器 330 解决当从读取数据管道 108 接收数据段时和当将数据段传送给固态存储控制器 104 的其他部分或传送给请求设备 155 时之间的差异。输出缓冲器 330 还允许数据总线以比读取数据管道 108 能够支持的速率高的速率接收来自读取数据管道 108 的数据，以提升数据总线 204 运行的效率。

[0209] 在一种实施方式中，读取数据管道 108 包括媒体解密模块 332，该媒体解密模块 332 接收一个或多个来自 ECC 纠错模块 322 的加密过的请求包并在将一个或多个所述请求包发送给解包器 324 之前利用对于固态存储设备 102 唯一的加密密钥解密一个或多个所述请求包。通常，媒体解密模块 332 用以解密数据的加密密钥与媒体加密模块 318 使用的加密密钥一致。在另一种实施方式中，固态存储器 110 可具有两个或更多个分区且固态存储控制器 104 表现得好像有两个或更多个固态存储控制器 104（每个都在固态存储器 110 内的单独分区内运行）一样。在这种实施方式中，可对每个分区使用唯一的媒体加密密钥。

[0210] 在另一种实施方式中，读取数据管道 108 包括解密模块 334，该解密模块 334 在将数据段发送给输出缓冲器 330 之前解密由解包器 324 格式化的所述数据段。采用与读请求一起接收的加密密钥解密所述数据段，所述读请求初始化恢复由读取同步缓冲器 328 接收的请求包。解密模块 334 可利用与用于第一包的读请求一起接收的加密密钥解密第一包，然后可利用不同的加密密钥解密第二包或可将第二包未经解密地传送给读取数据管道 108 的下一级。通常，解密模块 334 使用与媒体解密模块 332 用以解密请求数据包的加密密钥不同的加密密钥解密数据段。当包与非秘密性加密临时值一起存储时，该临时值与加密密

钥一起使用以解密数据包。加密密钥可接收自客户端 114、计算机 112、密钥管理器或管理固态存储控制器 104 使用的加密密钥的其他设备。

[0211] 在另一种实施方式中,读取数据管道 108 包括解压缩模块 336,该解压缩模块 336 解压缩由解包器 324 格式化的数据段。在优选实施方式中,解压缩模块 336 使用存储在包头和容器标签中的一个或两个中的压缩信息以选择补充程序,压缩模块 312 使用该补充程序来压缩数据。在另一种实施方式中,解压缩模块 336 所使用的解压缩程序由请求解压缩的数据段确定。在另一种实施方式中,解压缩模块 336 根据每个对象类型或对象类基础的默认设置选择解压缩程序。第一对象的第一包可以能够废除默认解压缩程序设置,具有相对的对象类和对象类型的第二对象的第二包可采用默认解压缩程序,而具有相同的对象类和对象类型的第三对象的第三包可不经过解压缩。

[0212] 在另一种实施方式中,读取数据管道 108 包括读取程序模块 338,该读取程序模块 338 包括一个或多个在读取数据管道 108 内的用户可定义功能。读取程序模块 338 具有与写入程序模块 310 类似的特点并允许用户提供自定义功能给读取数据管道 108。读取程序模块 338 可位于图 3 中所示的位置、可位于读取数据管道 108 内的其他位置、或者可包括读取数据管道 108 内多个位置的多个部分。此外,在读取数据管道 108 内的多个不同位置可有多个独立运行的读取程序模块 338。本领域技术人员会认识到读取数据管道 108 内的读取程序模块 338 的其他形式。正如写入数据管道,读取数据管道 108 的级可重新排序,本领域技术人员会认识到读取数据管道 108 内的级的其他排列顺序。

[0213] 固态存储控制器 104 包括控制和状态寄存器 340 和相应的控制队列 342。控制和状态寄存器 340 和控制队列 342 有助于控制并按顺序排列与在写入和读取数据管道 106、108 内处理的数据相关联的命令和子命令。例如,封包器 302 中的数据段可具有一个或多个在与 ECC 发生器 304 关联的控制队列 342 内的相应控制命令或指令。当数据段被封包时,可在封包器 302 内执行一些指令或命令中。当从数据段建立的、最新形成的数据包被传送给下一级时,其他命令或指令可通过控制和状态寄存器 340 直接传送给下一个控制队列 342。

[0214] 可同时将命令和指令加载到控制队列 342 上以将包转发给写入数据管道 106,同时,由于每个管道级要执行各自的包,因此每个管道级读取合适的命令或指令。类似地,可同时将命令和指令加载到控制队列 342 上以从读取数据管道 108 请求包,而且,由于每个管道级要执行各自的包,因此每个管道级读取合适的命令或指示。本领域技术人员会认识到控制和状态寄存器 340 和控制队列 342 的其他特征和功能。

[0215] 固态存储控制器 104 和 / 或固态存储设备 102 还可包括内存库交错控制器 344、同步缓冲器 346、存储总线控制器 348 及多路转换器 (“MUX”) 350,这些设备相对于图 4A 和图 4B 描述。

[0216] 内存库交错

[0217] 图 4A 是根据本发明的位于固态存储控制器 104 内的内存库交错控制器 344 一种实施方式 400 的示意性框图。内存库交错控制器 344 连接到控制和状态寄存器 340 并通过 MUX350、存储总线控制器 348 和同步缓冲器 346 连接到存储 I/O 总线 210 和存储控制总线 212 上,这在下文中有所描述。内存库交错控制器 344 包括读取代理 402、写入代理 404、擦除代理 406、管理代理 408、读取队列 410a-n、写入队列 412a-n、擦除队列 414a-n、用于固态存储器 110 中的内存库 214 的管理队列 416a-n、内存库控制器 418a-n、总线仲裁器 420 和

状态 MUX422, 这些设备在下文中描述。存储总线控制器 348 包括具有重映射模块 430 的映射模块 424、状态捕捉模块 426 和 NAND 总线控制器 438, 这些设备在下文中描述。

[0218] 内存库交错控制器 344 将一条或多条命令送往内存库交错控制器 344 中的两个或更多个队列, 并在固态存储器 110 的内存库 214 之间协调存储在队列中的命令的执行, 以使得第一类型的命令在一个内存库 241a 上执行而第二类型的命令在第二内存库 214b 上执行。所述一条或多条命令按命令类型分别送入队列中。固态存储器 110 的每个内存库 214 在内存库交错控制器 344 内具有相应的队列集, 且每个队列集包括每个命令类型的队列。

[0219] 内存库交错控制器 344 在固态存储器 110 的内存库 214 之间协调存储在队列中的命令的执行。例如, 第一类型的命令在在一个内存库 241a 上执行而第二类型的命令在第二内存库 214b 上执行。通常, 命令类型和队列类型包括读取和写入命令和队列 410、412, 但是还可包括存储媒介指定的其他命令和队列。例如, 在图 4A 所描述的实施方式中, 擦除和管理队列 414、416 被包括在其中且适于闪存、NRAM、MRAM、DRAM、PRAM 等。

[0220] 对于其他类型的固态存储器 110, 可包括其他类型的命令和相应的队列而不脱离本发明的范围。FPGA 固态存储控制器 104 的灵活性允许存储媒介的灵活性。如果将闪存换成另一种固态存储类型, 可改变内存库交错控制器 344、存储总线控制器 348 和 MUX350 以适应媒介类型而不显著地影响数据管道 106、108 和其他固态存储控制器 104 运行。

[0221] 在图 4A 所描述的实施方式中, 对每个内存库 214 来说, 内存库交错控制器 344 包括: 用于从固态存储器 110 读取数据的读取队列 410、用于将命令写入固态存储器 110 的写入队列 412、用于擦除固态存储器中的擦除块的擦除队列 414、用于管理命令的管理队列 416。内存库交错控制器 344 还包括相应的读取、写入、擦除和管理代理 402、404、406、408。在另一种实施方式中, 控制和状态寄存器 340 和控制队列 342 或类似元件在没有内存库交错控制器 344 的情况下为了发送给固态存储器 110 的内存库 214 的数据而将命令排队。

[0222] 在一种实施方式中, 代理 402、404、406、408 将预定用于特定内存库 214a 的合适类型的命令送到内存库 214a 的修正队列。例如, 读取代理 402 可接收用于内存库 -1 214b 的读命令并将所述读命令送到内存库 -1 读取队列 410b。写入代理 404 可接收将数据写入固态存储器 110 的内存库 -0 214a 的写入命令并然后会将所述写入命令发送给内存库 -0 写入队列 412a。类似地, 擦除代理 406 可接收擦除命令以擦除内存库 -1 214b 中的擦除块并然后会将所述擦除命令传送给内存库 -1 擦除队列 414b。管理代理 408 通常接收管理命令、状态请求及其类似消息, 如复位命令或读取内存库 214(如内存库 -0 214a) 的配置寄存器的请求。管理代理 408 将所述管理命令发送给内存库 -0 管理队列 416a。

[0223] 代理 402、404、406、408 通常还监测队列 410、412、414、416 的状态并当队列 402、404、406、408 满、接近满、丧失功能时, 发送状态、中断或其他消息。在一种实施方式中, 代理 402、404、406、408 接收命令并生成相应的子命令。在一种实施方式中, 代理 402、404、406、408 通过控制和状态寄存器 340 接收命令并生成相应的子命令, 所述子命令被转发给队列 410、412、414、416。本领域技术人员会认识到代理 402、404、406、408 的其他功能。

[0224] 队列 410、412、414、416 通常接收命令并存储所述命令直到所述命令被要求传送给固态存储器内存库 214。在典型的实施方式中, 队列 410、412、414、416 是先进先出(“FIFO”)寄存器或以 FIFO 运行的类似组件。在另一种实施方式中, 队列 410、412、414、416 按与数据、重要性或其他标准相匹配的顺序来存储命令。

[0225] 内存库控制器 418 通常接收来自队列 410、412、414、416 的命令并生成合适的子命令。例如，内存库 -0 写入队列 412a 可接收将数据包的页写入内存库 -0214a 的命令。内存库 -0 控制器 418a 可在合适的时间接收写入命令并可为每个存储在写入缓冲器 320 中的数据包生成一个或多个写入子命令（将要被写入内存库 -0 214a 的页中）。例如，内存库 -0 控制器 418a 可生成验证内存库 -0 214a 和固态存储阵列 216 状态的命令、选择写入一个或多个数据包的合适位置的命令、清除位于固态存储阵列 216 内的输入缓冲器的命令、将一个或多个数据包传送所述输入缓冲器的命令、将输入缓冲器放到选定位置中的命令、检验数据被正确编程的命令，并且如果发生程序故障，则一次或多次地中断主控制器 224、重试写入同一物理地址并重试写入不同的物理地址。此外，与实例中的写入命令一起，存储总线控制器 348 会将一条或多条命令乘以每条存储 I/O 总线 210a-n 从而翻倍，而所述命令的逻辑地址映射到用于存储 I/O 总线 210a 的第一物理地址，并映射到用于存储 I/O 总线 210a 的第二物理地址，下面将详细描述。

[0226] 通常，总线仲裁器 420 选自内存库控制器 418 并从内存库控制器 418 的输出队列提取子命令，并且将这些子命令以最优化内存库 214 性能的序列形式发给存储总线控制器 348。在另一种实施方式中，总线仲裁器 420 可响应高级中断并修改普通选择标准。在另一种实施方式中，主控制器 224 可通过控制和状态寄存器 340 控制总线仲裁器 420。本领域技术人员会认识到总线控制器 420 可控制和交错从内存库控制器 418 传送到固态存储器 110 的命令序列。

[0227] 通常，总线仲裁器 420 协调来自内存库控制器 418 适当的命令和命令类型所需的相关数据的选择，并将所述命令和数据发送给存储总线控制器 348。总线仲裁器 420 通常还将命令发送给存储控制总线 212 以选择合适的内存库 214。对于闪存或其他具有异步、双向串行的存储 I/O 总线 210 的固态存储器 110 而言，一次只能传送一条命令（控制信息）或数据集。例如，当将写入命令或数据通过存储 I/O 总线 210 传送给固态存储器 110 时，读取命令、读取的数据、擦除命令、管理命令或其他状态命令不能在存储 I/O 总线 210 上传输。例如，当从存储 I/O 总线 210 读取数据时，不能向固态存储器 110 写入数据。

[0228] 例如，在内存库 -0 的写操作期间，总线仲裁器 420 选择在其队列顶部具有写入命令或一系列写入子命令的内存库 -0 控制器 418a，所述一系列写入子命令使得存储总线控制器 348 执行后继的序列。总线仲裁器 420 将写入命令转发给存储总线控制器 348，该存储总线控制器 348 通过下列方式建立了写入命令：通过存储控制总线 212 选择内存库 -0 214a、发送清除与内存库 -0 214a 关联的固态存储元件 110 的输入缓冲器的命令、发送验证与内存库 -0 214a 关联的固态存储元件 216、218、220 的状态的命令。然后，存储总线控制器 348 通过包含了物理地址存储 I/O 总线 210 传送写入命令，该物理地址如同映射自逻辑擦除块地址一样包括用于每个单独的物理擦除固态存储元件 216a-m 的逻辑擦除块地址。然后，存储总线控制器 348 通过多路转换器 350 将写入缓冲器经写入同步缓冲器多路复用到存储 I/O 总线 210 并使写入数据流向合适的页。当所述页写满时，然后，存储总线控制器 348 促使与内存库 -0 214a 关联的固态存储元件 216a-m 将输入缓冲器编入固态存储元件 216a-m 的内存单元。最终，存储总线控制器 348 验证状态以确保所述页被正确编程。

[0229] 读操作与上文的写操作实例类似。在读操作期间，通常，总线仲裁器 420 或内存库交错控制器 344 的其他组件接收数据和相应状态信息并将数据发送给读取数据管道

108,同时将状态信息发送给控制和状态寄存器 340。通常,从总线仲裁器 420 传送给存储总线控制器 348 的读数据命令会促使多路转换器 350 将读数据通过存储 I/O 总线 210 传送给读取数据管道 108 并通过状态多路转换器 422 向控制和状态寄存器 340 发送状态信息。

[0230] 总线仲裁器 420 协调不同的命令类型和数据存取模式,使得在任意给定的时间内,在总线上只有合适的命令类型或对应数据。如果总线仲裁器 420 已选择了写入命令,且写入子命令和对应数据正在被写入固态存储器 110,总线仲裁器 420 不会允许在存储 I/O 总线 210 存在其他命令类型。有利地是,总线仲裁器 420 使用定时信息(如预定的命令执行时间)以及接收到的关于内存库 214 状态的信息,以协调总线上不同命令的执行,这样做的目标是最小化或消除总线的停工时间。

[0231] 通过总线仲裁器 420 的主控制器 224 通常使用存储在队列 410、412、414、416 中的命令的预定完成时间以及状态信息,使得在一个内存库 214a 上执行与命令关联的子命令时,而在其他内存库 241b-n 上执行其他命令的其他子命令。当内存库 214a 完全执行完一条命令时,总线仲裁器 420 将其他命令传给内存库 214a。总线仲裁器 420 还可与协调存储在队列 410、412、414、416 的命令一起协调不存储在队列 410、412、414、416 的其他命令。

[0232] 例如,可发出擦除命令以擦除固态存储器 110 内的一组擦除块。执行擦除命令可消耗比执行写入或读取命令多 10 到 1000 倍的时间,或消耗比执行程序命令多 10 到 100 倍的时间。对于 N 个内存库 214,内存库交错控制器可将擦除命令分割为 N 条命令,每条命令擦除内存库 214a 的虚拟擦除块。当内存库 -0 214a 执行擦除命令时,总线仲裁器 420 可选择在其他内存库 214b-n 上执行的其他命令。总线仲裁器 420 还可与其他组件(如存储总线控制器 348、主控制器 224 等)一起工作以在总线之间协调命令的执行。利用总线仲裁器 420、内存库控制器 418、队列 410、412、414、416、和内存库交错控制器的代理 402、404、406、408 协调命令的执行可显著的提升性能(相比于其他没有内存库交错功能的固态存储系统)。

[0233] 在一种实施方式中,固态控制器 104 包括一个内存库交错控制器 344,该内存库交错控制器 344 为固态存储器 110 的所有存储元件 216、218、220 提供服务。在另一种实施方式中,固态控制器 104 内存库包括用于每个存储元件行 216a-m、218a-m、220a-m 的交错控制器 344。例如一个内存库交错控制器 344 服务存储元件的一行 SSS 0.0-SSS 0.N 216a、218a、220a,第二内存库交错控制器 344 服务存储元件的第二行 SSS 1.0-SSS 1.N 216b、218b、220b,等等。

[0234] 图 4B 是示出了根据本发明的位于固态存储设备 104 内的内存库交错控制器 344 的一种替代实施方式 401 的示意性框图。图 4B 所示实施方式中描述的组件 210、212、340、346、348、350、402-430 大体上与相对于图 4A 描述的内存库交错装置 400 类似,除了下述不同点:每个内存库 214 包括单独的队列 432a-n 及用于内存库的(如内存库 -0 214a)读取命令、写入命令、擦除命令、管理命令等被传送给内存库 214 的单独队列 432a。在一种实施方式中,队列 432 是 FIFO。在另一种实施方式中,队列 432 可具有以不同于存储的顺序的顺序从队列 432 中提取的命令。在另一种替代实施方式(未示出)中,读取代理 402、写入代理 404、擦除代理 406 和管理代理 408 可结合成单个代理,所述单个代理将命令分配给合适的队列 432a-n。

[0235] 在另一种替代的实施方式(未示出)中,命令存储在单独的队列中,其中,可以以

不同于存储的顺序的顺序从队列中提取命令,从而使得内存库交错控制器 344 在余下的内存库 214b-n 上执行。本领域技术人员会轻易地认识到其他能够在一个内存库 214a 上执行命令而在其他内存库 214b-n 上执行其他命令的队列配置和类型。

[0236] 特定存储组件

[0237] 固态存储控制器 104 包括同步缓冲器 346,该同步缓冲器 346 从固态存储器 110 发送和接收的命令和状态消息。同步缓冲器 346 位于固态存储时钟域和本地总线时钟域之间的边界上,并提供缓冲以解决时钟域差异。同步缓冲器 346、写入同步缓冲器 308 和读取同步缓冲器 328 可独立地或共同运作以缓冲数据、命令、状态消息等等。在优选实施方式中,同步缓冲器 346 所处的位置使得跨越时钟域的信号数量最少。本领域技术人员会认识到:时钟域间的同步可任意运行在固态存储设备 102 的其他位置,以优化设计实施方案的某些方面。

[0238] 固态存储控制器 104 包括存储总线控制器 348,该存储总线控制器 348 解释和翻译用于发送给或读取自固态存储器 110 的数据的命令并基于固态存储器 110 的类型接收自固态存储器 110 的状态消息。例如,存储总线控制器 348 可针对不同的存储类型、不同性能特点、不同制造商的存储器等而具有不同的定时要求。存储总线控制器 348 还将控制命令发送给存储控制总线 212。

[0239] 在优选实施方式中,固态存储控制器 104 包括 MUX350,该 MUX350 包括多路转换器 350a-n 的阵列,其中,每个多路转换器用于固态存储阵列 110 的一行。例如,多路转换器 350a 与固态存储元件 216a、218a、220a 关联。MUX350 通过存储总线控制器 348、同步缓冲器 346 和内存库交错控制器 344 将来自写入数据管道 106 的数据和来自存储总线控制器 348 的命令经存储 I/O 总线 210 路由至固态存储器 110,并将来自固态存储器 110 的数据和状态消息经存储 I/O 总线 210 路由至读取数据管道 108 和控制和状态寄存器 340。

[0240] 在优选实施方式中,固态存储控制器 104 包括用于固态存储元件的每一行的(如 SSS 0.1 216a、SSS 0.2 218a、SSS 0.N 220a)的 MUX350。MUX350 将来自写入数据管道 106 的数据和发送给固态存储器 110 的命令通过存储 I/O 总线 210 结合起来,并将需要由读取数据管道 108 处理的数据从命令中分离出来。存储在写入缓冲器 320 中的包通过用于固态存储元件的每一行(SSS x.0 to SSSx.N 216、218、220)的写入缓冲器 308 由写入缓冲器外的总线传给用于固态存储元件的每一行(SSS x.0 to SSS x.N 216、218、220)的 MUX350。MUX350 从存储 I/O 总线 210 接收命令和读取数据。MUX350 还将状态消息传给存储总线控制器 348。

[0241] 存储总线控制器 348 包括映射模块 424。映射模块 424 将擦除块的逻辑地址映射到擦除块的一个或多个物理地址。例如,每个内存库 214a 具有 20 个存储元件的阵列(如 SSS 0.0 至 SSS M.0 216)的固态存储器 110 可具有映射到擦除块的 20 个物理地址的特定擦除块的逻辑地址(每个存储元件有一个物理地址)。由于平行访问存储元件,所以位于存储元件 216a、218a、220a 的行中的每个存储元件中的同一位置的擦除块会分享物理地址。为了选择一个擦除块(如在存储元件 SSS 0.0 216a 中)代替行(如在存储元件 SSS 0.0、0.1, ..., 0.N 216a、218a、220a 中)中的所有擦除块,可选择一个内存库(在这种情况下为内存库 -0 214a)。

[0242] 这种用于擦除块的逻辑到物理的映射是有好处的,这是由于如果一个擦除块已损

坏或不可访问,所述映射可改为映射到另一擦除块。当一个元件的擦除块出错时,这种方法减少了失去整个虚拟擦除块的损失。重映射模块 430 将擦除块的逻辑地址的映射改为虚拟擦除块的一个或多个物理地址(遍布存储元件的阵列)。例如,虚拟擦除块 1 可映射到存储元件 SSS 0.0 216a 的擦除块 1、映射到存储元件 SSS 1.0 216b 的擦除块 1、... 和映射到存储元件 M.0 216m,虚拟擦除块 2 可映射到存储元件 SSS 0.1 218a 的擦除块 2、映射到存储元件 SSS1.1 218b 的擦除块 2、... 和映射到存储元件 M.1 218m,等等。

[0243] 如果存储元件 SSS 0.0 216a 的擦除块 1 损坏、由于损耗遇到错误或由于一些原因不能被使用,重映射模块 430 可将从逻辑到物理的映射改为指向虚拟擦除块 1 的擦除块 1 的逻辑地址的映射。如果存储元件 SSS 0.0 216a 的空闲擦除块(将其称为擦除块 221)可用且当前并未被映射,重映射模块 430 可改变虚拟擦除块 1 的映射为映射到指向存储元件 SSS 0.0 216 的擦除块 221,而继续指向存储元件 SSS 1.0 216b 的擦除块 1、存储元件 SSS 2.0(未示出)的擦除块 1、... 和指向存储元件 M.0 216m。映射模块 424 或重映射模块 430 可按固定顺序映射擦除块(虚拟擦除块 1 到存储元件的擦除块 1,虚拟擦除块 2 到存储元件的擦除块 2,等等)或可按基于其他一些标准的顺序映射存储元件 216、218、220 的擦除块。

[0244] 在一种实施方式中,可按访问时间分组擦除块。按访问时间分组、均衡命令执行的时间(如将数据编入或写入指定擦除块的页)可平均命令补齐,从而使得在虚拟擦除块的擦除块之间执行的命令不会由于最慢的擦除块而被限制。在另一种实施方式中,可按损耗程度、运行状况来分组擦除块。本领域技术人员会认识到当映射或重映射擦除块时需要考虑的其他问题。

[0245] 在一种实施方式中,存储总线控制器 348 包括状态捕捉模块 426,该状态捕捉模块 426 接收来自固态存储器 110 的状态消息并将该状态消息发送给状态 MUX422。在另一种实施方式中,当固态存储器 110 为闪存时,存储总线控制器 348 包括 NAND 总线控制器 428。NAND 总线控制器 428 将命令从读取和写入数据管道 106、108 传送给固态存储器 110 中的正确位置,并根据所述闪存的特点协调命令执行的时间,等等。如果固态存储器 110 为另一种类型的固态存储器,则将 NAND 总线控制器 428 替换为针对存储类型的总线控制器。本领域技术人员会认识到 NAND 总线控制器 428 的其他功能。

[0246] 流程图

[0247] 图 5A 是根据本发明的在固态存储设备 102 内采用数据管道管理数据的方法 500 的一种实施方式的示意性流程图。方法 500 始于步骤 502,输入缓冲器 306 接收一个或多个将要被写入固态存储器 110 的数据段(步骤 504)。通常来说,所述一个或多个数据段包括对象的至少一部分,但也可以是整个对象。封包器 302 可创建一个或多个对象指定包以及对象。封包器 302 为每个包添加包头,所述包头通常包括包的长度和对象内包的序列号。封包器 302 接收一个或多个存储在输入缓冲器 306 的数据或元数据段(步骤 504),并通过创建一个或多个大小适于固态存储器 110 的包来封包所述一个或多个数据或元数据段(步骤 506),其中,每个包包括一个包头和来自一个或多个段的数据。

[0248] 通常,第一包包括对象标识符,该对象标识符确定对象,为了该对象而创建包。第二包可包括具有信息的包头,该信息由固态存储设备 102 用于关联第二包和第一包中确定的对象,该包头还具有在对象内定位第二包的偏移信息和数据。固态存储设备控制器 202 管理内存库 214 和包流向的物理区域。

[0249] ECC 发生器 304 接收来自封包器 302 的包并为数据包生成的 ECC(步骤 508)。通常,在包和 ECC 块之间没有固定关系。ECC 块可包括一个或多个包。包可包括一个或多个 ECC 块。包可始于 ECC 块内的任意位置并可在 ECC 块内的任意位置结束。包可始于第一 ECC 块内的任意位置并可在相继的 ECC 块中的任意位置结束。

[0250] 写入同步缓冲器 308 在将 ECC 块写入固态存储器 110 之前缓冲分布在对应 ECC 块中的包 (步骤 510), 然后固态存储控制器 104 在考虑到时钟域差异的适当的时间写入数据 (步骤 512), 方法 500 终止于步骤 514。写入同步缓冲器 308 位于本地时钟域和固态存储器 110 时钟域的边界上。注意到为方便起见, 方法 500 描述了接收一个或多个数据段并写入一个或多个数据包, 但通常接收数据段流或组通常, 若干包括完整固态存储器 110 的虚拟页的 ECC 块被写入固态存储器 110。通常, 封包器 302 接收某个大小的数据段并生成另一大小的包。这必然需要数据或元数据段或数据或元数据段的部分结合起来, 以形成将段的所有数据捕捉进包的数据包。

[0251] 图 5 是示意性流程图, 示出了根据本发明的用于服务器内 SAN 的方法的一种实施方式。方法 501 始于步骤 552, 存储通信模块 162 促进第一存储控制器 152a 和位于第一服务器 112a 外部的至少一个设备之间的通信 (步骤 554)。第一存储控制器 152a 和外部设备之间的通信独立于第一服务器 112a。第一存储控制器 152a 位于第一服务器 112a 内部, 并且第一存储控制器 152a 控制至少一个存储设备 154a。第一服务器 112a 包括与第一服务器 112a 和第一存储控制器 152a 搭配使用的网络接口 156a。服务器内 SAN 模块 164 服务存储请求 (步骤 556), 方法 501 结束于步骤 558。所述服务器内 SAN 模块利用网络协议和 / 或总线协议服务所述存储请求 (步骤 556)。服务器内 SAN 模块 164 服务独立于第一服务器 112a 的所述存储请求 (步骤 556), 并且服务请求接收自客户端 114、114a。

[0252] 图 6 是根据本发明的在固态存储设备 102 内采用数据管道管理数据的方法 600 的再一种实施方式的示意性流程图。方法 600 始于步骤 602, 输入缓冲器 306 接收一个或多个将要被写入固态存储器 110 的数据或元数据段 (步骤 604)。封包器 302 为每个包添加包头, 所述包头通常包括对象内包的长度。封包器 302 接收一个或多个存储在输入缓冲器 306 中的段 (步骤 604), 并通过创建一个或多个大小适于固态存储器 110 的包来封包所述一个或多个段 (步骤 606), 其中每个包包括包头和来自一个或多个段的数据。

[0253] ECC 发生器 304 接收来自封包器 302 的包并生成一个或多个用于包的 ECC 块 (步骤 608)。写入同步缓冲器 308 在将 ECC 块写入固态存储器 110 之前缓冲分布在对应 ECC 块中的包 (步骤 610), 然后固态存储控制器 104 在考虑到时钟域差异的合适的时间写入数据 (步骤 612)。当从固态存储器 110 请求数据时, 包括一个或多个数据包的 ECC 块被读入读取同步缓冲器 328 并被缓冲 (步骤 614)。通过存储 I/O 总线 210 接收包的 ECC 块。由于存储 I/O 总线 210 是双向, 当读取数据时, 写操作、命令操作等被停止。

[0254] ECC 纠错模块 322 接收暂存在读取同步缓冲器 328 中的请求包的 ECC 块, 并在必要时修正每个 ECC 块中的错误 (步骤 616)。如果 ECC 纠错模块 322 确定在 ECC 块中存在一个或多个错误并且错误可利用 ECC 一并修正, ECC 纠错模块 322 修正 ECC 块中的错误 (步骤 616)。如果 ECC 纠错模块 322 确定探测到的错误不可用 ECC 修正, 则 ECC 纠错模块 322 发送中断。

[0255] 解包器 324 在 ECC 纠错模块 322 修正任何错误之后接收请求包 (步骤 618) 并通

过检查和删除每个包的包头解包所述包（步骤 618）。对齐模块 326 接收经过解包的包、删除多余的数据、并采用与请求数据段的设备兼容的形式按对象的数据段重新格式化所述数据（步骤 620）。输入缓冲器 330 接收经过解包的请求包，并在包传送给请求设备 155 之前缓冲包（步骤 622），方法 600 终止于步骤 624。

[0256] 图 7 是示意性流程图，示出了根据本发明的在固态存储设备 102 内利用内存库交错管理数据的方法 700 的一种实施方式。方法 700 始于步骤 702，内存库交错控制器 344 将一条或多条命令传给两个或多个队列 410、412、414、416（步骤 704）。通常，代理 402、404、406、408 根据命令类型将所述命令传给队列 410、412、414、416（步骤 704）。队列 410、412、414、416 的每个集包括用于每个命令类型的队列。内存库交错控制器 344 在内存库 214 之间协调存储在队列 410、412、414、416 的所述命令的执行（步骤 706），以使得第一类型的命令在一个内存库 214a 上执行，而第二类型的命令在第二内存库 214b 上执行，方法 700 结束于步骤 708。

[0257] 存储空间的恢复

[0258] 图 8 是示意性框图，示出了根据本发明的在固态存储设备 102 中收集垃圾的装置 800 的一种实施方式。装置 800 包括顺序存储模块 802、存储部选择模块 804、数据恢复模块 806 及存储部恢复模块 808，这些模块将在下文中描述。在其他实施方式中，装置 800 包括垃圾标记模块 812 和擦除模块 810。

[0259] 装置 800 包括顺序存储模块 802，该顺序存储模块 802 将数据包顺序地写入存储部内的页。无论是新的包还是修改过的包，这些包都按顺序存储。在这种实施方式中，通常不将修改过的包写回其先前存储的位置。在一种实施方式中，顺序存储模块 802 将包写入存储部的页中的第一位置，然后写入该页中的下一个位置，并继续写入下一个位置和再下一个位置，直到该页被写满。然后，顺序存储模块 802 开始填充所述存储部中的下一页。这个过程一直持续到所述存储部被写满。

[0260] 在优选实施方式中，顺序存储模块 802 开始将包写入内存库（内存库 -0214a）的存储元件（如 SSS 0.0 到 SSS M.0 216）中的存储写入缓冲器。当所述存储写入缓冲器写满时，固态存储控制器 104 使得所述存储写入缓冲器中的数据被编入内存库 214a 的存储元件 216 中的指定页。然后，另一个内存库（如内存库 -1 214b）被选定，并且当第一内存库 -0 214a 编程所述指定页时，顺序存储模块 802 开始将包写入内存库 214b 的存储元件 218 的存储写入缓冲器。当内存库 214b 的存储写入缓冲器写满时，该存储写入缓冲器中的内容被编入每个存储元件 218 中的另一指定页。这个过程是有效率的，这是因为当一个内存库 214a 编程页时，可填充另一个内存库 214b 的存储写入缓冲器。

[0261] 所述存储部包括固态存储设备 102 中的固态存储器 110 的一部分。通常，所述存储部为擦除块。对于闪存来说，擦除块上的擦除操作通过为每个单元充电将 1 写入所述擦除块中的每一位。相比于始于全为 1 的位置的程序操作，这是一个冗长过程，并且，当数据被写入时，一些位通过给被写为 0 的单元放电而改变为 0。然而，当固态存储器 110 不是闪存时或固态存储器 110 具有擦除周期消耗的时间和其他操作（如读取或编程）消耗的时间差不多的闪存时，所述存储部可不需要被擦除。

[0262] 正如此处所使用的，存储部在大小上与擦除块等同，但可（或可不）被擦除。当在此处使用擦除块时，擦除块可指存储元件（如 SSS 0.0 216a）内指定大小的特定区域，

并通常包括一定数量的页。当“擦除块”与闪存结合使用时，擦除块通常是在写入之前被擦除的存储部。当“擦除块”与“固态存储器”一起使用时，擦除块可（或可不）被擦除。正如此处所使用的，擦除块可包括一个擦除块或擦除块组，存储元件（如 SSS 0.0 到 SSS M.0 216a-n）的每一行都具有该擦除块组中的一个擦除块，擦除块或擦除块组在此处还可被称为虚拟擦除块。当擦除块指与所述虚拟擦除块关联的逻辑构建时，所述擦除块在此处可被称为逻辑擦除块（“LEB”）。

[0263] 通常，按照处理的顺序顺序地存储所述包。在一种实施方式中，当使用写入数据管道 106 时，顺序存储模块 802 按照包从写入数据管道 106 出来的顺序存储包。这种顺序可能是由于下述原因：来自请求设备 155 的数据段与读取自另一存储部的有效数据包（正如在下述的恢复操作期间从存储部恢复数据一样）混合。将恢复的、有效的数据包重路由到写入数据管道可包括如上文中相对于图 3 的固态存储控制器 104 描述的垃圾收集器旁路 316。

[0264] 装置 800 包括选择恢复的存储部的存储部选择模块 804。选择恢复的存储部可以使顺序存储模块 802 将所述存储部重新用于写入数据，因此将所述恢复的存储部添加到存储池中，或者所述存储部被重新用于在确定下述条件后从所述存储部中恢复有效数据：所述存储部失效、不可靠、应该被刷新、或其他将所述存储部暂时地或永久地移出所述存储池的理由。在另一种实施方式中，存储部选择模块 804 通过识别具有大量无效数据的存储部或擦除块来选择恢复的存储部。

[0265] 在另一种实施方式中，存储部选择模块 804 通过识别具有低额损耗的存储部或擦除块来选择恢复的存储部。例如，识别具有低额损耗的存储部或擦除块可包括识别无效数据少、擦除重复的次数少、位出错率低或程序计数低（缓冲器中一页数据写入所述存储部中的页的次数少；程序计数可从下列情况开始被测量：制造设备时、所述存储部最近一次被擦除时、其他任意事件发生时及这些情况的组合）的存储部。存储部选择模块 804 还可使用上述参数中的任意组合或其他参数以确定具有低额损耗的存储部。通过确定具有低额损耗的存储部来选择恢复的存储部可有助于发现未充分利用的存储部，还可由于损耗均衡而被恢复，等等。

[0266] 在另一种实施方式中，存储部选择模块 804 通过识别具有高额损耗的存储部或擦除块来选择恢复的存储部。例如，识别具有高额损耗的存储部或擦除块来选择恢复的存储部包括识别擦除重复次数多、位出错率高、具有不可恢复的 ECC 块或程序计数高的存储部。存储部选择模块 804 还可使用上述参数的任意组合或其他参数以确定具有高额损耗的存储部。通过确定具有高额损耗的存储部来选择恢复的存储部可有助于发现被过度使用的存储部，还可通过利用擦除周期刷新所述存储部而被恢复等等，或者使所述存储部像不能使用那样不提供服务。

[0267] 装置 800 包括数据恢复模块 806，该数据恢复模块 806 从选定为恢复的存储部中读取有效数据包、将所述有效数据包与其他将要由顺序存储模块 802 顺序地写入的数据包排队并升级具有由顺序存储模块 802 写入的有效数据的新物理地址的索引。通常，所述索引为对象索引，该对象索引将对象的数据对象标识符映射到形成包的位置的物理地址，所述数据对象存储在固态存储器 110 中。

[0268] 在一种实施方式中，装置 800 包括存储部恢复模块 808，该存储部恢复模块 808 为使用或再使用而准备所述存储部并将所述存储部标记为对顺序存储模块 802 可用，以在数

据恢复模块 806 完成从所述存储部中复制有效数据之后顺序地写入数据包。在另一种实施方式中，装置 800 包括存储部恢复模块 808，该存储部恢复模块 808 将选定为恢复的存储部标记为无法存储数据。通常，这是由于存储部选择模块 804 识别具有高额损耗的存储部或擦除块来选择恢复的存储部，从而使得所述存储部或擦除块没有条件被用于可靠的数据存储。

[0269] 在一种实施方式中，装置 800 位于固态存储设备 102 的固态存储设备控制器 202 内。在另一种实施方式中，装置 800 控制固态存储设备控制器 202。在另一种实施方式中，装置 800 的一部分位于固态存储设备控制器 202 内。在另一种实施方式中，由数据恢复模块 806 升级的对象索引也位于固态存储设备控制器 202 内。

[0270] 在一种实施方式中，所述存储部为擦除块，并且装置 800 包括擦除模块 810，该擦除模块 810 在数据恢复模块 806 完成从所述选定的擦除块中复制有效数据包之后并在存储部恢复模块 808 将所述擦除块标记为可用之前，擦除选定为恢复的擦除块。对于闪存和其他擦除操作消耗的时间比读取或写入操作消耗的时间长得多的固态存储器来说，在使数据块可以写入新数据之前擦除所述数据块有助于高效的操作。当固态存储器 110 布置在内存库 2_14 内时，擦除模块 810 的擦除操作可在一个内存库上执行，而另一个内存库可执行读取、写入或其他操作。

[0271] 在一种实施方式中，装置 800 包括垃圾标记模块 812，该垃圾标记模块 812 将存储部中的数据包识别为无效，以响应指示所述数据包不再有效的操作。例如，如果数据包被删除，垃圾标记模块 812 可将所述数据包识别为无效。读 - 修改 - 写操作是用于将数据包识别为无效的另一种方法。在一种实施方式中，垃圾标记模块 812 可通过升级索引将所述数据包识别为无效。在另一种实施方式中，垃圾标记模块 812 可通过存储另一数据包将所述数据包识别为无效，所述另一数据包指示无效的数据包已经被删除。这种方法是有利的，这是由于在固态存储器 110 中存储所述数据包已被删除的信息允许对象索引重建模块 272 或类似模块重建具有项的对象索引，所述项指示所述无效的数据包已经被删除。

[0272] 在一种实施方式中，装置 800 可被用于在清洗命令之后填充数据的虚拟页中的剩余部分，以提升整体的性能，其中，所述清洗命令使数据停止流入写入数据管道 106，直到写入数据管道 106 为空且所有的包已被永久地写入非易失性固态存储器 110。这具有以下好处：降低了需要的垃圾收集的量、减少了用于擦除存储部的时间并减少了编程虚拟页所需的时间。例如，可仅在准备将一个小包写入固态存储器 110 的虚拟页内时，接收清洗命令。编程这个几乎为空的页可能会引起下述结果：需要立即恢复浪费的空间；导致所述存储部内的有效数据被当作垃圾不必要的收集；及擦除、恢复所述存储空间并将所述存储空间返回到可用空间池以被顺序存储模块 802 写入。

[0273] 将所述数据包标记为无效而不是实际上擦除无效的数据包是有效率的，这是因为，如上所述，对于闪存和其他类似存储器来说，擦除操作消耗相当长的时间。允许垃圾收集系统（如装置 800 中所述的）在固态存储器 110 内自主地运行提供了一种将擦除操作与读取、写入或其他更快的操作分开的方法，从而使得固态存储设备 102 能比其他许多固态存储系统或数据存储设备运行得快得多。

[0274] 图 9 是示意性流程图，示出了根据本发明的用于存储恢复的方法 900 的一种实施方式。方法 900 始于步骤 902，顺序存储模块 802 将数据包顺序地写入存储部（步骤 904）。

所述存储部是固态存储设备 102 中的固态存储器 110 的一部分。通常，存储部为擦除块。所述数据包源于对象，而且所述数据包按处理的顺序被顺序地存储。

[0275] 存储部选择模块 804 选择恢复的存储部（步骤 906），并且数据恢复模块 806 从选定为恢复的存储部中读取有效数据包（步骤 908）。通常，有效数据包为未被标记为擦除、删除或其他一些无效数据标识符的数据包，所述数据包被视为有效或“好”的数据。数据恢复模块 806 将有效数据包与其他预定由顺序存储模块 802 顺序地写入的数据包排队（步骤 910）。数据恢复模块 806 升级具有由顺序存储模块 802 所写入的数据的新物理地址的索引（步骤 912）。所述索引包括从数据包的物理地址到对象标识符的映射。这些数据包存储在固态存储器 110 中，并且所述对象标识符对应于所述数据包。

[0276] 在数据恢复模块 806 完成从所述存储部复制有效数据后，存储部恢复模块将选定为恢复的存储部标记为对顺序存储模块 802 可用（步骤 914），以顺序地写入数据包，方法 900 结束于步骤 916。

[0277] 渐进式 RAID

[0278] 图 10 是示意性框图，示出了根据本发明的用于渐进式 RAID 和前端分布式 RAID 的系统 1600 的一种实施方式。系统 1600 包括 N 个存储设备 150 和 M 个奇偶校验 - 镜像存储设备 1602，一个或多个客户端可通过计算机网络 116 访问存储设备 150 和奇偶校验 - 镜像存储设备 1602。N 个存储设备 150 和奇偶校验 - 镜像存储设备 1602 可位于一个或多个服务器 112 内。存储设备 150、服务器 112、计算机网络 116 和客户端 114 大体上与上文描述的类似。奇偶校验 - 镜像存储设备 1602 通常与 N 个存储设备 150 类似或相同，并且通常被指定为用于条带的奇偶校验 - 镜像存储设备 1602。

[0279] 在一种实施方式中，N 个存储设备 150 和 M 个奇偶校验 - 镜像存储设备 1602 被包括在一个服务器 112 内或者可通过一个服务器 112 被访问，并可以通过系统总线、SAN 或其他计算机网络 116 联网在一起。在另一种实施方式中，N 个存储设备 150 和 M 个奇偶校验 - 镜像存储设备 1602 位于多台服务器 112a-n+m 内或者可通过多台服务器 112a-n+m 被访问。例如，存储设备 150 和奇偶校验 - 镜像存储设备 1602 可以是上文中相对于图 1C 的系统 103 和图 5B 的方法 501 描述的服务器内 SAN 的一部分。

[0280] 在一种实施方式中，奇偶校验 - 镜像存储设备 1602 存储存储在渐进式 RAID 中的条带的所有奇偶校验数据段。在另一种优选实施方式中，分配给渐进式 RAID 的存储设备集 1604 中的存储设备 150 被分配为用于特定条带的奇偶校验 - 镜像存储设备 1602，这种分配是轮换的，从而所述奇偶校验数据段在 N+M 个存储设备 150 之间为每个条带轮换。这种实施方式通过将单个存储设备 150 分配为用于每个条带的奇偶校验 - 镜像存储设备 1602 提供了性能上的优势。通过轮换奇偶校验 - 镜像存储设备 1602，与计算和存储奇偶校验数据段有关的开销可以是分散的。

[0281] 在一种实施方式中，存储设备 150 为固态存储设备 102，每个存储设备 150 都具有关联的固态存储器 110 和固态存储控制器 104。在另一种实施方式中，每个存储设备 150 包括固态存储控制器 104，并且关联的固态存储器 110 作为用于其他花费少、性能低的存储器（如磁带存储器或硬盘驱动器）的缓存。在另一种实施方式中，服务器 112 中的一个或多个包括将存储请求发送给渐进式 RAID 的一个或多个客户端 114。本领域技术人员会认识到可为渐进式 RAID 配置的具有 N 个存储设备 150 和一个或多个奇偶校验 - 镜像存储设备 1602

的其他系统配置。

[0282] 图 11 是示意性框图,示出了根据本发明的用于渐进式 RAID 的装置 1700 的一种实施方式。在不同的实施方式中,装置 1700 包括存储请求接收器模块 1702、条带化模块 1704、奇偶校验 - 镜像模块 1706、奇偶校验进展模块 1708、奇偶校验更替模块 1710、镜像集模块 1712、更新模块 1714、具有直接客户端响应模块 1718 的镜像恢复模块 1716、预整合恢复模块 1720、后整合恢复模块 1722、数据重建模块 1724 和奇偶校验重建模块 1726,这些模块将在下文中描述。模块 1702-1726 被描述位于服务器 112 内,但是模块 1702-1726 的一些功能或全部功能还可分布在多个服务器 112、存储控制器 152、存储设备 150、客户端 114 等设备之内。

[0283] 装置 1700 包括接收存储数据的请求的存储请求接收器模块 1702,其中,所述数据是文件的数据或对象的数据。在一种实施方式中,所述存储请求是对象请求。在另一种实施方式中,所述存储请求是块存储请求。在一种实施方式中,所述存储请求不包括数据,但包括命令,存储设备 150 和奇偶校验 - 镜像存储设备 1602 可使用该命令以从客户端或从其他源 DAM 或 RDMA 数据。在另一种实施方式中,所述存储请求包括由于所述存储请求而将要被存储的数据。在另一种实施方式中,所述存储请求包括一条能够将数据存储在所述存储设备集 1604 中的命令。在另一种实施方式中,所述存储请求包括多条命令。本领域技术人员会认识到适于渐进式 RAID 存储数据的其他存储请求。

[0284] 所述数据存储在装置 1700 可访问的位置。在一种实施方式中,所述数据在随机存取存储器 (“RAM”) 中可用,所述随机存取存储器如客户端 114 或服务器使用的 RAM。在另一种实施方式中,所述数据存储在硬盘驱动器、磁带存储器或其他大容量存储器中。在一种实施方式中,所述数据被配置为对象或文件。在另一种实施方式中,所述数据被配置为作为对象或文件的一部分的数据块。本领域技术人员会认识到作为所述存储请求的目标的所述数据的其他形式和位置。

[0285] 装置 1700 包括为数据计算条带模式的条带化模块 1704。所述条带模式包括一个或多个条带,其中,每个条带包括 N 个数据段的集。通过,条带中数据段的数量取决于分配给所述 RAID 群组的存储设备 150 的数量。例如,如果采用 RAID5,一个存储设备被指定为奇偶校验 - 镜像存储设备 1602a 以为特定的条带存储奇偶校验数据。如果四个存储设备 150a、150b、150c、150d 被分配给所述 RAID 群组,条带在除所述奇偶校验数据段外还会具有四个数据段。条带化模块 1704 将条带的 N 个数据段写入 N 个存储设备 150a-n,从而使得所述 N 个数据段中的每一个被写入分配给所述条带的存储设备 150 的集 1604 中的不同的存储设备 150a、150b、…… 150n。本领域技术人员会领会到可被分配给用于特定 RAID 级别的 RAID 群组的存储设备 150 的不同组合,并会领会到创建条带模式和将数据分割成每条带 N 个数据段的方法。

[0286] 装置 1700 包括奇偶校验 - 镜像模块 1706,该奇偶校验 - 镜像模块 1706 将所述条带的 N 个数据段的集写入存储设备集 1604 中的一个或多个奇偶校验 - 镜像存储设备 1602,其中,奇偶校验 - 镜像存储设备 1602 是除 N 个存储设备 150 之外的设备。然后,所述 N 个数据段可用于后续计算奇偶校验数据段。奇偶校验 - 镜像模块 1706 将 N 个数据段的集复制到奇偶校验 - 镜像存储设备 1602(这通常比存储所述 N 个数据段需要更少的时间),而不是立即计算所述奇偶校验数据段。一旦所述 N 个数据段存储在奇偶校验 - 镜像存储设备

1602,如果 N 个存储设备 150 中的一个不可用,所述 N 个数据段就可被读取或可被用于恢复数据。读取数据还具有 RAID 0 配置的优点,这是由于全部所述 N 个数据段一起从一个存储设备(如 1602a)中获取。对于不止一个的奇偶校验 - 镜像存储设备(如 1602a、1602b),奇偶校验 - 镜像模块 1706 将所述 N 个数据段复制到每个奇偶校验 - 镜像存储设备 1602a、1602b。

[0287] 装置 1700 包括奇偶校验进展模块 1708,该奇偶校验进展模块 1708 为所述条带计算一个或多个奇偶校验数据段,以响应存储整合操作。由所述 N 个数据段计算出来的所述一个或多个奇偶校验数据段存储在奇偶校验 - 镜像存储设备 1602 上。奇偶校验进展模块 1708 在一个或多个奇偶校验 - 镜像存储设备 1602 中的每一个上存储一个奇偶校验数据段。所述存储整合操作旨在在一个或多个奇偶校验 - 镜像存储设备 1602 中的至少一个上至少恢复存储空间和 / 或数据。例如,存储整合操作可以是上文中相对于图 8 和图 9 的装置 800 和方法 900 描述的在固态存储设备 102 上的数据垃圾收集。所述存储整合操作还可包括用于硬盘驱动器的碎片整理操作或其他整理数据以增加存储空间的类似操作。正如此处所使用的,所述存储整合操作还可包括恢复数据的操作(例如,如果存储设备 150 不可用,从错误中恢复数据,或由于其他读取数据的原因而从奇偶校验 - 镜像存储设备 1602 中恢复数据)。在另一种实施方式中,当奇偶校验 - 镜像存储设备 1602 不那么繁忙时,奇偶校验生成模块 1708 容易地计算所述奇偶校验数据段

[0288] 有利地是,通过延迟计算和存储条带的所述奇偶校验数据段,奇偶校验 - 镜像存储设备 1602 上的所述 N 个数据段可用于读取所述数据段、恢复数据、重建存储设备 150 上的数据,直到奇偶校验 - 镜像存储设备 1602 上需要更多的存储空间或其他需要存储整合操作的原因。然后,奇偶校验进展模块 1708 可独立于存储请求接收器模块 1702、条带化模块 1704 或奇偶校验 - 镜像模块 1706 像后台操作一样运行。本领域技术人员会轻易地认识到延迟计算奇偶校验数据段的其他理由,其中,延迟计算所述奇偶校验数据段作为渐进式 RAID 操作的一部分。

[0289] 在一种实施方式中,模块 1702-1708 的功能中(接收存储数据的请求、计算条带模式并将 N 个数据段写入 N 个存储设备、将 N 个数据段的集写入奇偶校验 - 镜像存储设备、计算奇偶校验数据段)的一些或全部在下述设备中实现:存储设备集 1604 的存储设备 150、客户端 114 和第三方 RAID 管理设备。所述第三方 RAID 管理设备可以是服务器 112 或其他计算机。

[0290] 在一种实施方式中,装置 1700 包括奇偶校验更替模块 1710,该奇偶校验更替模块 1710 为每个条带变更被分配为一个或多个用于所述条带的奇偶校验 - 镜像存储设备 1602 的存储设备集 1604 中的存储设备 150。如上文中相对于图 10 的系统 1600 所描述的,通过轮换用于奇偶校验 - 镜像存储设备(用于条带)的存储设备 150,不同奇偶校验数据段的计算工作分布在存储设备集 1604 的存储设备 150 之间。

[0291] 在另一种实施方式中,存储设备集 1604 是第一存储设备集,并且装置 1700 包括镜像集模块 1712,该镜像集模块 1712 创建除第一存储集 1604 之外的一个或多个附加存储设备集,从而使得所述一个或多个附加存储设备集中的每一个至少包括关联的条带化模块 1704,该条带化模块 1704 将所述 N 个数据段写入所述一个或多个附加存储集中的每一个集中的 N 个存储设备。在一种相关的实施方式中,所述一个或多个附加存储设备集中的每

一个集包括关联的用于存储所述 N 个数据段的集的奇偶校验 - 镜像模块 1706 和用于计算一个或多个奇偶校验数据段的奇偶校验进展模块 1708。在镜像集模块 1712 创建一个或多个镜像存储设备集的情况下, RAID 可以是嵌套的 RAID(如 RAID 50)。在这种实施方式中, RAID 级可从 RAID 10(其中, 数据被条带化和镜像化)渐进到 RAID50 或 RAID60(其中, 为每个存储设备集 1604 计算和存储奇偶校验数据段)。

[0292] 在一种实施方式中, 装置 1700 包括更新模块 1714。更新模块 1714 通常在奇偶校验 - 镜像存储设备 1602 上的 N 个数据段还未渐进为奇偶校验数据段的情况下使用。更新模块 1714 接收已更新的数据段, 其中, 所述已更新的数据段对应于 N 个存储设备 150 上存储的 N 个数据段中的现有数据段。更新模块 1714 将所述已更新的数据段复制到存储所述现有数据段的所述条带的存储设备 150, 并将所述已更新的数据段复制到所述条带的一个或多个奇偶校验 - 镜像存储设备 1602。更新模块 1714 用所述已更新的数据段替换存储在 N 个存储设备 150a-n 中的存储设备 150 上的所述现有数据段, 并用所述已更新的数据段替换存储在一个或多个奇偶校验 - 镜像存储设备 1602 上的相应的所述现有数据段。

[0293] 在一种实施方式中, 替换数据段包括将所述数据段写入存储设备 150 并然后将相应的数据段标记为对后续垃圾收集无效。这种实施方式的一个实例被描述用于固态存储器 110 和上文中相对于图 8 和图 9 描述的垃圾收集装置。在另一种实施方式中, 替换数据段包括用已更新的数据段覆盖现有数据段。

[0294] 在一种实施方式中, 存储设备集 1604 是第一存储设备集, 装置 1700 包括镜像恢复模块 1716, 当第一存储集 1604 中的存储设备 150 不可用时, 该镜像恢复模块 1716 恢复存储在第一存储集 1604 中的存储设备 150 上的数据段。所述数据段是从包含所述数据段的拷贝的镜像存储设备中恢复的。所述镜像存储设备包括一个或多个存储设备 150 的集的存储所述 N 个数据段的拷贝的一个存储设备。

[0295] 在另一种实施方式中, 镜像恢复模块 1716 为响应来自客户端 114 的读取所述数据段的读取请求而恢复所述数据段。在另一种相关的实施方式中, 镜像恢复模块 1716 还包括直接客户端响应模块 1718, 该直接客户端响应模块 1718 将请求的数据段从所述镜像存储设备发送给客户端 114。在这种实施方式中, 所述请求的数据段被复制到客户端 114, 从而客户端 114 不需要等到所述数据段被恢复就将所述数据段传送到客户端 114。

[0296] 在一种实施方式中, 装置 1700 包括预整合恢复模块 1720, 该预整合恢复模块 1720 为响应读取数据段的请求而恢复存储在存储集 1604 的存储设备 150 上的所述数据段。在这种实施方式中, 存储设备 150 不可用, 并且所述数据段是先于奇偶校验进展模块 1708 在一个或多个奇偶校验 - 镜像存储设备 1602 上生成所述一个或多个奇偶校验数据段, 从奇偶校验 - 镜像存储设备 1602 恢复的。

[0297] 在另一种实施方式中, 装置 1700 包括后整合恢复模块 1722, 该后整合恢复模块 1724 恢复存储在存储集的存储设备 150 上的数据段。在这种实施方式中, 存储设备 150 不可用, 并且所述数据段是在奇偶校验进展模块 1708 生成所述一个或多个奇偶校验数据段之后, 利用存储在一个或多个奇偶校验 - 镜像存储设备 150 上的一个或多个奇偶校验数据段恢复的。例如, 后整合恢复模块 1722 利用奇偶校验数据段和可用的 N 个存储设备 150 上的可用的数据段重新创建丢失的数据段。

[0298] 在一种实施方式中, 装置 1700 包括数据重建模块 1724, 该数据重建模块 1724 在重

建操作期间将恢复的数据段存储在替代存储设备上,其中,所述恢复的数据段与存储在不可用的存储设备 150 上的不可用数据段相匹配。不可用的存储设备 150 是存储设备集 1602 中的 N 个存储设备 150 中的一个。通常,所述重建操作发生在存储所述不可用数据段的存储设备 150 出现故障以后。所述重建操作是将数据段恢复到所述替代存储设备上,以匹配先前存储在不用可存储设备 150 上的数据段。

[0299] 可为所述重建操作而从数个来源恢复所述数据段。例如,如果匹配的数据段驻留在奇偶校验 - 镜像存储设备 1602 上,所述数据段可在渐进之前从奇偶校验 - 镜像存储设备 1602 恢复。在另一个实例中,所述数据段可从包含所述不可用数据段的拷贝的镜像存储设备中恢复。通常,如果所述恢复的数据段不驻留在一个或多个奇偶校验 - 镜像存储设备 1602 上,所述数据段是从所述镜像存储设备恢复的,但是,即使匹配的数据段在镜像存储设备上可用,所述数据也可从所述镜像存储设备恢复。

[0300] 在另一个实例中,如果所述恢复的数据段不驻留在奇偶校验 - 镜像存储设备 1602 或所述镜像存储设备中,由一个或多个奇偶校验数据段和所述 N 个数据段中的可用数据段再次生成再生数据段。通常,丢失的数据段仅在其不以某种形式存在于另一个存储设备 150 上时才会再生。

[0301] 在另一种实施方式中,装置 1700 包括奇偶校验重建模块 1726,该奇偶校验重建模块 1726 在奇偶校验重建操作中在替代存储设备上重建恢复的奇偶校验数据段,其中,所述恢复的奇偶校验数据段与存储在不可用的奇偶校验 - 镜像存储设备上的不可用奇偶校验数据段相匹配。所述不可用的奇偶校验 - 镜像存储设备是一个或多个奇偶校验 - 镜像存储设备 1602 中的一个。所述奇偶校验重建操作将奇偶校验数据段恢复到替代存储设备以匹配先前存储在不可用奇偶校验 - 镜像存储设备上的奇偶校验数据段。

[0302] 为了在所述重建操作中再生所述恢复的奇偶校验数据段,用于重建的数据可以来自不同的源。在一个实例中,利用存储在第二存储设备 150 集的奇偶校验 - 镜像存储设备 1602(存储所述条带的镜像拷贝)上的奇偶校验数据段恢复所述恢复的奇偶校验数据段。当镜像拷贝可用时,利用镜像奇偶校验数据段是可取的,这是由于不需要重新计算所述恢复的奇偶校验数据段。在另一个实例中,如果所述 N 个数据段在 N 个存储设备上可用,则由存储在 N 个存储设备 150 中的一个上的所述 N 个数据段再次生成所述恢复的奇偶校验数据段。通常,当单一故障发生在正在被重建的奇偶校验 - 镜像存储设备 1602 上时,所述 N 个数据段在 N 个存储设备 150 上可用。

[0303] 在另一个实例中,如果 N 个数据段中的一个或多个在第一存储设备集 1604 的 N 个存储设备 150 上不可用并且匹配的奇偶校验数据段在第二存储设备 150 集上不可用,则由第二存储设备 150 集的一个或多个存储设备 150(存储所述 N 个数据段的拷贝)再次生成所述恢复的奇偶校验数据段。在又一种实施方式中,由可用数据段和不匹配的奇偶校验数据段再次生成所述恢复的奇偶校验数据段,而不考虑这些数据段在一个或多个存储设备 150 集中的位置。

[0304] 在奇偶校验 - 镜像存储设备在存储设备集 1604 中的存储设备 150 之间更替的情况下,通常,数据重建模块 1724 和奇偶校验重建模块 1726 结合在一起工作以在重建的存储设备 150 上重建数据段和奇偶校验数据段。当第二奇偶校验 - 镜像存储设备 1602b 可用时,数据重建模块 1724 和奇偶校验重建模块 1726 能够在存储设备集 1604 的两个存储设备

150、1602 出现故障后重建两个存储设备。在奇偶校验 - 镜像存储设备 1602 还未渐进到创建奇偶校验 - 镜像数据段的情况下, 数据段或存储设备 150 的恢复速度比下列事件之后的数据段或存储设备 150 的恢复速度快 : 奇偶校验 - 镜像存储设备 1602 已经渐进、计算并存储了用于条带的奇偶校验数据段和用于计算所述奇偶校验数据段的奇偶校验 - 镜像存储设备 1602 上的 N 个数据段已经被删除。

[0305] 图 12 是示意性框图, 示出了根据本发明的利用渐进式 RAID 更新数据段的装置 1800 的一种实施方式。通常, 装置 1800 涉及 RAID 群组, 其中, 一个或多个奇偶校验 - 镜像存储设备 1602 已经渐进并且包括奇偶校验数据段 (不包括用以创建所述奇偶校验数据段的所述 N 个数据段)。装置 1800 包括更新接收器模块 1802、更新复制模块 1804、奇偶校验更新模块 1806, 这些模块在下文中描述。装置 1800 的模块 1802-1806 被描述位于服务器 112 内, 但也可位于存储设备 150、客户端内部或位于设备的任意组合的内部, 或者分布在数个设备之间。

[0306] 条带、数据段、存储设备 150、存储设备集 160、奇偶校验数据段、和一个或多个奇偶校验 - 镜像存储设备 1602 大体上类似于上文中相对于图 11 的装置 1700 描述的条带。装置 1800 包括更新接收器模块 1802, 该更新接收器模块 1802 接收已更新的数据段, 其中, 所述已更新的数据段对应于现有条带的现有数据段。在另一种实施方式中, 更新接收器模块 1802 还可接收多个更新信息并可一起或分别处理所述更新信息。

[0307] 装置 1800 包括更新复制模块 1804, 该更新复制模块 1804 将已更新的数据段复制到存储相应的现有数据段的存储设备 150, 并将所述已更新的数据段复制到一个或多个对应于所述现有条带的奇偶校验 - 镜像存储设备 1602。在另一种实施方式中, 更新复制模块 1804 将所述已更新的数据段复制到奇偶校验 - 镜像存储设备 1602 或存储所述现有数据段的存储设备 150, 并然后验证所述已更新的数据段的拷贝被转发给其他设备 1602、150。

[0308] 装置 1800 包括奇偶校验更新模块 1806, 该奇偶校验更新模块 1806 为响应存储整合操作而为所述现有条带的一个或多个奇偶校验 - 镜像存储设备计算一个或多个已更新的奇偶校验数据段。所述存储整合操作类似于上文中相对于图 11 的装置 1700 描述的存储整合操作。所述存储整合操作旨在利用一个或多个已更新的奇偶校验数据段在一个或多个奇偶校验 - 镜像存储设备 1602 上至少恢复存储空间和 / 或数据。通过等待更新一个或多个奇偶校验数据段, 更新可以被推迟到更合适的时候或等到需要整合存储空间的时候。

[0309] 在一种实施方式中, 由所述现有奇偶校验数据段、所述更新的数据段和所述现有数据段计算所述已更新的奇偶校验数据段。在一种实施方式中, 所述现有数据段在为生成所述更新的奇偶校验数据段而读取所述现有数据段之前被保持在一个位置。这种实施方式的一个好处是 : 可将与复制所述现有数据段到奇偶校验 - 镜像存储设备 1602 或其他生成所述更新的奇偶校验数据段的位置有关的开销推迟到必要的时候。这种实施方式的一个不足是 : 如果保持所述现有数据段的存储设备 150 出现故障, 在生成所述已更新的奇偶校验数据段之前必须恢复所述现有数据段。

[0310] 在另一种实施方式中, 当 N 个存储设备 150a-n 中的存储所述现有数据段的存储设备 150 接收所述更新的数据段的拷贝时, 所述现有数据段被复制到奇偶校验 - 镜像存储设备 1602。然后, 存储所述现有数据段, 直到所述存储整合操作。在另一种实施方式中, 如果所述存储整合操作发生在触发计算所述已更新的奇偶校验数据段的存储整合操作之前, 则

所述现有数据段被复制到奇偶校验 - 镜像存储设备 1602，以响应 N 个存储设备 150a-n 中的存储所述现有数据段的存储设备 150 上的存储整合操作。后一种实施方式是有利的，这是因为直到存储所述现有数据段的存储设备 150 上的或奇偶校验 - 镜像存储设备 1602 上的存储整合操作需要才复制所述现有数据段。

[0311] 在一种实施方式中，由所述现有奇偶校验数据段、所述已更新的数据段和增量数据段计算出所述更新的奇偶校验数据段，其中，所述增量数据段按所述更新的数据段和所述现有数据段之间的差异生成。通常，生成增量数据段是更新所述奇偶校验数据段中的部分解决方案或中间步骤。生成增量数据段是有利的，这是因为所述增量数据段可以被高度压缩并可以在传送之前被压缩。

[0312] 在一种实施方式中，在为了生成所述已更新的奇偶校验数据段而读取所述增量数据段之前，所述增量数据段存储在存储所述现有数据段的存储设备上。在另一种实施方式中，当存储所述现有数据段的存储设备 150 接收所述已更新的数据段的拷贝时，所述增量数据段被复制到奇偶校验 - 镜像存储设备 1602。在另一种实施方式中，所述增量数据段被复制到奇偶校验 - 镜像存储设备 1602，以响应存储所述现有数据段的存储设备 150 上的存储整合操作。正如复制所述现有数据段一样，后一种实施方式是有利的，这是因为直到所述现有数据段上的存储整合操作之前或触发计算所述已更新的奇偶校验数据段的另一个存储整合操作之前，才移动增量数据文件。

[0313] 在不同的实施方式中，模块 1802、1804、1806 的操作的一部分或全部（即接收已更新的数据段、复制所述已更新的数据段和计算所述已更新的奇偶校验数据段）发生在下述设备上：存储设备集 1604 的存储设备 150、客户端 114 或第三方 RAID 管理设备。在另一种实施方式中，独立于所述更新接收器模块 1802 的操作和更新复制模块 1804 的操作进行所述存储整合操作。

[0314] 图 13 是示意性流程图，示出了根据本发明的利用渐进式 RAID 管理数据的方法 1900 的一种实施方式。方法 1900 始于步骤 1902，存储请求接收器模块 1702 接收存储数据的请求（步骤 1904），其中，所述数据是文件的数据或对象的数据。条带化模块 1704 为所述数据计算条带模式并将所述 N 个数据段写入 N 个存储设备 150（步骤 1906）。所述条带模式包括一个或多个条带。每个条带包括 N 个数据段的集，其中，所述 N 个数据段中的每一个被写入分配给所述条带的存储设备集 1604 中不同的存储设备 150。

[0315] 奇偶校验 - 镜像模块 1706 将所述条带的 N 个数据段的集写入存储设备集 1604 中的一个或多个奇偶校验 - 镜像存储设备 1602（步骤 1908）。一个或多个奇偶校验 - 镜像存储设备是除 N 个存储设备 150 之外的设备。奇偶校验进展模块 1708 确定是否有等待中的存储整合操作（步骤 1910）。如果奇偶校验进展模块 1708 确定没有等待中的存储整合操作，方法 1900 返回并再次确定是否有等待中的存储整合操作。在其他实施方式中，存储请求接收器模块 1702、条带化模块 1704 和奇偶校验 - 镜像模块 1706 继续接收存储请求、计算条带模式和存储数据段。

[0316] 如果奇偶校验进展模块 1708 确定没有等待中的存储整合操作（步骤 1910），奇偶校验进展模块 1708 为所述条带计算奇偶校验数据段（步骤 1912）。由存储在奇偶校验 - 镜像存储设备 1602 上的 N 个数据段计算所述奇偶校验数据段。奇偶校验进展模块 1708 将所述奇偶校验数据段存储在奇偶校验 - 镜像存储设备 1602 上（步骤 1912），方法 1900 结束于

步骤 1914。所述存储整合操作的执行独立于接收存储 N 个数据段的请求（步骤 1904）、将 N 个数据段写入 N 个存储设备（步骤 1906）或将 N 个数据段写入一个或多个奇偶校验 - 镜像存储设备（步骤 1908）。所述存储整合操作旨在至少恢复奇偶校验 - 镜像存储设备 1602 上的存储空间或数据。

[0317] 图 14 是示意性流程图，示出了根据本发明的利用渐进式 RAID 更新数据段的方法 2000 的一种实施方式。方法 2000 始于步骤 2002，更新接收器模块 1802 接收已更新的数据段（步骤 2004），其中，所述已更新的数据段对应于现有条带的现有数据段。更新复制模块 1804 将所述已更新的数据段复制到存储相应的现有数据段的存储设备 150 和对应于所述现有条带的一个或多个奇偶校验 - 镜像存储设备 1602（步骤 2006）中。

[0318] 奇偶校验更新模块 1806 确定是否存储整合操作在等待中（步骤 2008）。如果奇偶校验更新模块 1806 确定没有等待中的存储整合操作（步骤 2008），奇偶校验更新模块 1806 等待存储整合操作。在一种实施方式中，方法 2000 返回并接收其他已更新的数据段（步骤 2004），并复制所述已更新的数据段（步骤 2006）。如果奇偶校验更新模块 1806 确定有等待中的存储整合操作（步骤 2008），奇偶校验更新模块 1806 为所述现有条带的一个或多个奇偶校验 - 镜像存储设备计算一个或多个已更新的奇偶校验数据段（步骤 2010），方法 2000 结束于步骤 2012。

[0319] 前端分布式 RAID

[0320] 传统的 RAID 系统被配置为与 RAID 控制器一起使用，所述 RAID 控制器具有如下功能：接收数据、为所述数据计算条带模式、将所述数据分割为数据段，计算奇偶校验条带、将所述数据存储在存储设备上、更新所述数据段等等。当一些 RAID 控制器允许一些功能成为分布式的功能时，由 RAID 控制器管理的存储设备不直接与客户端通信以存储在 RAID 中条带化的数据。用于 RAID 过程的替代存储请求和数据通过所述 RAID 存储控制器。

[0321] 要求所述 RAID 控制器接触所有将要被存储在 RAID 中的数据是没有效率的，这是因为这种方法产生了数据流瓶颈。这个问题在读 - 修改 - 写处理期间尤为突出，其中，RAID 群组中的全部驱动器的带宽和性能被消耗，而实际上仅更新了子集。此外，被指定用于由所述 RAID 控制器管理的数据的存储设备中的区域通常用于 RAID 群组并且不能独立地被访问。通过客户端访问存储设备 150 通常通过分区存储设备 150 来实现。当使用分区时，支持普通存储访问的分区不用于 RAID，而分配给 RAID 群组的分区不支持普通数据存储访问。为全域性地优化效用而超额预定分区的方案不仅复杂而且更加难以管理。此外，分配给一个 RAID 群组的存储空间不能通过多于一个的 RAID 控制器访问，除非一个被指定为主控制器，而其他 RAID 控制器作为从机，除非主 RAID 控制器未被激活、丧失功能等等。

[0322] 典型的 RAID 控制器还在 RAID 群组的存储设备 150 之外生成奇偶校验数据段。这可能是无效率的，这是因为奇偶校验数据段通常在生成之后被发送给存储设备 150 以便于存储，这需要 RAID 控制器的计算能力。追踪奇偶校验数据段的位置和更新信息还必须在 RAID 控制器内完成而不是在自主地在存储设备 150 上完成。

[0323] 如果独立的 RAID 控制器断开连接，当有必要确保所述数据保持在可用状态时，RAID 控制器通常互相交叉连接并交叉连接至驱动器，和 / 或像成套设备一样镜像化，但这样使数据可用性管理的花费昂贵且难以管理，还显著地降低了存储子系统的可靠性。

[0324] 需要一种用于前端分布式 RAID 系统、装置和方法，所述前端分布式 RAID 允许在每

个数据段、每个对象、每个文件或类似基础上使用 RAID，所述前端分布式 RAID 无需 RAID 控制器和位于客户端和存储设备之间的 RAID 控制器对。在这种系统、装置和方法中，RAID 群组可被创建用于一个数据段、对象或文件，该 RAID 群组还可在一个存储设备群组中由一个 RAID 控制器管理，而第二 RAID 控制器可被创建用于包含第一 RAID 群组的一些相同的存储设备的另一个数据段、对象或文件。RAID 控制功能可分布在客户端 114、第三方 RAID 管理设备之间，或分布在多个存储设备 150 之间，前端分布式 RAID 系统、装置和方法还可将命令发送给 RAID 群组的存储设备 150 并可允许存储设备 150 通过直接存储器存取（“DMA”）或远程 DMA（“RDMA”）直接访问和复制数据。

[0325] 图 16 是示意性框图，示出了根据本发明的可被前端分布式 RAID 访问的系统 1600 的一种实施方式。上文中对图 16 中相对于渐进式 RAID 描述的组件进行的说明也可应用到前端分布式 RAID。对于前端分布式 RAID，存储设备集 1604 形成 RAID 群组并包括自主运行且能够独立地通过网络 116 或一个或多个冗余网络 116 接收和服务来自客户端 114 的存储请求的存储设备 150。

[0326] 在存储设备集 1604 中的存储设备 150 之中，一个或多个存储设备 150 被指定为用于条带的奇偶校验 - 镜像存储设备 1602。通常，一个或多个奇偶校验 - 镜像存储设备 1602 的功能大体上类似于其他存储设备 150。在典型的配置中，指定的奇偶校验 - 镜像存储设备 1602 在存储设备集 1604 的存储设备 150 之间变更，奇偶校验 - 镜像存储设备 1602 实质上具有与其他存储设备 150 一样的特点，这是由于奇偶校验 - 镜像存储设备 1602 也必须像非奇偶校验 - 镜像存储设备一样运行。类似的特点是关于上述的 RAID 群组内的操作和用于客户端 114 独立通信的自主操作。在不同的实施方式中，存储设备集 1604 的存储设备 150 可在其他方面（不涉及所述 RAID 环境下的功能）不同。

[0327] 存储设备集 1604 的存储设备 150 可以是独立的、可以是在一个或多个服务内成组的、可每一个驻留在一个服务器 112 内、可通过一个或多个服务器 112 被访问，等等。一个或多个客户端 114 可驻留在包括一个或多个存储设备 150 的服务器 112 内、可驻留在独立的服务器 112 内、可驻留在通过一个或多个计算网络 116 访问存储设备 150 的计算机、工作站、笔记本电脑等设备内，或位于类似设备内。

[0328] 在一种实施方式中，网络 116 包括系统总线，并且存储设备集 1604 中的一个或多个存储设备 150、1602 通过所述系统总线通信。例如，系统总线可以是 PCI-e 总线、串行高级技术附件（“串行 ATA”）总线、并行 ATA 或类似总线。在另一种实施方式中，所述系统总线是外部总线，如小型计算机系统接口（“SCSI”）、火线、光纤通道、USB、PCIe-AS、无限带宽或类似总线。本领域技术人员会意识到具有存储设备 150 的其他系统 1600 配置，其中，存储设备 150 不仅自主运行，还能够独立地通过一个或多个网络 116 接收和服务来自客户端 114 存储请求。

[0329] 图 15 是示意性框图，示出了根据本发明的用于前端分布式 RAID 的装置 2100 的一种实施方式。在不同的实施方式中，装置 2100 包括存储请求接收器模块 2102、条带化关联模块 2104、奇偶校验 - 镜像关联模块 2106、存储请求发送器模块 2108、前端奇偶校验生成模块 2110、奇偶校验更替模块 2118、数据段恢复模块 2112、数据重建模块 2114、奇偶校验更替模块 2116 和对等通信模块 2120，这些模块在下文中描述。在不同的实施方式中，装置 2100 可被包括在下列设备中：存储设备 150（如固态存储设备 102）、存储设备控制器 152（如固

态存储控制器 104)、服务器 112、第三方 RAID 管理设备等等，或者装置 2100 分布在不止一个的组件之间。

[0330] 装置 2100 包括存储请求接收器模块 2102，该存储请求接收器模块 2102 接收将数据存储在存储设备集 1604 中的存储请求。所述数据可以是文件或对象的一部分，或者可以是整个文件或对象。文件可包括任意信息块或用于存储信息的源，其中，这些块或源可用于计算机程序。文件可包括由处理器访问的任意数据结构。文件可包括数据库、文本串、计算机编码等等。对象通常是用于面向对象的编程的数据结构并且可包括具有(或不具有)数据的结构。在一种实施方式中，对象是文件的子集。在另一种实施方式中，对象独立于文件。在任何情况下，对象和文件在此处被定义为包括数据、数据结构、计算机编码和其他存储在存储设备上的信息的全部集。

[0331] 存储设备集 1604 包括形成 RAID 群组的自主存储设备 150，存储设备 150 自主地通过一个或多个网络 116 接收来自客户端 114 的存储请求。存储设备集 1604 的自主存储设备 150 中一个或多个被指定为用于条带的奇偶校验 - 镜像存储设备 1602。来自其他客户端的其他存储请求可存储在第二存储设备集上，其中，所述第二存储设备集可像第一存储设备集 1604 一样包括一个或多个相同的存储设备 150(和奇偶校验 - 镜像存储设备 1602)。为两个存储设备集 1604 所共用的存储设备 150 可在其内具有分配为存储空间的重叠部分。

[0332] 装置 2100 包括为所述数据计算条带模式的条带化关联模块 2104。所述条带模式包括一个或多个条带。每个条带包括 N 个数据段的集。条带的 N 个数据段还可包括一个或多个空数据段。条带化关联模块 2104 将 N 个数据段中的一个数据段与被分配给所述条带的存储设备集 1604 中的 N 个存储设备 150a-n 中的一个关联。在一种实施方式中，条带化关联模块 2104 利用将要被发送给存储设备 150 的存储请求将数据段与存储设备 150 关联，该存储请求指令存储设备获取对应于来自发送所述存储请求的客户端 114 的数据段的数据。

[0333] 在另一种实施方式中，所述存储请求大体上与所述数据段的数据无关。大体上与数据无关意味着所述存储请求一般来说不包括作为所述存储请求的主题的数据，但可包括可能是数据的一部分的字符、字符串等等。例如，如果所述数据包括一串重复的、相同的字符(如一串 0 字符)，所述存储请求可包括所述数据包括一串 0 字符的指示而并不包括所述数据中的所有零字符。本领域技术人员会认识到发送存储请求而不发送数据的主体但同时仍然允许少量的或单个实例的某些字符或字符串存在于所述存储请求中的其他方法。所述存储请求包括命令，该命令允许 N 个存储设备 150a-n 利用 DMA 或 RDMA 操作或类似操作检索所述数据。

[0334] 在另一种实施方式中，条带化关联模块 2104 通过在将要被发送给存储设备 150 的存储请求中识别数据段的数据将所述数据段与存储设备 150 关联。识别所述数据段的数据可包括数据段标识符、数据段位置或地址、数据段长度或其他允许存储设备 150 判定哪个数据包括所述数据段的信息。

[0335] 在一种实施方式中，条带化关联模块 2104 在存储请求中将数据段与存储设备 150 关联，以使得客户端 114 能够在广播中发送包括所述数据段的数据，从而每个存储设备 150 能够存储关联的数据段并丢弃对应于未被分配给存储设备 150 的数据段的数据。在另一种实施方式中，条带化关联模块 2104 在存储请求中可能是通过为每个数据段分配地址将数据段与存储设备 150 关联，以使得客户端 114 可在组播中发送包括所述数据段的数据，从而

每个存储设备 150 能够存储关联的数据段并丢弃对应于未被分配给存储设备 150 的数据段的数据。本领域技术人员会认识到用于条带化关联模块 2104 将数据段与存储设备 150 关联,从而将一个或多个数据段通过下述方式传给一个或多个存储设备的其他方法:广播、组播、单播、任意播等。

[0336] 在一种相关的实施方式中,条带化关联模块 2104 在存储请求中将数据段与存储设备 150 关联,以使得客户端 114 能够广播、组播、单播(等)所述存储请求,并且每个存储设备 150 能够接收来自客户端 114 的涉及与存储设备 150 关联的所述数据段的存储请求的一部分,还能够丢弃不涉及与存储设备 150 关联的一个或多个数据段的存储请求的那部分。

[0337] 在另一种实施方式中,由存储请求接收器模块 2102 接收的所述存储请求包括作为所述存储请求主题的数据,并且条带化关联模块 2104 通过准备用于包括数据段的存储设备 150 的存储请求将数据段与存储设备 150 关联。条带化关联模块 2104 可运行在下列设备内:客户端 114、第三方 RAID 管理设备、存储设备 150、1602,等等。

[0338] 装置 2100 包括奇偶校验 - 镜像关联模块 2106,该奇偶校验 - 镜像关联模块 2106 将 N 个数据段的集与存储设备集 1604 中的一个或多个奇偶校验 - 镜像存储设备 1602 关联。一个或多个奇偶校验 - 镜像存储设备 1602 是除 N 个存储设备 150a-n 之外的设备。在一种实施方式中,奇偶校验 - 镜像关联模块 2106 将 N 个数据段的集与每个奇偶校验 - 镜像存储设备 1602 关联,从而每个奇偶校验 - 镜像存储设备 1602 能够为了生成奇偶校验数据段而接收并存储条带的 N 个数据段。在另一种实施方式中,奇偶校验 - 镜像关联模块 2106 将条带的数据段与每个奇偶校验 - 镜像存储设备 1602 关联,从而奇偶校验 - 镜像存储设备 1602a-m 充当存储在 N 个存储设备 150a-n 中的 N 个数据段的镜像。

[0339] 在不同的实施方式中,奇偶校验 - 镜像关联模块 2106 利用单个存储请求、多个存储请求或上文中相对于条带化关联模块 2104 描述的其他关联技术(如为了 DMA、RDMA、广播、组播而设立奇偶校验 - 镜像存储设备 1602 的存储请求,或将 N 个数据段包括在存储请求中)将 N 个数据段的集与一个或多个奇偶校验 - 镜像存储设备 1602 关联。奇偶校验 - 镜像关联模块 2106 可运行在下列设备内:客户端 114、第三方 RAID 管理设备、存储设备 150、1602,等等。

[0340] 装置 2100 包括存储请求发送器模块 2108,该存储请求发送器模块 2108 将一个或多个存储请求发送给存储设备集 1604 中的每个存储设备,每个存储请求能够将与接收所述存储请求的存储设备 150、1602 关联的一个或多个数据段存储在存储设备 150、1602 上。在一种实施方式中,每个存储请求不包括作为所述存储请求的主题的数据。在另一种实施方式中,每个存储请求使得存储设备集 1604 的 N 个存储设备 150 和奇偶校验 - 镜像存储设备 1602 能够利用 DMA 或 RDMA 下载关联的数据段的数据。在另一种实施方式中,存储请求包含足够的信息以从来自客户端 114 的广播中挑选用于关联的数据段的相关存储请求或相关数据。在另一种实施方式中,存储请求包括关联的数据段的数据。

[0341] 在一种实施方式中,每个存储请求识别作为条带的存储设备集 1604 的一部分的存储设备 150、1602。通过包括识别存储设备集 1604 的存储设备 150、1602 的步骤,如果充当主机的存储设备 150 出现故障,另一个存储设备 150 可接管主机以管理 RAID 数据。在另一种实施方式中,当存储设备断开连接时,识别存储设备集 1604 使得自主存储设备 150、

1602 能够恢复数据，并当替代存储设备被附加到存储设备集 1604 内时，使得自主存储设备 150、1602 能够独立于客户端重建数据。在另一种实施方式中，识别存储设备集 1604 的存储设备 150、1602 代表了用于传送数据段或存储请求的组播组。识别信息可与存储在存储设备集 1604 的存储设备 150、1602 上的、用于对象或文件的元数据一起存储。

[0342] 在一种实施方式中，装置 2100 包括前端奇偶校验生成模块 2110，当奇偶校验 - 镜像关联模块 2106 将 N 个数据段的集与一个或多个奇偶校验 - 镜像存储设备 1602 中的每一个关联时，该前端奇偶校验生成模块 2110 独立于客户端 114 为所述条带计算奇偶校验数据段，并将所述奇偶校验数据段存储在奇偶校验 - 镜像存储设备 1602 上。由提供给奇偶校验 - 镜像存储设备 1602 的 N 个数据数据段的集计算所述奇偶校验数据段。当存储设备集 1604 包括了不止一个奇偶校验 - 镜像存储设备 1602 时，前端奇偶校验生成模块 2110 通常生成不同的奇偶校验数据段，从而存储设备集 1604 中的两个或更多个存储设备 150、1602 可出现故障，并且奇偶校验数据段信息允许恢复不可用的数据段或奇偶校验数据段。

[0343] 在另一种实施方式中，当运行在存储设备集 1604 的存储设备 150 中和 / 或第三方 RAID 管理设备中时，前端奇偶校验生成模块 2110 计算所述奇偶校验数据段。例如，独立于客户端 114 的、发送所述存储请求的服务器 112 可计算所述奇偶校验数据段。在另一种实施方式中，前端奇偶校验生成模块 2110 运行在奇偶校验 - 镜像存储设备内以计算所述奇偶校验数据段。例如，奇偶校验 - 镜像存储设备 1602 中的存储控制器 152 可充当用于由存储设备集 1604 形成的 RAID 群组的主存储控制器。

[0344] 在另一种实施方式中，前端奇偶校验生成模块 2110 计算所述奇偶校验数据段并将计算出的奇偶校验数据段发送给开成镜像的第二存储设备集中的一个或多个附加奇偶校验 - 镜像存储设备 1602。这种实施方式是有利的，这是因为与计算奇偶校验数据段有关的开销只需要一次，而不需要为每个存储设备集 1604 执行开销，这样做的额外好处是减少了网络 116 的数据流量。

[0345] 在一种实施方式中，装置 2100 还可包括数据段恢复模块 2112，如果存储设备 150 不可用并且接收到读取不可用数据段或包括不可用数据段的数据的请求，数据段恢复模块 2112 恢复存储在存储设备集 1604 的存储设备 150 上的数据段。利用存储设备集 1604 的可用存储设备 150 上的数据段、奇偶校验数据段和存储设备集 1604 的可用存储设备 150、1602 上的数据段的结合恢复所述数据段，或者从包括所述数据段的拷贝的镜像存储设备中恢复所述数据段。通常，镜像存储设备是存储 N 个数据段的拷贝的存储设备集的一个存储设备 150。数据段恢复模块 2112 可运行并恢复来自下述设备中的不可用数据段：存储设备 150、奇偶校验 - 镜像存储设备 1602、第三方 RAID 管理设备、镜像存储设备等等。

[0346] 在另一种实施方式中，装置 2100 包括数据重建模块 2114，该数据重建模块 2114 在重建操作中将恢复的数据段存储在替代存储设备 150 上。例如，如果存储设备 150 由于出现故障、失去同步性等原因而变得不可用，数据重建模块 2114 可重建存储设备 150 以替换不可用的存储设备 150。在一种实施方式中，重建的存储设备 150 是已经可用的源存储设备 150。

[0347] 所述恢复的数据段与存储在存储设备集 1604 的不可用存储设备 150 上的不可用数据段匹配。所述重建操作通常将一个或多个数据段和奇偶校验数据段恢复到替代存储设备 150 上，从而使其与先前存储在不可用存储设备 150 上的数据段和奇偶校验数据段相匹

配。

[0348] 在一种实施方式中,所述恢复的数据段利用存储设备集 1604 的可用存储设备 150 上的可用数据段被恢复用于重建操作。在另一种实施方式中,所述恢复的数据段利用来自一个或多个奇偶校验 - 镜像存储设备 1602 的奇偶校验数据段和存储设备集 1604 的可用存储设备 150 上的可用数据段的结合被恢复用于重建操作。在另一种实施方式中,所述恢复的数据段利用读取自奇偶校验 - 镜像存储设备 1602 的匹配数据段被恢复用于重建操作。在又一种实施方式中,所述恢复的数据段利用来自镜像存储设备的匹配数据段被恢复用于重建操作。数据重建模块 2114 能够运行并存储接收自下述设备的数据段 : 客户端 114、第三方 RAID 管理设备、存储设备 150、1602、镜像存储设备等等。

[0349] 在另一种实施方式中,装置 2100 包括奇偶校验重建模块 2116,该奇偶校验重建模块 2116 在重建操作中在替代存储设备 1602 上重建恢复的奇偶校验数据段。重建操作大体上与上文中相对于数据重建模块 2114 描述的重建操作类似。奇偶校验重建模块 2116 类似于数据重建模块 2114 运行,除了奇偶校验重建模块 2116 重建奇偶校验数据段。恢复的奇偶校验数据段与存储在分配给条带的不可用奇偶校验 - 镜像存储设备 1602 上的不可用奇偶校验数据段相匹配。

[0350] 在不同的实施方式中,通过下述方法恢复所述奇偶校验数据段 : 复制存储在镜像存储设备集的奇偶校验 - 镜像存储设备 1602 上的所述奇偶校验数据段、从存储设备集 1604 的奇偶校验 - 镜像存储设备 1602 复制所述奇偶校验数据段 (如果与不可用奇偶校验数据段一致)、利用存储在存储设备集 1604 的可用存储设备 150、1602 和包含数据段的拷贝的镜像存储设备上的 N 个数据段中的一个或多个和奇偶校验数据段生成所述奇偶校验数据段,等等。数据重建模块 2114 可驻留在下列设备上并运行和存储恢复的数据段 : 客户端 114、第三方 RAID 管理设备、存储设备 150、镜像存储设备等等。

[0351] 有利地是,装置 2100 并不限于在存储设备 150、1602 中将数据存储至用于此处描述的前端分布式 RAID 操作的分区。作为替代的是,自主存储设备 (如 150a) 可独立地接收来自客户端 114 的将经 RAID 或未经 RAID 的数据存储在存储设备 150a 的一个或多个区域中,存储设备 150a 依然可被条带化关联模块 2104、奇偶校验 - 镜像关联模块 2106 和前端奇偶校验生成模块 2110 用于存储数据。

[0352] 在一种实施方式中,由存储请求接收器模块 2102 接收的或由存储请求发送器模块 2108 发送的一个或多个存储请求识别包括条带的存储设备集 1604 的存储设备 150。有利地是,如果主控制器丧失功能,在存储请求中识别存储设备集 1604 的存储设备 150 有助于备份 RAID 控制器运行。例如,如果存储设备集 1604 的存储设备 150 在存储请求中被识别并且位于奇偶校验 - 镜像存储设备 1602 中的所述主控制器不可用,另一个奇偶校验 - 镜像存储设备 1602 或 N 个存储设备 150a-n 中的另一个可以成为所述主控制器。

[0353] 在一种实施方式中,装置 2100 包括奇偶校验更替模块 2118,该奇偶校验更替模块 2118 为每个条带变更被分配为用于所述条带的奇偶校验 - 镜像存储设备 1602 的存储设备集 1604 中的存储设备 150。奇偶校验更替模块 2118 的优点已在上文中描述。在另一种实施方式中,存储设备集 1604 的存储设备 150 形成对等群组,并且装置 2100 包括对等通信模块 2120,该对等通信模块 2120 在存储设备集 1604 的存储设备 150、1602 内发送并接收存储请求。对等通信模块 2120 还可在存储设备集 1604 外部的对等设备中发送并接收存储请

求。

[0354] 在一种优选实施方式中,所述存储请求是通过利用装置 2100 的模块 2102-2120 在存储设备 1604 的存储设备 150、1602 之间条带化对象的数据来存储对象的对象请求。在另一种实施方式中,存储设备集 1604 的自主存储设备 150、1602 中的一个或多个被分配到第一 RAID 群组中用于第一对象或文件的至少一部分,并被分配到第二 RAID 群组中用于第二对象或文件的至少一部分。例如,一个存储设备 150a 可以是用于一个或多个条带的存储设备集 1604 的主 RAID 控制器,而第二存储设备 150b 可以是用于包括了存储设备集 1604 的一些或全部存储设备 150 的 RAID 群组的主 RAID 控制器。有利地是,装置 2100 允许灵活的分组存储设备 150、1602 以形成不同客户端 114 的 RAID 群组。

[0355] 图 16 是示意性流程图,示出了根据本发明的用于前端分布式 RAID 的方法 2200 的一种实施方式。方法 2200 始于步骤 2202,存储请求接收器模块 2102 接收将数据存储在存储设备集 1604 的存储设备 150a-n 中的存储请求(步骤 2204)。条带化关联模块 2104 计算用于所述数据的条带模式(步骤 2206)并将 N 个数据段中的每一个数据段与 N 个存储设备 150a-n 中的一个关联(步骤 2208)。

[0356] 奇偶校验 - 镜像关联模块 2106 将 N 个数据段的集与一个或多个偶校验 - 镜像存储设备 1602 关联(步骤 2210)。存储请求发送器模块 2108 将一个或多个存储请求发送给存储设备集 1604 中的每一个存储设备 150、1602(步骤 2212)。每个存储请求足以将与接收所述存储请求的存储设备 150 关联的一个或多个数据段存储在存储设备 150 上。然后,所述数据的数据段像被所述存储请求指令一样利用 DMA、RDMA、广播、组播等技术被传送给存储设备集 1604 的存储设备 150、1602。可选择地,前端奇偶校验生成模块 2110 为条带计算奇偶校验数据段(步骤 2214),方法 2200 结束于步骤 2216。

[0357] 共享的、前端、分布式 RAID

[0358] 图 10 是示意性框图,示出了根据本发明的充当用于共享的、前端分布式 RAID(除了渐进式 RAID 和前端分布式 RAID 之外)的系统 1600 的一种实施方式。上文中对图 16 中相对于渐进式 RAID 和前端分布式 RAID 描述的组件的说明同样适用于共享的前端分布式 RAID。正如前端分布式 RAID 一样,存储设备集 1604 形成 RAID 群组并包括存储设备 150,存储设备自主运行并能够独立地通过网络 116 接收并服务来自客户端 114 的存储请求。

[0359] 对于共享的、前端分布式 RAID,系统 1600 包括两个或更多个客户端 114,从而两个或更多个客户端 114 中的每一个发送涉及相同数据的存储请求。所述存储请求是并发的,这是由于所述存储请求的到达使得在另一个存储请求到达之前,一个存储请求还没有完成。存储设备集 1604 的存储设备 150 之中的一个或多个被指定为用于条带的奇偶校验 - 镜像存储设备 1602。通常,一个或多个奇偶校验 - 镜像存储设备 1602 的功能大体上类似于其他存储设备 150。

[0360] 在典型的配置中,指定的奇偶校验 - 镜像存储设备 1602 在存储设备集 1604 的存储设备 150 之间轮换,奇偶校验 - 镜像存储设备 1602 实质上具有与其他存储设备 150 相同的特点,这是因为奇偶校验 - 镜像存储设备 1602 也必须像非奇偶校验 - 镜像存储设备一样运行。类似的特点是相对于上述的 RAID 群组内的操作和用于客户端 114 独立通信的自主操作。在不同的实施方式中,存储设备集 1604 的存储设备 150 可在其他方面(不涉及所述 RAID 环境下的功能)不同。

[0361] 存储设备集 1604 的存储设备 150 可以是独立的、可以是在一个或多个服务内成组的、可每一个驻留在一个服务器 112 内、可通过一个或多个服务器 111 被访问，等等。一个或多个客户端 114 可驻留在包括一个或多个存储设备 150 的服务器 112 内、可驻留在独立的服务器 112 内、可驻留在通过一个或多个计算网络 116 访问存储设备 150 的计算机、工作站、笔记本电脑等设备内，或位于类似设备内。

[0362] 在一种实施方式中，网络 116 包括系统总线，并且存储设备集 1604 中的一个或多个存储设备 150、1602 通过所述系统总线通信。例如，系统总线可以是 PCI-e 总线、串行高级技术附件（“串行 ATA”）总线、并行 ATA 或类似总线。在另一种实施方式中，所述系统总线是外部总线，如小型计算机系统接口（“SCSI”）、火线、光纤通道、USB、PCIe-AS、无限带宽或类似总线。本领域技术人员会意识到具有存储设备 150 的其他系统 1600 配置，其中，存储设备 150 自主运行，并能够独立地通过网络 116 接收和服务来自客户端 114 存储请求。

[0363] 图 17 是示意性框图，示出了根据本发明的用于共享的、前端分布式 RAID 的装置 2300 的一种实施方式。在不同的实施方式中，装置 2300 包括多存储请求接收器模块 2302、条带化模块 2304、奇偶校验 - 镜像模块 2306、定序器模块 2308、主验证模块 2310、主确定模块 2312、主错误模块 2314、奇偶校验生成模块 2316 和奇偶校验更替模块 2318，这些模块在下文中描述。

[0364] 装置 2300 包括多存储请求接收器模块 2302，该多存储请求接收器模块 2302 接收来自至少两个客户端 114 的至少两个存储请求，以将数据存储在存储设备集 1604 的存储设备 150 中。所述数据包括文件的数据或对象的数据。与装置 2300 有关的存储请求中的每一个至少具有一部分共有数据，此外，所述存储请求是并发存储请求，这是由于所述存储请求的到达使得在另一个存储请求到达之前，一个存储请求还没有完成。这些并发的存储请求具有损坏前端分布式 RAID 系统 1600 的共有数据的风险。在一种实施方式中，所述并发的存储请求可来自一个客户端 114。在另一种实施方式中，所述并发的存储请求来自两个或更多个客户端 114。

[0365] 多存储请求可更新存储在存储设备集 1604 的存储设备 150 上的一个或多个数据段，其中，条带化模块 2304 将预先存储的数据条带化为存储在存储设备集 1604 的存储设备 150 上的数据段。在一种实施方式中，存储请求将所述数据第一次写入 RAID 群组。在这种情况下，所述数据通常会存在于其他位置，并可通过一个或多个服务器 114 访问，然后，一个存储请求将所述数据复制到 RAID 群组，而另一个存储请求同时访问所述数据。

[0366] 多存储请求可包括一个更新存储在存储设备集 1604 的存储设备 150 上的一个或多个数据段的请求，还可包括目标为至少一部分共有数据的一个或多个读取请求。如果更新请求没有完成，则存储设备集 1604 的存储设备 150 返回的读取请求可由预先存在和损坏所述数据的已更新数据的结合组成。

[0367] 装置 2300 包括条带化模块 2304，该条带化模块 2304 计算（为每个并发存储请求）用于所述数据的条带模式，并将 N 个数据段写入存储设备集 1604 的 N 个存储设备 150a-n。所述条带模式包括一个或多个条带，每个条带包括 N 个数据段的集。N 个数据段中的每一个被写入存储设备集 1604 中的不同的存储设备 150 并被分配给所述条带。装置 2300 包括奇偶校验 - 镜像模块 2306，该奇偶校验 - 镜像模块 2306（为每个并发存储请求）将所述条带的 N 个数据段的集写入被指定为奇偶校验 - 镜像存储设备 1602 的存储设备集 1604 中的存

储设备 150。奇偶校验 - 镜像存储设备 1602 是除 N 个存储设备 150a-n 之外的设备。

[0368] 条带化模块 2304 还用于计算一个或多个存储设备 150a-n 的一致性, 其中, 一个或多个作为文件或对象的一部分的数据段读取自一个或多个存储设备 150a-n。

[0369] 装置 2300 包括定序器模块 2308, 该定序器模块 2308 确保来自第一客户端 114 的第一存储请求完成之后才执行来自第二客户端的第二存储请求, 其中, 至少两个并发存储请求包括第一和第二存储请求。在其他实施方式中, 定序器模块 2308 确保所述第一存储请求完成之后才执行两个或多个其他并发存储请求。有利地是, 定序器模块 2308 有助于并发存储请求的顺序执行, 从而避免损坏数据。在一种实施方式中, 定序器模块 2308 通过下述方法协调并发存储请求的执行: 利用所述的存储请求必须访问所述数据的主控制器、利用锁止系统、两阶段提交或本领域技术人员熟知的其他方法。下文描述了定序器模块 2308 使用的一些方法。

[0370] 在一种实施方式中, 定序器模块 2308 通过下述方法确保所述第一存储请求完成之后才执行并发存储请求: 接收来自存储设备集 1604 的存储设备 150 中的每一个的应答, 其中, 存储设备 150 在执行第二存储请求之前与一起接收第一存储请求和存储请求。通常, 应答确认存储请求已完成。在一种实施方式中, 存储设备 150 中受所述存储影响的每一个设备被写入, 且在定序器模块 2308 开始执行第二存储请求之前从每个存储设备 150 接收应答。

[0371] 在一种实施方式中, 完成存储请求可包括执行单个存储设备(如 150a)上的等待中的第二存储请求的一部分之前, 完成指令单个存储设备(如 150a)的第一存储请求的一部分。定序器模块 2308 可独立地验证存储设备 150 上的存储请求的一部分是否完成。在这种实施方式中, 写入涉及第一存储请求的数据段直到所述第一存储请求的所有数据段被完成后才需要被延迟。定序器模块 2308 可协调发生在存储设备集 1604 的存储设备 150 上的不同请求的执行以确保所述数据不被损坏。

[0372] 在一种实施方式中, 在条带化模块 2304 和奇偶校验 - 镜像模块 2306 都将与存储请求有关的所述数据段写入存储设备集 1604 的存储设备 150 后, 接收存储请求完成的应答。在另一种实施方式中, 在条带化模块 2304 和奇偶校验 - 镜像模块 2306 都将与存储请求有关的所述数据段写入存储设备集 1604 的存储设备 150 且存储设备 150、1602 中的每一个都确认所述数据段已经被写入后, 接收存储请求完成的应答。

[0373] 在一种实施方式中, 定序器模块 2308 通过在首先到达的并发请求之间选择存储请求选择用于执行的第一存储请求。在另一种实施方式中, 定序器模块 2308 通过选择具有较早的时间戳的存储请求选择用于执行的第一存储请求。在另一种实施方式中, 定序器模块 2308 通过使用一些选择标准选择存储请求选择用于执行的第一存储请求。例如, 定序器模块 2308 可选择以某种方式被请求客户端 114 标记为高优先级的存储请求、可选择来自优先客户端 114 的存储请求, 等等。本领域技术人员会认识到定序器模块 2308 可利用一些选择标准选择第一存储请求的其他方法。

[0374] 在一种实施方式中, 多存储请求接收器模块 2302、条带化模块 2304、奇偶校验 - 镜像模块 2306 和定序器模块 2308 是主控制器(未示出)的部分, 该主控制器控制并服务所述并发存储请求。所述主控制器的全部或部分可驻留并运行在下述设备内: 客户端 114、第三方 RAID 管理设备、存储设备集 1604 的存储设备 150 或存储设备 150 的存储控制器 152。

通过使用主控制器用于为所述数据执行服务请求,定序器模块 2308 可获悉指令所述数据的存储请求并可随后识别并发存储请求,然后还可将所述并发存储请求按上述要求排序:存储在存储设备集 1604 的存储设备 150 上的数据不会被损坏。本领域技术人员会认识到控制服务指令所述数据的存储请求的主控制器的其他实现方式。

[0375] 在另一种实施方式中,所述主控制器是两个或更多个能够服务来自一个或多个客户端 114 的所述并发存储请求的主控制器的群组的一部分,其中,所述存储请求是指令存储在存储设备集 1604 的存储设备 150 上的所述数据。例如,主控制器可为第一客户端 114 服务存储请求,而第二主控制器可为第二客户端 114 服务存储请求。第一和第二客户端 114 都可访问存储在存储设备集 1604 的存储设备 150 上的数据,因此允许并发存储请求。一个主控制器可以是存储设备 150a 的一部分,而其他主控制器可以是第二存储设备 150b 的一部分。在另一种实施方式中,第一主控制器可以是第一存储设备集 1604a 一部分,而第二主控制器可以是镜像存储设备集 1604b 的一部分。

[0376] 在主控制器是访问存储设备集 1604 的存储设备 150 的主控制器群组的一部分的情况下,装置 2300 可包括主验证模块 2310,该主验证模块 2310 在执行接收到的存储请求之前确认服务所述接收到的存储请求的主控制器正在控制先于一个或多个并发存储请求的执行的所述存储请求的执行。在这种实施方式中,其他主控制器接收所述并发存储请求,并且服务请求至少有一部分数据与其他主控制器接收到的并发存储请求相同。

[0377] 例如,主控制器可接收存储请求,然后,主验证模块 2310 可在所述存储请求执行前轮询其他主控制器以验证所述主控制器仍然是用于所述存储请求的数据的主控制器。验证的一部分包括验证所述主控制器之间能够通信,从而指定的主控制器在所述存储请求执行之前被验证。这种方法可在前端 RAID 控制器被指定为主控制器而另一个前端 RAID 控制器被指定为备用控制器的情况下是有利的。在另一个实例中,主控制器可接收从文件或对象读取数据段的存储请求,然后主验证模块 2310 可轮询其他主控制器,从而验证没有进行中的文件或对象的更新。在另一个实例中,主控制器可使用主验证模块 2310 以获取控制用于所述存储请求的数据。

[0378] 一种验证所述主控制器仍然是用于执行所述存储请求的主控制器的方法是:使用三路轮询方案,其中两个设备 / 控制器必须能够用于投票选出用于存储请求的主控制器,以便于继续轮询。这种方案使用对竞争成为主控制器的控制来说是第三方的设备(未示出),并且保留哪一个控制器被分配为主控制器的记录。这个主验证设备可以是另一个控制器、服务器上的客户端 114 等等,并能够与群组中可充当主控制器的控制器通信。然后,主验证模块 2310 的一部分可驻留在所述主验证设备内,而主验证模块 2310 的一部分位于每个控制器内。

[0379] 在一个实例中,系统 1600 包括第一前端分布式 RAID 控制器(“第一控制器”)、第二前端分布式 RAID 控制器(“第二控制器”),其中每一个控制器都可以是主控制器和分享的主验证设备。第一和第二控制器和主验证设备之间都可相互通信。主验证模块 2310 可将第一控制器指定为主控制器而把第二控制器指定为用于存储在存储设备集 1604 的存储设备 150 上的数据的备用控制器,并且主验证模块 2310 可将主控制器的信息存储在控制器和主验证设备上。只要保持第一控制器、第二控制器和主验证设备之间的通信,主验证模块 2310 就能够确认第一控制器为主控制器。

[0380] 如果第一（主）控制器接收存储请求，第二（备用）控制器变得不可用或与第一控制器和主验证设备的通信丢失，主验证模块 2310 能够通过主验证设备和第一（主）控制器之间的通信验证所述第一控制器仍然是主控制器，并且由于第一控制器和主验证设备都确认所述第控制器确实是主控制器，主验证模块 2310 可允许存储请求继续进行。由第二（备份）控制器接收的存储请求不会继续进行，这是由于第二控制器通过主验证模块 2310 识别到其不是主控制器。

[0381] 另一方面，如果第一（主）控制器不可用或不能与第二（备份）控制器和主验证设备通信，并且第二（备份）控制器接收存储请求，主验证模块 2310 能够识别到第二控制器和主验证模块不能与第一控制器通信，并且主验证模块 2310 能够指定第二（备份）控制器为主控制器，存储请求也得以继续进行。然后，主控制器指定的改变被记录在第二控制器上。

[0382] 如果第一控制器是操作性的，并与第二控制器和所述主验证设备完全断开通信，用于第一控制器接收的数据的任何存储请求将不会被执行。如果通信恢复，第一控制仍然不会执行存储请求，这是由于所述第二控制器和所述主验证模块都将第二控制器识别为主控制器。当然，这种主控制器指定可以被重置。本领域技术人员会认识到分配和重新分配主控制器指定给主控制器中的一个。

[0383] 如果主验证设备不可用且第一存储控制器接收存储请求，主验证模块 2310 运行在第一和第二控制器上的部分能够验证所述第一控制器是主控制器，并且存储请求可继续进行。如果所述第二控制器接收存储请求，主验证模块 2310 运行在第一和第二控制器上的部分能够验证所述第一控制器是主控制器，并且存储请求不会再继续进行。在其他实施方式中，不止两个的控制器是轮询方案的一部分。本领域技术人员会认识到主验证模块 2310 能够在执行存储请求之前验证控制器是主控制器的其他方法。

[0384] 在另一种实施方式中，装置 2300 包括主确定模块 2312。在发送存储请求之前，主确定模块 2312 将主确定请求发送给主控制器群组。然后，主控制器群组识别哪一个控制器被指定为用于存储请求的主控制器，并将标识主控制器的响应发回主确定模块 2312。主确定模块 2312 为所述存储请求接收主控制器的标识符并指令所述请求设备将存储请求发送给指定的主控制器。在一种实施方式中，主确定模块 2312 位于并运行在客户端 114 内。在另一种实施方式中，主确定模块 2312 位于第三方 RAID 管理设备内并在其内执行请求。在另一种实施方式中，主确定模块 2312 位于存储设备 150 内。在另一种实施方式中，主确定模块 2312 分布在两个或更多个存储设备 150 之间。

[0385] 在又一种实施方式中，装置 2300 包括返回错误指示的主错误模块 2314。在一种实施方式中，如果由主控制器控制的多存储请求接收器模块 2302 接收到不由主控制器控制的存储请求，主错误模块 2314 返回错误指示。

[0386] 在另一种实施方式中，如果主确定模块 2312 或主验证模块 2310 在所述存储请求执行完成时确定主控制器不再是确定的主控制器，主错误模块 2314 返回错误指示。这种实施方式通常发生在当主控制器开始执行存储请求并与群组中的其他主控制器的通信丢失时，或者发生在轮询方案中的与其他主控制器和主验证设备的通信丢失时。在另一种实施方式中，如果由主控制器控制的多存储请求接收器模块 2302 接收不由所述主控制器控制的存储请求，主错误模块 2314 返回错误指示。

[0387] 在另一种实施方式中，主控制器控制传送给一个或多个次级主控制器的存储请求。每个所述次级主控制器控制用于存储在存储设备集 1604 的存储设备 150、1602 上的数据的存储请求。在另一种实施方式中，控制所述次级主控制器的主控制器也是用于指令存储在存储设备集 1604 的存储设备 150、1602 上的数据的存储请求的次级主控制器。

[0388] 在另一种实施方式中，主控制器控制传送给一个或多个次级主控制器的存储请求，并且每个所述次级主控制器控制用于存储在对所述次级主控制器来说唯一的存储设备集的存储设备 150 上的数据的存储请求。装置 2300 是灵活的，从而任何主控制器都能够成为相对于其他作为次级主控制器的控制器的主控制器。一些次级主控制器能够存储设备集 1604，而其他次级主控制器能够控制不同的存储设备集。在另一种实施方式中，主控制器可以是奇偶校验 - 镜像存储设备 1602 或 N 个存储设备 150a-n 中的一个。

[0389] 在另一种实施方式中，当所述主控制器离线或不能确定其是指定的主控制器时，次级主控制器可以成为主控制器。本领域技术人员会认识到用于在一个或多个次级主控制器之间分配或重新分配主控制器指定的各种静态和动态的方法。

[0390] 在一种优选实施方式中，装置 2300 包括奇偶校验生成模块 2316，该奇偶校验生成模块 2316 为所述条带计算奇偶校验数据段并将所述奇偶校验数据段存储在奇偶校验 - 镜像存储设备 1602 上。由奇偶校验 - 镜像存储设备 1602 上的 N 个数据段的集计算奇偶校验数据段。这种实施方式通常通过 RAID5、RAID6 或其他 RAID 级别（但通常不包括 RAID0、RAID1、RAID10 等等）实现。

[0391] 在另一种优选实施方式中，装置 2300 包括奇偶校验更替模块 2318，该奇偶校验更替模块 2318 为每个条带变更被分配为一个或多个用于所述条带的奇偶校验 - 镜像存储设备 1602 的存储设备集 1604 中的存储设备 150。轮换每个条带的奇偶校验数据段提升了性能。奇偶校验更替模块 2318 可与条带化模块 2304 一起使用，以计算一个或多个存储设备 150a-n 之间的一致性，作为文件或对象的一部分的一个数据段从一个或多个存储设备 150a-n 中读取、写入或更新。

[0392] 不同的模块 2302-23018 的功能可一起在单个主控制器中实现，或者可分布在下述设备之间：一个或多个客户端 114、第三方 RAID 管理设备和一个或多个存储设备 150、1602。本领域技术人员会认识到此处描述的功能是分布式的不同的实施方式。

[0393] 图 18 是示意性流程图，示出了根据本发明的用于共享的、前端分布式 RAID 的方法 2400 的一种实施方式。方法 2400 始于步骤 2402，多存储请求接收器模块 2302 接收来自至少两个客户端 114 的至少两个存储请求（步骤 2404），以读取数据或将数据存储在存储设备集 1604 中的存储设备 150。所述数据来自文件，或者是对象的数据，并且每个所述存储请求具有至少一部分共有的数据，并且，所述存储请求是并发存储请求，这是由于所述存储请求的到达使得在一个存储请求到达之前，两个存储请求中的另一个存储请求还没有完成。条带化模块 2304 为所述数据计算条带模式（步骤 2406），其中，所述条带模式包括一个或多个条带并且每个条带包括 N 个数据段的集。条带化模块 2304 还读取条带的 N 个数据段，或将条带的 N 个数据段写入存储设备集 1604 中的 N 个存储设备 150a-n（步骤 2408），其中，N 个数据段中的每一个都被写入或读取自独立的存储设备 150。

[0394] 当所述存储请求是写入操作时，奇偶校验 - 镜像模块 2306 将所述条带的 N 个数据段的集写入存储设备集 1604 中的一个或多个奇偶校验 - 镜像存储设备 1602（步骤 2306），

其中，奇偶校验 - 镜像存储设备 1602 是除 N 个存储设备 150a-n 之外的设备。奇偶校验 - 镜像模块 2306 还读取存储在奇偶校验 - 镜像存储设备 1602 中的数据段或奇偶校验数据段 (2410)。定序器模块 2308 确保来自第一客户端 114 的第一存储请求完成后才执行来自第二客户端 114 的存储请求。方法 2400 结束于步骤 2416。第一和第二存储请求是并发存储请求。

[0395] 本发明可采用其他指定形式实施而不脱离本发明的宗旨或本质特点。描述的实施方式在各个方面被视为仅仅是示例性而不是限制性的。因此，本发明的范围由附属的权利要求确定，而不是由上述说明书确定。在本发明的权利要求的含义和等价范围内的所有改变被包含在本发明的保护范围内。

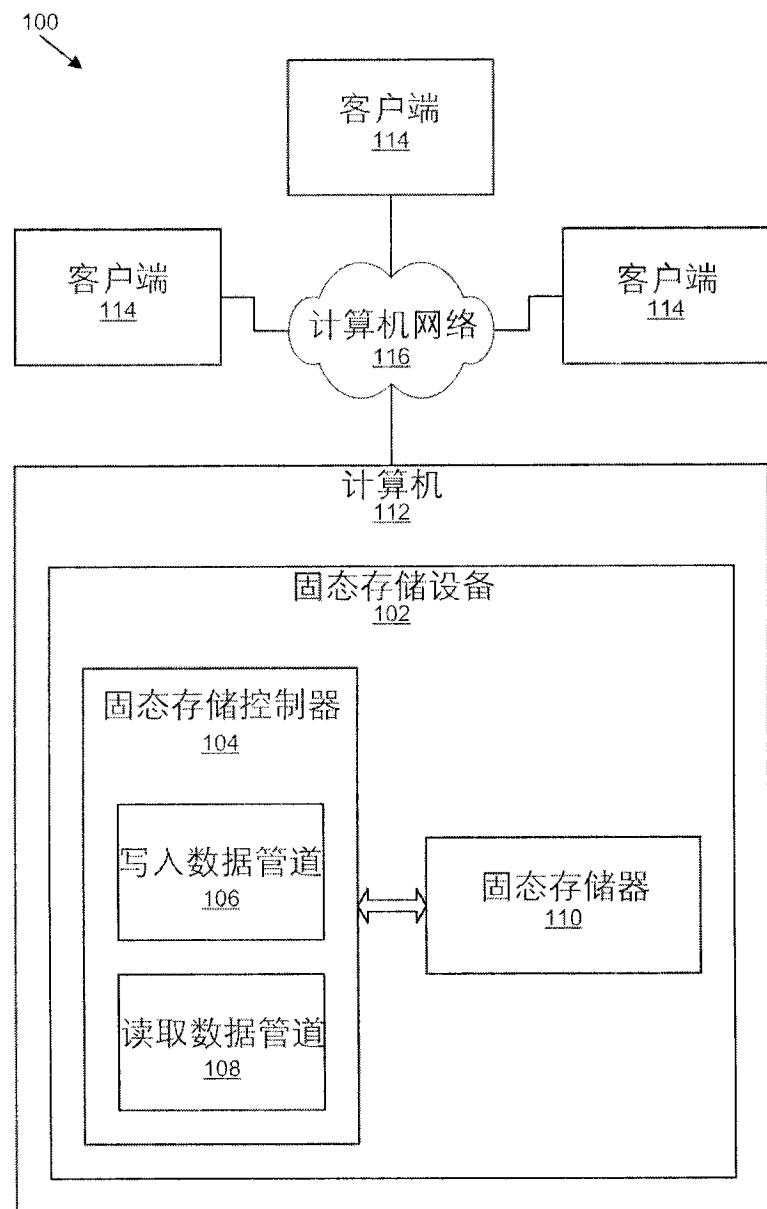


图 1A

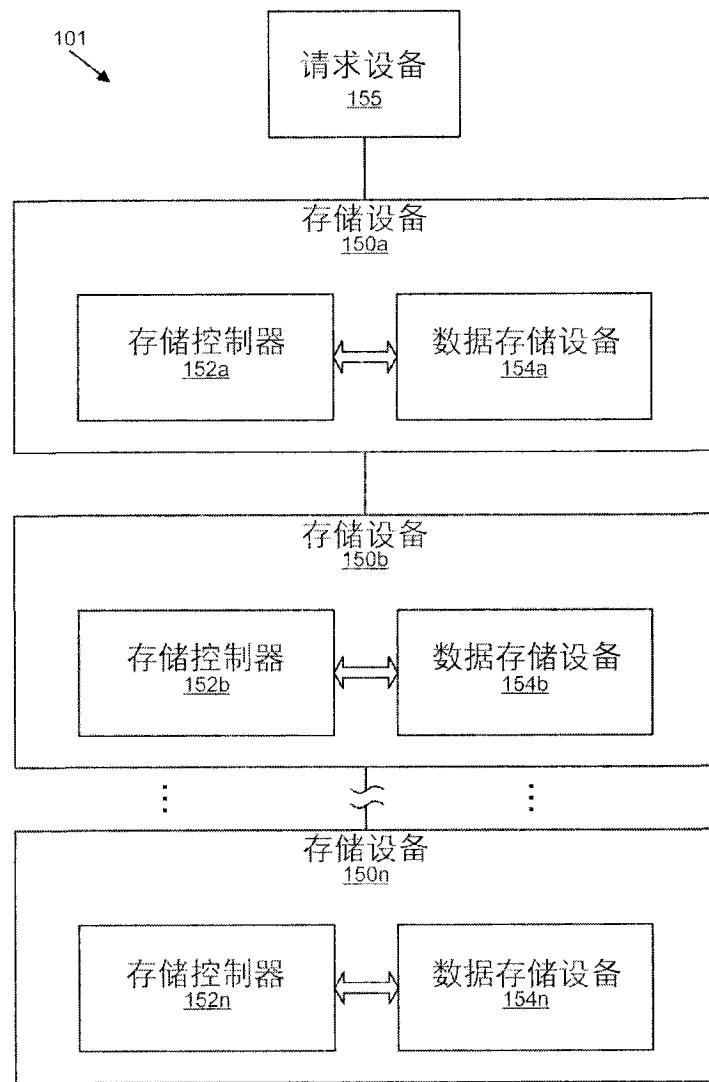


图 1B

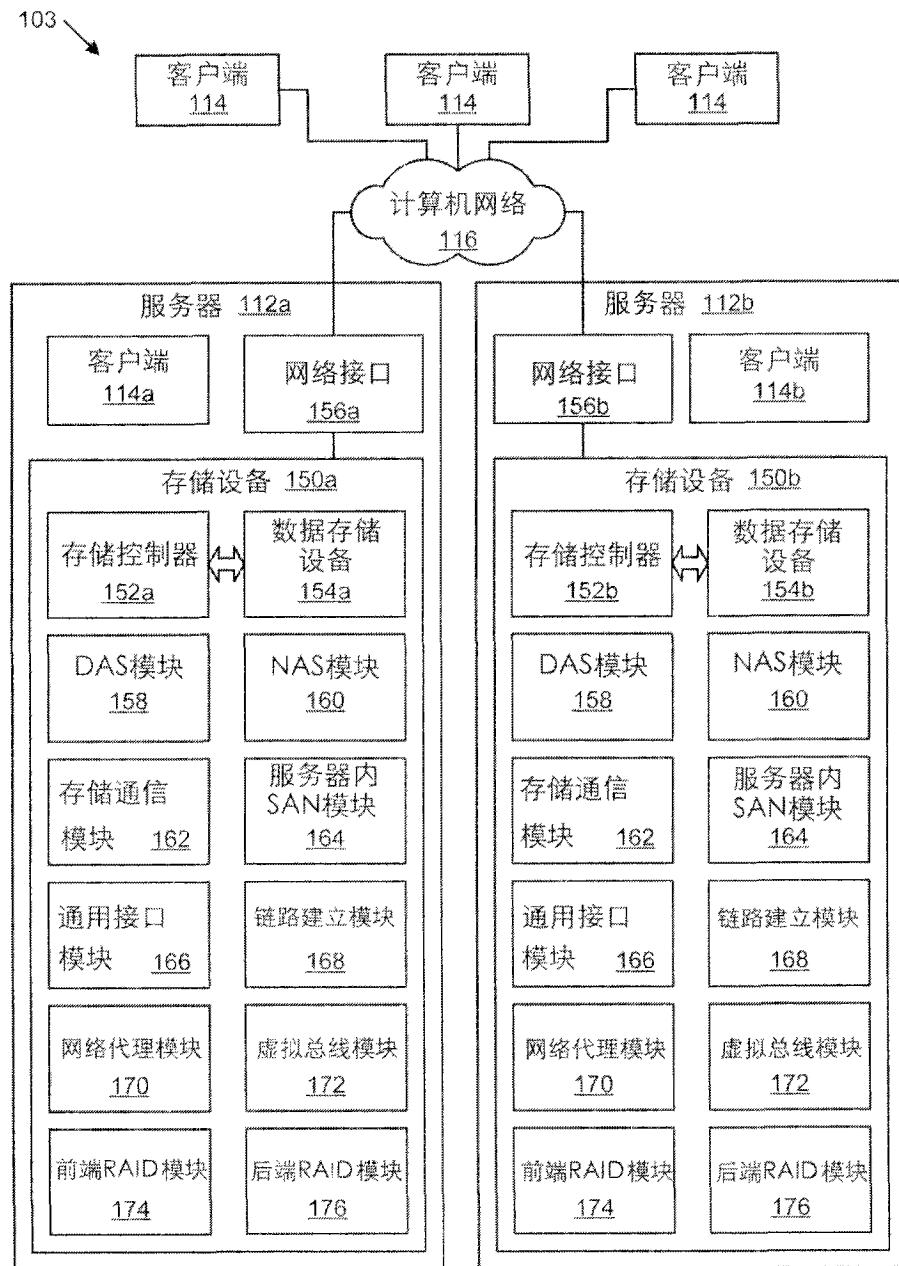


图 1C

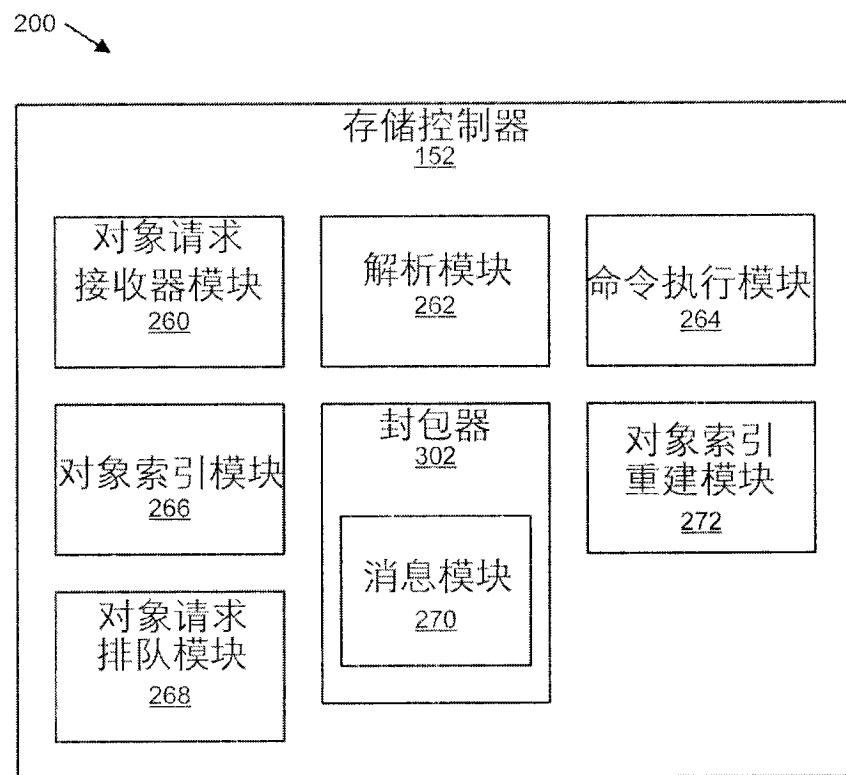


图 2A

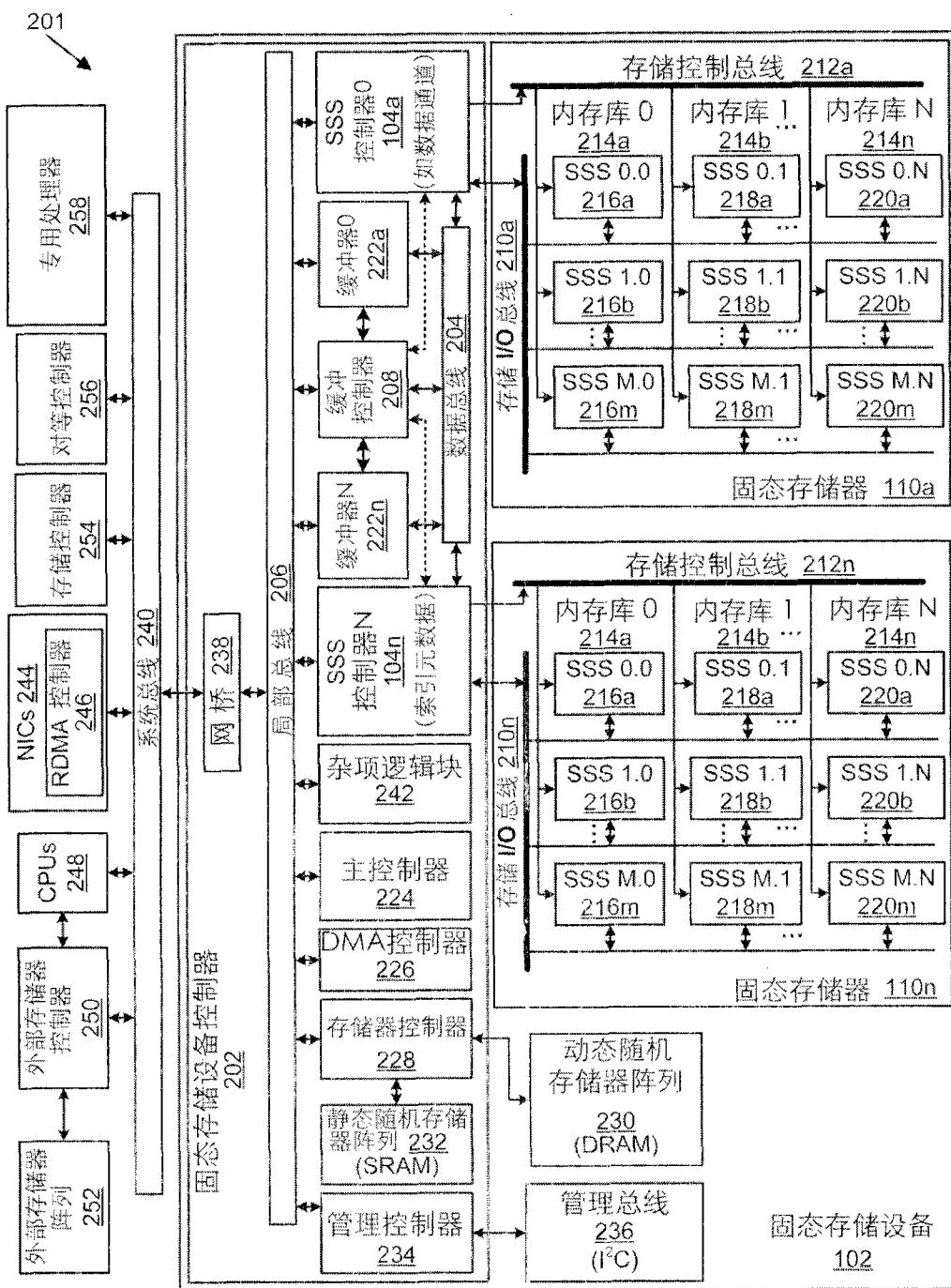


图 2B

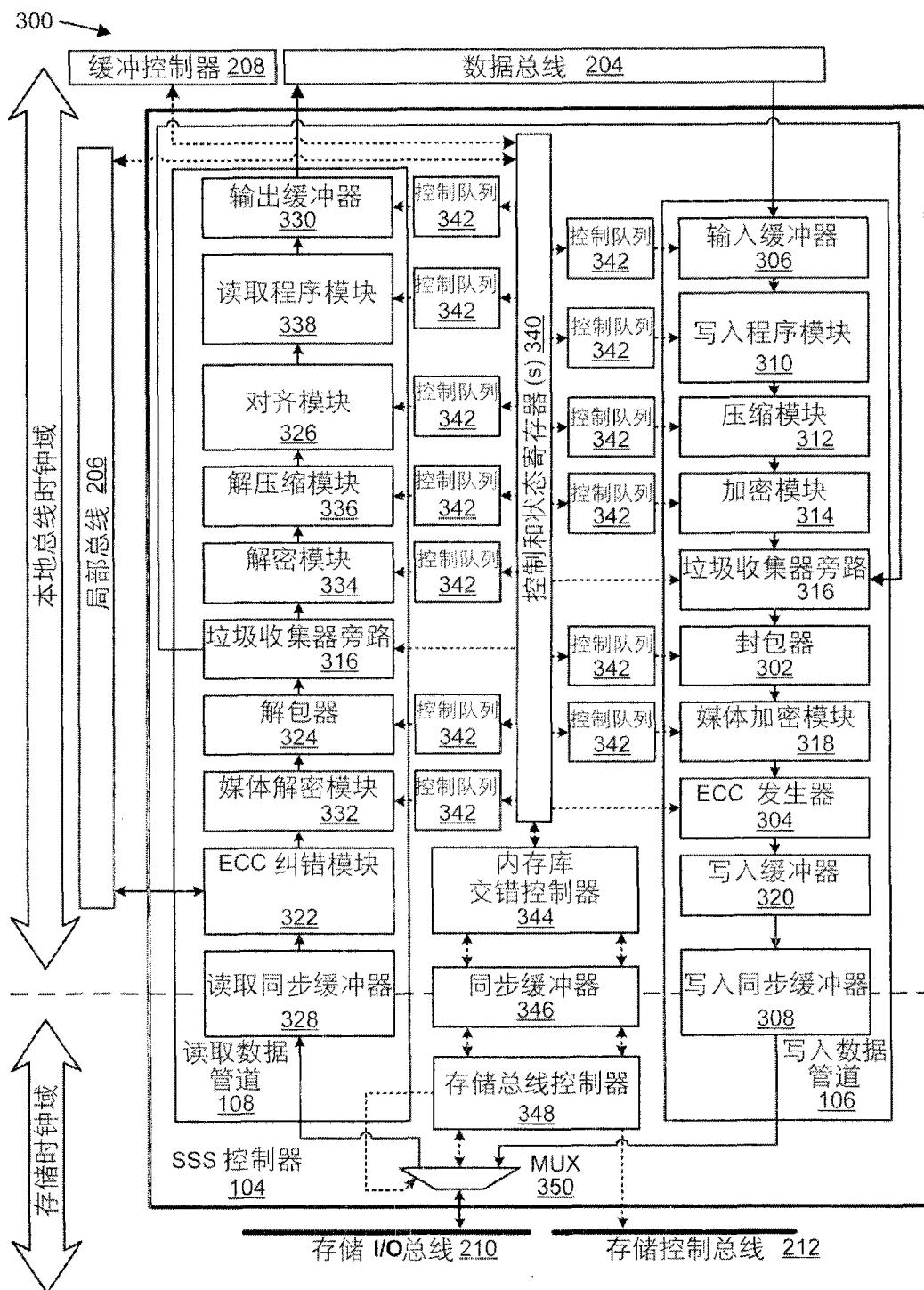


图 3

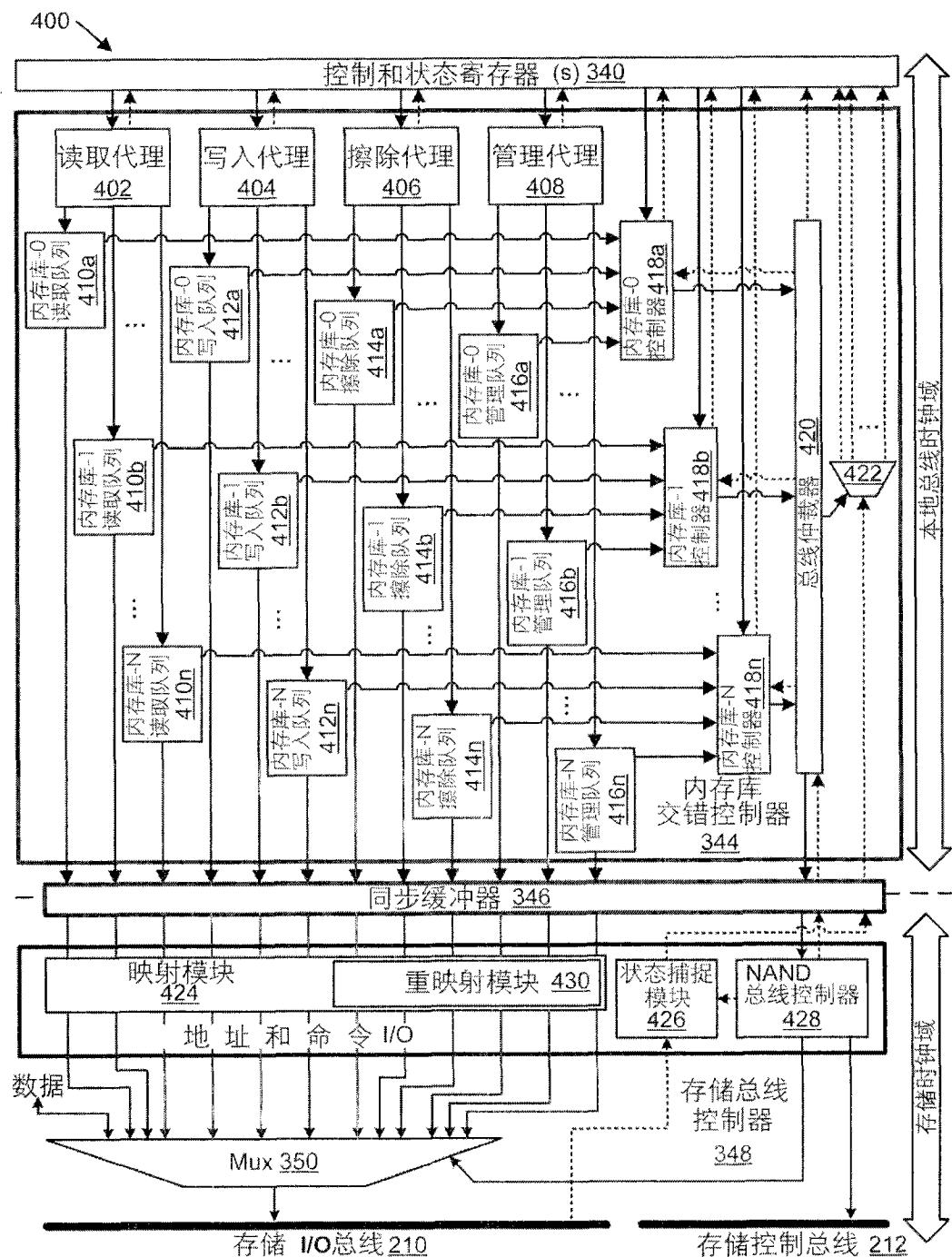


图 4A

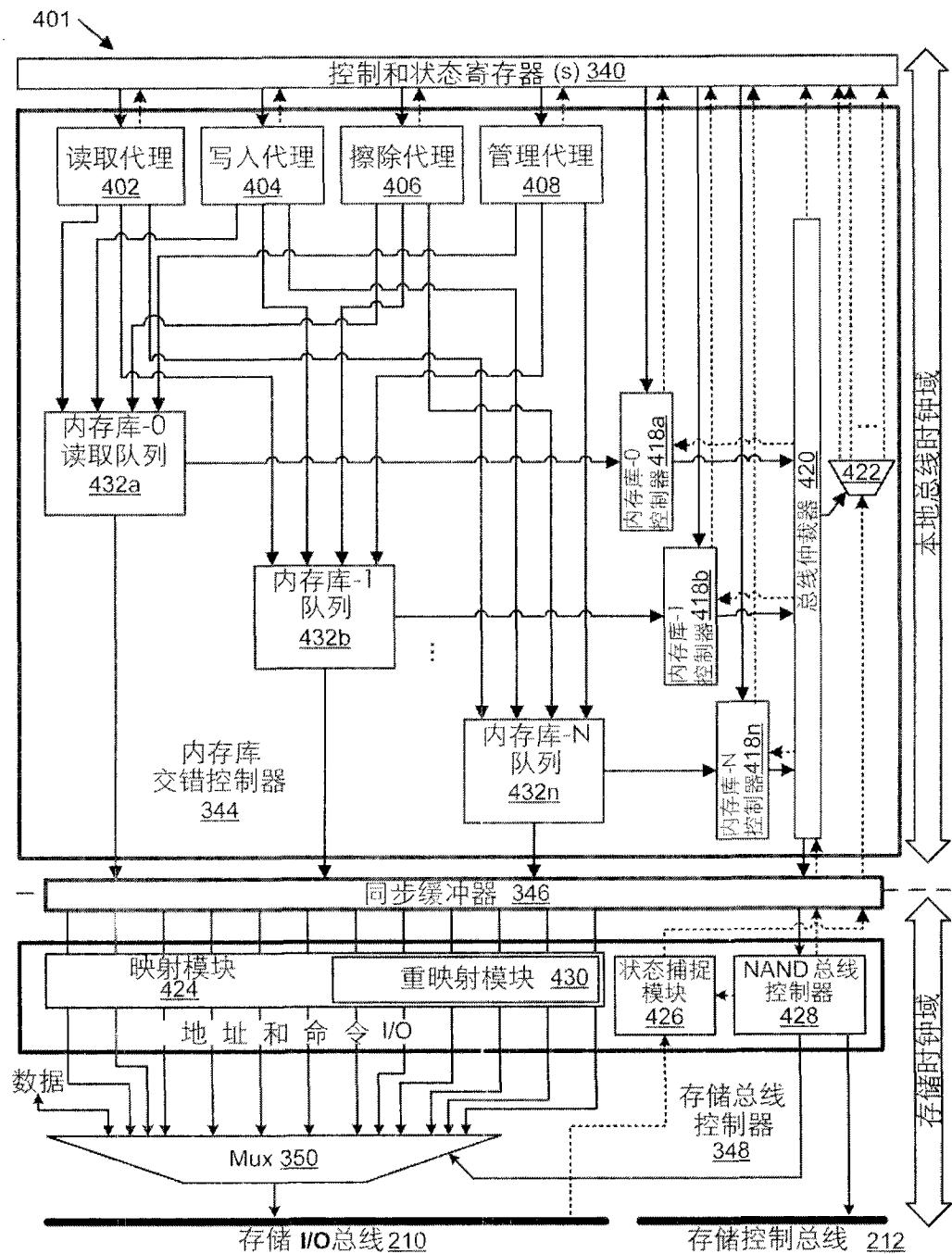


图 4B

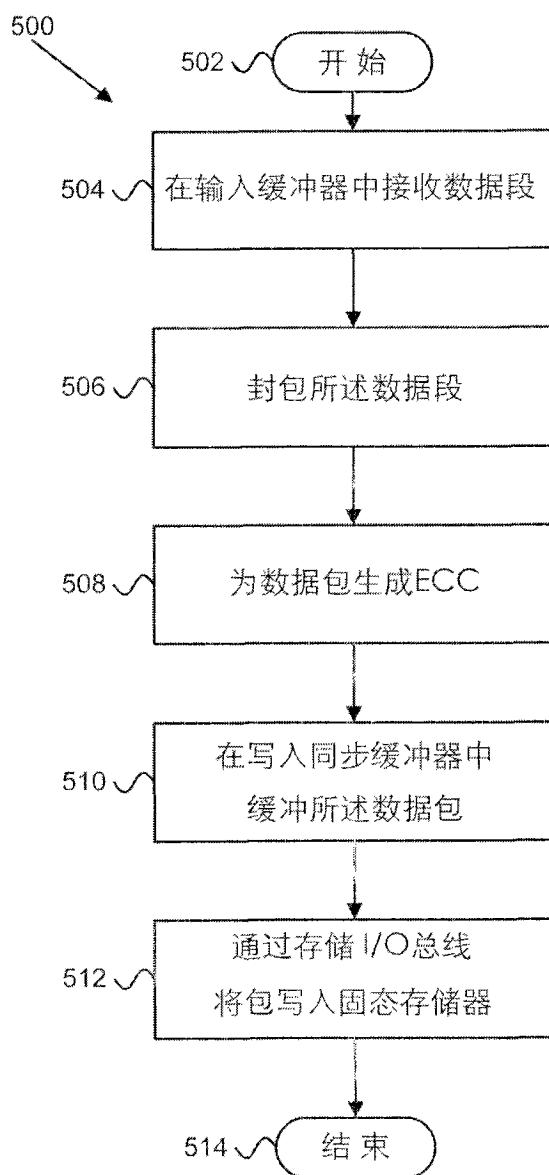


图 5A

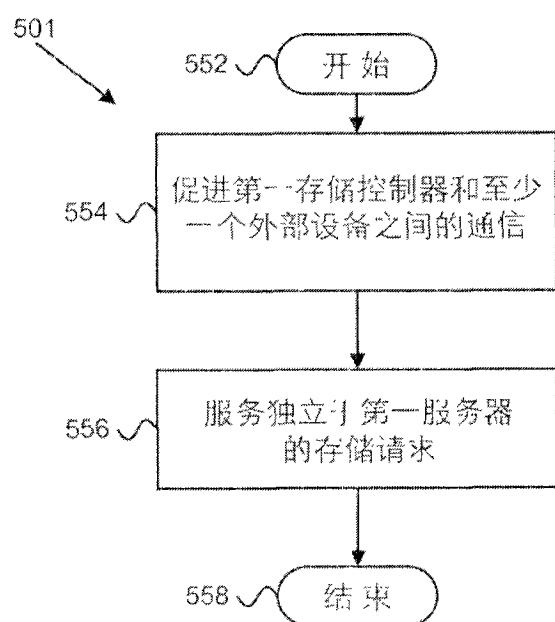


图 5B

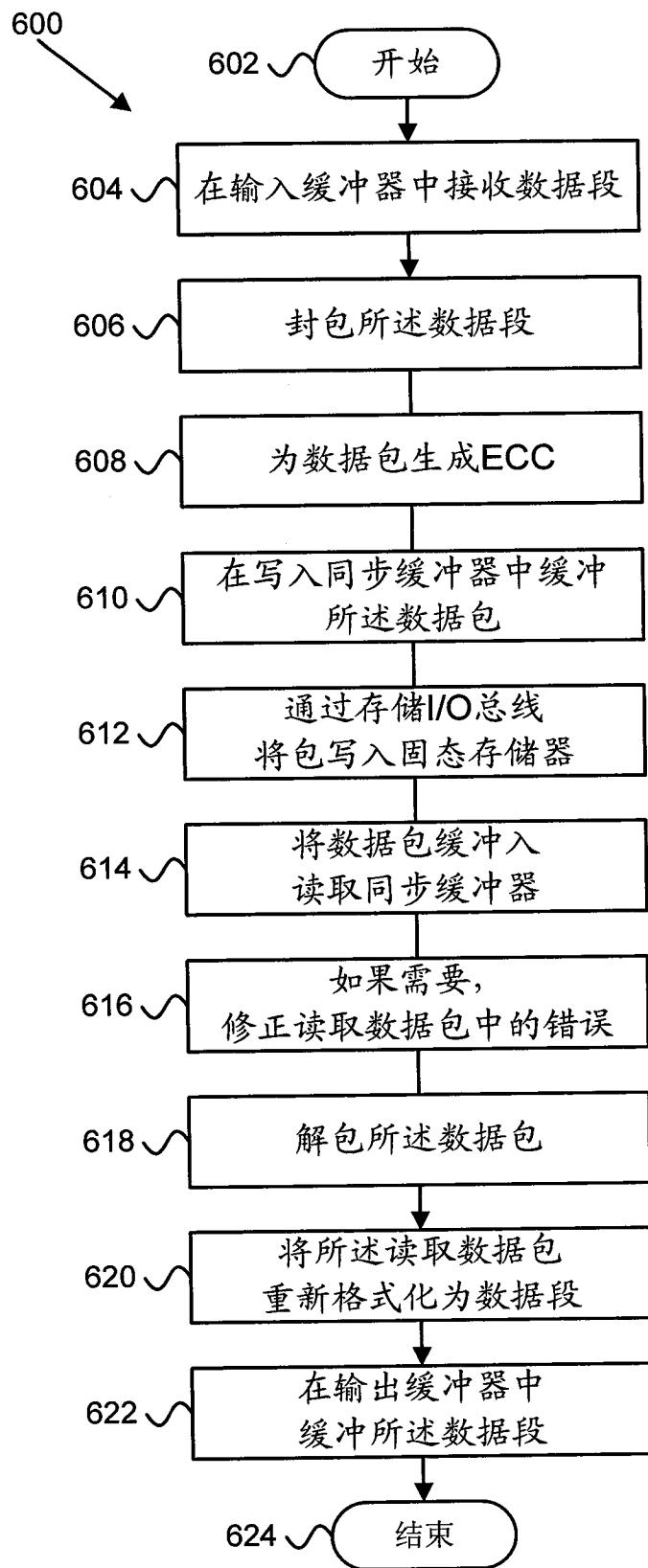


图 6

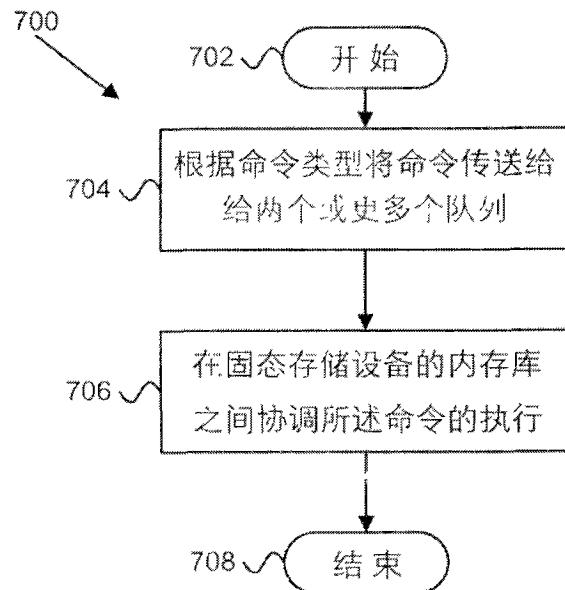


图 7

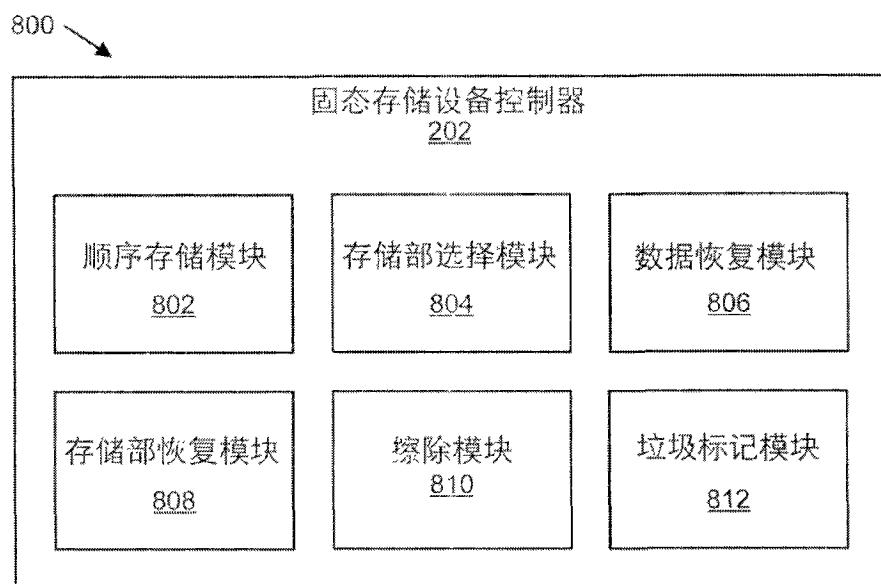


图 8

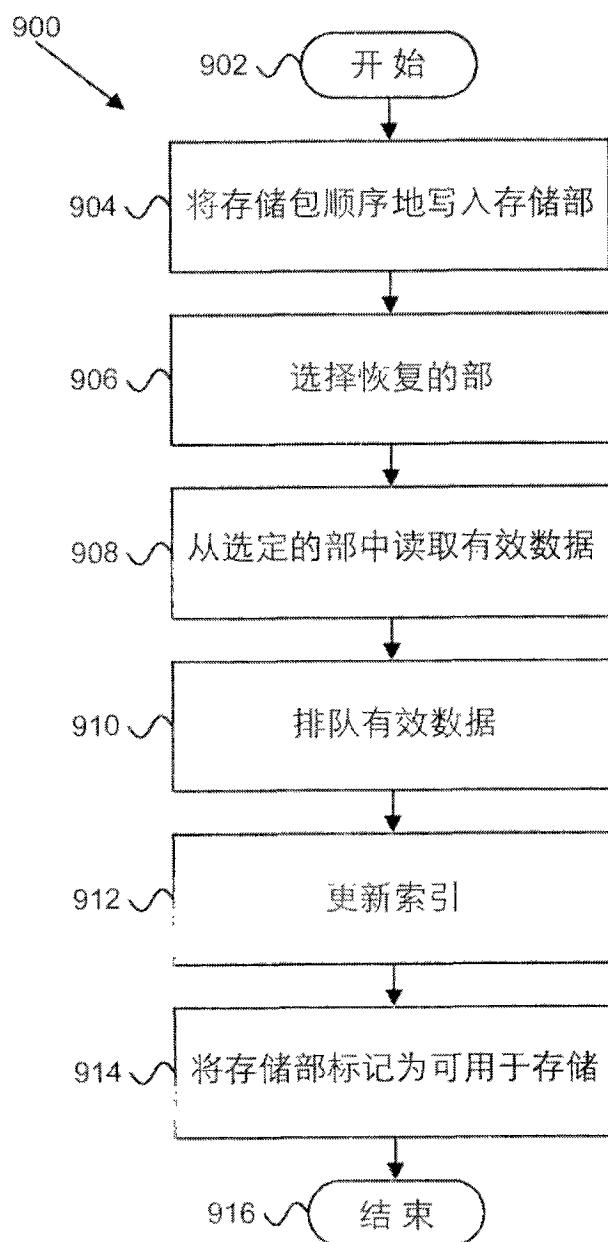


图 9

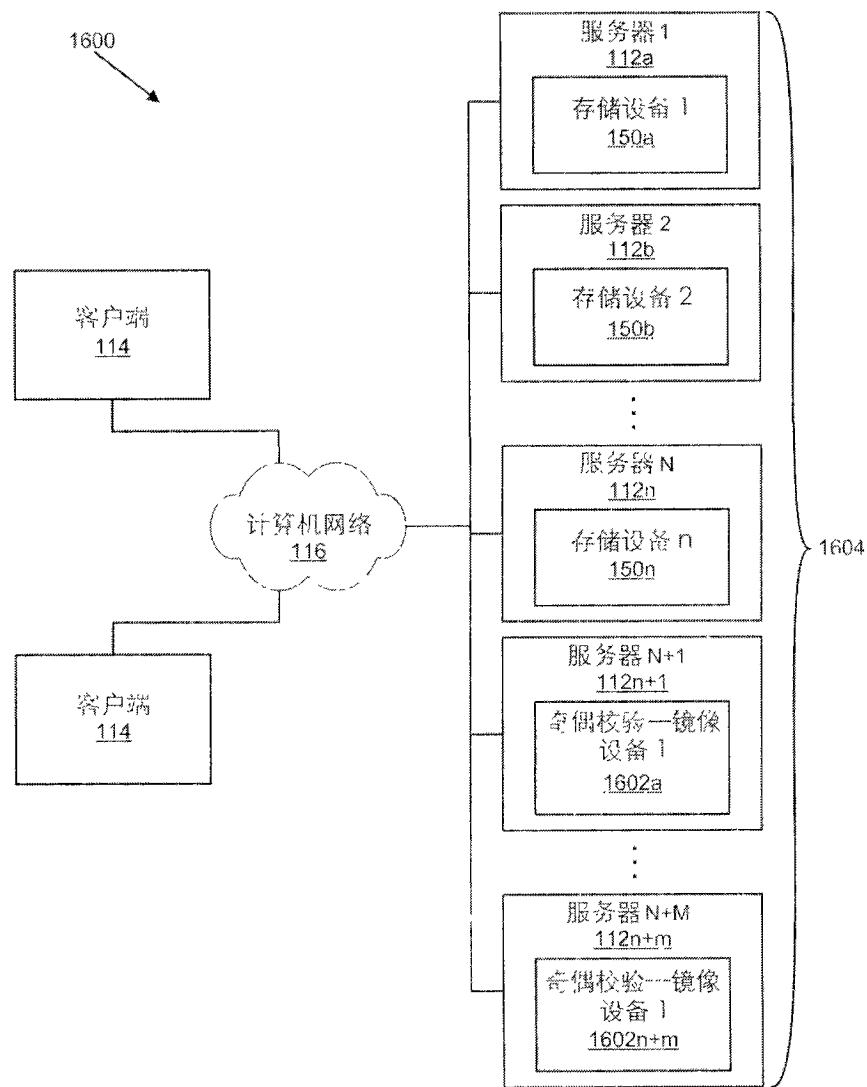


图 10



图 11

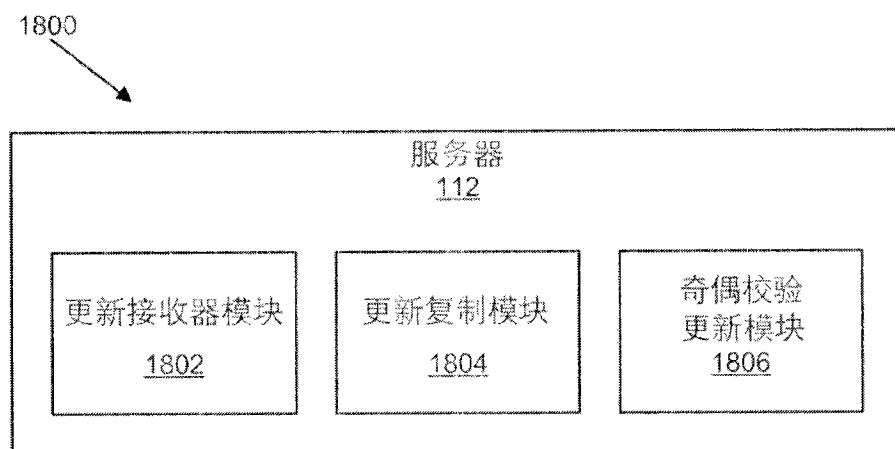


图 12

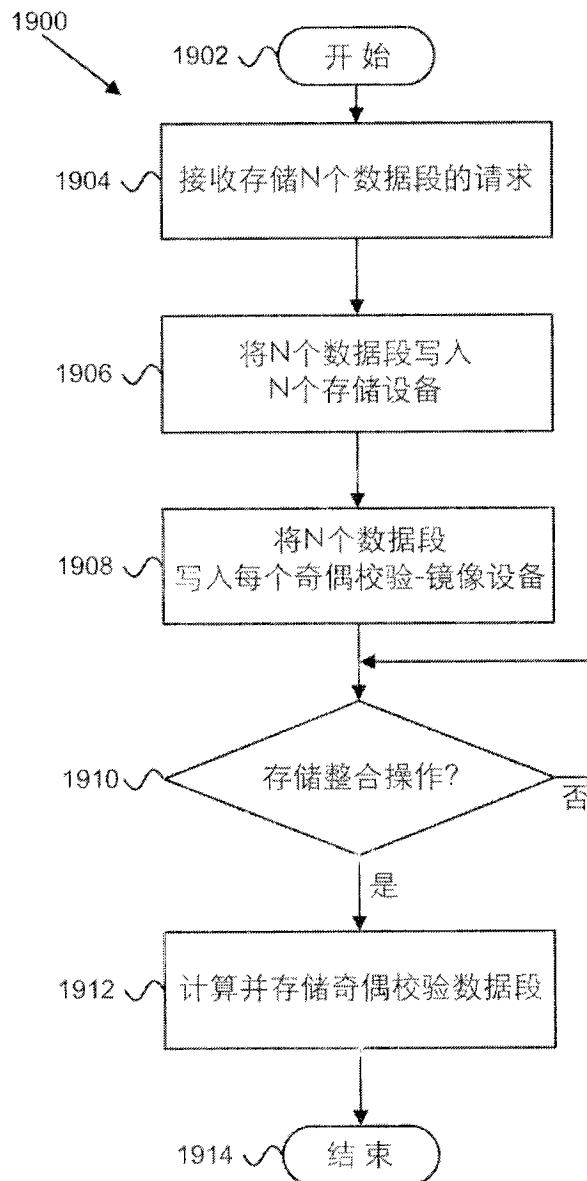


图 13

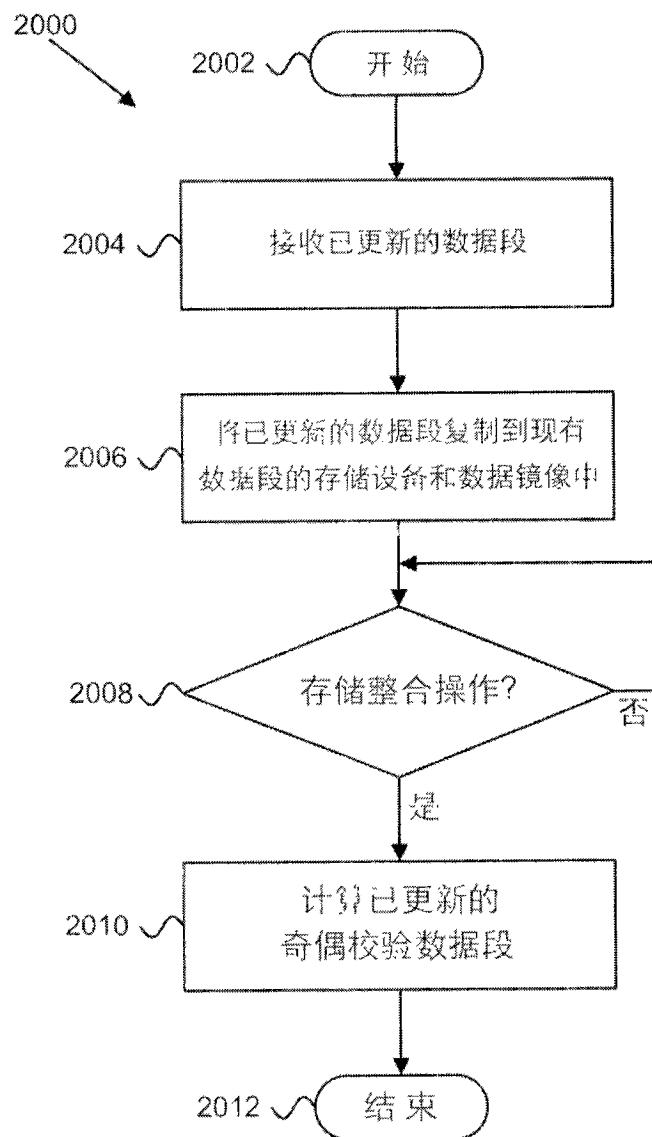


图 14

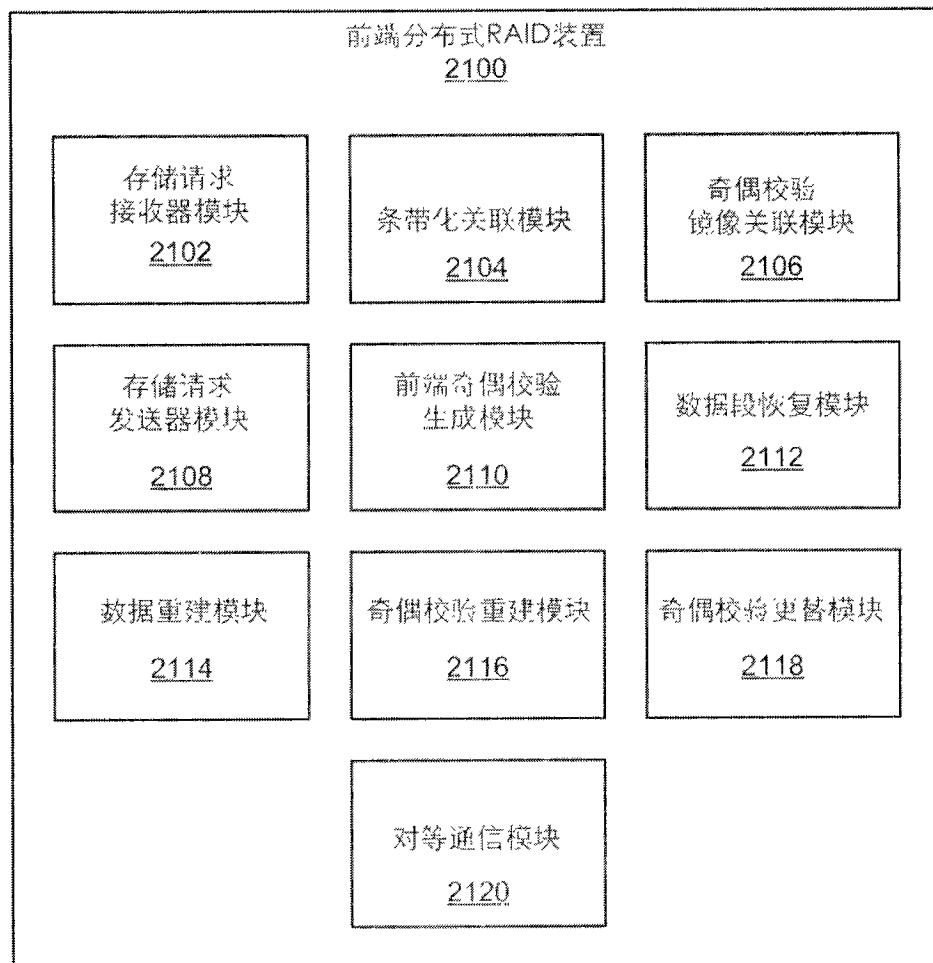


图 15

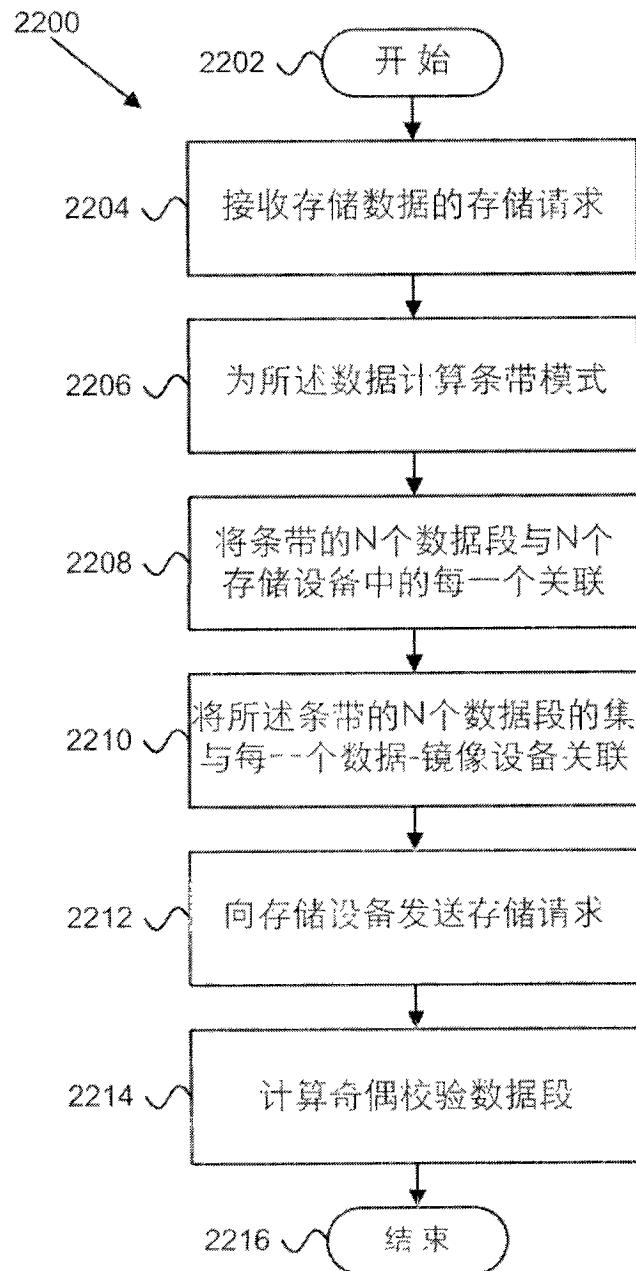


图 16

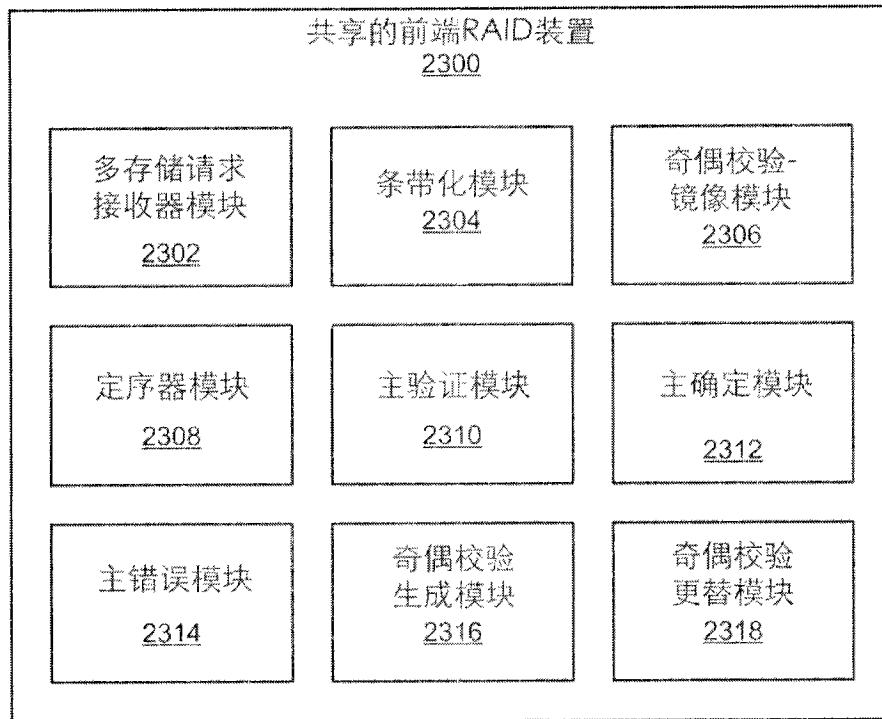


图 17

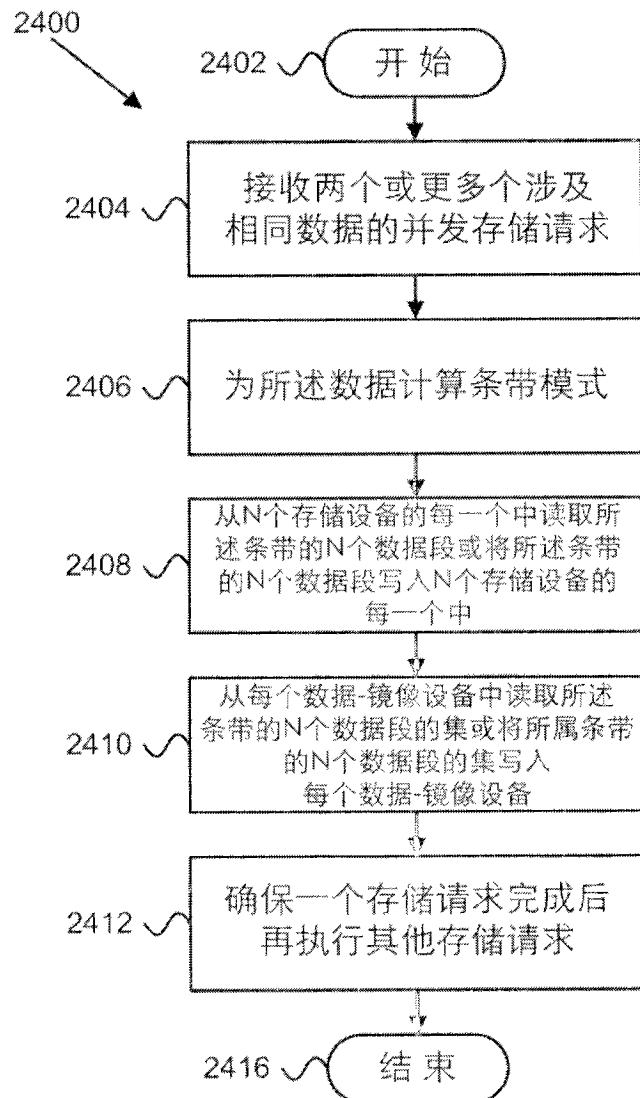


图 18