



[12] 发明专利申请公开说明书

[21] 申请号 200410062138.0

[43] 公开日 2004年12月8日

[11] 公开号 CN 1553346A

[22] 申请日 2004.7.2

[21] 申请号 200410062138.0

[30] 优先权

[32] 2003.7.2 [33] US [31] 60/483,926

[71] 申请人 普安科技股份有限公司

地址 台湾台北县

[72] 发明人 刘宁一 李泽涵 施明文 王源辉
包崇华

[74] 专利代理机构 北京市柳沈律师事务所

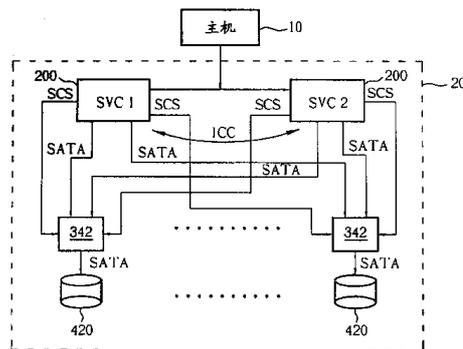
代理人 蒲迈文 黄小临

权利要求书 15 页 说明书 35 页 附图 45 页

[54] 发明名称 冗余外部储存虚拟化计算机系统

[57] 摘要

本发明提供一种冗余外部储存虚拟化计算机系统，其包含有：一主机，用来发出输出入请求；一冗余外部储存虚拟化控制器对，耦接至该主机，用以执行输出入操作，以响应该主机所发出的输出入请求；以及多个物理储存装置，用来提供储存空间给该计算机系统。其中，每一个物理储存装置经由点对点序列讯号连结耦接至冗余外部储存虚拟化控制器对，冗余外部储存虚拟化控制器对包含有一第一与一第二储存虚拟化控制器，其皆耦接至主机，在冗余外部储存虚拟化控制器对中，当第二储存虚拟化控制器离线时，第一储存虚拟化控制器将接替该第二储存虚拟化控制器原先所执行的功能。



1. 一计算机系统，包含有：

一主机，用来发出输出入请求；

5 一冗余外部储存虚拟化控制器对，用于执行输出入操作以响应该主机发出的输出入请求，其包括有耦接至该主机的一第一与一第二外部储存虚拟化控制器；以及

10 一组物理储存装置包含至少一物理储存装置，用来提供该计算机系统储存空间，该组至少一物理储存装置中的至少一成员，包括有一物理储存装置经由一点对点序列讯号连结耦接于该冗余储存虚拟化控制器对；

其中，当该冗余储存虚拟化控制器对中的一个储存虚拟化控制器未上线或者上线后又离线，则该冗余储存虚拟化控制器组中的另一个储存虚拟化控制器将自动地接替该冗余储存虚拟化控制器对中原先该个储存虚拟化控制器原先执行的功能。

15 2. 如权利要求 1 所述的计算机系统，其中该点对点序列讯号连结为一序列先进技术接取接口 (SATA) 输出入装置连结。

20 3. 如权利要求 1 或 2 所述的计算机系统，其中对该物理储存装置中至少一个而言，该计算机系统还包含有一存取控制开关，其耦接于每个该物理储存装置与该冗余储存虚拟化控制器对之间，用以选择切换该物理储存装置至该冗余储存虚拟化控制器对在该第一及该第二储存虚拟化控制器之间的连接。

4. 如权利要求 1 或 2 所述的计算机系统，其中，该冗余储存虚拟化控制器对中，每一该储存虚拟化控制器还包含有：

25 一中央处理电路，用于执行输出入操作以响应于该主机的该输出入请求；

至少一输出入装置连结控制器，耦接于该中央处理电路；

至少一主机端输出入装置连结端口，设置于该至少一输出入装置连结控制器的一个中，用来耦接至该主机；以及

30 至少一装置端输出入装置连结端口，设置于该至少一输出入装置连结控制器的一个中，用来经由该点对点序列讯号连结耦接至该至少一物理储存装置。

5. 如权利要求 4 所述的计算机系统, 其中该主机端输出装置连结端口中的一个与该装置端输出装置连结端口中的一个设置于同一个该输出装置连结控制器中。

6. 如权利要求 4 所述的计算机系统, 其中该主机端输出装置连结端口中的一个与该装置端输出装置连结端口中的一个设置于不同的该输出装置连结控制器中。

7. 一冗余储存虚拟化子系统, 用来提供一主机储存空间, 该冗余储存虚拟化子系统包含有:

一冗余外部储存虚拟化控制器对, 用来执行输出操作以响应于由该主机发出的输出请求, 其包括有用于耦接至该主机的一第一与一第二外部储存虚拟化控制器; 以及

一组物理储存装置包含至少一物理储存装置, 用来提供该主机储存空间, 该组至少一物理储存装置中至少一成员, 包括有一物理储存装置经由一点对点序列讯号连结耦接于该冗余储存虚拟化控制器对;

其中, 当该冗余储存虚拟化控制器对中的一个储存虚拟化控制器未上线或者上线后又离线, 则该冗余储存虚拟化控制器对中的另一个储存虚拟化控制器将自动地接替该冗余储存虚拟化控制器对中原先该个储存虚拟化控制器原先执行的功能。

8. 如权利要求 7 所述冗余储存虚拟化子系统, 其中该点对点序列讯号连结为一序列先进技术接取接口输出装置连结。

9. 如权利要求 7 或 8 所述的冗余储存虚拟化子系统, 其还包含有一附加于该冗余储存虚拟化子系统的可拆卸盒, 用以设置该至少一物理储存装置的一个于其中。

10. 如权利要求 7 或 8 所述的冗余储存虚拟化子系统, 其中对该物理储存装置中每一个而言, 该子系统还包含有一存取控制开关, 其耦接于该物理储存装置与该冗余储存虚拟化控制器对之间, 用以选择切换该物理储存装置至该冗余储存虚拟化控制器对在该第一及该第二储存虚拟化控制器间的连接。

11. 如权利要求 10 所述的冗余储存虚拟化子系统, 其还包含有一附加于该冗余储存虚拟化子系统的可拆卸盒, 用以设置该至少一物理储存装置的一个与该存取控制开关于其中。

12. 如权利要求 10 所述的冗余储存虚拟化子系统, 其中耦接于该物理储存装置与该冗余储存虚拟化控制器对之间的该存取控制开关, 可选择性地使该物理储存装置的序列讯号, 在该存取控制开关为第一配接状态时, 配接往返于该第一储存虚拟化控制器, 在该存取控制开关为第二配接状态时, 配接往返于该第二储存虚拟化控制器。

13. 如权利要求 12 所述的冗余储存虚拟化子系统, 其中还包含一存取所有权仲裁机制, 其设置于该储存虚拟化控制器对与该存取控制开关之间, 用于控制该存取控制开关的配接状态。

14. 如权利要求 13 所述的冗余储存虚拟化子系统, 其中该存取所有权仲裁机制包含有一存取所有权检测机制, 用以判定是否存取所有权为该储存虚拟化控制器其中之一所拥有。

15. 如权利要求 13 所述的冗余储存虚拟化子系统, 其中该存取所有权仲裁机制还包含有一存取所有权授予机制, 用于当该储存虚拟化控制器其中之一请求存取所有权时, 授予该存取所有权。

16. 如权利要求 10 所述的冗余储存虚拟化子系统, 还包含有:

一合作机制, 用以使该冗余储存虚拟化控制器对共同控制该存取控制开关的配接状态;

一监视机制, 用以使该储存虚拟化控制器对中每一该储存虚拟化控制器得以监视该储存虚拟化控制器对中另一个储存虚拟化控制器的状态; 以

20 及

一状态控制机制, 用以使该储存虚拟化控制器对中的每个储存虚拟化控制器在独立于该储存虚拟化控制器对中另一个储存虚拟化控制器的状态下, 得以强制取得该存取控制开关的完全的控制。

17. 如权利要求 7 或 8 所述的冗余储存虚拟化子系统, 其中, 在该冗余储存虚拟化控制器对中, 每个该储存虚拟化控制器还包含有:

一中央处理电路, 用于执行输出操作以响应于该主机的该输出请求;

至少一输出装置连结控制器, 耦接于该中央处理电路;

至少一主机端输出装置连结端口, 设置于该至少一输出装置连结控制器的一个中, 用来耦接至该主机; 以及

至少一装置端输出装置连结端口, 设置于该至少一输出装置连结

控制器的一个中，用来经由该点对点序列讯号连结耦接至该至少一物理储存装置。

5 18. 如权利要求 17 所述的冗余储存虚拟化子系统，其中该主机端输出装置连结端口中的一个与该装置端输出装置连结端口中的一个设置于同一个该输出装置连结控制器中。

19. 如权利要求 17 所述的冗余储存虚拟化子系统，其中该主机端输出装置连结端口中的一个与该装置端输出装置连结端口中的一个设置于不同的该输出装置连结控制器中。

10 20. 如权利要求 17 所述的冗余储存虚拟化子系统，其中还包括有一逻辑介质单元，经由该主机端输出装置连结端口中的一第一端口呈现于该主机上，该逻辑介质单元亦通过该主机端输出装置连结端口中的一第二端口冗余地呈现至该主机。

15 21. 如权利要求 20 所述的冗余储存虚拟化子系统，其中该第一主机端输出装置连结端口与该第二主机端输出装置连结端口为该冗余储存虚拟化控制对中同一储存虚拟化控制器上的输出装置连结端口。

20 22. 如权利要求 20 所述的冗余储存虚拟化子系统，其中该第一主机端输出装置连结端口为该冗余储存虚拟化控制对中的一个储存虚拟化控制器上的一输出装置连结端口；以及该第二主机端输出装置连结端口为该冗余储存虚拟化控制对中的另一个储存虚拟化控制器上的输出装置连结端口。

23. 如权利要求 20 所述的冗余储存虚拟化子系统，其中该第一主机端输出装置连结端口与该第二主机端输出装置连结端口耦接于相同的主机端输出装置连结上。

25 24. 如权利要求 23 所述的冗余储存虚拟化子系统，其中该第一主机端输出装置连结端口与该第二主机端输出装置连结端口经由一开关电路耦接至该同一的主机端输出装置连结上。

25. 如权利要求 20 所述的冗余储存虚拟化子系统，其中该第一主机端输出装置连结端口与该第二主机端输出装置连结端口各耦接至不同的主机端输出装置连结上。

30 26. 如权利要求 17 所述的冗余储存虚拟化子系统，其中至少一该主机端输出装置连结端口为下列的一个：于目标模式(target mode)时的支

持点对点连结的光纤信道，于目标模式时支持公用回路连结的光纤信道，于目标模式时支持专用回路连结的光纤信道，操作于目标模式的并列小型计算机系统接口(并列 SCSI)，操作于目标模式的支持因特网小型计算机系统接口(iSCSI)协议的以太网络，操作于目标模式的序列附加小型计算机系统接口(Serial-Attached SCSI, SAS)，以及操作于目标模式时的序列先进技术接取接口(SATA)。

27. 如权利要求 7 或 8 所述的冗余储存虚拟化子系统，其中还包括有一控制器间通讯信道，设置于该冗余储存虚拟化控制器对间，用以传递状态同步信息。

10 28. 如权利要求 27 所述的冗余储存虚拟化子系统，其中该控制器间通讯信道为一现存的输出装置连结，藉此，控制器间通讯交换是与输出请求以及关联数据一起多任务传输。

29. 如权利要求 27 所述的冗余储存虚拟化子系统，其中该控制器间通讯信道为一专用信道，及其主要功能为交换该状态同步信息。

15 30. 如权利要求 27 所述的冗余储存虚拟化子系统，其中该控制器间通讯信道为下列的一个：光纤信道，序列先进技术接取接口(SATA)信道，并列小型计算机系统接口(并列 SCSI)信道，以太网络，序列附加小型计算机系统接口(SAS)信道，以及集成电路间接口(I2C)通道。

20 31. 如权利要求 7 或 8 所述的冗余储存虚拟子系统，其中该冗余储存虚拟化控制器对可用以执行输出请求复位路径传送功能。

32. 如权利要求 7 或 8 所述的冗余储存虚拟子系统，其中该冗余储存虚拟化控制器对可用以执行物理储存装置存取所有权转移功能。

25 33. 如权利要求 7 或 8 所述的冗余储存虚拟化子系统，其中该冗余储存虚拟化控制器对的至少一成员储存虚拟化控制器包含有至少一扩充端口，经由一多装置装置端输出装置连结，耦接至一包含有至少一物理储存装置的第二组物理储存装置。

34. 如权利要求 33 所述的冗余储存虚拟化子系统，其中包含有至少一该扩充端口的一组扩充端口的各个成员彼此耦接在一起，且通过一开关电路连接至该第二组物理储存装置。

30 35. 如权利要求 33 所述的冗余储存虚拟化子系统，其中包含有至少一该扩充端口的一组扩充端口的各个成员彼此耦接在一起，且未通过中介

电路直接连接至该第二组物理储存装置。

36. 如权利要求 33 所述的冗余储存虚拟化子系统, 其中一包含有至少二该扩充端口的扩充端口组形成一冗余扩充端口组, 用以互相执行输出请求复位路径传送功能, 藉以使得正常是通过该冗余扩充端口组中的一第一成员端口而传送至一物理储存装置的输出请求可复位路径递送而通过该冗余扩充端口组中的一第二成员端口。

37. 如权利要求 33 所述的冗余储存虚拟化子系统, 其中该第二组物理储存装置中的一成员具有一对冗余端口, 以及该对冗余端口中的一成员端口耦接至一包含至少一该扩充端口的扩充端口组。

38. 如权利要求 37 所述的冗余储存虚拟化子系统, 其中通过该第二组物理储存装置中该成员的该冗余端口, 输出请求复位路径传送功能得以被执行, 藉以使得正常是经由该冗余端口对中的一第一成员端口而被传送给一物理储存装置的输出请求可复位路径递送而经由该冗余端口对中的一第二成员端口至该物理储存装置。

39. 如权利要求 38 所述的冗余储存虚拟化子系统, 其中一包含有至少二该扩充端口的扩充端口组形成一冗余扩充端口组, 用以互相执行输出请求复位路径传送功能, 藉以使正常是经由该冗余扩充端口组中的一第一成员端口而传送至一物理储存装置的输出请求可复位路径递送而经由该冗余扩充端口组中一第二成员端口。

40. 如权利要求 37 所述的冗余储存虚拟化子系统, 其中该物理储存装置冗余端口对中每个成员端口耦接至不同的扩充端口组, 该扩充端口组包含有至少一该扩充端口。

41. 如权利要求 37 所述的冗余储存虚拟化子系统, 其中该冗余物理储存装置端口对的该成员端口与该包含至少一该扩充端口的扩充端口组经由一开关电路互相耦合在一起。

42. 如权利要求 41 所述的冗余储存虚拟子系统, 其中该扩充端口组包含有一第一与一第二扩充端口次组, 以形成一对互补次组, 该每一次组包含有至少一成员扩充端口。

43. 如权利要求 42 所述的冗余储存虚拟化子系统, 其中藉由该开关电路而实现的连结讯号线开关机制之一, 为耦接该对互补次组的该第一次组至该物理储存冗余端口对的一第一成员端口, 以及耦接该对互补次组的

该第二次组至该物理储存冗余端口对的一第二成员端口。

44. 如权利要求 42 所述的冗余储存虚拟化子系统, 其中藉由该开关电路而实现的连结讯号线开关机制之一, 为耦接该对互补次组的二个次组至该物理储存冗余端口对的一第一成员端口。

5 45. 如权利要求 42 所述的冗余储存虚拟化子系统, 其中藉由该开关电路而实现的连结讯号线开关机制之一, 为耦接该对互补次组的该第一次组至该物理储存冗余端口对的一第一成员端口。

46. 如权利要求 42 所述的冗余储存虚拟化子系统, 其中该开关电路实现一连结讯号线的开关机制, 该机制支持下列各种安排:

10 (1) 耦接该对互补次组的该第一次组至该物理储存装置冗余端口对的一第一成员端口, 以及耦接该对互补次组的该第二次组至该物理储存冗余端口对的一第二成员端口;

(2) 耦接该对互补次组的二个次组至该物理储存冗余端口对的该第一成员端口;

15 (3) 耦接该对互补次组的二个次组至该物理储存冗余端口对的该第二成员端口;

(4) 耦接该对互补次组的该第一次组至该物理储存冗余端口对的该第一成员端口;

20 (5) 耦接该对互补次组的该第二次组至该物理储存冗余端口对的该第二成员端口;

(6) 耦接该对互补次组的该第二次组至该物理储存冗余端口对的该第一成员端口; 以及

(7) 耦接该对互补次组的该第一次组至该物理储存冗余端口对的该第二成员端口。

25 47. 如权利要求 37 所述的冗余储存虚拟化子系统, 其中该冗余物理储存装置端口对中的该成员端口与该包含至少一扩充端口的扩充端口组未经中介电路直接耦接在一起。

30 48. 如权利要求 37 所述的冗余储存虚拟化子系统, 其中该冗余储存虚拟化控制器对中的一成员储存虚拟化控制器还包含有至少二该扩充端口, 以形成一冗余扩充端口组。

49. 如权利要求 48 所述的冗余储存虚拟化子系统, 其中该冗余扩充

端口组中的一第一扩充端口与一第二扩充端口分别耦接至该包含至少一物理储存装置的第二组物理储存装置的一成员物理储存装置的冗余物理储存装置端口对中不同的成员端口。

50. 如权利要求 48 所述的冗余储存虚拟化子系统, 其中该冗余扩充端口组中的一第一扩充端口与一第二扩充端口皆耦接至该包含至少一物理储存装置的第二组物理储存装置的一成员物理储存装置的冗余物理储存装置端口对中相同的成员端口。

51. 如权利要求 50 所述的冗余储存虚拟化子系统, 其中该第一该第二扩充端口未经中介电路而直接连接至该包含至少一物理储存装置的第二组物理储存装置的一成员物理储存装置的冗余物理储存装置端口对中相同的成员端口。

52. 如权利要求 37 所述的冗余储存虚拟化子系统, 其还包含有:

一第一扩充端口组, 其包含至少一设置于该冗余储存虚拟化控制对中的第一储存虚拟化控制器上的扩充端口; 以及

一第二扩充端口组, 其包含至少一设置于该冗余储存虚拟化控制对中的第二储存虚拟化控制器上的扩充端口;

其中该第一扩充端口组与该第二扩充端口组一起形成一冗余扩充端口组对。

53. 如权利要求 52 所述的冗余储存虚拟化子系统, 其中该第一扩充端口组与该第二扩充端口组分别耦接至包含有至少一物理储存装置的该第二组物理储存装置中每一该物理储存装置的冗余物理储存装置端口对中不同的成员端口。

54. 如权利要求 52 所述的冗余储存虚拟化子系统, 其中该第一扩充端口组与该第二扩充端口组皆耦接至包含有至少一物理储存装置的该第二组物理储存装置中每一该物理储存装置的冗余物理储存装置端口对中相同的成员端口。

55. 如权利要求 33 所述的冗余储存虚拟化子系统, 其中至少一该扩充端口为下列的一个: 光纤信道, 并列小型计算机系统接口(并列 SCSI), 序列先进技术接取接口(SATA), 以太网络, 以及序列附加小型计算机系统接口(SAS)。

56. 如权利要求 7 或 8 所述的冗余储存虚拟化子系统, 其中该物理储

存装置为一序列先进技术接取接口物理储存装置。

57. 如权利要求 7 或 8 所述的冗余储存虚拟化子系统，其中该物理储存装置为一并列先进技术接取接口物理储存装置。

58. 如权利要求 7 或 8 所述的冗余储存虚拟化子系统，其中该冗余储存虚拟化控制器对的每一该储存虚拟化控制器包含有一状态定义电路，用以迫使该冗余储存虚拟化控制器对中的另一个储存虚拟化控制器的外部连接讯号线进入一预设状态。

59. 如权利要求 7 或 8 所述的冗余储存虚拟化子系统，其中该冗余储存虚拟化控制器对中每一该储存虚拟化控制器包含有一自杀电路，用以迫使其自身的外部连接讯号线进入一预设状态。

60. 一种应用于一冗余储存虚拟化控制器对中的外部储存虚拟化控制器，包含有：

一中央处理电路，用以执行输出操作以响应一主机的输出请求；
至少一输出装置连结控制器，耦接于该中央处理电路；

至少一主机端输出装置连结端口，设置于该至少一输出装置连结控制器的一个中，用来耦接至该主机；以及

至少一装置端输出装置连结端口，设置于该至少一输出装置连结控制器的一个中，用来耦接至至少一物理储存装置并与其执行点对点序列讯号传递；

其中，该外部储存虚拟化控制器，在该冗余储存虚拟化控制器对中的另一个外部储存虚拟化控制器未上线或者上线后又离线时，将自动地接替该另一个外部储存虚拟化控制器原先所执行的功能。

61. 如权利要求 60 所述的储存虚拟化控制器，其中该主机端输出装置连结端口中的一个与该装置端输出装置连结端口中的一个设置于同一个该输出装置连结控制器中。

62. 如权利要求 60 所述的储存虚拟化控制器，其中该主机端输出装置连结端口中的一个与该装置端输出装置连结端口中的一个设置于不同的该输出装置连结控制器中。

63. 如权利要求 60 所述的储存虚拟化控制器，其中该至少一装置端输出装置连结端口中之一包含有一序列先进技术接取接口连结端口，用以经由一序列先进技术接取接口输出装置连结与该至少一物理储存装

置的一个连接。

64. 如权利要求 60 或 63 所述的储存虚拟化控制器，还包含有一检测机制，用以察觉该另一个储存虚拟化控制器是否为离线或失效的状态。

5 65. 如权利要求 60 或 63 所述的储存虚拟化控制器，还包含有一状态定义电路，用以迫使该冗余储存虚拟化控制器对中的另一个储存虚拟化控制器的外部连接讯号线进入一预设状态。

66. 如权利要求 60 或 63 所述的储存虚拟化控制器，还包含有一自杀电路，用以迫使其自身的外部连接讯号线进入一预设状态。

10 67. 如权利要求 60 或 63 所述的储存虚拟化控制器，其中该功能包含：将原本该另一个储存虚拟化控制器所呈现给主机并使主机可使用的可存取资源呈现给主机并使主机可使用，以及将该储存虚拟化控制器自身所呈现给主机并使主机可使用的可存取资源呈现给主机并使主机可使用。

15 68. 如权利要求 60 或 63 所述的储存虚拟化控制器，其中还包含有一存取所有权仲裁机制，用以提供决定该储存虚拟化控制器对那一个储存虚拟化控制器拥有存取所有权。

69. 如权利要求 68 所述的储存虚拟化控制器，其中该存取所有权仲裁机制包含有一存取所有权检测机制，用以判定是否存取所有权为该储存虚拟化控制器所拥有。

20 70. 如权利要求 68 所述的储存虚拟化控制器，其中该存取所有权仲裁机制包含有一存取所有权授予机制，用于当该储存虚拟化控制器其中的一请求存取所有权时，授予存取所有权。

71. 如权利要求 60 或 63 所述的储存虚拟化控制器，还包含有：

25 一合作机制，用以与该储存虚拟化控制器对中的另一个储存虚拟化控制器一起去合作控制一存取控制开关的配接状态；
一监视机制，用以使该储存虚拟化控制器对中的该个储存虚拟化控制器得以监视该储存虚拟化控制器对该另一个储存虚拟化控制器的状态；
以及

30 一状态控制机制，用以使该储存虚拟化控制器对中的该储存虚拟化控制器在独立于该储存虚拟化控制器对中的该另一个储存虚拟化控制器的状态下，得以强制取得该存取控制开关的完全的控制。

72. 如权利要求 60 或 63 所述的储存虚拟化控制器，其中还包含有一

控制器间通讯端口,用以传递该冗余储存虚拟化控制器对中的该个储存虚拟化控制器与该另一个储存虚拟化控制器间的状态同步信息。

73. 如权利要求 72 所述的储存虚拟化控制器,其中该控制器间通讯端口为一现存的输出装置连结,藉此,控制器间通讯交换与输出请求以及关联数据一起多任务传输。

74. 如权利要求 72 所述的储存虚拟化控制器,其中该控制器间通讯端口为一专用端口,及其主要功能为交换该状态同步信息。

75. 如权利要求 72 所述的储存虚拟化控制器,其中该控制器间通讯端口为下列的一个: 光纤信道, 序列先进技术接取接口 (SATA), 并列小型计算机系统接口 (并列 SCSI), 以太网络, 序列附加小型计算机系统接口 (SAS), 以及集成电路间接口 (I2C)。

76. 如权利要求 60 或 63 所述的储存虚拟化控制器,其中该储存虚拟化控制器可用以执行输出请求复位路径传送功能。

77. 如权利要求 60 或 63 所述的储存虚拟化控制器,其中该储存虚拟化控制器可用以执行物理储存装置存取所有权转移功能。

78. 如权利要求 60 或 63 所述的储存虚拟化控制器,还包含有一扩充端口,用以经由一多装置装置端输出装置连结耦接至一包含有至少一物理储存装置的第二组物理储存装置。

79. 如权利要求 60 或 63 所述的储存虚拟化控制器,其中,至少一该主机端输出装置连结端口为下列的一个: 于目标模式时支持点对点连结的光纤信道,于目标模式时支持公用回路连结的光纤信道,于目标模式时的支持专用回路连结的光纤信道,操作于目标模式的并列小型计算机系统接口 (并列 SCSI), 操作于目标模式时支持因特网小型计算机系统接口 (iSCSI) 协议的以太网络,操作于目标模式的序列附加小型计算机系统接口 (SAS), 以及操作于目标模式时的序列先进技术接取接口 (SATA)。

80. 一种于一具有一第一与一第二外部储存虚拟化控制器并配置为一冗余储存虚拟化控制器对的计算机系统中执行储存虚拟化的方法,该方法包含:

响应该计算机系统的一主机端发出的输出请求,以该冗余储存虚拟化控制器对中的一个储存虚拟化控制器执行输出操作,并使用点对点序列讯号传递方式存取该计算机系统至少一物理储存装置; 以及

当在该冗余储存虚拟化控制器对中的该个储存虚拟化控制器未上线或上线后又离线时,则由该冗余储存虚拟化控制器对中的另一个储存虚拟化控制器执行该输出操作以响应该主机发出的该输出请求,并使用点对点序列讯号传递方式存取该计算机系统中的该至少一物理储存装置。

5 81. 如权利要求 80 所述的方法,其中该点对点序列讯号传输遵循 SATA 协议。

82. 如权利要求 80 或 81 所述的方法,其中当该一储存虚拟化控制器未上线或者上线后又离线时,则该另一个储存虚拟化控制器将自动接替该个储存虚拟化控制器原先所执行的功能。

10 83. 如权利要求 82 所述的方法,其中该功能包含:将原本该个储存虚拟化控制器所呈现给主机并使主机可使用的可存取资源呈现给主机并使主机可使用,以及将该另一个储存虚拟化控制器自身所呈现给主机并使主机可使用的可存取资源呈现给主机并使主机可使用。

15 84. 如权利要求 80 或 81 所述的方法,其还包含有提供一复位路径传送机制,用以使该冗余储存虚拟化控制器对得以执行输出请求复位路径传送的功能。

85. 如权利要求 84 所述的方法,其中该输出请求复位路径传送功能由下列步骤所执行:

20 该冗余储存虚拟化控制对中的一请求启动者传送一输出请求给该冗余储存虚拟化控制器对中一存取权所有者;以及

该存取权所有者执行该请求启动者所传送的该输出请求。

86. 如权利要求 85 所述的方法,其中还包含有在该存取权所有者与该物理储存装置间传递而与该输出请求相关的讯息与数据的至少一部份,被代传至该请求发动者。

25 87. 如权利要求 80 或 81 所述的方法,其还包含有下列步骤:

提供一存取控制开关,其耦接于该至少一物理储存装置中的一与该外部冗余储存虚拟化控制器对之间,用以选择性地使该物理储存装置的序列讯号,于该存取控制开关为第一配接状态时,配接往返于该第一储存虚拟化控制器,于该存取控制开关为第二配接状态时,配接往返于该第二储存虚拟化控制器。

30

88. 如权利要求 80 或 81 所述的方法,还包含有提供一存取所有权转

移机制，用以使该冗余储存虚拟化控制器对中该个储存虚拟化控制器与该另一个储存虚拟化控制器合作移转存取所有权往返于其间。

89. 如权利要求 88 所述的方法，其中该储存虚拟化控制对间与存取所有权转移相关的信息交换，是通过控制器间通讯的一部份进行。

5 90. 如权利要求 88 所述的方法，其中该存取所有权转移机制包含下列步骤：

 (a) 存取权所有者决定释出该存取所有权并移转所有权给该冗余储存虚拟化控制器对的另一个储存虚拟化控制器；

10 (b) 该存取权所有者释出该存取所有权，因此其不再为存取权所有者；以及

 (c) 该另一个储存虚拟化控制器取得该存取所有权，并成为该物理储存装置的新的存取权所有者。

15 91. 如权利要求 90 所述的方法，其中该 (a) 决定释出该存取所有权与转移所有权给该另一个储存虚拟化控制器的步骤是由取得一存取请求者角色的该另一个储存虚拟化控制器所触发，其是藉由发出一存取权请求指示给该存取权所有者以请求存取所有权转移而触发。

20 92. 如权利要求 90 所述的方法，其中，还包含有以下的步骤：在存取权所有者决定释出该存取所有权并移转所有权给该另一个储存虚拟化控制器之后，以及在该存取权所有者释出该存取所有权之前，该存取权所有者将新的输出请求放入队列中使其之后再执行，并且做完所有正在进行的输出请求。

 93. 如权利要求 90 所述的方法，其中该存取权所有者释出该存取所有权的步骤，包含有改变一耦接于该储存虚拟化控制器对与该物理储存装置之间的存取控制开关的状态。

25 94. 如权利要求 90 所述的方法，其中该存取请求者取得该存取所有权的步骤，包含有改变一耦接于该储存虚拟化控制器对与该物理储存装置之间的存取控制开关的状态。

30 95. 一种执行储存虚拟化于一具有一第一与一第二外部储存虚拟化控制器而配置为一冗余储存虚拟化控制器对的计算机系统的方法，该方法包含：

 响应自该计算机系统的一主机端发出的输出请求，以该冗余储存

虚拟化控制器对中的一个储存虚拟化控制器执行输出操作以存取该计算机系统中至少一物理储存装置；以及

5 当在该冗余储存虚拟化控制器对中的该个储存虚拟化控制器未上线或上线后又离线时，则藉由该冗余储存虚拟化控制器对中的另一个储存虚拟化控制器执行该输出操作以响应该主机发出的该输出请求，以存取该计算机系统中该至少一物理储存装置；

10 其中，通过一冗余输出装置连结端口对执行输出请求复位路径传送功能，此冗余输出装置连结端口对包括：一位于该第一储存虚拟化控制器的第一输出装置连结端口，以及一位于该第二储存虚拟化控制器的第二输出装置连结端口。

96. 如权利要求 95 所述的方法，其中当该个储存虚拟化控制器未上线或上线后又离线时，该另一个储存虚拟化控制器将接替该个储存虚拟化控制器原先执行的功能。

15 97. 如权利要求 95 所述的方法，其中该输出请求复位路径传送功能由下列步骤所执行：

该冗余储存虚拟化控制器对中的一请求启动者传送一输出请求给该冗余储存虚拟化控制器对中的一存取权所有者；以及

该存取权所有者执行该请求启动者所传送的该输出请求。

20 98. 如权利要求 97 所述的方法，其中还包含有于该存取权所有者与该物理储存装置间传递而与该输出请求关联的讯息与数据的至少一部份，被代传至该请求启动者。

99. 如权利要求 95 所述的方法，其还包含有提供一存取所有权转移机制，用以使该冗余储存虚拟化控制器对中该个储存虚拟化控制器与该另一个储存虚拟化控制器合作移转存取所有权往返于其间。

25 100. 如权利要求 99 所述的方法，其中该储存虚拟化控制器对间与存取所有权转移相关的信息交换，是通过控制器间通讯的一部份进行。

101. 如权利要求 99 所述的方法，其中该存取所有权转移机制包含下列步骤：

30 (a) 存取权所有者决定释出该存取所有权并移转所有权给该冗余储存虚拟化控制器对中的另一个储存虚拟化控制器；

(b) 该存取权所有者释出该存取所有权，因此其不再为存取权所有

者；以及

(c) 该另一个储存虚拟化控制器取得该存取所有权，并成为该物理储存装置的新的存取权所有者。

5 102. 如权利要求 101 所述的方法，其中该(a)决定释出该存取所有权与转移所有权给该另一个储存虚拟化控制器的步骤是由取得一存取请求者角色的该另一个储存虚拟化控制器所触发，其是藉由发出一存取权请求指示给该存取权所有者以请求存取所有权转移而触发。

10 103. 如权利要求 101 所述的方法，还包含有以下的步骤：在存取权所有者决定去释出该存取所有权并移转所有权给该另一个储存虚拟化控制器之后，以及在该存取权所有者释出该存取所有权之前，该存取权所有者将新的输出请求放入队列中使其之后再执行，并且做完所有正在进行的输出请求。

冗余外部储存虚拟化计算机系统

5 技术领域

本发明涉及一冗余外部储存虚拟化计算机系统 (redundant external storage virtualization computer system) 相关, 特别是涉及一种利用点对点序列讯号连结作为主要装置端输出入 (IO) 装置连结的冗余外部储存虚拟化计算机系统。

10

背景技术

所谓储存虚拟化 (storage virtualization) 是一种将物理储存空间虚拟化的技术, 其是将物理储存装置 (PSD, physical storage devices) 的不同区段结合成可供一主机系统存取使用的逻辑储存体 (logical storage entity) - 在此称为「逻辑介质单元」(LMU, logical media unit)。
15 该技术主要用于磁盘阵列 (RAID) 储存虚拟化, 经由此磁盘阵列的技术, 可将较小物理储存装置结合成为容量较大、可容错、高效能的逻辑介质单元。

储存虚拟化控制器 (SVC, storage virtualization controller) 的主要目的是将物理储存介质的各区段的组合映像 (map) 形成一主机系统可见的逻辑介质单元。由该主机系统发出的输出入 (IO) 请求于接收后会被先被剖析并解译, 且相关的操作及数据会被编译成物理储存装置的输出入请求。这个过程可以是间接地, 例如运用快取、延迟 (如: 回写 (write-back))、
20 预期 (anticipate) (先读 (read-ahead))、群集 (group) 等操作来加强效能及其它的操作特性, 因而一主机输出入请求并不一定是一对一的方式直接对应于物理储存装置输出入请求。
25

外部 (或可称为独立式 (stand-alone)) 储存虚拟化控制器为一种经由输出入接口连接于主机系统的储存虚拟化控制器, 且其可连接至位于主机系统外部的装置, 一般而言, 外部储存虚拟化控制器通常是独立于主机进行运作。

30 外部 (或独立式) 直接存取磁盘阵列控制器 (external direct-access RAID controller) 是外部储存虚拟化控制器的一个例子。磁盘阵列控制器

是将一个或多个物理储存装置的区段组合以构成逻辑介质单元，而它们的构成方式由所采用的特定磁盘阵列型态 (RAID level) 决定，其所构成的逻辑介质单元对于主机系统而言，为可连续寻址的，以使每一逻辑介质单元可被利用。典型地，一个单一的磁盘阵列控制器 (single RAID controller) 可支持多种磁盘阵列型态，因此，不同的逻辑介质单元可以由物理储存装置的各个区段藉由不同的磁盘阵列型态而以不同的方式组合而成，所组合成的各个不同的逻辑介质单元则具有各该磁盘阵列型态的特性。

另一个外部储存虚拟化控制器的例子是 JBOD (Just a Bunch of Drives) 模拟控制器。JBOD 为「仅是一捆盘机」的缩写，是一组物理直接存取储存装置，并经由一个或多个多装置输出入装置连结信道 (multiple-device I/O device interconnect channel) 直接连接于一主机系统上。但使用点对点输出入装置连结连接至该主机系统的物理储存装置 (如 SATA 硬盘、PATA 硬盘等)，无法通过直接连结而构成如前述的 JBOD 系统，因为这些直接存取储存装置并不允许多个装置直接连接至输出入装置信道。至于智能型的 JBOD 仿真器，是藉由将输出入请求映像到物理直接存取储存装置的方式，而用来仿真多个多装置输出入装置连结直接存取储存装置，而其中该物理直接存取储存装置是个别地经由点对点输出入装置连结信道连接至 JBOD 仿真器。

另一个外部储存虚拟化控制器的例子为一种外部磁带备份子系统用的控制器。

将一对外部储存虚拟化控制器配置成一冗余对的主要动机是为了即使是在单个储存虚拟化控制器发生故障或是失效的情形下，主机依旧可以连续不中断的执行数据存取工作，此是可利用在该储存虚拟化控制器中加入一功能以使得其中一个控制器发生障碍或完全失能的情形下另一个控制器可接管其工作而实现。

在装置端，对受该储存虚拟化控制器管理的所有物理储存装置，此两个控制器必须皆能对其进行存取，而不论此物理储存装置原本是否被指定由其所管理。在主机端，则必须使每一个储存虚拟化控制器，即便是在它的同伴储存虚拟化控制器 (mate) 处于原本就没上线或上线后又因故下线的情况下 (例如像是故障/失效、维护操作等所造成的下线)，具有将所有

可供存取的资源呈现于主机且使该资源可被主机所利用的能力，这些可存取的资源也包括原来就指定由另一个储存虚拟化控制器所管理的资源。

在上述装置端，一代表性的实施方式，是采用多启动者 (multiple-initiator)、多装置 (multiple-device) 类型的装置端输出入装置连结 (如：光纤、并列小型计算机系统接口 (并列 SCSI, parallel small computer system interface)，而所有的装置端输出入装置连结皆连接至此两个储存虚拟化控制器，所以此二储存虚拟化控制器中任一个皆可存取连接于装置端输出入装置连结上的任何一个物理储存装置。当此二储存虚拟化控制器皆为在线操作时，每个物理储存装置将由其中一储存虚拟化控制器或另一储存虚拟化控制器管理，至于由谁管理通常是由使用者的设定或系统组态来决定，例如：对一由物理储存装置的磁盘阵列 (RAID) 组合所构成的逻辑介质单元，此逻辑介质单元中的所有物理储存装置，将由此逻辑介质单元所指定的特定储存虚拟化控制器所管理。

一典型的装置端实施方案，包含有多装置输出入装置连结，其连接至该主机与该二储存虚拟化控制器，并就每一个连结而言，每个虚拟储存控制器将呈现它自己所独有的一组映像至逻辑介质单元的装置识别码 (device ID)。当其中一个特定的储存虚拟化控制器未上线或是下线时，则在线的储存虚拟化控制器会在主机端连结上呈现两组装置识别码，其中一组为自己所拥有，另一组为在正常状态下是指定给其同伴 (mate) 的，且此在线的储存虚拟化控制器亦将映像逻辑介质单元至这些识别码，其映像的方式与此二储存虚拟化控制器皆在线且完全可操作时所采用的映射方式相同。在此一实施方案中，即使当一储存虚拟化控制器未上线时，主机并不需要特殊功能，来从一个装置/路径转换至另一个，即可继续对所有逻辑介质单元进行存取，这种实施方式通常可以称为「透明冗余 (transparent redundancy)」。

冗余储存虚拟化控制器的组态分为两类，第一类是主动-待命模式 (active-standby)，在此模式中，其中一个储存虚拟化控制器对储存虚拟化子系统中的所有逻辑介质单元的所有输出入请求进行呈现、管理及处理，而另一储存虚拟化控制器将仅是处于待命状态 (stand by)，而于主动储存虚拟化控制器发生障碍或失能时，随时接替主动储存虚拟化控制器。第二种是主动-主动模式 (active-active)，在此模式中，此两个储存虚拟

化控制器对同时呈现于此储存虚拟化子系统中的各种逻辑介质单元的输入请求进行呈现、管理及处理。在主动-主动模式中，上述二储存虚拟化控制器一直都准备在另一个储存虚拟化控制器因故障(malfunction)而导致发生障碍或失能的情况下接管对方。主动-主动模式，通常提供较好的效能，因为其两个储存虚拟化控制器的资源(例如：中央处理器(CPU, central processing unit)时间、内部总线频宽...等)与单一储存虚拟化控制器相比可负荷较多的输出请求服务。

在冗余储存虚拟化系统中的另外一个基本要件就是每个储存虚拟化控制器需能监视另一个的状态。此可利用一介于此二储存虚拟化控制器间的控制器间通讯信道(ICC, inter-controller communications channel)来完成，是利用此通道来交换该二储存虚拟化控制器的操作状态。此通讯信道可以是专用的，其唯一的作用就是交换与此冗余储存虚拟化子系统的操作相关的参数与数据。或者此通讯信道是一个或多个主机端或装置端的输出装置连结，经由此种连结，这些操作上的参数、数据交换可连同在这些连结上的主机-储存虚拟化控制器或装置-储存虚拟化控制器输出请求所关联的数据一起被多任务传输。

此外，冗余储存虚拟化系统的另一个重要要件是一个储存虚拟化控制器要能使另一个储存虚拟化控制器“完全失能”(completely incapacitate)，如此，此储存虚拟化控制器才能不受干扰的完全接替另一个储存虚拟化控制器。举例来说，为了让此一存活的储存虚拟化控制器承接它同伴的身分，其可能需要承接该离线的储存虚拟化控制器其原先呈现于主机端输出装置连结上的装置识别码，这接下来，将要求该离线的储存虚拟化控制器释出对这些识别码的控制。

上述的“失能作用”要件，典型地是藉由拉启该离线控制器的重置讯号线来实现，拉启重置讯号线将导致所有外部连接讯号线进入一预设状态，以排除对该存活的储存虚拟化控制器可能的干扰。为了实现前述，一种常用的方法是在储存虚拟化控制器之间连结重置讯号线，如此一来一储存虚拟化控制器就能重置另一储存虚拟化控制器。另外一种方法是利用储存虚拟化控制器建立一自我检测的能力的方式来实现，当储存虚拟化控制器自己发生故障时，其可以藉由拉启自我的重置讯号线来自杀(例如：其包含一监视定时器(watchdog timer)，若于一预设区间，在储存虚拟化控制

器上所执行的程序对此监视定时器的轮询发生失败时，此监视定时器将拉启一重置讯号)，而使所有外部连接讯号线进入一预设状态而可排除对存活的储存虚拟化控制器干扰的可能性。

5 传统上，储存虚拟化采用并列 SCSI 或是光纤输出装置连结作为主要装置端输出装置连结，用来连接物理储存装置与储存虚拟化控制器对，其中并列 SCSI 或是光纤都是多装置输出装置连结，多装置输出装置连结是一种允许多个输出装置被直接连接至单一或多个主机系统的输出装置连结，这表示其无须任何额外的装置外介入主动电路(请注意典型的光纤信道仲裁循环(FC-AL) JBOD 虽具有装置外主动电路，但此电路的目的并非为了赋予互相连结功能，而是为适应可能发生在直接存取储存装置上的故障或是抽换可能在输出装置连结上造成毁灭性损坏的直接存取储存装置)。多装置输出装置连结中常见的例子是光纤信道 FC-AL 与
10 并列 SCSI。多装置输出装置连结的频宽是由藉该连结而互连的所有主机及所有装置所共享。

15 请参阅图 1，图 1 为一已知冗余外部储存虚拟化计算机系统的方块示意图。请注意，其中主机端输出装置连结允许一储存虚拟化控制器接替其同伴，藉由接收其同伴正常时呈现于此连结上的输出装置连结的识别码以及以其同伴相同的方式映像逻辑介质单元至这些识别码。另外需注意的是，该装置端输出装置连结允许该二储存虚拟化控制器存取所有连接
20 至此装置端输出装置连结上的物理储存装置，举例来说：一典型的输出装置连结，可能用于主机端或装置端，为并列 SCSI 或光纤 FC-AL 多启动者、多装置输出装置连结，因此，操作在目标模式(即装置模式)下的此二储存虚拟化控制器都连接至该主机上的一单一连结，且允许操作在启动者模式的此二储存虚拟化控制器连同多个装置皆连结至装置端。然而图
25 1 中所示的架构的缺点是，当单一个物理储存装置故障，视其故障的本质为何，将可能导致整个装置端输出装置连结失效而使所有其它连接至此相同连结上的物理储存装置无法被存取。

图 2 显示一种改进方案，可有效避免由于一导致单一装置端连结失败的故障而使连接于该相同装置端输出装置连结的其它物理储存装置的
30 存取中断，其是藉由使用双端口物理储存装置以及对每一个物理储存装置增加一额外的连结而实现。用这种方法，单一装置端输出装置连结的阻

断(可能是因一物理储存装置上的连结控制 IC 故障而造成), 将不会造成其它连接于此相同连结上的其它物理储存装置无法存取, 因为连接至每个物理储存装置的第二连结将能被用来不受干扰地存取这些物理储存装置。

图 2 所示的架构还具有一优点, 就是输出请求负载能被分散于该冗余装置端连结间, 因而, 与图 1 所示的每组物理储存装置“单一”连结的结构相比, 图 2 藉此可有效加倍装置端输出装置连结子系统的整体频宽。在此情况下, 一特选的装置端输出装置连结, 典型地为光纤 FC-AL, 因为目前市售光纤 FC-AL 物理储存装置的双端口性质, 以及此光纤协议的组成允许一启动者(例如储存虚拟化控制器)去决定在不同连结上的那一些连结识别码对应至相同的物理储存装置。

虽然图 2 中所示的架构在面对装置端输出装置连结失效的问题上, 远比图 1 中所示的确实更为强健, 但其仍存有连接于一物理储存装置的故障使得连接至其双端口端口对上的两条输出装置连结同时当机的可能性。若发生此情况, 再一次, 连接在相同连结对上的其它物理储存装置的存取将被中断。且在由物理储存介质的标准单一冗余磁盘阵列(RAID)组合所构成的逻辑介质单元(如 RAID5)中, 这将是一场大灾难, 因为其可造成在此组合中的多个物理储存装置离线而导致整个逻辑储存单元离线。

发明内容

本发明的主要目的是提供一种冗余外部储存虚拟化计算机系统, 此计算机系统使用点对点序列讯号传输作为主要装置端输出装置连结, 以解决上述问题。

本发明披露一种冗余外部储存虚拟化计算机系统, 包含有: 一主机, 用来发送输出请求(I/O request); 一冗余储存虚拟化控制器对, 其耦接至主机, 用来执行输出操作以响应主机的输出请求; 多个物理储存装置用来提供此计算机系统储存空间。其中每个物理储存装置以点对点序列讯号连结, 耦接至冗余储存虚拟化控制器对, 此冗余储存虚拟化控制器对包含有耦接至主机的第一与第二储存虚拟化控制器。在冗余储存虚拟化控制器对中, 当第一储存虚拟化控制器未上线或是未运作时, 第二储存虚拟化控制器将接替此第一储存虚拟化控制器原先所执行的功能。在本发明的一实施例中此点对点序列讯号连结为一序列先进技术接取接口(SATA,

serial advanced technology attachment) 输出装置连结。

本发明的优点之一是在所提供的冗余外部储存虚拟化计算机系统中使用 SATA 为主要装置端输出装置连结，且每个物理储存装置都有其专用连结至该储存虚拟化控制器对。

- 5 本发明另一优点是该 SATA 输出装置连结不仅保护信息的有效负载数据的部份，还可保护控制信息。

本发明进一步披露了一种冗余储存虚拟化子系统，用来提供一主机储存空间，该冗余储存虚拟化子系统包含有：一冗余外部储存虚拟化控制器对，用来执行输出操作以响应于由该主机发出的输出请求，其包括有耦接至该主机的一第一与一第二外部储存虚拟化控制器；以及一组物理储存装置包括至少一物理储存装置，用来提供该主机储存空间，该组物理储存装置中的至少一成员包括有一物理储存装置经由点对点序列讯号连结耦接于该冗余储存虚拟化控制器对；其中，当该冗余储存虚拟化控制器对中的一个储存虚拟化控制器未上线或者上线后又下线，则该冗余储存虚拟化控制器对中的另一个储存虚拟化控制器将自动地接替该冗余储存虚拟化控制器对中该个储存虚拟化控制器原先执行的功能。

本发明还披露了一种应用于一冗余储存虚拟化控制器对中的外部储存虚拟化控制器，包含有：一中央处理电路，用以执行输出操作以响应一主机的输出请求；至少一输出装置连结控制器，耦接于该中央处理电路；至少一主机端输出装置连结端口，设置于该至少一输出装置连结控制器的一个中，用来耦接至该主机；以及至少一装置端输出装置连结端口，设置于该至少一输出装置连结控制器的一个中，用来耦接至至少一物理储存装置并执行的执行点对点序列讯号传递；其中一个该外部储存虚拟化控制器，在另一个外部储存虚拟化控制器未上线或者上线后又离线时，将接替该冗余储存虚拟化控制器组中该另一个外部储存虚拟化控制器原先所执行的功能。

本发明还披露了一种在一具有一第一与一第二外部储存虚拟化控制器并配置为一冗余储存虚拟化控制器对的计算机系统中执行储存虚拟化方法，该方法包含：响应该计算机系统的一主机端发出的输出请求，以该冗余储存虚拟化控制器组中的一个储存虚拟化控制器执行输出操作，并使用点对点序列讯号传递方式至该计算机系统内的至少一物理储存

装置；以及当在该冗余储存虚拟化控制器组中的该个储存虚拟化控制器未上线或上线后又下线，则藉由该冗余储存虚拟化控制器对中的另一个储存虚拟化控制器执行该输出操作以响应该主机发出的该输出请求，并使用点对点序列讯号传递方式存取该计算机系统中该至少一物理储存装置。

- 5 本发明还披露了一种执行储存虚拟化于一具有一第一与一第二外部储存虚拟化控制器而配置为一冗余储存虚拟化控制器对的计算机系统的方法，该方法包含：响应自该计算机系统的一主机端发出的输出请求，以该冗余储存虚拟化控制器对中的一个储存虚拟化控制器执行输出操作以存取该计算机系统中至少一物理储存装置；以及当在该冗余储存虚拟化控制器组中的该一储存虚拟化控制器未上线或上线后又下线时，则藉由该冗余储存虚拟化控制器组中另一个的储存虚拟化控制器执行该输出操作以响应该主机发出的该输出请求，以存取该计算机系统中该至少一物理储存装置；其中以一冗余输出装置连结端口对执行输出请求复位路径传送功能，此冗余输出装置连结端口对包括：一位于该第一储存虚拟化控制器的第一输出装置连结端口，以及一位于该第二储存虚拟化控制器的第二输出装置连结端口。
- 10
- 15

- 前述本发明所提供的储存虚拟化的方法，其执行过程可以藉由软件程序完成，因此本发明可以以计算机语言撰写程序后再加载一计算机可读取记录介质中，该记录介质可以是 IC 芯片、硬盘、光盘或其它可记录软件程序的物品。
- 20

附图说明

- 图 1 为一传统冗余外部储存虚拟化计算机系统的方块图。
- 图 2 为另一传统冗余外部储存虚拟化计算机系统的方块图。
- 25 图 3 为依据本发明的一冗余外部储存虚拟化计算机系统的方块图。
- 图 4 为利用一条讯号线来控制状态的存取控制开关的方块图。
- 图 5 为一利用两条讯号线来控制状态的存取控制开关的方块图。
- 图 6 为依据本发明的储存虚拟化控制器的方块图。
- 图 7 为图 6 中所示的中央处理电路的一实施例的方块图。
- 30 图 8 为图 7 中所示的中央处理芯片组/奇偶性引擎的一实施例的方块图。

- 图 9 为图 6 中所示的 SATA 输出装置连结控制器的方块图。
- 图 10 为图 9 中所示的 PCI-X 转 SATA 控制器的方块图。
- 图 11 为图 10 中所示的 SATA 端口的方块图。
- 图 12 为一符合 SATA 协议的传输架构的示意图。
- 5 图 13 为一符合 SATA 协议的第一帧信息结构的示意图。
- 图 14 为一符合 SATA 协议中第二帧信息结构的示意图。
- 图 15 为一转换程序的流程图。
- 图 16 为使用二元讯号对存取所有权仲裁机制进行转换程序的流程图。
- 图 17 为使用另一个二元讯号对存取所有权仲裁机制进行转换程序的
- 10 时序图。
- 图 18 为图 17 中所示的转换程序的流程图。
- 图 19 为当储存虚拟化控制器组中有一储存需拟化控制器故障时的强制转换程序的流程图。
- 图 20 为一输出请求路径的流程图。
- 15 图 21 为一冗余储存虚拟化控制器连结扩充端口实施例的方块示意图。
- 图 22 为一显示如何使用硬件开关而完成可切换连结的方块图。
- 图 23 显示一种依赖硬件讯号检测以启动此开关状态的变更的电路设计。
- 图 24 显示一种从第一储存虚拟化控制器 (SVC1) 与第二储存虚拟化控
- 20 制器 (SVC2) 各自取得输入讯号 C1 与 C2, 以触发此开关状态的变更的电路设计。
- 图 25 显示一种图 23 与图 24 中所示电路的混合型电路的示意图。
- 图 26 为一冗余储存虚拟化控制器连结冗余扩充端口实施例方块示意图。
- 25 图 27 为另一冗余储存虚拟化控制器连结冗余扩充端口实施例的方块示意图。
- 图 28 为利用一硬件开关去连结如图 27 所示的连接于二储存虚拟化控制器的二个输出装置连结的实施例的方块图。
- 图 29 显示一种依赖硬件讯号检测以启动图 28 所示的硬件开关的开关
- 30 状态变更的电路设计的示意图。
- 图 30 显示一种从第一储存虚拟化控制器 (SVC1) 与第二储存虚拟化控

制器(SVC2)各自取得输入讯号 C1 与 C2, 以触发图 28 所示的硬件开关的开关状态变更的电路设计的示意图。

图 31 显示一种图 29 与图 30 中所示电路的混合型电路。

5 图 32 为在冗余储存虚拟化控制器连结扩充端口上输出请求路径的流程图。

图 33 为一冗余外部储存虚拟化计算机系统的示意图, 其包含每个储存虚拟化控制器中的两个个别的主机端端口, 连接至两个完全分开的主机端输出装置连结以及主机端口。

10 图 34 为利用一开关电路完成图 33 中所示的主机端连结的方块示意图。

图 35 为一冗余外部储存虚拟化计算机系统的方块示意图, 其中每一储存虚拟化控制器上都有一主机端端口, 用以连接到一主机端输出装置连结与主机端口。

图 36 为可拆卸冗余 SATA 物理储存装置盒的方块图。

15 图 37 为图 36 所示的可拆卸盒中特有的印刷电路板的较详细的方块图。

图 38 为一可拆卸冗余 PATA 物理储存装置盒的方块图。

图 39 为图 38 所示的可拆卸盒中特有的印刷电路板的较详细的方块图。

20 图 40 为图 4 所示控制开关活动的真值表。

图 41 为图 5 所示控制开关活动的真值表。

图 42 为当发生故障情况时, 用以改变循环通联路径的图表。

图 43 为图 29 中所示的电路的真值表。

图 44 为图 30 中所示的电路的真值表。

25 图 45 为图 31 中所示的电路的真值表。

附图符号说明

10	主机	20	SVS
200	SVC	220	主机端输出装置连结控制器
236	RCC 连结控制器	240	中央处理电路

242	CPU	244	中央处理芯片组 / 奇偶性引擎
246	ROM	248	NVRAM
262	XOR 引擎	264	XOR FIFO
280	存储器	300	SATA 输出入装置连结控制器
310	PCI-X 转 SATA 控制器	312	PCI-X 接口
314	解 / 多任务仲裁器	316	组态电路
318	总线接口	340	开关电路
342	存取控制开关	350	LCD 模块
360	EMS	420	物理储存装置
600	SATA 端口	620	DMA 寄存器
630	超集寄存器	640	指令区块寄存器
650	控制区块寄存器	660	双端口先进先出缓冲器
670	DMA 控制器	680	PIO
690	传输层	700	连结层
710	物理层	910	CPU 接口
920	存储器接口	922	CM FIFO
924	除错码产生电路	926	除错码修正电路
930、 932	PCI 接口	934 、 936	PM FIFO
940	X-总线接口	950	PM 总线
980	锁相回路	982	计时控制器
984	内部寄存器	986	UART

具体实施方式

请参阅图 3, 图 3 为本发明的一实施例的方块示意图, 此系统包含有一主机 10 以及一冗余储存虚拟化子系统 20 (SVS, redundant storage virtualization subsystem)。冗余储存虚拟化子系统 20 包含有一冗余储存虚拟化控制器对 (包括第一与第二储存虚拟化控制器 (SVC1, SVC2) 200, 与多个物理储存装置 420。此冗余储存虚拟化子系统的架构包含有专用的点对

点输出入装置连结，用来连接所有的物理储存装置 420 至此二储存虚拟化控制器 200，其中储存虚拟化控制器 200 可为一磁盘阵列控制器或是一个 JBOD 仿真器。

5 虽然图 3 中所示仅有一主机 10 与一储存虚拟化子系统 20 相互连接，实际应用时可用多个主机 10 连接一个储存虚拟化子系统 20，或是一主机 10 连接多个储存虚拟化子系统 20，或是多个主机 10 连接多个储存虚拟化子系统 20。主机 10 可为一主机计算机，如一服务器系统、工作站、个人计算机系统或是其它相关计算机等，而且主机 10 也可为另一储存虚拟化控制器。

10 为了使上述两个控制器 200 可存取相同的物理储存装置 420，本实施例中在储存虚拟化控制器 200 与物理储存装置 420 之间的装置端输出入装置连结路径上插入一存取控制开关 342。因为点对点连结的本质，每次仅有一储存虚拟化控制器 200 能启动存取一个物理储存装置，即该特定物理储存装置在当时所被指定到的储存虚拟化控制器 200。而另一个与此物理
15 储存装置 420 有关的储存虚拟化控制器 200 则停留在待命状态，其连接此特定的物理储存装置的输出入装置连结则被禁能(disabled)。每个储存虚拟化控制器 200 有一讯号线与存取控制开关 342 连接以控制此存取控制开关 342，此开关 342 决定那一个储存虚拟化控制器连结被接通至物理储存装置 420。

20 如图 4 所示，存取控制开关 342 外的这些讯号线可以被牵成一单一控制讯号线 CS，其依据图 40 所示的真值表(truth table)来控制存取控制开关 342 的状态。另外，亦可如图 5 所示，存取控制开关 342 被设计成接收两条控制输入线 CS1, CS2，而因为从两个储存虚拟化控制器 200 发出的讯号有四种可能的组合，故每个储存虚拟化控制器 200 送出的控制输入线
25 CS1, CS2 是依据图 41 中所示的真值表来决定此存取控制开关 342 的状态。

请注意，在冗余储存虚拟化子系统中任何主动组件或群组位于一热插拔(hot swappable)单元中，所以若一组件发生失效时，此子系统并不需要关机即可更换此组件，此种热插拔单元一般被称为现场可更换单元(FRU, field replaceable unit)。因此，主动组件，如：物理储存装置与存取
30 控制开关，很自然地应设置于热插拔单元中，且将物理储存装置与存取控制开关放置于同一热插拔单元是有其意义的，因为如果缺少任何一方，剩

下的另一方也不能独立运作。因此，存取控制开关与物理储存装置典型地设置于可拆卸的物理储存装置盒(physical storage device canister)中，图 36 与图 38，即显示了此种设计的方块示意图。

在一实施方案中，在此储存虚拟化子系统 20 中的所有的物理储存装置 420 可组合形成一物理储存装置阵列 400，以及所有存取控制开关 342 可组成一开关电路 340，图 6 即为此种实施方案的一个例子，图 6 为本发明中连接至主机 10 及物理储存装置阵列 400 的储存虚拟化控制器 200 的一实施例方块图。此实施例中，第一储存虚拟化控制器(SVC1) 200 包含有一主机端输出入装置连结控制器 220、一中央处理电路(CPC, central processing circuit) 240、一存储器 280、一 SATA 输出入装置连结控制器 300 以及一冗余控制器通讯(RCC, redundant controller communicating)连结控制器 236。此处虽以分开的功能方块描述，但在实际应用时，部份甚至全部的功能方块(functional block)皆可整合在一单一芯片上。例如：RCC 连结控制器 236 能与主机端输出入装置连结控制器 220 整合为一单芯片 IC。

主机端输出入装置连结控制器 220 连接至主机 10 及中央处理电路 240，用来作为第一储存虚拟化控制器(SVC1) 200 及主机 10 之间的接口及缓冲，其可接收由主机 10 传来的输出入请求和相关数据，并且将其转换和/或映像至中央处理电路 240。主机端输出入装置连结控制器 220 可以包含有一个或多个用来耦接于主机 10 的主机端端口。此处所提及的端口的类型可以为：光纤信道支持 fabric 连结(fibre channel supporting fabric)、点对点连结、公用回路连结和/或专用回路连结于目标模式，操作于目标模式的并行小型计算机系统接口(并行 SCSI, parallel small computer system interface)、支持因特网 SCSI(iSCSI, internet SCSI) 协议且操作于目标模式的以太网络，操作于目标模式的序列附加 SCSI(SAS, serial-attached SCSI)，以及操作于目标模式的序列先进技术接取接口(SATA, serial advanced technology attachment)。

当中央处理电路 240 接收到来自主机端输出入装置连结控制器 220 的主机输出入请求时，中央处理电路 240 会将此输出入请求剖析，并且执行某些操作以响应此输出入请求，以及将所请求的数据和/或报告和/或信息，由第一储存虚拟化控制器 200 经由主机端输出入装置连结控制器 220

5 传送至主机 10。将主机 10 传入的输入请求剖析之后，若所收到的为一读取请求且一个或多个操作被执行以为响应时，中央处理电路 240 会由内部或由存储器 280 中或藉由此二种方式取得所请求的数据，并将这些数据传送至主机 10。若所请求的数据无法于内部取得或并不存在于存储器 280，该读取请求将会经由 SATA 输出装置连结控制器 300 及该开关电路 340 10 发送至物理储存装置阵列 400，然后这些所请求的数据将由物理储存装置阵列 400 传送至存储器 280，之后再经由主机端输出装置连结控制器 220 传送到主机 10。当由主机 10 传入的写入请求 (write request) 传达至中央处理电路 240 时，于写入请求被剖析并执行一个或多个操作后，中央处理电路 240 通过主机端输出装置连结控制器 220 接收从主机 10 传入的数据，将其储存在存储器 280 中。对于同步或异步装置操作两者，数据皆经由中央处理电路 240 传送至物理储存装置阵列 400。当该写入请求为一回写请求 (write back request)，输出做完报告 (IO complete report) 会先被传送至主机 10，而后中央处理电路 240 才会执行实际的写入操作；而 15 当该写入请求为一完全写入请求 (write through request)，则输出做完报告会在数据已实际写入物理储存装置阵列 400 后才被传送至主机 10。存储器 280 连接于中央处理电路 240，其作为一缓冲器，用来缓冲传送于主机 10 及物理储存装置阵列 400 之间通过中央处理电路 240 的数据。实际应用时，存储器 280 可以是动态随机存取存储器 (DRAM, dynamic random 20 access memory)，或更特别地，该 DRAM 亦可为同步动态随机存取存储器 (SDRAM, synchronous dynamic random access memory)。

25 SATA 输出装置连结控制器 300 为介于中央处理电路 240 及物理储存装置阵列 400 间的装置端输出装置连结控制器，用来作为储存虚拟化控制器 200 及物理储存装置阵列 400 间的接口及缓冲。SATA 输出装置连结控制器 300 接收由中央处理电路 240 传入的输入请求及相关数据，并将其映像和/或传送至物理储存装置阵列 400。为了符合 SATA 协议的规范，SATA 输出装置连结控制器 300 会将经由中央处理电路 240 传入的数据及控制讯号再格式化，并且将这些数据及讯号传送至物理储存装置阵列 400。

30 在本实施例中，可于中央处理电路 240 上附接一箱体管理服务电路 360 (EMS circuitry, enclosure management service circuitry)，作为一容置物理储存装置阵列 400 箱体的管理电路。然而储存虚拟化子系统 20

亦有其它的配置方式，例如可依各种不同产品的功能设计而定，而将箱体管理服务箱体管理服务电路 360 省略，或是将箱体管理服务箱体管理服务电路 360 整合在中央处理电路 240 中。

在本实施例中，在第一储存虚拟化控制器 (SVC1) 200 上的 RCC 连结控制器 236 用来连接中央处理电路 240 到第二储存虚拟化控制器 (SVC2) 200。除此之外，SATA 输出装置连结控制器 300 经由开关电路 340 连接至物理储存装置阵列 400，开关电路 340 亦连接至第二储存虚拟化控制器 (SVC2) 200。在此一架构中，第二储存虚拟化控制器 (SVC2) 200 可附接于第一储存虚拟化控制器 (SVC1) 200，且物理储存装置阵列 400 可藉由开关电路 340，而被此两个储存虚拟化控制器 200 所存取。更甚者，由主机 10 发出的控制/数据讯号可从中央处理电路 240 通过 RCC 连结控制器 236 传送给第二储存虚拟化控制器 (SVC2) 200 或更进一步地传送给一第二物理储存装置阵列 (图中未示)。

请参阅图 7，图 7 为中央处理电路 240 的一实施例，其中包含有 CPU 芯片组/奇偶性引擎 224 (CPU chipset/parity engine)，一中央处理器 242 (CPU)，一只读存储器 246 (ROM, read only memory)，一非易失性随机存取存储器 248 (NVRAM, non-volatile random access memory)，一液晶显示模块 350 (liquid crystal display module, LCD module)，及一箱体管理服务电路 360。其中该 CPU 242 可为，例如，一 Power PC CPU，而 ROM 246 可为一闪存，用来储存基本输入/输出系统 (BIOS) 和/或其它程序。NVRAM 248 用来储存该物理储存装置阵列输出操作执行状态的相关信息，以备输出操作尚未做完前发生不正常电源关闭时，作检验使用。LCD 模块 350 则是用来显示子系统的操作状态，箱体管理服务电路 360 用来控制该物理储存装置阵列的电源及进行其它的管理。ROM 246，NVRAM 248，LCD 模块 350 及箱体管理服务电路 360 皆经由一 X-总线 (X-bus) 连结至 CPU 芯片组/奇偶性引擎 224。另外，该 NVRAM 248 及该 LCD 模块 350 为可选择项目，在本发明的另一种配置中可以省略不设。

图 8 为本发明中 CPU 芯片组/奇偶性引擎 224 的一实施例，CPU 芯片组/奇偶性引擎 224 包含有奇偶性引擎 260，CPU 接口 910，存储器接口 920，周边组件连结 (PCI, peripheral component interconnect) 接口 930、932，X-Bus 接口 940，及主要存储器 (PM, primary memory) 总线 950，其中 PM

总线 950, 举例而言, 为一 64-bit, 133Mhz 总线, 且连接至奇偶性引擎 260、CPU 接口 910、存储器接口 920、PCI 接口 930、932、X-Bus 接口 940 上, 用以于其间通联数据讯号及控制讯号。

由主机端输出装置连结控制器 220 所发出的数据及控制信号经由
5 PCI 接口 930, 传送至 PM 先进先出缓冲器 934 (PM FIFO) 中缓冲, 再进入 CPU 芯片组/奇偶性引擎 224。其中连结至主机端输出装置连结控制器 220 的 PCI 接口 930 可为, 举例而言, 64-bit, 66Mhz 的频宽。于 PCI 从属周期 (PCI slave cycle) 中, PCI 接口 930 拥有 PM 总线 950 (PM Bus), 使 PM 先进先出缓冲器 934 中的数据及控制信号被传送至存储器接口 920
10 或是 CPU 接口 910。

由 PM Bus 950 传至 CPU 接口 910 的数据及控制信号, 而后会传送至 CPU 242 进行进一步的处理, 而 CPU 接口 910 及 CPU 242 间的沟通管道则可为, 举例而言, 64-bit 数据传输线及 32-bit 地址线来进行。此数据及控制信号会经由一频宽为 64-bit, 133Mhz 的 CPU 至存储器先进先出缓冲器 922 (CM FIFO, CPU to memroy FIFO), 传送至存储器接口 920。
15

在 CM 先进先出缓冲器 922 及存储器接口 920 之间, 有一除错码产生电路 924 (ECC circuit, error correction code circuit), 用以产生一 ECC 码, 而其产生的方式可为, 举例而言, 将 8-bit 的数据以异或 (XOR) 运算后, 产生一单一位的 ECC 码。接下来, 存储器接口 920 将数据及 ECC
20 码储存在存储器 280 中。该存储器 280 可为, 举例而言, SDRAM。而存储器 280 中的数据经过除错码修正电路 926 (ECC correction circuit), 并与除错码产生电路 924 中的 ECC 码作比较, 最后再被传送到 PM Bus 950, 其中除错码修正电路 926 是用来进行单一位自动修正 (1-bit auto-correction) 及多位检错 (multi-bit error detecting)。

奇偶性引擎 260 响应于 CPU 242 的指示, 来执行一特定磁盘阵列型态的奇偶性功能。当然, 在一些特定的条件下, 比如说 RAID0, 奇偶性引擎 260 可以关掉而不执行奇偶性功能。在图 8 所示的实施例, 奇偶性引擎 260 包含有一 XOR 引擎 262 经由 XOR 先进先出缓冲器 (XOR FIFO) 264 而连接至 PM Bus 950。该 XOR 引擎 262, 举例而言, 可对一给定的地址及长度
30 的存储器位置来执行 XOR 运算。

锁相回路 980 (PLL, phase locked loop) 是用于在相关讯号间维持适

当的相移(phase shift)。而计时控制器 982 (timer controller)是用来提供各种不同时钟及讯号的时间基准。内部寄存器 984 (internal register)是用来暂存CPU芯片/奇偶性引擎 224的状态,及控制PM Bus 950中的数据流动,而一对通用异步收发器(UART ,universal asynchronous receiver and transmitter,)功能方块 986则是用作CPU芯片/奇偶性引擎 224对外的接口,且该接口规格为RS232。

在实际应用时,PCI接口 930,932可代换为周边组件连结扩充(PCI-X, peripheral component interconnect extended)接口,或者是以周边组件连结快捷(PCI Express)接口取代PCI接口 930, 932。

10 请参考图9,图9为图6中SATA输出装置连结控制器300的方块图,在本实施例中,SATA输出装置连结控制器300包含有两个PCI-X转SATA控制器310(PCI-X to SATA controller)。图10为图9中PCI-X转SATA控制器310的方块图,其中每个PCI-X转SATA控制器310包含有一连接至中央处理电路240的PCI-X接口312,一连接至PCI-X接口312的译码/多任务仲裁器314(Dec/Mux arbiter),以及八个连接至Dec/Mux仲裁器314的SATA端口600。PCI-X接口312包含有一连接至Dec/Mux仲裁器314的总线接口318,以及一用来储存PCI-X转SATA控制器310组态的组态电路316(configuration circuit)。Dec/Mux仲裁器314将在PCI-X接口312与多个SATA端口600间进行仲裁,且执行自PCI-X接口312至SATA端口600的交易(transaction)的地址译码。而数据及控制讯号将经由此PCI-X转SATA控制器310的SATA端口600,被传送至物理储存装置420。在实际应用中,PCI-X转SATA控制器310可由PCI转SATA控制器取代,而在PCI转SATA控制器中,PCI-X接口312可由一PCI接口取代。同样地,在其它的实施例中,PCI-X转SATA控制器310可由一PCI Express转SATA控制器取代,而在PCI Express转SATA控制器中,PCI-X接口312系由一PCI Express接口取代。

30 接下来请参考图11,图11为图10中SATA端口600的一实施例方块图。如图11中所示,SATA端口600包含有一超集寄存器630(superset register),一指令区块寄存器640(command block register),一控制区块寄存器650(control block register),以及一直接存储器存取寄存器620(DMA register),前述所有的寄存器系经由Dec/Mux仲裁器314,

而连接至 PCI-X 接口 312 的总线接口 318。经由填值于上述的寄存器中，数据得以通过一由直接存储器存取控制器 670 所控制的双端口先进先出缓冲器 660，而在 Dec/Mux 仲裁器 314 与传输层 690 (transport layer) 间传输。数据传送至传输层 690 后，被再格式化成为帧信息结构 (FIS, frame information structure)，并传送到连结层 700 (link layer)。

5 连结层 700 稍后将帧信息结构再格式化成为帧 (frame)，以加入帧起始信息 (SOF, start of frame)，循环冗余校验码 (CRC, cyclic-redundancy check code)，帧结束信息 (EOF, end of frame) 等，并将其以 8b/10b 编码方式转译成 8b/10b 编码的字符而实现，并将其传送到物理层 710 (PHY layer)。

10 物理层 710 经由一对差动讯号线 (differential signal lines)——传输线 LTX+ 及 LTX-——送出讯号至物理储存装置 420，并经由另一对差动讯号线——接收线 LRX+ 及 LRX-——接收来自物理储存装置 420 的讯号，其中各组的两条讯号线，例如 LTX+ 及 LTX-，同时个别传送以一参考电压 V_{ref} 为准的正负电压的讯号 TX+/TX-，例如 +V/-V 或是 -V/+V 的电压讯号，所以它们的电压差是 +2V 或是 -2V，如此一来便可增加讯号的品质。在 LRX+ 及 LRX- 接收在线也可以使用相同的方法接收讯号 RX+/RX-。

15 当一帧由物理层 710 传送至连结层 700，连结层 700 将用 8b/10b 编码的字符进行译码，并且除去 SOF，CRC，EOF 的部份，其中经由帧信息结构 FIS 计算得出的 CRC 将会被拿来与所接收到 CRC 作比较，用来确定所接收的信息的正确性。当传输层 690 接收到来自连结层 700 的 FIS 讯号，传输层 690 将会决定 FIS 的型式，并依照 FIS 的型式将 FIS 的内容传送到所指定的区域。

20 图 12 为符合 SATA 协议的传输结构，其中在序列线中通联的讯号为一连串使用 8b/10b 编码的字符，其最小单位为双字组 (double-word, 32 位)。每一个双字组的内容将被组合以提供低阶的控制信息，或是用以传送主机与相连结的装置间的信息，而在讯号在线传送的两种数据结构为基元 (primitive) 以及帧。

25 一基元是由一单一的双字组所组成，其为主机与装置间通讯信息中最简单的单位。当一基元中的字节在编码之后，其所产生的型样 (pattern) 便不容易被误解成其它型式的基元或是其它任意的型态。基元主要的用途

是传送实时 (real-time) 状态的信息, 这些信息是用来控制信息的传递以及协调主机及装置间的通讯。一基元的第一字节为一特别的字符。

一帧是由多个双字组所构成, 并以 SOF (Start of Frame) 基元开始, 以 EOF (End of Frame) 基元结束。在 SOF 基元之后为一使用者有效负载, 称之为帧信息结构 (FIS, frame information structure)。另外循环冗余校验码 (CRC) 为紧接在 EOF 基元之前的最后非基元双字组, 且 CRC 为依据 FIS 运算得来。另外, 介于 SOF 与 EOF 间可以有流程控制基元 HOLD 或是 HOLDA 用来调整数据流, 以达到速率匹配 (speed matching) 的目的。

传输层 690 用来建构 FIS 以用于传送, 以及是于自连结层 700 接收到 FIS 时将其分解。且该传输层 690 并不维护 ATA 指令或是先前的 FIS 内容的前后关系 (context)。当收到请求时, 传输层 690 会收集 FIS 内容, 并依照正确的顺序建构 FIS。FIS 的型态有很多种, 图 13 及 14 分别为其中之一。

请参阅图 13, 图 13 为一信息帧的示意图。如图中所示, 一直接存储器存取设定的 FIS 在字段 0 处包含有一标头 (HEADER), 而其第一字节 (字节 0) 则定义了该 FIS 的型态 (41h), 此 FIS 的型态则定义了此 FIS 其余的字段, 和定义它的全部长度为七个双字组。字节 1 中的位 D 则标示了该后续数据传送的方向, D 为 1 表示传送端至接收端, D 为 0 则表示接收端至传送端。字节 1 中的位 I 为一中断位 (interrupt bit), 而位 R 为一保留位, 且设为 0。直接存储器存取缓冲器识别码的高/低字段 (DMA buffer identifier high/low field, 字段 2 和字段 1), 则分别标示了该主机存储器的直接存储器存取缓冲区域。直接存储器存取缓冲器偏移字段 (DMA buffer offset field, field 4), 为进入缓冲器内的字节偏移。直接存储器存取传送计数字段 (DMA transfer count field, field 5), 则为此装置所读取或写入的字节数量。

请参阅图 14, 图 14 为另一型态的信息帧的示意图。如图中所示, 一数据 FIS (DATA FIS) 于字段 0 处包含有一标头, 且该第一字节 (字节 0) 定义了该数据型 FIS 的型式 (46h), 而此 FIS 的型式则定义了其余的字段以及它的全长为 $n+1$ 双字组 (double-word)。还有, 字节 1 中的 R 位则为保留位, 并且设定为 0, 而字段 1 至 n 为双字组数据, 其中包含有要传送的数据。一单一数据 FIS 内的数据有上限。

请回顾在图 6 的实施例, 主机端输出装置连结控制器 220 及装置端
输出装置连结控制器 300 (SATA 输出装置连结控制器 300), 可使用相
同类型的 IC 芯片, 而其中主机端输出装置连结控制器 220 上的输出
装置连结端口的组态被设定为主机端的输出装置连结端口, 而装置端
5 输出装置连结控制器 300 中的输出装置连结端口的组态则被设定为装置
端的输出装置连结端口使用。另外, 亦可采用一单一芯片, 其组态可被
设定为同时包含有主机端输出装置连结端口及装置端输出装置连结
端口, 用以在同一时间分别耦接至主机 10 及物理储存装置阵列 400。

请回顾图 4、图 5, 通常, 图 4、图 5 中所述的存取控制开关 342 维持
10 在一状态而使被指定处理装置端输出请求的储存虚拟化控制器 200 配接
至物理储存装置 420 的状态, 而这些装置端输出请求为响应主机端输出
请求所产生的操作的结果。但是, 在某些情况下, 其必须暂时允许另一
个储存虚拟化控制器 200 来存取物理储存装置 420, 此情况可能出现的一
个组态的例子是, 储存虚拟化控制器 200 中之一被选定为主人 (Master),
15 当此储存虚拟化控制器 200 涉及某些物理储存装置相关管理功能时, 且此
管理功能由此储存虚拟化控制器 200 所执行 (例如: 监视物理储存装置 420
的健康状态或对物理储存装置 420 上的介质的某一个保留给储存虚拟控制
器组内部使用的区域进行存取), 同时, 一些可经由主机端输出装置连
20 结而可被主机 10 存取的逻辑介质单元, 则是指定给另一个储存虚拟化控
制器 200 作为处理主机端输出请求之用。在此情况下, 该储存虚拟化控
制器 200 彼此之间可以互相通讯 (如可能可以通过图 6 所示的 RCC 连结控
制器 236), 以决定一适当的机会, 使开关 342 的状态得以安全地被改变,
进而允许另一个储存虚拟化控制器 200 对此物理储存装置 420 进行存取,
而无须中断物理储存装置 420 正在进行中的存取。此步骤在其后将称之为
25 “物理储存装置存取权转移”。

图 15 为一转换程序的流程图, 其中每一储存虚拟化控制器 200 皆能
决定存取控制开关 342 目前的状态, 其是利用允许储存虚拟化控制器 200
读回控制讯号的状态, 或藉由使每一个储存虚拟化控制器 200 在存储器中
维护一存放目前状态的镜像而实现。一目前未配接至物理储存装置 420 却
30 要求存取物理储存装置 420 的储存虚拟化控制器 20 (称为存取权请求者,
access requester), 将发出一个请求给目前正配接至物理储存装置 420

的储存虚拟化控制器 200(称为存取权所有者, access owner), 使此存取控制开关 342 的状态得以被切换成允许此存取权请求者存取物理储存装置 420。当存取权所有者接收到此请求时, 此存取权所有者会等待一“适当”的时机, 来开始此转换程序, 此程序必须使任何正在进行中(pending)的
5 输出请求被做完、并将所有尚未开始执行的输出请求放入队列中(queueing)。当所有正在进行中的请求都被做完, 此存取权所有者就会更改它的开关控制讯号的状态, 以释出对物理储存装置 420 的存取能力, 然后发送一个应答讯号(acknowledgement)给存取权请求者, 告知它现在可以安全的存取物理储存装置 420 了。此时, 存取权请求者也更改其开关控制讯号的状态, 以取得可对物理储存装置进行存取的能力, 这使得转换程序完成, 同时此存取请求者将变成新的存取权所有者。
10

此时, 新的存取权所有者是可自由的依意愿去发出输出请求给物理储存装置 420, 并且新的存取权所有者可保留此所有权, 直到原本的存取权所有者请求将存取权切换回去, 并且进行上述相同的步骤。或者, 在
15 一些“适当”的时机, 例如, 所有当时要发的输出请求皆已执行完毕时, 此新的存取权所有者可以自动地开始切回(switch-back)程序, 此是藉由更改其存取开关控制讯号的状态, 以释出对物理储存装置 420 的存取能力, 以及发出一主动发出的(unsolicited)应答讯号给原先的存取权所有者, 告知它新的存取权所有者已释出存取权, 此原先的存取权所有者现在可以
20 收回存取权了。

基本上来说, 不论是新的存取权所有者保留所有权直到原本的存取权所有者发出一“切回”请求, 或是自动转移所有权给原本的存取权所有者, 要采用何种方式, 可以是依系统预设而固定, 或是根据一些因素而动态决定, 如: 此二个储存虚拟化控制器相对的存取频率, 或比对保留所有权与
25 自动归还所有权给原本的存取权所有者此两方式在效能上相对的影响。

为实现上述转换程序, 有一些控制器间通讯的机制可供选用。在这里提出一种可能的通讯机制, 其为二元讯号对存取所有权仲裁机制(binary signal pair access ownership arbitration mechanism), 其中每一 SATA 输出装置连结均有一对“存取请求”讯号线, 对于每一条数字二元讯号
30 线, 都有一个储存虚拟化控制器 200 可设定/清除此讯号线(在此讯号在线的“主动”储存虚拟化控制器), 以及有另一储存虚拟化控制器 200 可

读取此讯号线的状态（“被动”储存虚拟化控制器）。此存取请求讯号线中之一以第一储存虚拟化控制器(SVC1) 200 为主动端、第二储存虚拟化控制器(SVC2) 200 为被动端，同时，另一存取请求讯号线以第一储存虚拟化控制器(SVC1) 200 为被动端、第二储存虚拟化控制器(SVC2) 200 为主动端。

5 在被动端，储存虚拟化控制器 200 读取其另一个储存虚拟化控制器的存取请求讯号的目前状态，以及从上回的读取后是否讯号线曾改变过状态。当读取时，则后者将被清除，亦即从上回的读取后是否讯号线曾发生过状态改变的纪录将会被清除。

在此一机制中，起初，一储存虚拟化控制器 200 拥有 SATA 输出装置连结的所有权以及相连于此输出装置连结的讯号线为拉启(assert)状态，当另一个储存虚拟化控制器 200 意欲取得所有权时，其成为一存取请求者并且拉启其“主动”讯号线，然后监视其“被动”讯号线，监看其状态的改变，此改变代表者此讯号线在一些时点曾被停止拉启(deassert)，意即，代表者存取权所有者应答其请求。此时，此发出请求的储存虚拟化控制器 200 可利用改变存取控制开关 342 的状态，来接管该 SATA 输出装置连结，所以该发出请求的储存虚拟化控制器 200 就配接到物理储存装置 420 上了。

10

15

另一方面，此存取权所有者会持续监视其被动讯号线是否有拉启，若检测此被动讯号线拉启后，就会在一个适当的时机，开始等待任何正在进行的输出请求，同时将任何新的输出请求放入队列中。当所有正在进行的输出请求皆做完，藉停止拉启其主动讯号线，以应答此存取控制请求。如果其希望归还存取控制，例如当有新的输出请求被排在队列中而要发出时，则重新拉启此主动讯号线。存取权请求者则是监视被动讯号线是否有状态上的改变，而不是监视一停止拉启状态，因为存取权所有者可能在停止拉启之后接着立刻拉启讯号线，而使存取权请求者根本没有机会检测到此停止拉启状态。图 16 为显示描述于前的流程图。

20

25

上述二元讯号对存取所有权仲裁机制的一变形以达到协调存取所有权转移，是利用一对硬件电路，此处称为存取所有权仲裁(AOA, access ownership arbitration)电路，每个储存虚拟化控制器一个，以间接控制存取控制开关控制讯号，而不是控制那些直接由储存虚拟化控制器来控制的讯号。

30

此二个存取所有权仲裁电路中之一的输出端连接并控制与其中一个
储存虚拟化控制器相关的存取控制开关控制讯号，而另一个存取所有权仲
裁电路的输出端连接并控制与另一储存虚拟化控制器相关的存取控制开
关控制讯号。除此之外，每一存取所有权仲裁电路将来自于此二储存虚拟
5 化控制器的存取所有权请求讯号 (AOR, access ownership request
signals) 作为输入端。当一储存虚拟化控制器未拥有存取权也没有请求
存取所有权时，它的存取所有权请求讯号会维持在停止拉启的状态。在此
状态下，与这个储存虚拟化控制器关联的存取所有权仲裁电路的输出讯号
为未活动。当此储存虚拟化控制器欲取得所有权时，其拉启它的存取所有
10 权请求讯号，如果此时另一个储存虚拟化控制器的存取所有权请求讯号未
活动，则此发出请求的储存虚拟化控制器关联的存取所有权仲裁电路将会
拉启它的输出讯号，从而拉启与发出请求的储存虚拟化控制器关联的存取
控制开关控制讯号。如果此时另一储存虚拟化控制器的存取所有权请求
讯号为活动中，则此发出请求的储存虚拟化控制器的存取所有权仲裁电路
15 输出仍维持停止拉启状态，直到另一个储存虚拟化控制器的存取所有权请求
讯号停止拉启后，此发出请求的储存虚拟化控制器的存取所有权仲裁电路
输出才会变为活动中。这发出请求的储存虚拟化控制器的存取所有权仲
裁电路的输出接着维持活动中，直到此发出请求储存虚拟化控制器的存取
所有权请求讯号停止拉启为止，且此活动不受另一储存虚拟化控制器的存取
20 所有权请求讯号所影响。基本上说，该二存取所有权仲裁电路可紧靠着存
取控制开关而放置，譬如像是与存取控制开关一奇偶性于该物理储存装置
盒中。

在这存取权仲裁机制中需要有一判定设备 (facility)，能在两个储存
虚拟化控制器同时拉启他们的存取所有权请求讯号的情况下，使一储存虚
25 拟化控制器来判定它是不是被授与存取所有权，以及使目前拥有存取所有
权的储存虚拟化控制器来判定另一储存虚拟化控制器何时请求存取所有
权。前者可藉由使一储存虚拟化控制器具有决定存取控制开关的状态的能
力而实现，而使一储存虚拟化控制器具有决定另一储存虚拟化控制器的存
取所有权请求讯号的状态的能力则可实现后者。然而，既然此两个判定是
30 在存取所有权转移过程期间不同的时间点做出的，它们可以合并成一个单
一设备，其是由每一储存虚拟化控制器提供单一数字二元讯号所构成，称

为另一个储存虚拟化控制器存取所有权请求讯号 (ASAOR , alternate storage virtualization controller access ownership request signal), 并且由储存虚拟化控制器上所执行的固件来读取此讯号的状态。正常状态下, 这讯号会反映另一储存虚拟化控制器的存取所有权请求讯号的状态, 然而, 当此储存虚拟化控制器取得存取所有权时, 它的 ASAOR 会被清除为不活动 (inactive), 而不受另一个储存虚拟化控制器的存取所有权请求讯号的状态影响, 且会维持此一状态直到此储存虚拟化控制器的固件读取, 其后, 它能回到正常状态下而反映另一个储存虚拟化控制器的存取所有权请求的状态。图 17 为此实施方案中不同讯号间相互影响的时序图。请参阅图 18, 图 18 为对应至图 17 的存取权仲裁机制的流程图。如图中所示, 在此一存取权仲裁机制中, 当一个储存虚拟化控制器意欲取得所有权时, 它会拉启它的 AOR 讯号, 然后开始监视它的 ASAOR 讯号。当此储存虚拟化控制器检测到它的 ASAOR 讯号为不活动时, 则知道其已经取得所有权且可以进行存取物理储存装置了。接着, 此储存虚拟化控制器会维持其 AOR 讯号为拉启的状态, 直到其意欲释出存取所有权为止, 亦即停止拉启其 AOR 讯号。上述整个过程中, 此储存虚拟化控制器想要维持存取所有权, 则它除了维持它的 AOR 讯号为拉启外, 它也监视它的 ASAOR 讯号是否拉启, 若它检测到它的 ASAOR 讯号拉启, 表示另一个储存虚拟化控制器希望取得所有权, 则在一个适当的时点, 它会开始等待所有正在进行中的输出请求做完, 同时将所有新的输出请求放入队列, 而当所有正在进行中的输出请求都做完时, 它才会释出所有权, 并且停止拉启它的 AOR 讯号。如果它希望存取所有权被归还, 如, 当有一些新的输出请求在队列中要被发出时, 则它会立刻重新拉启它的 AOR 讯号。

另一种可能的通讯机制是经由支持多重位和/或字节信息传送的通讯信道来传递存取所有权转移请求或是应答讯息。一组低成本、专用的通讯信道-例如集成电路间信道 (I2C, inter-IC) -可用来实现前述交换这些请求或应答讯息的目的。此外, 亦可利用现存的、用以允许在冗余对中的二个储存虚拟化控制器彼此互相沟通的控制器间通讯信道 (ICC) 来实现, 去交换这些存取所有权转移请求与应答讯息, 作为此二储存虚拟化控制器间交换的正常状态同步信息的一部分。这些控制器间通讯信道可以为光纤信

道、序列先进技术接取接口信道、并列小型计算机系统接口信道、以太网络以及序列附加小型计算机系统接口信道等。

一种强制存取所有权转换的情况是，当存取权所有者 200 故障而另一个储存虚拟化控制器 200 必须接替其功能时。图 19 显示在这种情况下转换存取所有权的程序。当另一个储存虚拟化控制器 200 检测到此故障储存虚拟化控制器 200 无法正常运作时，另一个储存虚拟化控制器 200 就会拉启故障储存虚拟化控制器的重置讯号 (reset signal)，使其能完全地失能，并迫使所有的外部讯号线进入预设状态。其中上述外部讯号线中之一为此故障储存虚拟化控制器 200 的存取控制开关控制讯号；在拉启此储存虚拟化控制器的重置讯号时，这控制讯号线即被设定成使存活的储存虚拟化控制器 200 能够配接至物理储存装置 420 的状态。在拉启此故障储存虚拟化控制器的重置讯号之后，该存活的储存虚拟化控制器 200 就会设定其存取控制开关控制讯号，使成为配接到物理储存装置 420 的状态，如此一来，即完成转换存取权的程序。

存取控制开关 342 将会持续此状态直到故障储存虚拟化控制器 200 被更换或是重新上线，并且其请求所有权转移。其中对每个储存虚拟化控制器 200 处于重置 (reset)、电源启动 (power-Up)、或初始化 (initialization) 期间时，其存取控制开关讯号线的状态维持在使此储存虚拟化控制器 200 本身不能够配接至物理储存装置的状态，以确保其不会因不慎迫使存取控制开关 342 进入一中断存取的状态，而干扰其它在线储存虚拟化控制器可能正在进行的物理储存装置的存取。

就正常状态下不具有物理储存装置 420 存取所有权的储存虚拟化控制器 200 而言，一处理临时性 (occasional) 存取需求的另一个方法是，令存取权所有者扮演一代理者，用来发送此请求存取的储存虚拟化控制器 200 的出入请求，此处请求存取的储存虚拟化控制器称为存取请求者 (access requester)，此方法中必须执行一称为“出入请求复位路径传送 (IO request rerouting)”的操作。在这种方法中，存取请求者必须转移所有必要的出入请求的信息给存取权所有者，用以建构一出入请求而发送给物理储存装置 420。除出入请求信息之外，在出入请求发送或执行期间或之前，存取请求者将会转移任何要被写入物理储存装置的有效负载数据给存取权所有者，且在出入请求执行后或执行期间，任何从

物理储存装置中读取的有效负载数据将会被代转回此存取请求者。在此输出请求执行做完，将会回传此操作做完的状态给存取请求者，此操作做完状态典型地为一指示操作成功或失败以及失败理由的信息。图 20 是描绘此一方法的流程图。

- 5 输出请求复位路径传送的功能可经由一冗余输出装置连结端口对而被执行，此输出装置连结端口对中的每个成员端口分别设置在冗余储存虚拟化控制器对中各储存虚拟化控制器上，但却通过输出装置连结连接至相同的物理储存装置。另外此冗余输出装置连结端口对的一个成员端口可以是一个装置端输出装置连结端口，或是一个混合输出装置
- 10 连结端口。此混合输出装置连结端口，对一些输出操作是扮演一装置端输出装置连结端口，以及对其他输出操作是扮演一主机端输出装置连结端口。在这一点上，此冗余输出装置连结端口对中的成员端口可包含点对点输出装置连结端口，例如：SATA 连结，或者是，其可包含多
- 15 装置输出装置连结端口，例如：光纤信道或并列 SCSI 连结。此外，如此的输出请求复位路径传送的功能也可以适用于利用此冗余储存虚拟化控制器对的互补扩充端口来执行输出请求。此冗余储存虚拟化控制器对的互补扩充端口将在下文中结合图 21 至图 32 详细解说。

相较于为使两个储存虚拟化控制器 200 都能存取每一个物理储存装置 420 而来来回回地实际转移存取所有权，此输出请求复位路径传送具有两个优点。首先，因 SATA 协议的本质，在此连结上从停止而未活动状态 (down state) 至准备好可活动状态 (up state) 时，其要求一相当长的启动期间 (bring-up)，这可能会使得在存取权请求者从存取权所有者处接收到所有权时到所有权请求者能真正开始启动物理储存装置存取之间会有一显著的延迟。其次，使 SATA 接口停摆随即再使其准备好的过程，可能会

20 需要造成任何一端的 SATA 接口电路进入异常状况处理的状态，而因为异常状况处理程序的测试通常不会像正常状况程序的测试一样的彻底，因此，有时候错误可能会出现而干扰此连结去成功的再启动 (re-bring-up)。为了减少此一风险，最好试着去减少任何一端需解译与处理异常状况的产生，在此例子中，包括减少存取所有权需转移的情况。

25 所有的装置端输出装置连结都是 SATA 的“纯”SATA 储存虚拟化控

- 30 制器有一个限制，就是它的可连结的物理储存装置的数目受限于可包装在

一单一储存虚拟化控制器当中的装置端输出装置连结的数目，而 SATA 的规格当中，讯号线的最大长度仅限于 1.5 公尺，以致于连接到一储存虚拟化控制器的物理储存装置一定要靠得够近，使讯号线的长度不超过 1.5 公尺。由于这些限制，SATA 储存虚拟化子系统只能提供最多 16 个 SATA 物理储存装置 5 的连接。所以一纯 SATA 储存虚拟化子系统无法像光纤 FC-AL 储存虚拟化子系统一样，拥有经由同一组装置端输出装置连结的外接扩充机箱连接至最多为 250 个物理储存装置的扩充性。

为了克服以上的限制，本发明的储存虚拟化控制器上选择性地可包含一个或多个扩充用装置端多装置输出装置连结 (expansion device-side multiple-device I/O device interconnect)，在此称为装置端扩充端口，如并列 SCSI 或是光纤 FC-AL 或是支持 iSCSI 的以太网或是 SAS 等，而 10 这些连结可允许外接扩充机箱 (chassis)。这些机箱可为由物理储存装置所构成的原始型 JBODs (native JBODs)，是直接连接到连结上而不需要介于其中的转换电路，也可是一仿真该原始型 JBODs 的智能型 JBOD 仿真子系统。此智能型 JBOD 仿真子系统使用 SATA 或 PATA 物理储存装置组合及 15 单一或者冗余储存虚拟化控制器所构成，其中的储存虚拟化控制器是用来提供将连接 JBOD 子系统与主要储存虚拟化子系统 (primary storage virtualization subsystem) 的多装置端输出装置连结协议，转换到连接 JBOD 储存虚拟化控制器与其所管理的物理储存装置的装置端输出装置 20 连结 (SATA 与 PATA) 协议。当然，扩充端口亦可以是 SATA 扩充端口，惟其有连结长度与物理储存装置数量上的限制，一如前文所述一般。

本发明提出三种装置端扩充端口布线实施例以供选择。图 21 为其中一实施例，其中在一个储存虚拟化控制器上的每一个装置端扩充端口是与 25 在另一储存虚拟化控制器上的互补物 (complement，亦即互补扩充端口) 相连系，此处称为冗余储存虚拟化控制器互连扩充端口 (redundant SVC interconnected expansion port) 实施方案。此会使这两个储存虚拟化控制器在正常操作期间共享此装置端连结以及其所提供的频宽，并且允许这两个储存虚拟化控制器完全存取每个储存单位端口。其更进一步地，允许任一储存虚拟化控制器保有对所有储存单位的完全存取，此包含原先指定 30 给另一个储存虚拟化控制器的储存单位，即便是在该另一个储存虚拟化控制器故障的情况下。图 22 显示硬件开关如何被使用而实现此种具有循环

形式(loop-style)多装置输出入装置连结(如光纤 FC-AL)的扩充端口的可切换连结。在正常操作期间,所有开关的状态为“0”,从而,在此连结上的两储存虚拟化控制器可与储存单元连结。若第一储存虚拟化控制器(SVC1)故障,M2会设为“1”,且维持M1被清除为“0”,从而,绕过第一

5 储存虚拟化控制器(SVC1),并自第二储存虚拟化控制器(SVC2)至储存单元建立一直接连接。若是第二储存虚拟化控制器(SVC2)故障,则讯号M1会设为“1”(M2为随意讯号,“don't care”),藉此,绕过第二储存虚拟化控制器(SVC2),并且自第一储存虚拟化控制器(SVC1)至储存单元建立一直接连接。此开关活动可藉由一个硬件讯号检测电路(SDC, hardware

10 signal detection circuit)启动,此SDC检测是不是有一个有效的讯号呈现在S1或S2。或者此一开关活动亦可经由此两个储存虚拟化控制器其中之一检测到另一个储存虚拟化控制器故障时启动。

图23显示一电路设计,此电路设计为依赖硬件讯号检测以启动此开关状态的变更。图24显示一电路设计,此电路设计是从第一储存虚拟化控制器(SVC1)与第二储存虚拟化控制器(SVC2)各自取得输入讯号C1与

15 C2,以触发此开关状态的变更。在这种实施方案中,每一个控制讯号在其对应的储存虚拟化控制器离线时将被强迫进入一“清除(0)”状态(图中未显示此对应的电路),以避免若此控制讯号是三态(Tri-State)或是浮动(Floating)且因而被接着的电路所解释为“设定(1)”状态时可能引起的

20 后果。图25显示图23与图24的混合型,其是藉由硬件讯号检测或自储存虚拟化控制器的输入讯号来支持启动开关状态的变更,以提供比只有单独一种还大的弹性。

图26描绘一强化的实施例,在此称为冗余储存虚拟化控制器互连冗余扩充端口(redundant SVC interconnected redundant expansion port)

25 实施例。其中包含有成对的冗余扩充端口,而不是相互间独立的扩充端口,以用来使在提供服务给一个储存虚拟化控制器的扩充端口的一个连结断开或故障时,不会导致此储存虚拟化控制器对连接于此连结上的储存单元的存取能力全部丧失。在此一架构下,冗余对的每一端口分别与连接于此连结上的每个双端口物理储存装置的一端口相连接,或者是,与一双端口

30 储存虚拟化子系统的一端口相连接,此双端口储存虚拟化子系统仿真连接至此连结的多个物理储存装置,举例来说,其可为一JBOD仿真储存虚

拟化子系统。若其中一储存虚拟化控制器上的一扩充端口或连结故障，则输出请求会通过另一个扩充端口/连结复位路径传送。

图 27 描绘出另一可能的实施例，其中，一个储存虚拟化控制器上的每一扩充端口在另一个储存虚拟化控制器上有一冗余互补物 (redundant complement, 亦即冗余互补扩充端口)，此在一个储存虚拟化控制器上的扩充端口以及其在另一个储存虚拟化控制器上的冗余互补扩充端口是以下列方式连接至每一双端口储存单元的两端口上：一储存虚拟化控制器的扩充端口连接至一双端口对中的一端口，而其在另一储存虚拟化控制器上的互补扩充端口连结至此双端口对中的另一端口。此互补扩充端口并没有互连而是藉由每一储存单元具有双端口的性质来实现冗余的目的。然而，仅仅是储存单元的双端口性质，在面临其中一储存虚拟化控制器的扩充端口故障，或一连接储存虚拟化控制器的一扩充端口与一储存单元的连结故障时，并不足以支持二储存虚拟化控制器的存取皆得以维持的这种冗余功能。为实现此种目的，需要提供一些机制用以使输出请求复位路径传送，而在一储存虚拟化控制器自有的扩充端口或连接它与储存装置的连结故障时，改经由连接另一个储存虚拟化控制器至储存装置的连结。

图 28 描绘一实施方案，此实施方案为如果执行在此储存虚拟化子系统与储存单元之间的连结部分断掉或故障发生时，使用硬件开关去连接如图 27 所示的连接两个储存虚拟化控制器与储存单元的两个输出装置的连结。在储存虚拟化子系统与储存单元之间的连结一般为机箱间的缆线 (cable)，故特别易于断掉。如图 28 中所示，在正常操作期间，也就是，所有连结上的操作都适当地运作时，所有开关的状态皆为“0”。自扩充端口所发出的讯号直接被递送至储存单元中相对应的端口，此处称为预设路径，因此，在每一个储存虚拟化控制器的冗余对中的每一互补扩充端口/连结都可完全独立运作且不受干扰。

当由于连接至储存单元端口 2 的连结断掉或者因储存单元端口 2 自己本身的故障，使得所有的输出讯号必须变成从两个储存虚拟化控制器扩充端口传递至储存单元端口 1 时，M1 与 M2 系设为“1”，然而 M3 维持在“清除(0)”，并且 M4、M5 为随意值。当所有输出讯号自此二储存虚拟化控制器的扩充端口传递至储存单元端口 2 时，M1 与 M3 系设为“1”，而 M4 与 M5 维持在“清除(0)”，M2 则为随意值。若第一储存虚拟化控制器

(SVC1) 离线，而需直接从第二储存虚拟化控制器 (SVC2) 传递输出入讯号到储存单元端口 2 时，M1、M4 以及 M5 会维持在“清除(0)”，而 M2 与 M3 则为随意值。若，除了第一储存虚拟化控制器 (SVC1) 离线以外，且储存单元端口 2 的连结断掉或是储存单元端口 2 自己本身故障，则输出入讯号需要自第二储存虚拟化控制器 (SVC2) 传递至储存单元端口 1，此时藉由设定 M5 与 M2 为“1”，M1 维持为“清除(0)”，同时 M3 与 M4 为为随意值即可完成。相反地，若第二储存虚拟化控制器 (SVC2) 离线，而需直接从第一储存虚拟化控制器 (SVC1) 传递输出入讯号到储存单元端口 1 时，M2、M3 维持“清除(0)”，而 M1 与 M4 以及 M5 则为随意值。若，除了第二储存虚拟化控制器 (SVC2) 离线以外，亦有至储存单元端口 1 的连结断掉或可能是其自身的故障发生，则输出入讯号需要从第一储存虚拟化控制器 (SVC1) 传递至储存单元端口 2，此将藉由设定 M3 与 M4 为“1”，M5 维持为“清除(0)”，同时 M1 与 M2 为随意值即可做到。图 42 所示的表即概括此各种可能情况下此开关的设定。

此开关活动可藉由硬件讯号检测电路 (SDC, Hardware Signal Detection Circuit) 来启动，此 SDC 检测是否有一有效的讯号呈现在 S1 或 S2，或者亦可由此两个储存虚拟化控制器的一来启动，此是当储存虚拟化控制器检测到预设路径断掉或故障时启动此一开关活动。图 29 显示一电路设计，此电路设计依赖硬件讯号检测去启动此开关状态的变更 (此电路的真值表显示于图 43)。图中所示，R1 为一超控讯号 (override signal)，使得在第二储存虚拟化控制器 (SVC2) 和 / 或关联电路故障的情况下，第一储存虚拟化控制器 (SVC1) 得以强制取得该开关的状态的控制权。同样地，R2 亦为一超控讯号，使得在第一储存虚拟化控制器 (SVC1) 和 / 或关联的电路故障的情况下，第二储存虚拟化控制器 (SVC2) 得以强制取得该开关的状态的控制权。图 30 显示一电路设计，此电路设计从第一与第二储存虚拟化控制器 (SVC1, SVC2) 分别取得输入讯号 C1, C2，以触发此开关状态的变更 (此电路的真值表系显示于图 44)，在此例中，经由第一储存虚拟化控制器 (SVC1) 设定讯号 C1，表示其输出入请求要被递送到储存单元端口 2，以取代预设的储存单元端口 1；然而，经由 SVC2 设定讯号 C2，表示其输出入请求要被传送到储存单元端口 1，而非预设的储存单元端口 2。图 31 显示一种两者的混合型，其藉由硬件讯号检测或自二储存虚拟化控制器的

输入讯号两种方式支持启动开关状态变更，以提供比只有单独一种方式还大的弹性（此电路的真值表显示于图 45 中）。

5 对描绘于图 27 的组态上装置扩充端口的布线，尚有另一种完全无需任何的连结的方式可供选择。在此例中，为实现冗余的目的，经由改变输出请求的传递路径为自储存虚拟化控制器到另一个，且经储存虚拟化控制器间通讯连结而由存活的互补扩充端口/装置端连结送出，此内部储存虚拟化控制器沟通连结通常是用来使此二储存虚拟化控制器彼此的相互状态同步。

10 当一储存虚拟化控制器检测到一个连接到输出装置连结的储存单元无法再被存取，而此输出装置连结是与此储存虚拟化控制器的扩充端口的一连接，则不论是因为检测到此扩充端口/连结发生断掉/故障或是其它原因，此检测的储存虚拟化控制器会送出输出请求给另一个储存虚拟化控制器，使此另一个储存虚拟化控制器经由互补扩充端口/连结以及另一个储存单元端口而发送给相同的储存单元。在此输出请求执行期间，
15 任何与输出请求有关的数据/状态会在此二个储存虚拟化控制器间转移。如果此另一个储存虚拟化控制器的扩充端口/连结是在准备好可活动状态下且正常运作时，但是此另一个储存虚拟化控制器存取此储存单元仍然失败，则此储存单元就会被视为已失效或已被移除了。反之，若存取成功，则前述的存取能力丧失将被视为局限在原本的储存虚拟化控制器，并且
20 往后与此储存单元存取关联的输出请求会自动地复位传送路径至另一个储存虚拟化控制器，而经由互补扩充端口/连结发送出。在此期间，原本的储存虚拟化控制器经由其扩充端口/连结监视对此储存单元的可存取性，此功能一般是藉由周期地发送出内部产生的输出请求以检验此连结与储存单元的状态。如果在某一时点，此原本的储存虚拟化控制器发现
25 此储存单元现在可经由其扩充端口/连结而被存取，则停止输出请求复位传送路径至另一个储存虚拟化控制器，并且开始再一次地经由其自己的扩充端口/连结直接发送输出请求。图 32 即为上述操作的流程图。

此外，在一个储存虚拟化控制器中还有可能使用冗余主机端连结架构。在此种架构下，储存虚拟化控制器中包含有多个主机端连结端口时，
30 以将逻辑介质单元以相同的形式通过二个或更多的主机端连结呈现至主

机。此种设计架构的目的是，即使其中一条主机端连结或端口断掉、阻断或故障了，主机能仍维持对此逻辑介质单元进行存取。

图 33 描绘一冗余外部储存虚拟化计算机系统的示意图，其中在储存虚拟化控制器中的两个个别的主机端端口，连接至两个完全分开的主机端输出装置连结以及主机端口。在一储存虚拟化控制器上的每一个端口在其另一个储存虚拟化控制器上会有一个互补端口。在一代表性的主机端连结支持冗余的实施例中，每一储存虚拟化控制器将同一逻辑介质单元以相同的形式呈现至其二个主机端端口。

在正常操作下，主机能经由一个储存虚拟化控制器存取逻辑介质单元，其中该储存虚拟化控制器被配置为在一主机端连结上呈现此逻辑介质单元，这可能是冗余对其中一个或两个储存虚拟化控制器。如果其中一储存虚拟化控制器故障，则已经被此两个储存虚拟化控制器呈现至主机的逻辑介质单元将通过仍正常工作的储存虚拟化控制器维持存取，并且，在检测到由这些储存虚拟化控制器之一所处理的输出请求发生中断时，经由在主机上特别用途的多冗余路径 (multiple-Redundant-pathing) 功能的帮助，输出请求就会完全被递送到此正常工作的储存虚拟化控制器。

那些原本只经由现在故障中的储存虚拟化控制器而呈现至主机的逻辑介质单元，将立刻藉由正常工作的储存虚拟化控制器经由连接至主机的主机端连结而呈现至主机。对于这些逻辑介质单元以及连同所有重新指定的逻辑介质单元，此正常工作的储存虚拟化控制器能单藉藉由在每一连结上呈现所有逻辑介质单元而透明地接替主机输出请求的处理，其是以相同于此故障中的储存虚拟化控制器在其故障之前所做的方式进行。藉此种“透明接替 (transparent takeover)”，主机不需要特殊的功能去查觉储存虚拟化控制器故障以及由主机自己对输出请求复位路径传送以为响应。

除了储存虚拟化控制器的冗余架构以外，二组互补端口系也形成一冗余端口互补 (redundant port complement)。一主机中有两个独立端口使用两个个别的输出装置连结与这二个互补冗余端口组连接，则会有两条独立的路径至每个逻辑介质单元，且可经由该路径发出输出请求。若主机的一端口或一储存虚拟化控制器故障，或若输出装置连结断掉或阻断时，具有多冗余路径功能的主机即可通过另一冗余路径将输出请求复位

路径发送。另一种情况是，当两条路径都正常运作时，主机会在此二路径上选择发送输出请求的路径以求尽力平衡此二路径间的负载，此一技术称为负载平衡（Load Balancing）。

为了实现上述透明接替的功能，在每一储存虚拟化控制器上所形成的互补端口对中的每一端口对是物理上地相互连结。对于总线式的多装置输出装置连结而言，如：并列 SCSI，此连结仅仅由直接将装置布线在一起而构成，而不需任何中介电路。对于其它连结的型式，特殊的开关电路可用来实现上述所需的物理连结。图 34 显示一开关电路的例子，此开关电路能用来在光纤连结上实现此种连结，其中硬件讯号检测 SDC 被用来启动此开关状态的变更。在主机实行多冗余路径功能的架构中，有另一种主机端连结的结构，此结构仅需要较少数的连结即可实现类似的冗余特性，如图 35 中所示。需注意的是，其中用于将一储存虚拟化控制器连接该主机的主机端连结并未连结至另一个储存虚拟化控制器。在此一架构下，连结冗余藉由让每一逻辑介质单元，若可通过一个储存虚拟化控制器的主机端连结而可被主机存取，则亦通过另一个储存虚拟化控制器上的另一个主机端连结而可被主机存取的方式来实现。若其中一连结断掉、阻断或因其它原因故障，主机仍然能由另一个储存虚拟化控制器经由另一个连结而存取此逻辑介质单元。同样地，若储存虚拟化控制器中的一个故障，另一个储存虚拟化控制器能接替并且，再一次地主机仍然能由正常工作的储存虚拟化控制器经由另一个连结存取此逻辑介质单元。

经由上述说明可知，依据本发明的冗余储存虚拟化子系统的一实施态样如下，冗余储存虚拟化子系统包含两个冗余储存虚拟化控制器而形成一冗余对，该冗余储存虚拟化控制器对的至少一成员储存虚拟化控制器是包含有至少一扩充端口，可经由一多装置端输出装置连结而耦接至一包含有至少一物理储存装置的第二组物理储存装置；该第二组物理储存装置中的一成员具有一对冗余端口，且该对冗余端口中的一成员端口耦接至一包含至少一该扩充端口的扩充端口组；该冗余物理储存装置端口对的该成员端口与该包含至少一该扩充端口的扩充端口组经由一开关电路互相耦合在一起；且该扩充端口组包含有一第一与一第二扩充端口次组，以形成一对互补次组，该每一次组系包含有至少一成员扩充端口；该开关电路可实现一连结讯号线的开关机制而支持下列各种安排：（1）耦接该对互补次

组的该第一次组至该物理储存装置冗余端口对的一第一成员端口，以及耦
接该对互补次组的该第二次组至该物理储存冗余端口对的一第二成员端
口；（2）耦接该对互补次组的二个次组至该物理储存冗余端口对的该第
一成员端口；（3）耦接该对互补次组的二个次组至该物理储存冗余端口
5 对的该第二成员端口；（4）耦接该对互补次组的该第一次组至该物理储
存冗余端口对的该第一成员端口；（5）耦接该对互补次组的该第二次组
至该物理储存冗余端口对的该第二成员端口；（6）耦接该对互补次组的
该第二次组至该物理储存冗余端口对的该第一成员端口；以及（7）耦接
该对互补次组的该第一次组至该物理储存冗余端口对的该第二成员端口。

10 冗余 SATA 储存虚拟化子系统的一变化型为使用 PATA 物理储存装置
而不使用 SATA 物理储存装置。对每一个 PATA 物理储存装置，于存取控制
开关与此物理储存装置间，紧靠着 PATA 物理储存装置，需要安插一个 SATA
转 PATA 转换电路，且典型地是与此存取控制开关一起位于同一个现场可
15 更换单元中。此转换电路将 SATA 讯号及协议，转换成 PATA 讯号及协议，
并于相反方向时再转换回 SATA 讯号及协议。实际上，在短期来说，与 PATA
相比 SATA 磁盘的供应仍然短缺，且它的单价亦不算太便宜，因此使用 PATA
物理储存装置来替代 SATA 物理储存装置用于一 SATA 储存虚拟化子系统中
有其重要性。在此过渡期间，此种子系统让 PATA 物理储存装置可以替代
20 SATA 物理储存装置，消除了 SATA 物理储存装置供应上及成本上的顾虑。
在这样的子系统中通常可将转换电路与存取控制开关一起放置于存放物
理储存装置的可拆卸盒(removable canister)之中，因此，当后续有物理
储存装置或相关电路需要进行维修服务时，可以很容易地从系统上拆卸下
来。此外，藉由将转换电路设置于可拆卸盒当中，在 SATA 磁盘价格降低
至较可接受的程度时，原先装设 PATA 的可拆卸盒即可很方便的整个从系
25 统中移除，并将新的 SATA 物理储存装置及相关电路设置到系统上。

请参阅图 36 与图 38。图 36 为可拆卸冗余 SATA 物理储存装置盒的方
块图，图 37 为包括在此盒中特有的印刷电路板的较详细的方块图，存取
控制开关位于其中。图 38 为一可拆卸冗余 PATA 物理储存装置盒的方块图，
图 39 为包括在此盒中特有的印刷电路板的较详细的方块图，存取控制开
30 关与 SATA 转 PATA 转换电路位于其中。此两者中皆有一对 SATA 输出入装
置连结以及一组来自于二个储存虚拟化控制器的存取控制开关控制讯号，

连接进入一存取控制开关。主要不同的地方在于可拆卸 PATA 物理储存装置盒中多了一个 SATA 转 PATA 转换电路，这在可拆卸 SATA 物理储存装置槽中是没有的。

以上所述仅为本发明的较佳实施例，凡依本发明的权利要求所做的均
5 等变化与修饰，皆应属本发明专利的涵盖范围。

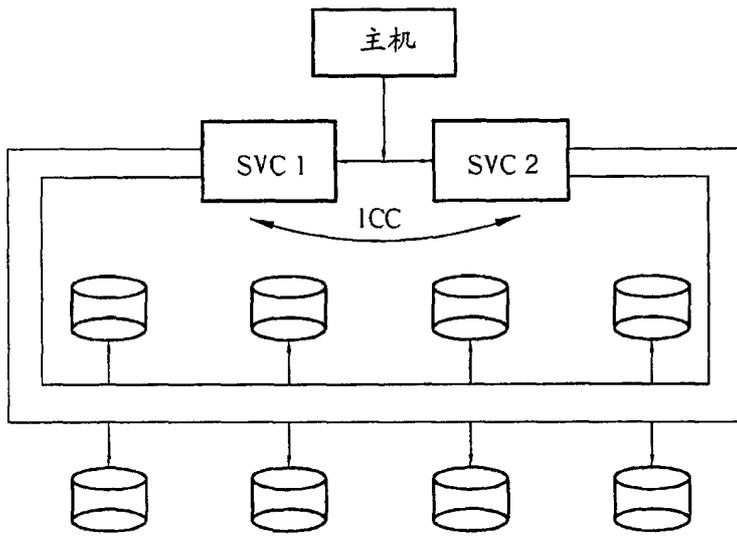


图 1

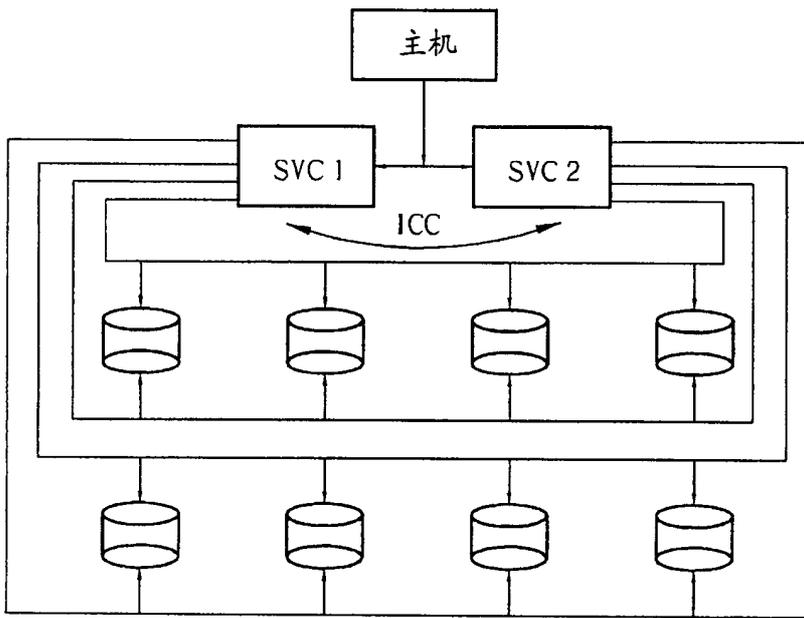


图 2

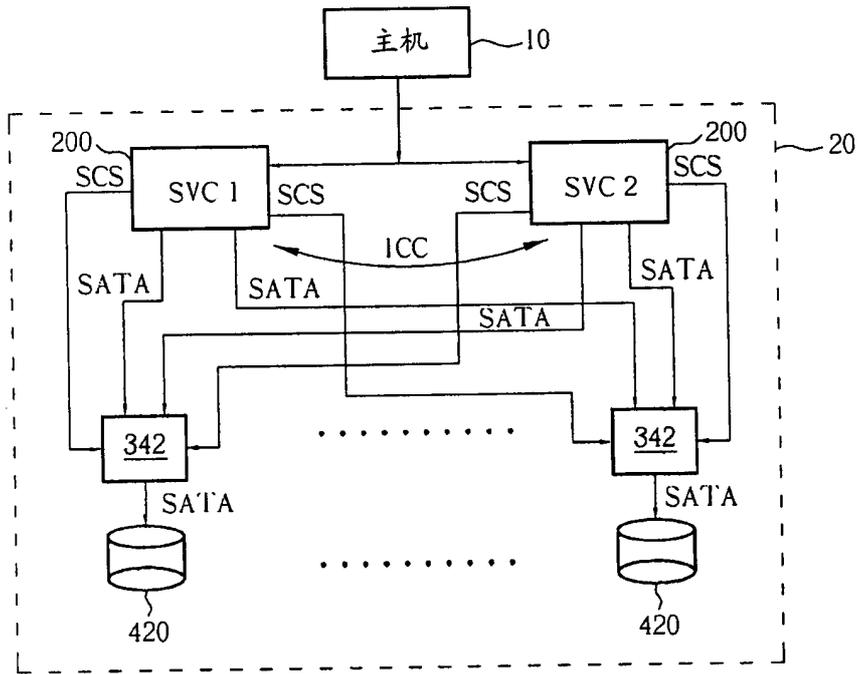


图 3

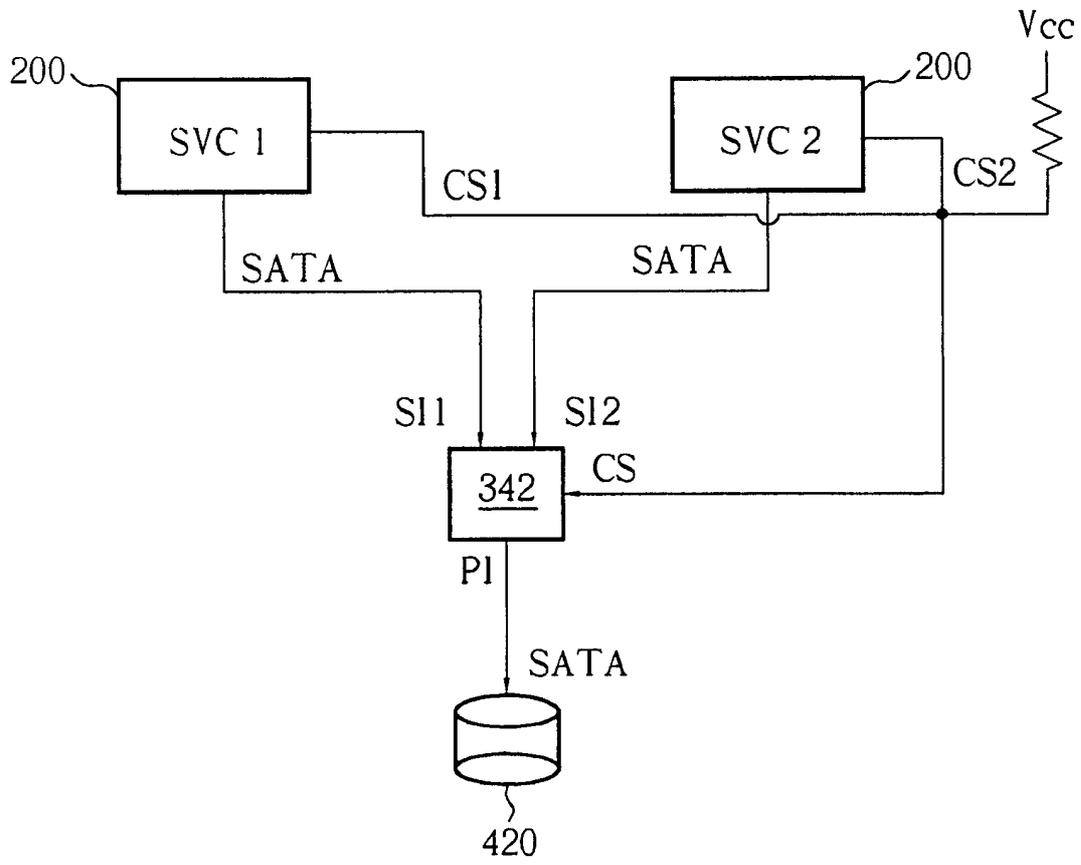


图 4

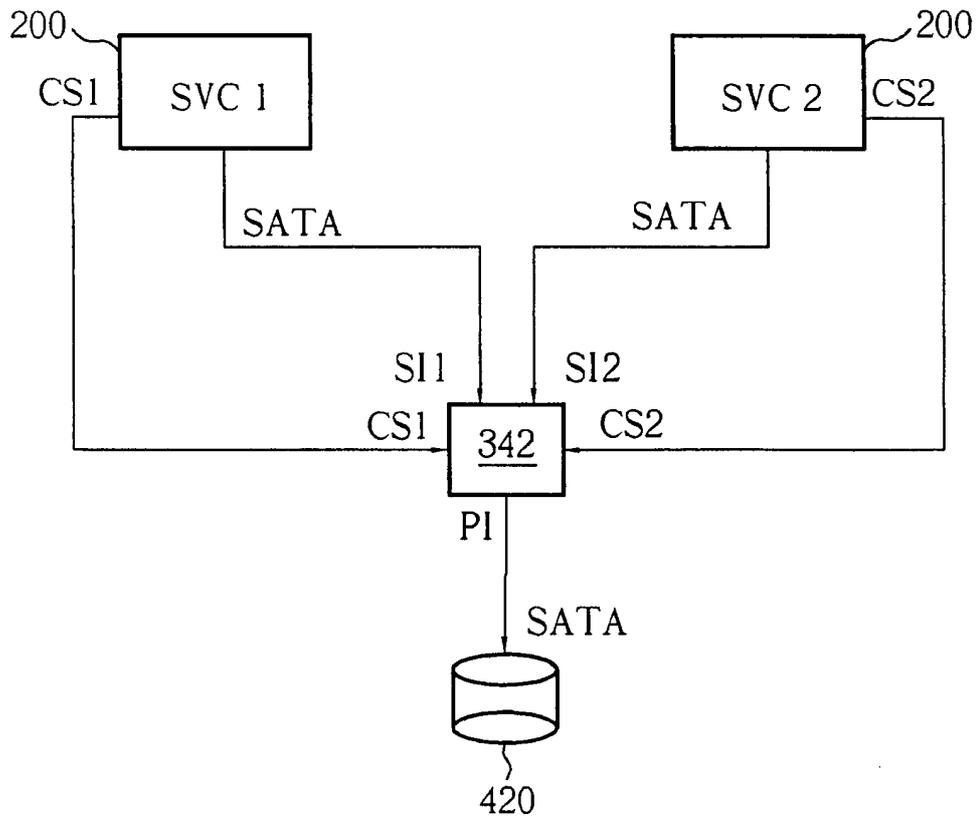


图 5

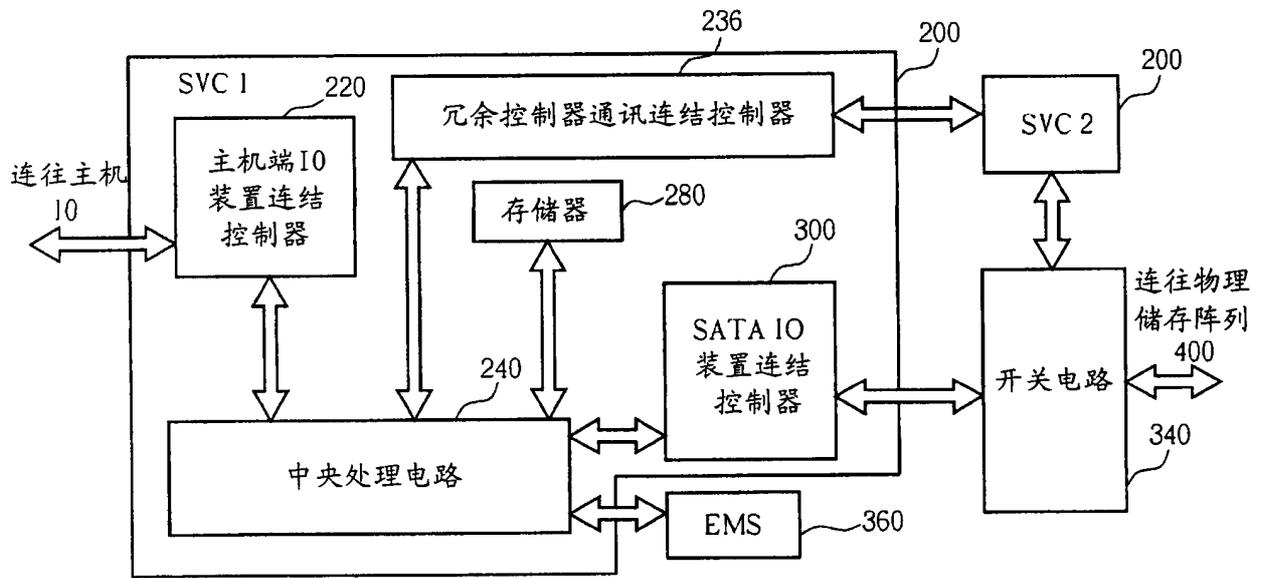


图 6

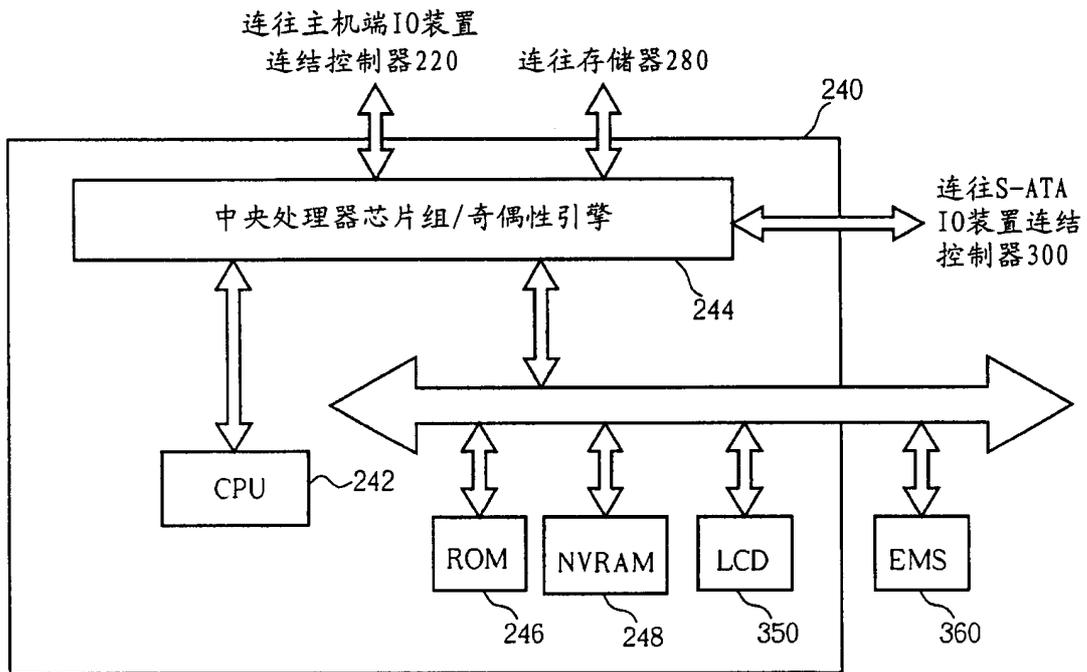


图 7

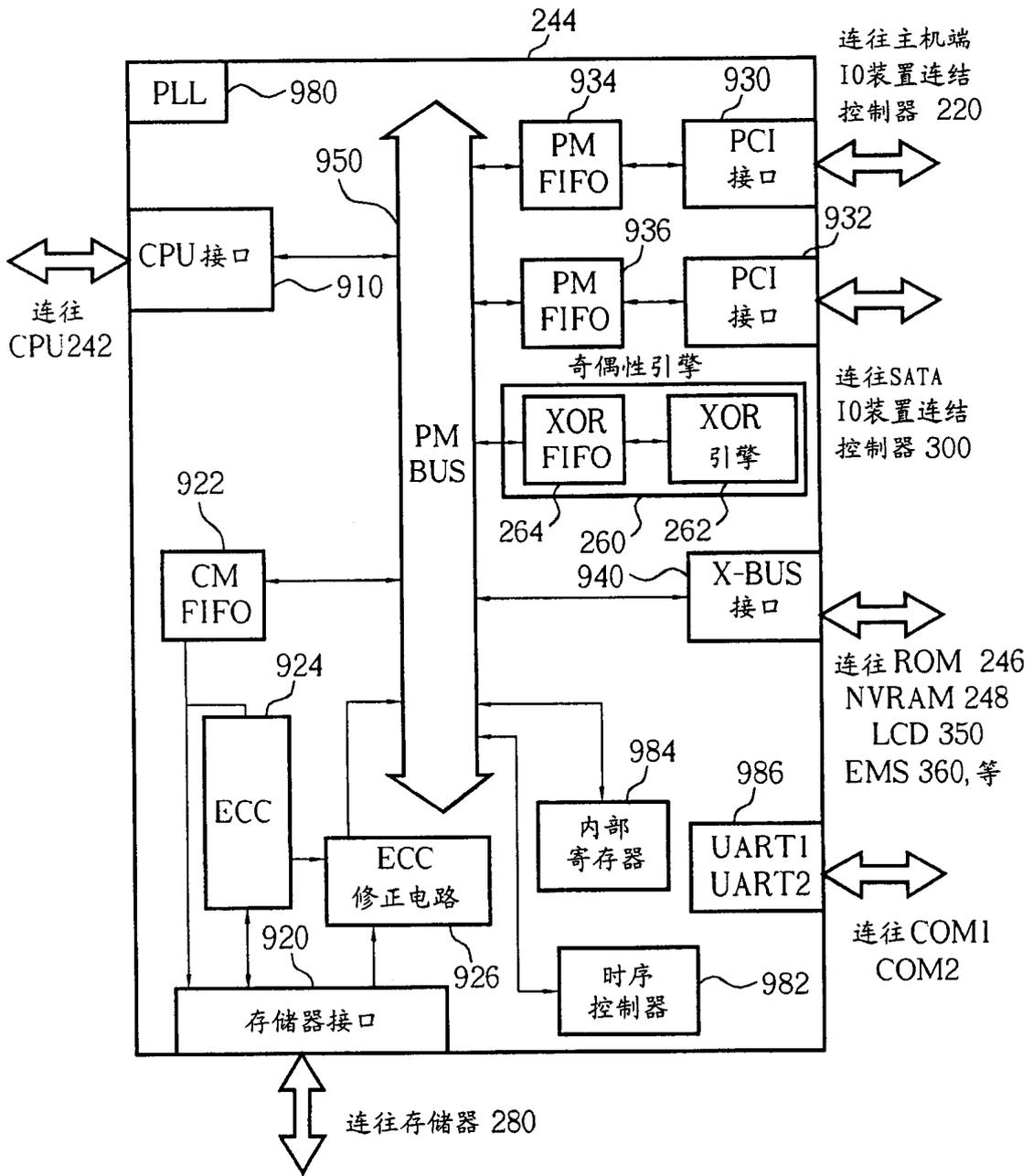


图 8

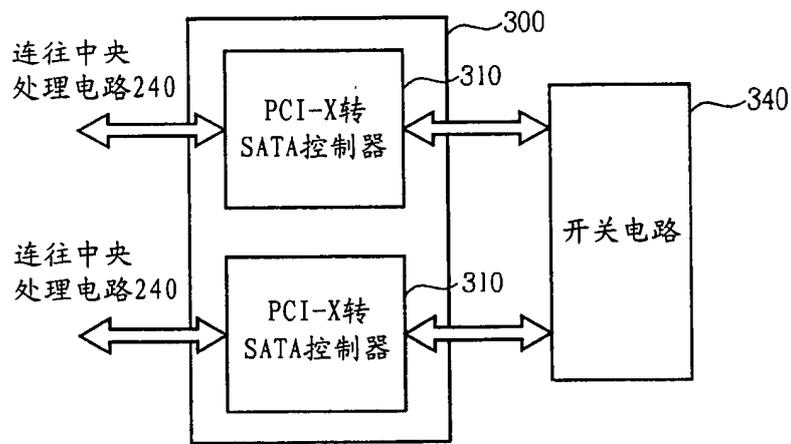


图 9

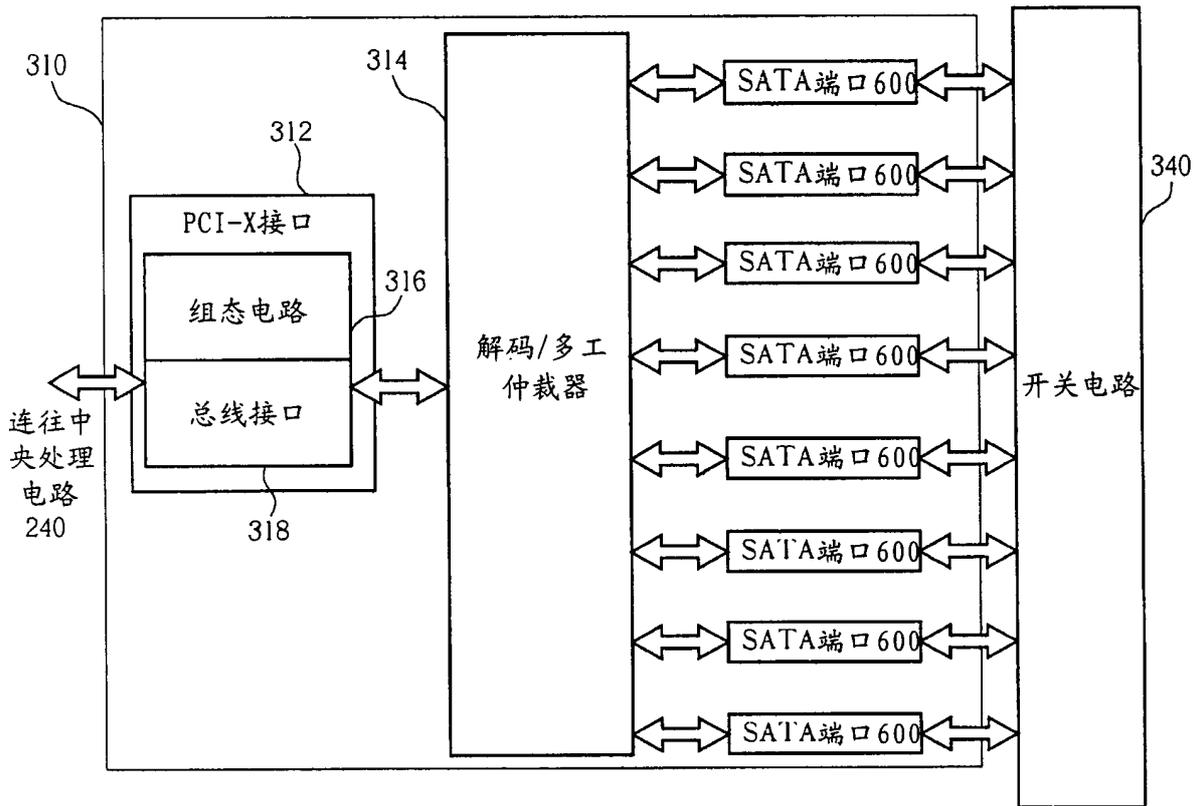


图 10

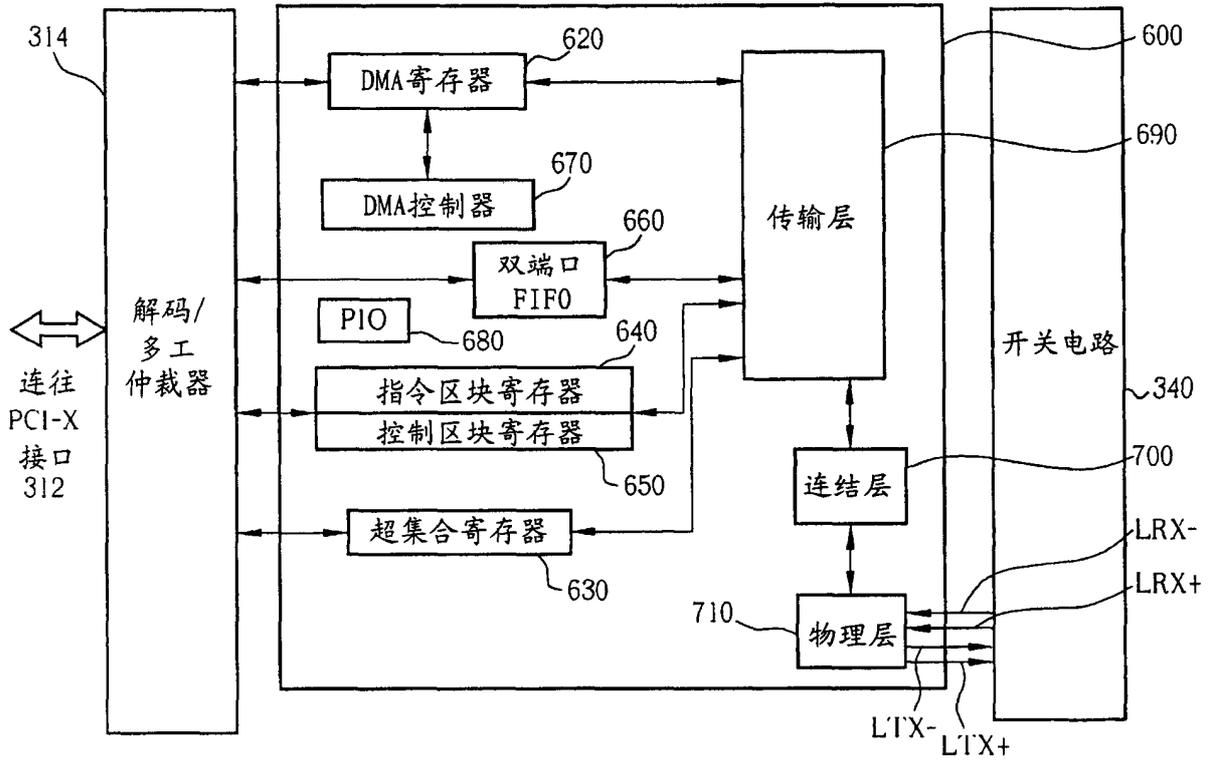


图 11

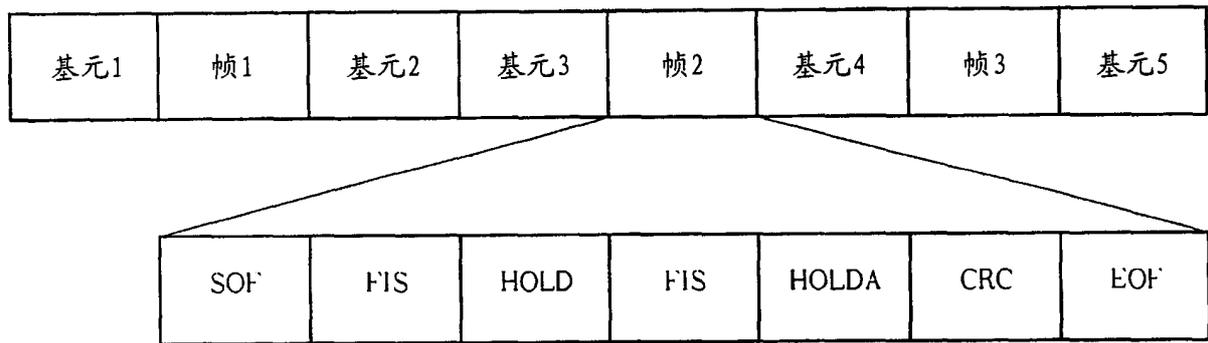


图 12

0			保留(0)				保留(0)			R I D		保留(0)			FIS 形态(41h)
1										DMA缓冲器识别码的低字段					
2										DMA缓冲器识别码的高字段					
3										保留(0)					
4										DMA缓冲器偏移字段					
5										DMA传送计数字段					
6										保留(0)					

图 13

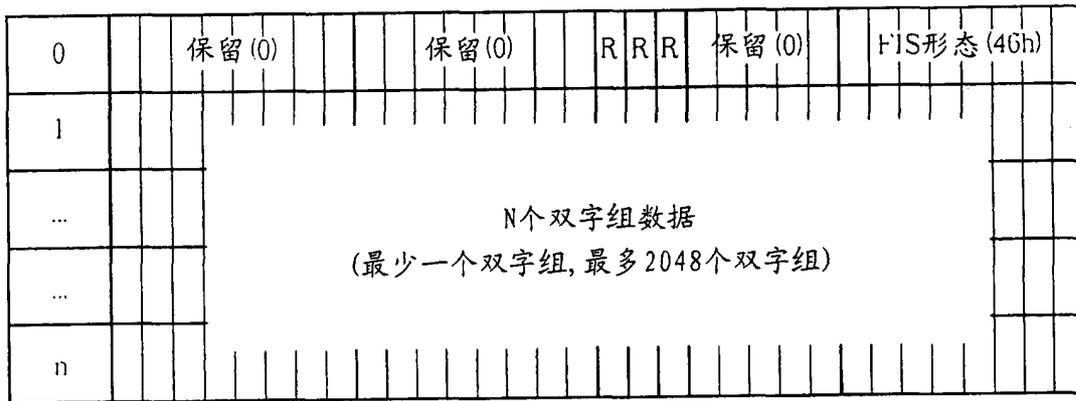


图 14

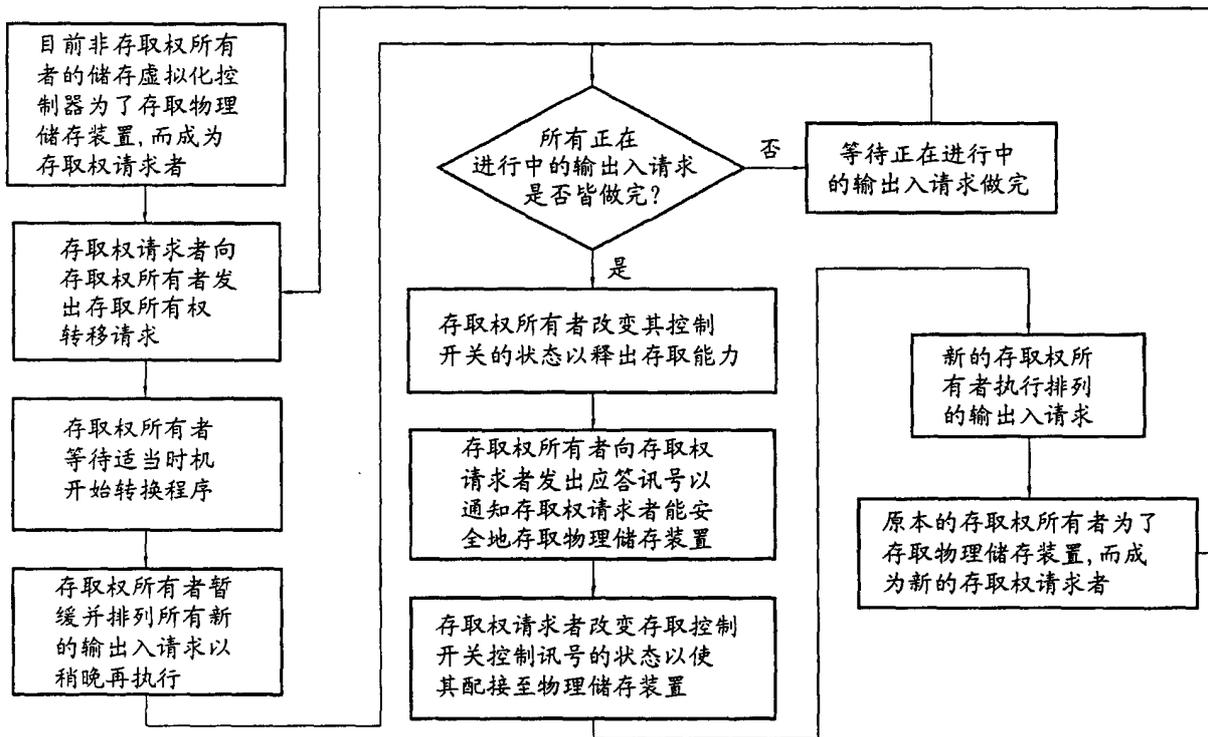


图 15

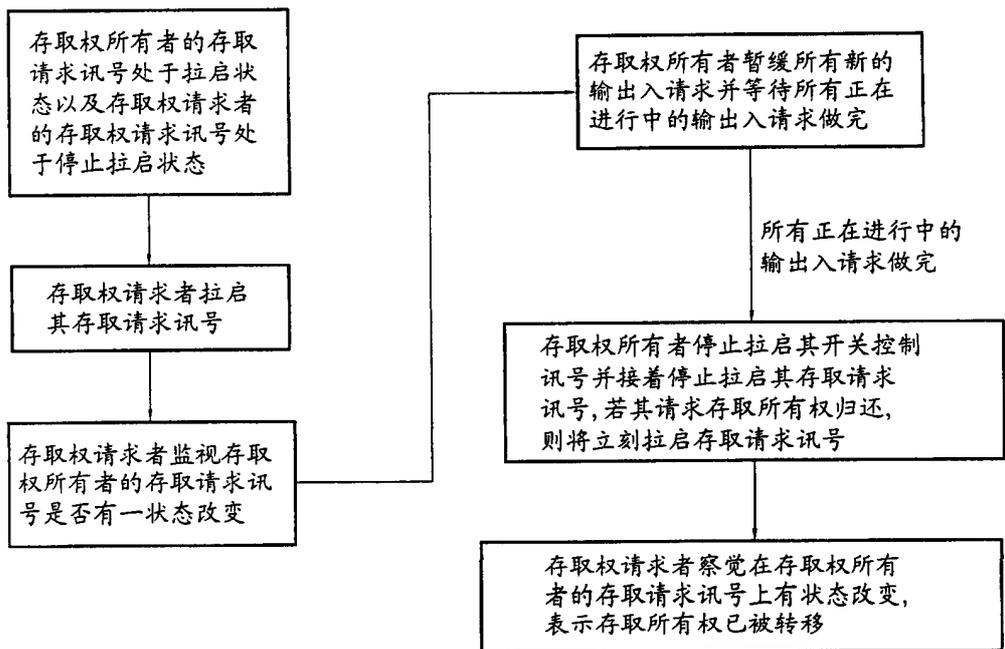


图 16

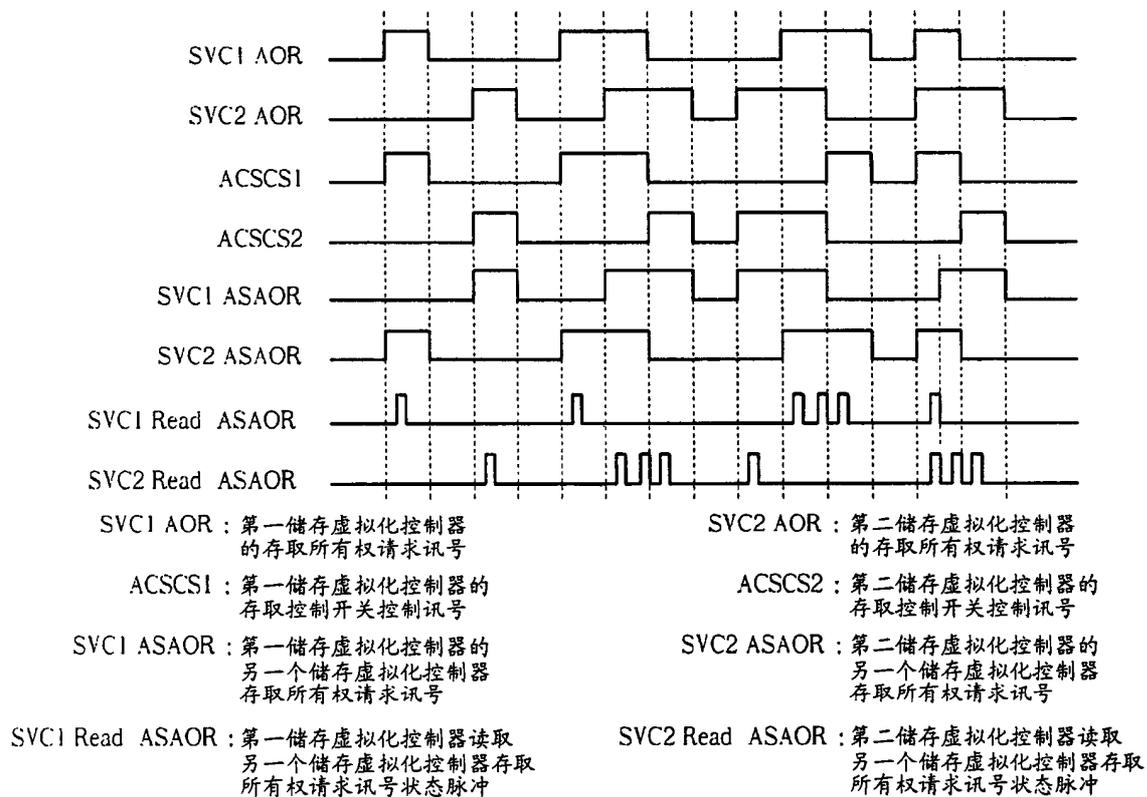


图 17

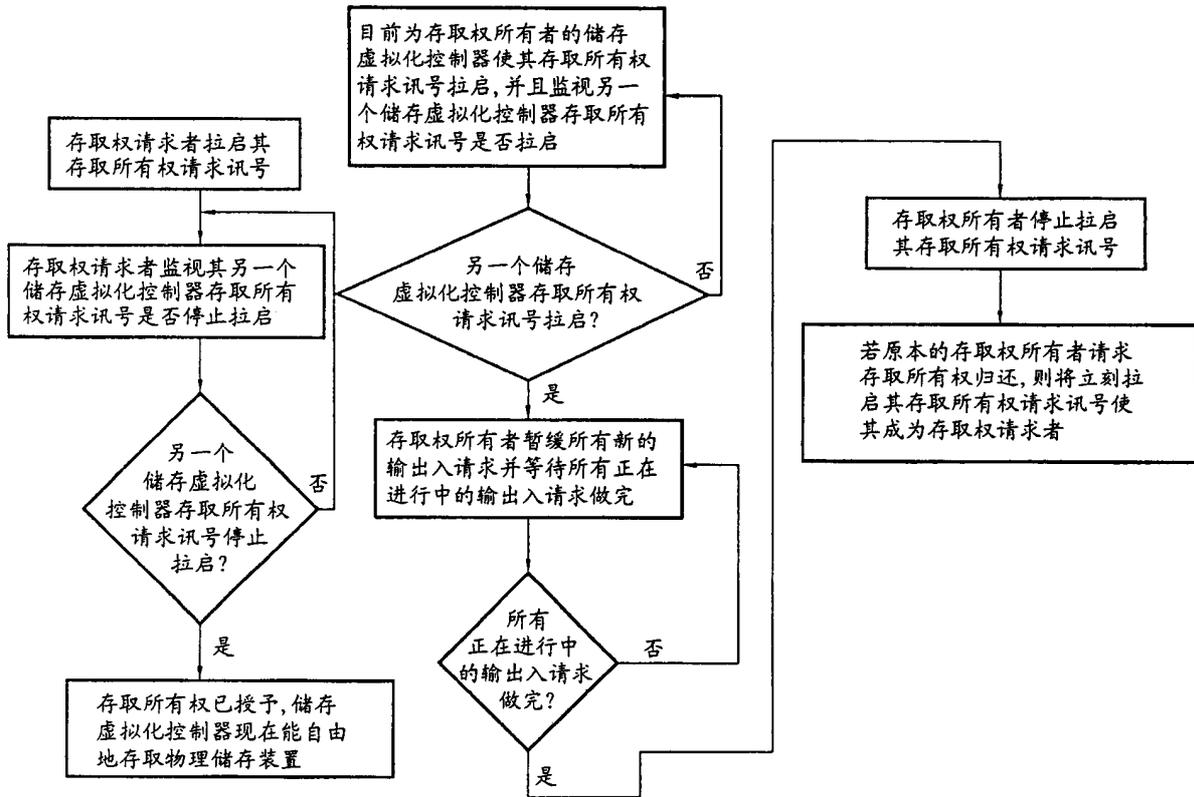


图 18

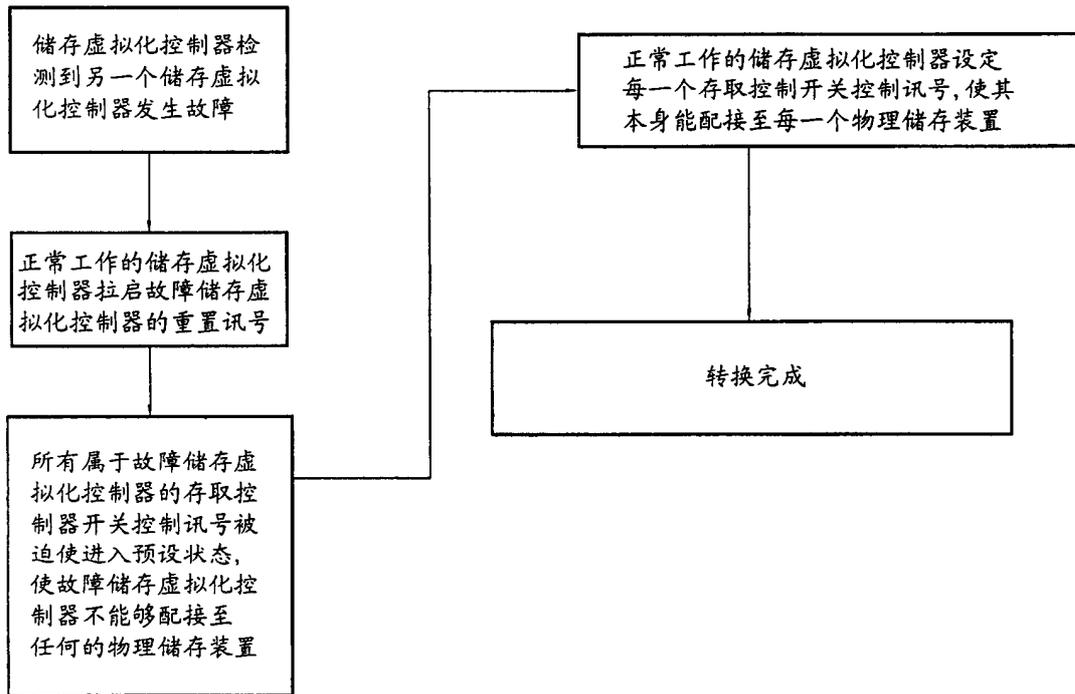


图 19

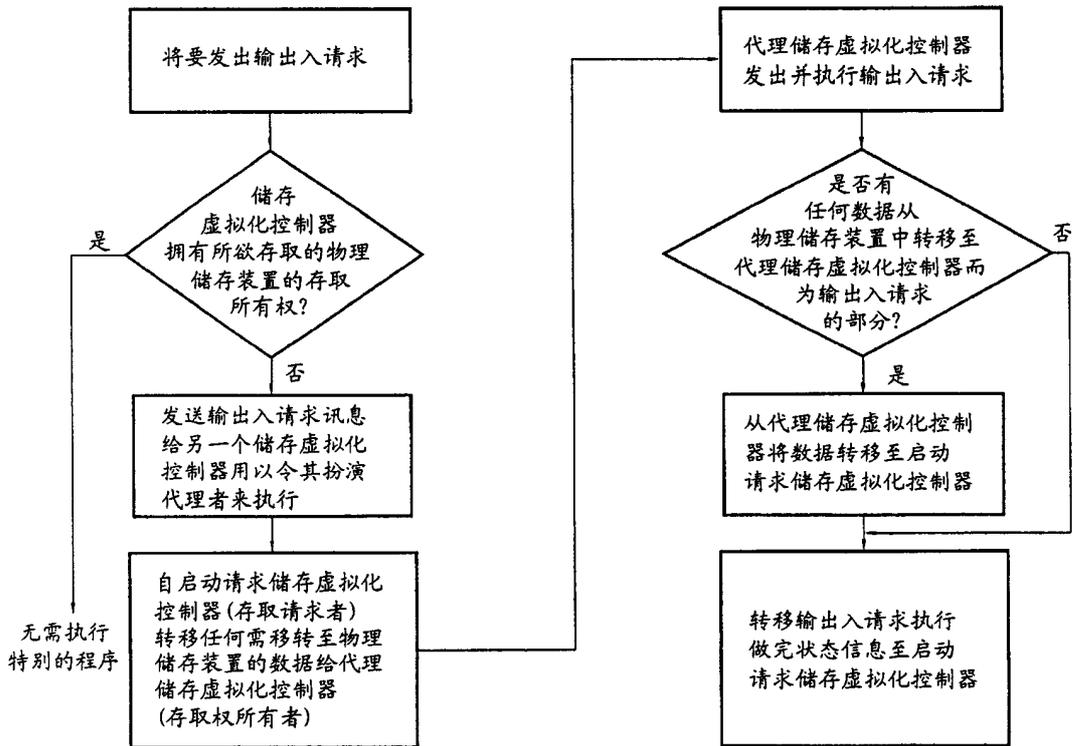


图 20

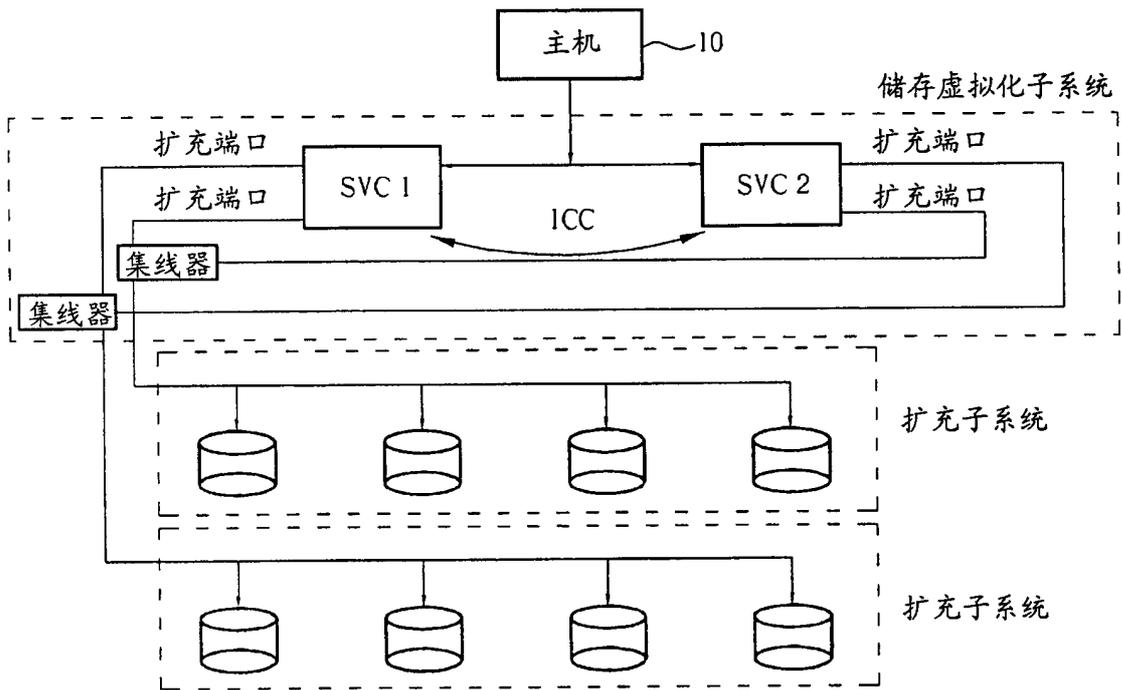


图 21

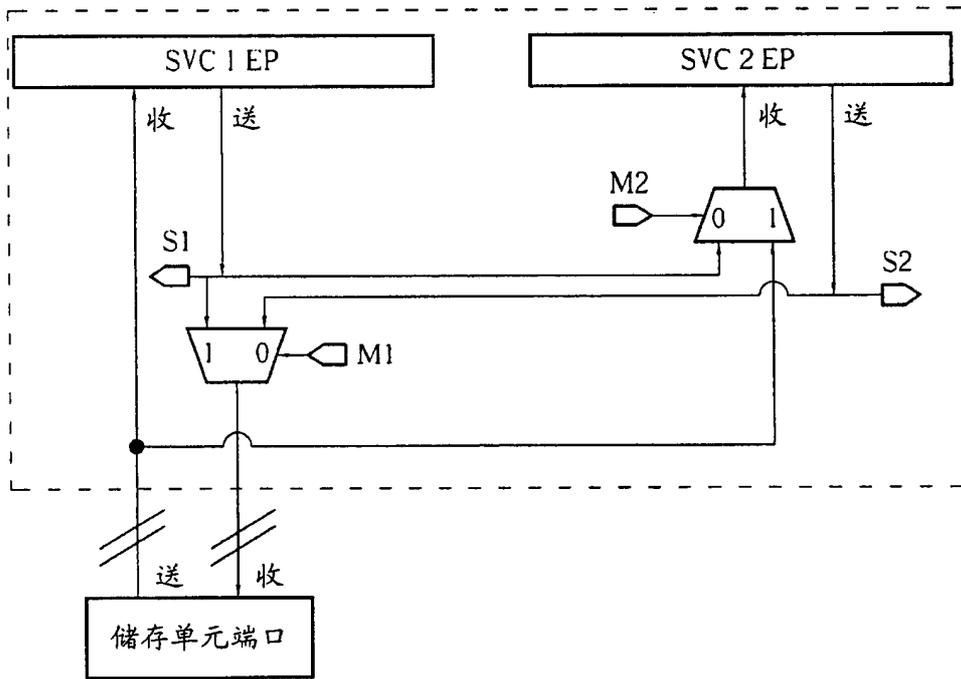


图 22

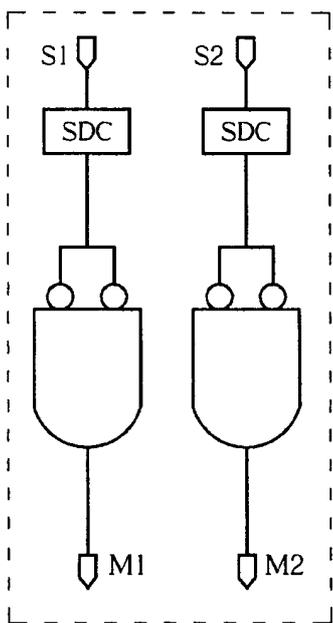


图 23

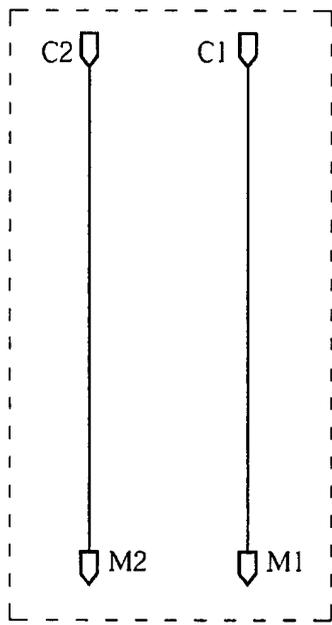


图 24

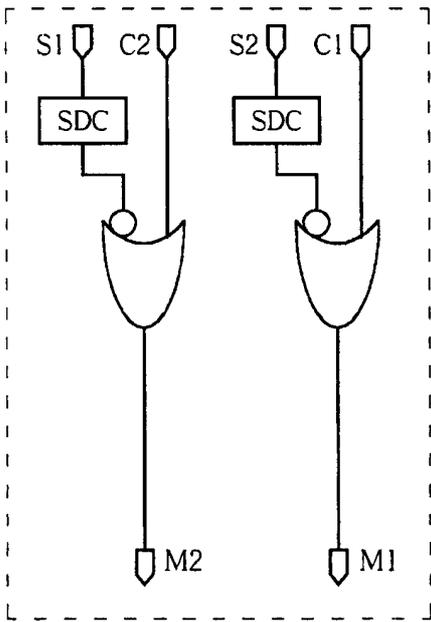


图 25

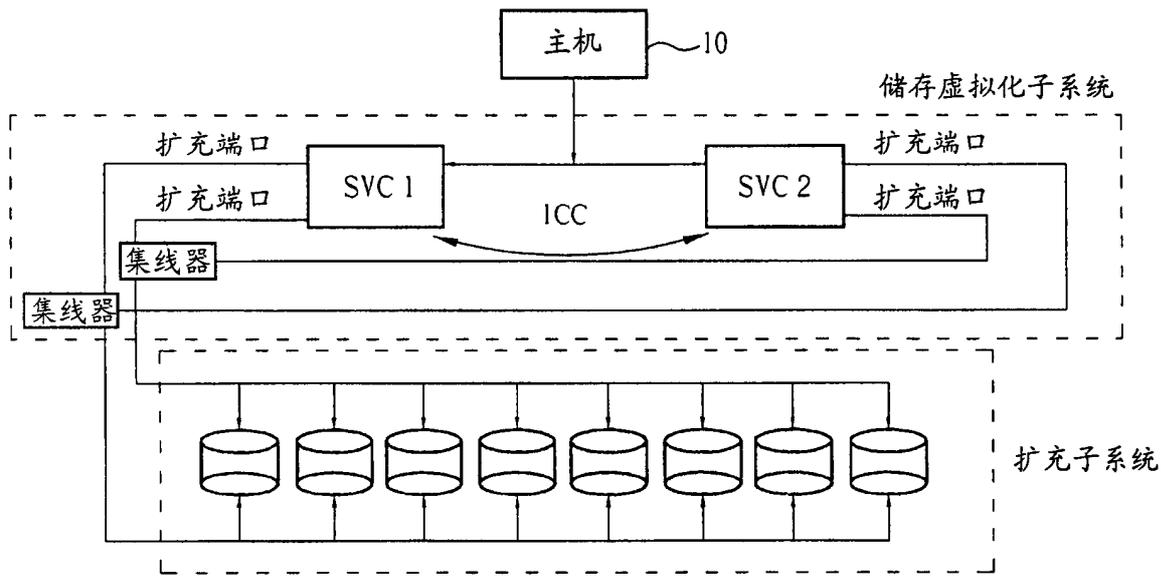


图 26

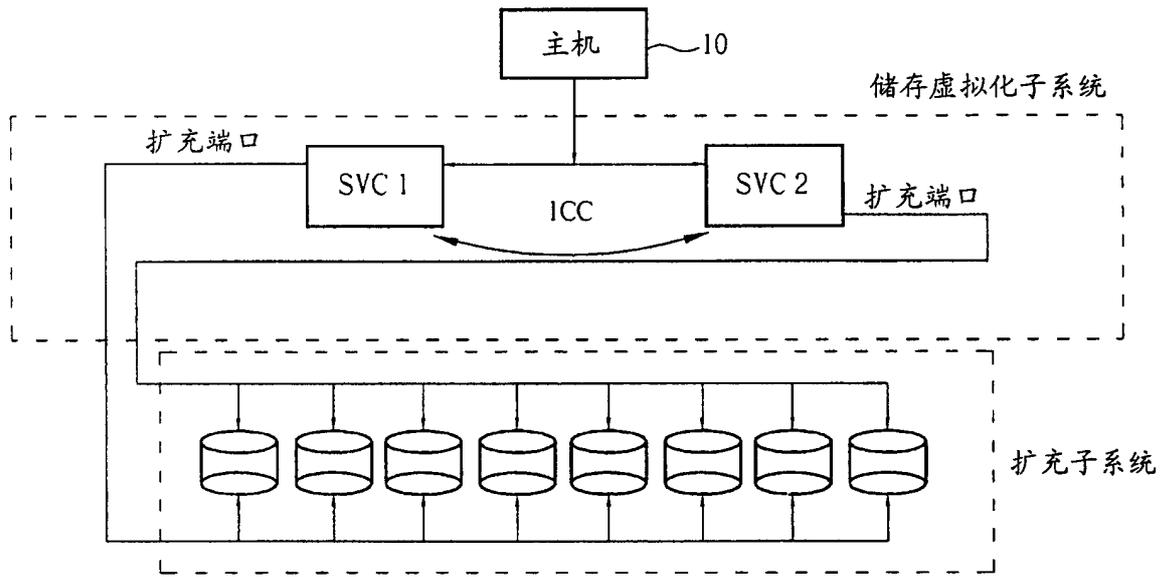


图 27

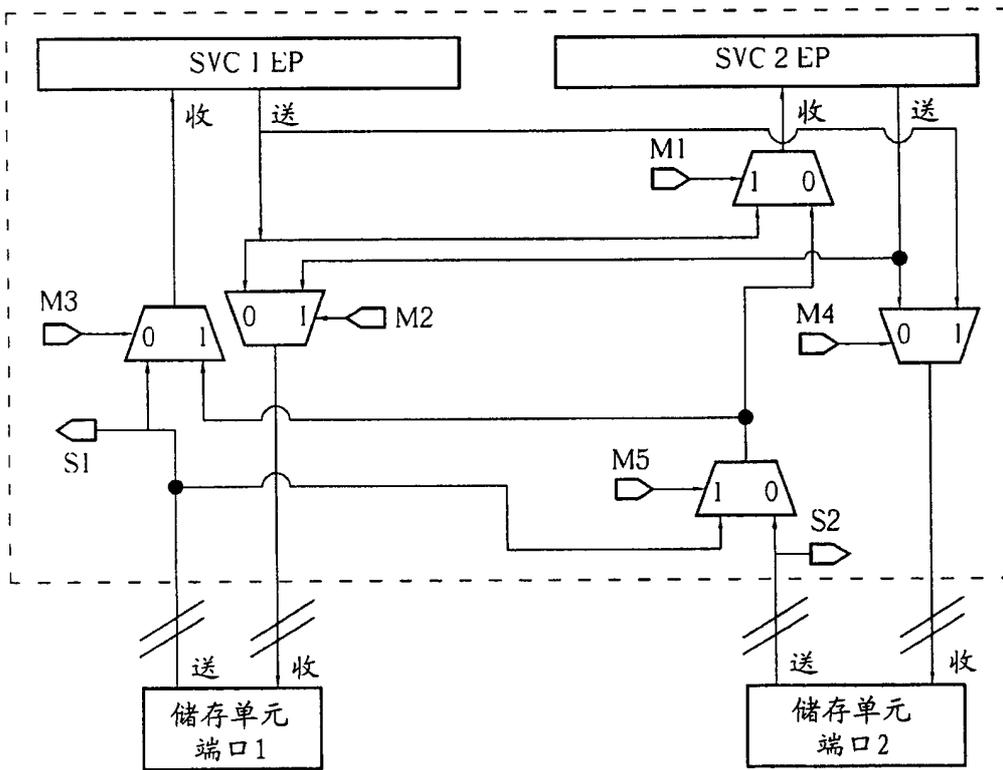


图 28

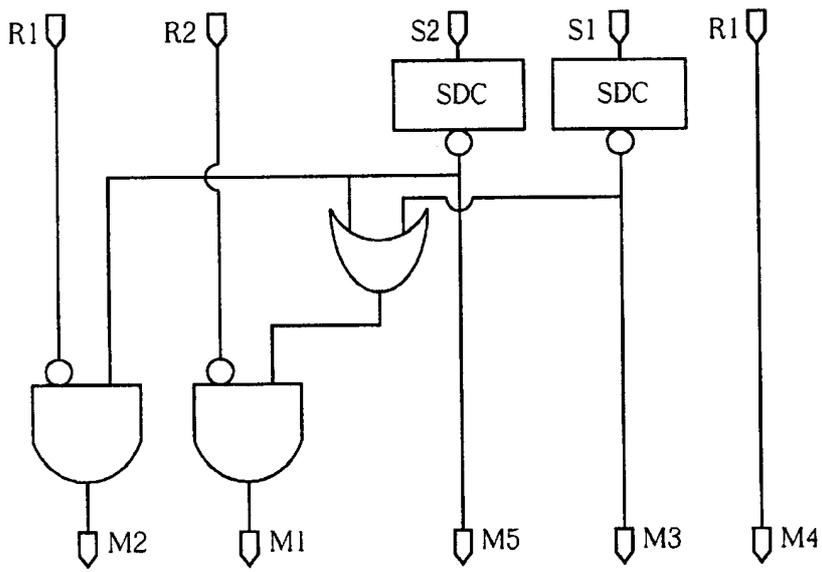


图 29

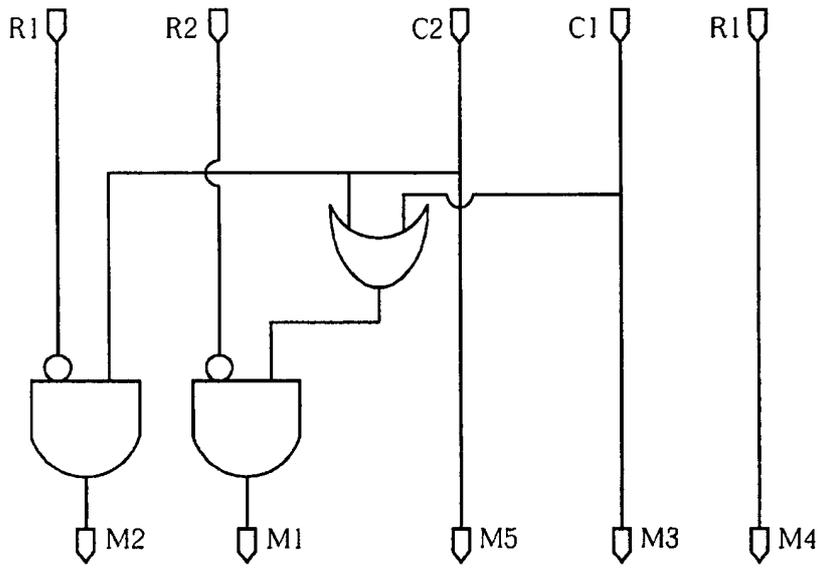


图 30

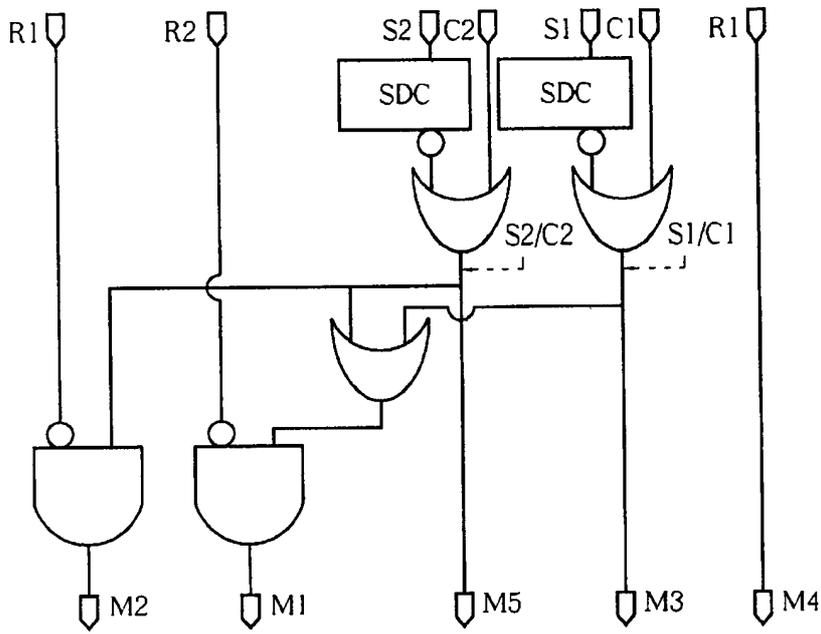


图 31

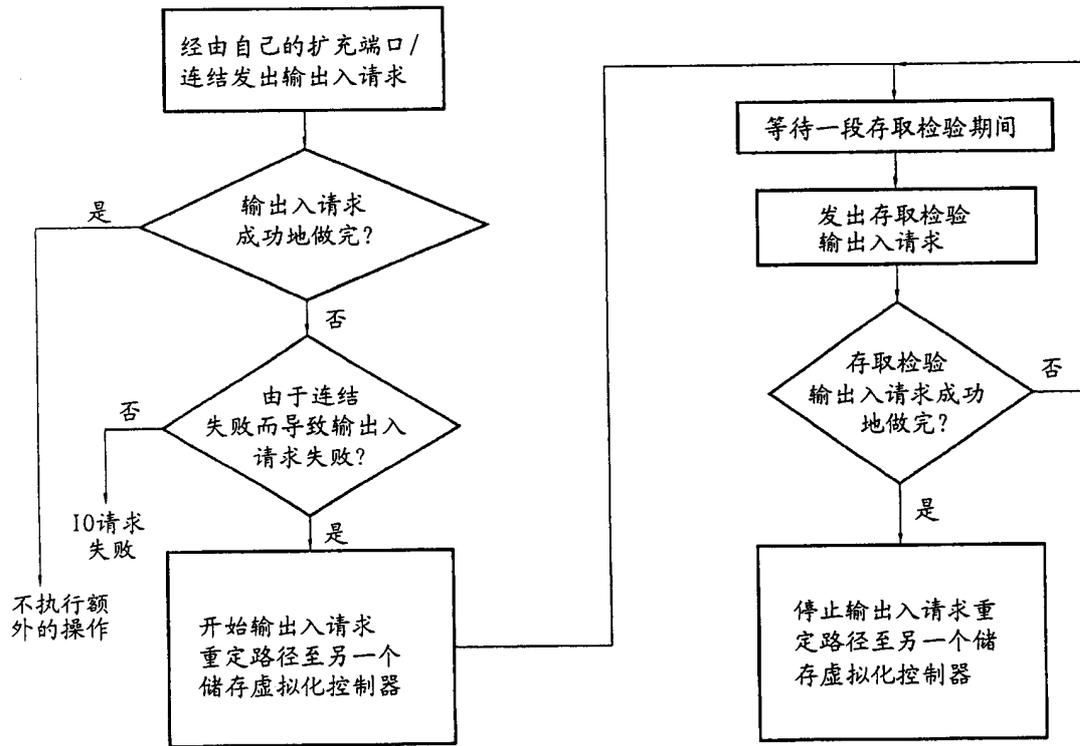


图 32

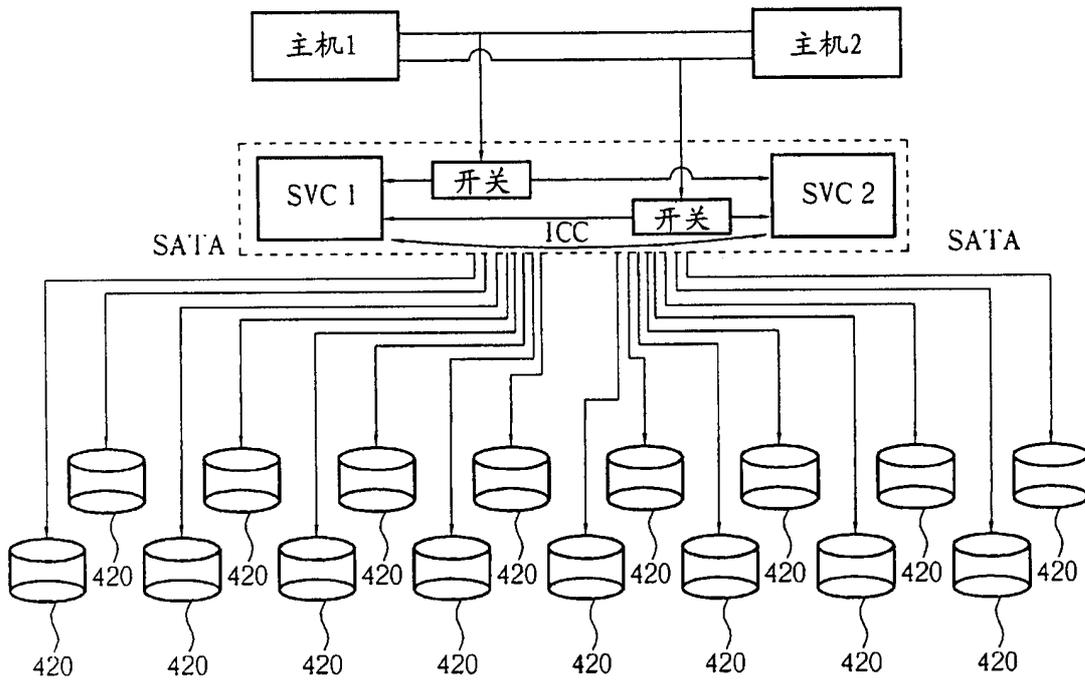


图 33

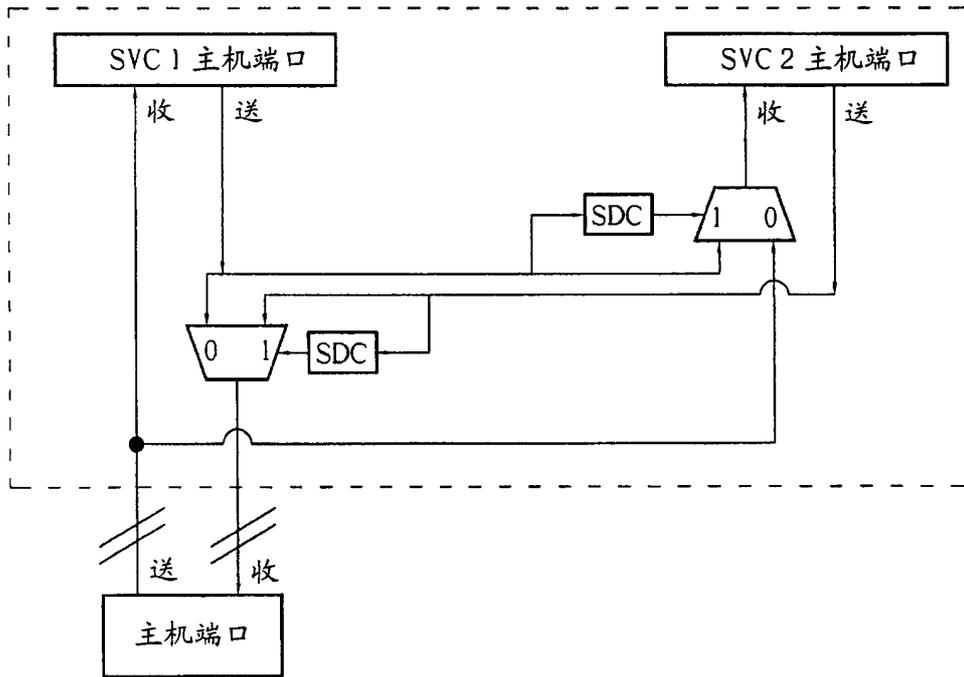


图 34

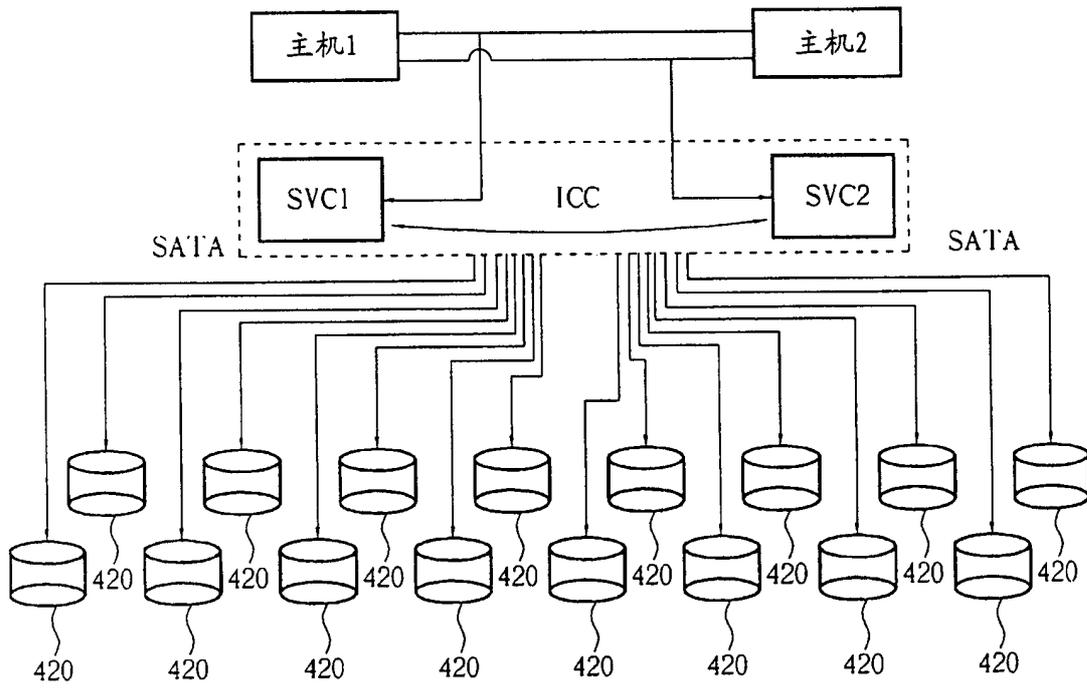


图 35

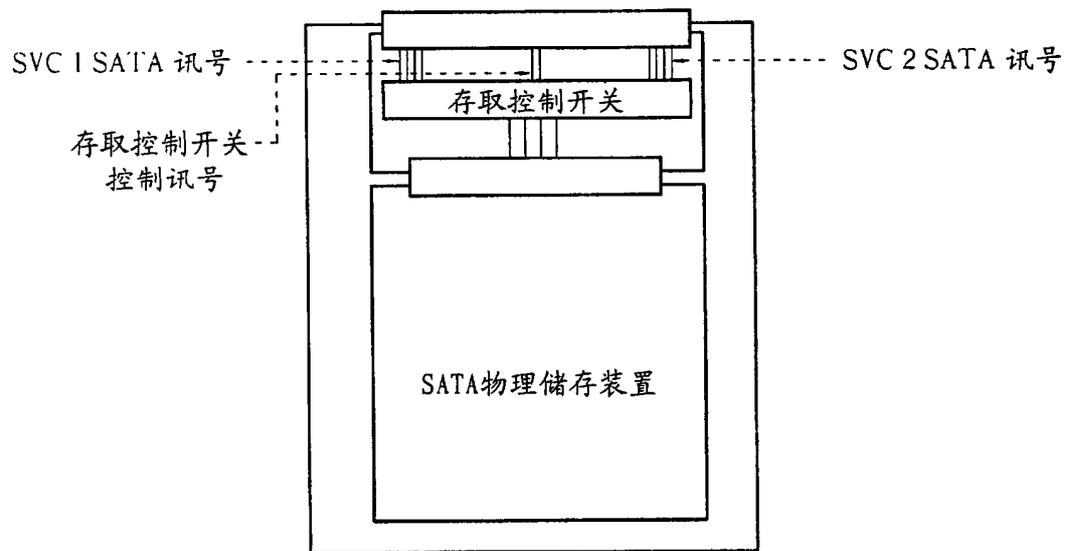


图 36

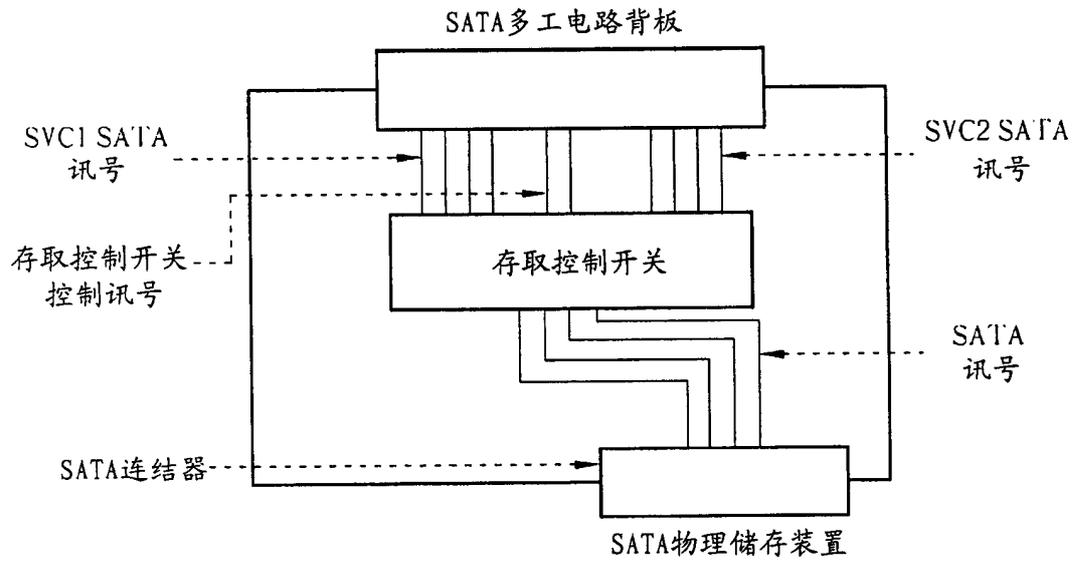


图 37

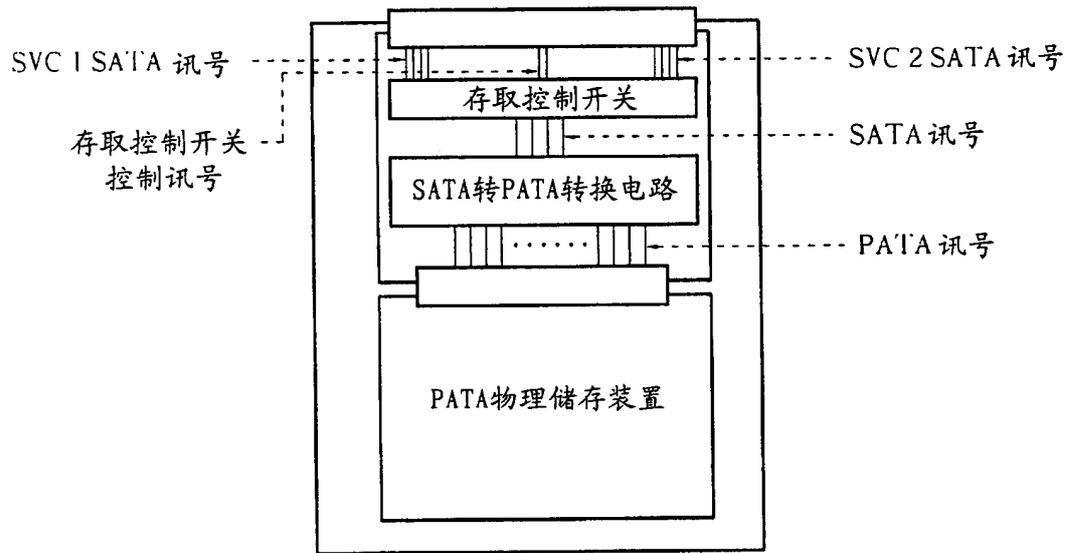


图 38

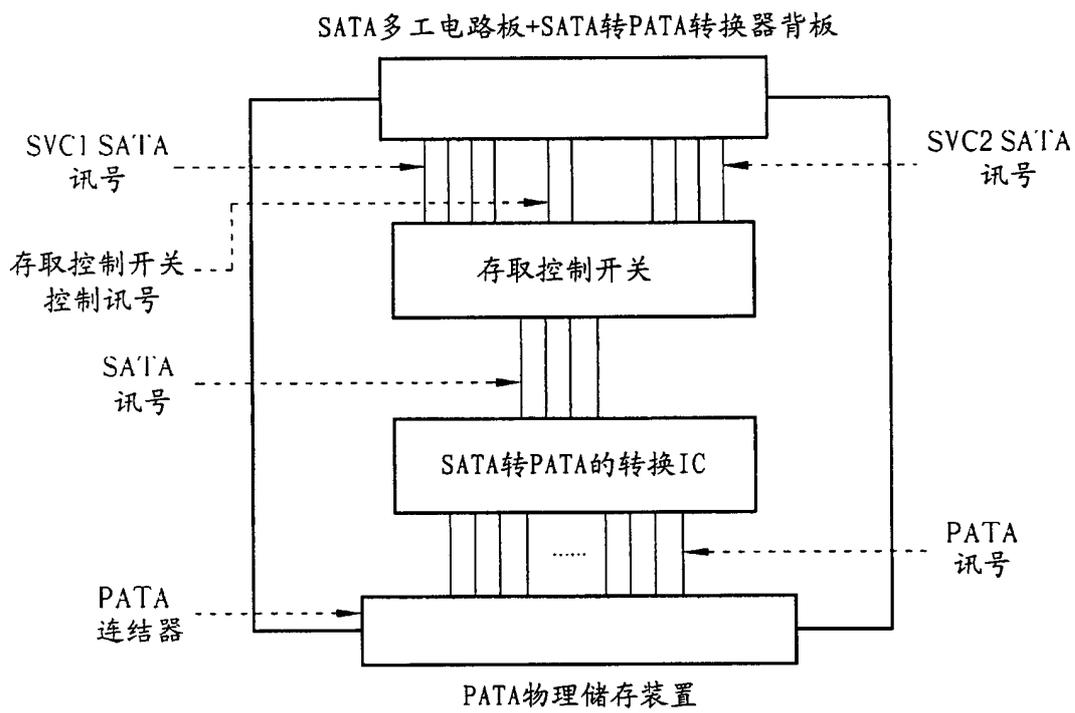


图 39

CS1	CS2	CS	SI1->PI	SI2->PI
0	0	0	导通	断路
0	Tri	0	导通	断路
Tri	0	0	导通	断路
Tri	Tri	1	断路	导通

图 40

CS1	CS2	SI1->PI	SI2->PI	注
0	0	Undef	Undef	*: 最新改变的CS1或CS2取得存取权
0	1	断路	导通	
1	0	导通	断路	
1	1	*注	*注	

图 41

Loop Connections	M1	M2	M3	M4	M5
$C1 \leftrightarrow P1, C2 \leftrightarrow P2$	0	0	0	0	0
$C1 \Rightarrow C2 \Rightarrow P1 \Rightarrow C1$	1	1	0	DC	DC
$C1 \Rightarrow C2 \Rightarrow P2 \Rightarrow C1$	1	DC	1	0	0
$C1 \leftrightarrow P1$	DC	0	0	DC	DC
$C1 \leftrightarrow P2$	DC	DC	1	1	0
$C2 \leftrightarrow P2$	0	DC	DC	0	0
$C2 \leftrightarrow P1$	0	1	DC	DC	1

Cn: SVCn, Pn: 储存单元端口 n; n=1, 2.

图 42

S1	S2	R1	R2	M1	M2	M3	M4	M5
Val	Val	0	0	0	0	0	0	0
Inv	Val	0	0	1	0	1	0	0
Val	Inv	0	0	1	1	0	0	1
Val	Val	1	0	0	0	0	1	0
Inv	Val	1	0	1	0	1	1	0
Val	Inv	1	0	1	0	0	1	1
Val	Val	0	1	0	0	0	0	0
Inv	Val	0	1	0	0	1	0	0
Val	Inv	0	1	0	1	0	0	1

图 43

C1	C2	R1	R2	M1	M2	M3	M4	M5
0	0	0	0	0	0	0	0	0
1	0	0	0	1	0	1	0	0
0	1	0	0	1	1	0	0	1
0	0	1	0	0	0	0	1	0
1	0	1	0	1	0	1	1	0
0	1	1	0	1	0	0	1	1
0	0	0	1	0	0	0	0	0
1	0	0	1	0	0	1	0	0
0	1	0	1	0	1	0	0	1

图 44

S1/C	S2/C	R1	R2	M1	M2	M3	M4	M5
0	0	0	0	0	0	0	0	0
1	0	0	0	1	0	1	0	0
0	1	0	0	1	1	0	0	1
0	0	1	0	0	0	0	1	0
1	0	1	0	1	0	1	1	0
0	1	1	0	1	0	0	1	1
0	0	0	1	0	0	0	0	0
1	0	0	1	0	0	1	0	0
0	1	0	1	0	1	0	0	1

图 45