[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 12/08 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200780046866.4

[43] 公开日 2009 年 10 月 21 日

[11] 公开号 CN 101563677A

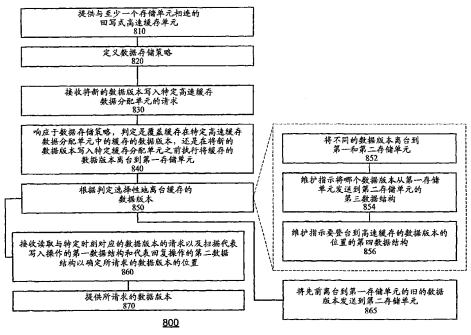
[22] 申请日 2007.12.19
[21] 申请号 200780046866.4
[30] 优先权
[32] 2006.12.20 [33] GB [31] 0625330.6
[86] 国际申请 PCT/EP2007/064213 2007.12.19
[87] 国际公布 WO2008/074830 英 2008.6.26
[85] 进入国家阶段日期 2009.6.18
[71] 申请人 国际商业机器公司
地址 美国纽约
[72] 发明人 M·法克特 S·菲恩布利特
G·拉登 D·H·洛仑兹
S·S·平特 P·K·塔-施马
[74] 专利代理机构 北京市中咨律师事务所
代理人 于 静 杨晓光
权利要求书 7 页 说明书 25 页 附图 16 页

[54] 发明名称

使用回写式高速缓存单元管理数据的系统、方法和计算机程序产品

[57] 摘要

一种管理数据的方法，所述方法包括：提供与至少一个存储单元相连的回写式高速缓存单元；接收将新的数据版本写入特定高速缓存数据分配单元的请求；响应于数据存储策略，判定是覆盖缓存在所述特定高速缓存数据分配单元中的缓存的数据版本，还是在将所述新的数据版本写入所述特定高速缓存数据分配单元之前执行将所述缓存的数据版本离台到第一存储单元；接收读取与特定时刻对应的数据版本的请求以及扫描代表写入操作的第一数据结构和代表回复操作的第二数据结构以确定所请求的数据版本的位置。



1. 一种管理数据的方法，所述方法包括:

提供第一级存储器单元和第二级存储器单元;

其中所述第一级存储器单元包括第一高速缓存单元和第二高速缓存单元并且其中所述第二级存储器单元包括第一存储单元和第二存储单元;

接收将新的数据版本写入所述第一高速缓存单元的特定高速缓存数据分配单元的请求;

将所述新的数据版本缓存在所述特定高速缓存数据分配单元处，以及响应于数据存储策略，将当前存储在所述特定高速缓存数据分配单元中的缓存的数据版本离台到所述第一存储单元和所述第二存储单元。

2. 如权利要求 1 中所述的方法，还包括维护指示要被登台到所述高速缓存的数据版本的位置的第四数据结构。

3. 如权利要求 1 或 2 中所述的方法，其中所述缓存和离台包括将缓存的数据版本从所述第一高速缓存单元离台到所述第一存储单元和所述第二存储单元。

4. 如权利要求 1 或 2 中所述的方法，其中所述缓存和离台包括判定将所述缓存的数据版本离台到所述第一和第二存储单元中的一个还是两个数据存储单元以及响应于所述判定，离台所述缓存的数据版本。

5. 如权利要求 1 或 2 中所述的方法，其中所述方法还包括从所述第一存储单元提供离台的数据版本以及从所述第二存储单元提供旧的数据版本。

6. 如权利要求 1 中所述的方法，其中所述方法包括接收读取与特定时刻对应的数据版本的请求以及扫描代表写入操作的第一数据结构和代表回复操作的第二数据结构以确定所请求的数据版本的位置。

7. 如权利要求 2 中所述的方法，其中响应于数据存储策略判定是否覆盖缓存的数据版本包括判定多版本高速缓存单元是否可以存储额外的数据版本。

8. 如权利要求 2 中所述的方法，还包括将不同的数据版本离台到单个存储单元。

9. 一种管理数据的方法，所述方法包括:

提供与至少一个存储单元相连的回写式高速缓存单元;

接收将新的数据版本写入特定高速缓存数据分配单元的请求;

响应于数据存储策略，判定是覆盖缓存在所述特定高速缓存数据分配单元中的缓存的数据版本，还是在将所述新的数据版本写入所述特定高速缓存数据分配单元之前执行将所述缓存的数据版本离台到第一存储单元;

接收读取与特定时刻对应的数据版本的请求以及扫描代表写入操作的第一数据结构和代表回复操作的第二数据结构以确定所请求的数据版本的位置。

10. 如权利要求 9 中所述的方法，还包括将不同的数据版本离台到第一和第二存储单元以及维护指示将哪个数据版本从所述第一存储单元发送到所述第二存储单元的第三数据结构。

11. 如权利要求 9 或 10 中所述的方法，还包括维护指示要被登台到所述高速缓存的数据版本的位置的第四数据结构。

12. 如权利要求 9 或 10 中所述的方法，还包括使用截听来提供所请求的数据版本的位置。

13. 如权利要求 9 中所述的方法，还包括将缓存的数据版本离台到第一存储单元以提供离台的数据版本以及将先前离台到所述第一存储单元的旧的数据版本发送到第二存储单元。

14. 如权利要求 13 中所述的方法，其中发送所述旧的数据版本包括生成包含发送所述旧的数据版本的计时的伪入口。

15. 如权利要求 13 或 14 中所述的方法，其中发送所述旧的数据版本包括将所述旧的数据版本登台到回写式高速缓存单元以及将所述旧的数据版本从所述回写式高速缓存单元离台到所述第二存储单元。

16. 如权利要求 13 中所述的方法，还包括响应于回复操作，将旧的数据版本从所述第二存储单元物理地复制到所述第一存储单元，同时更新指

示要被登台到所述高速缓存的数据版本的位置的所述第四数据结构。

17. 如权利要求 9 中所述的方法，其中响应于数据存储策略，判定是否覆盖缓存的数据版本包括判定多版本高速缓存单元是否可以存储额外的数据版本。

18. 如权利要求 9 中所述的方法，还包括将不同的数据版本离台到单个存储单元。

19. 一种包括计算机可用介质的计算机程序产品，所述计算机可用介质包含计算机可读程序，其中当所述计算机可读程序在计算机上执行时，可使所述计算机执行以下操作:

接收将新的数据版本写入特定高速缓存数据分配单元的请求；响应于数据存储策略，判定是覆盖缓存在所述特定高速缓存数据分配单元中的缓存的数据版本，还是在将所述新的数据版本写入所述特定高速缓存数据分配单元之前执行将所述缓存的数据版本离台到第一存储单元;

接收读取与特定时刻对应的数据版本的请求以及扫描代表写入操作的第一数据结构和代表回复操作的第二数据结构以确定所请求的数据版本的位置。

20. 如权利要求 19 中所述的计算机程序产品，其中当所述计算机可读程序在计算机上执行时，还导致所述计算机将不同的数据版本离台到第一和第二存储单元以及维护指示将哪个数据版本从所述第一存储单元发送到所述第二存储单元的第三数据结构。

21. 如权利要求 19 或 20 中所述的计算机程序产品，其中当所述计算机可读程序在计算机上执行时，还导致所述计算机维护指示要被登台到所述高速缓存的数据版本的位置的第四数据结构。

22. 如权利要求 19 或 20 中所述的计算机程序产品，其中当所述计算机可读程序在计算机上执行时，还导致所述计算机使用截听来提供所请求的数据版本。

23. 如权利要求 19 中所述的计算机程序产品，其中当所述计算机可读程序在计算机上执行时，还导致所述计算机将缓存的数据版本离台到第一

存储单元以提供离台的数据版本以及将先前离台到所述第一存储单元的旧的数据版本发送到第二存储单元。

24. 如权利要求 23 中所述的计算机程序产品，其中当所述计算机可读程序在计算机上执行时，还导致所述计算机生成包含发送所述旧的数据版本的计时的伪入口。

25. 一种包括计算机可用介质的计算机程序产品，所述计算机可用介质包含计算机可读程序，其中当所述计算机可读程序在计算机上执行时，可使所述计算机执行以下操作:

接收将新的数据版本写入第一高速缓存单元的特定高速缓存数据分配单元的请求;

将所述新的数据版本缓存在所述特定高速缓存数据分配单元处，以及响应于数据存储策略，将当前存储在所述特定高速缓存数据分配单元中的缓存的数据版本离台到第二级存储器单元;

其中所述新的数据版本和所述缓存的数据版本中的一个数据版本存储在属于同一存储器级别的两个不同存储器单元处。

26. 如权利要求 25 中所述的计算机程序产品，其中当所述计算机可读程序在计算机上执行时，还导致所述计算机将所述新的数据版本缓存在所述第一高速缓存单元和第二高速缓存单元处。

27. 如权利要求 25 或 26 中所述的计算机程序产品，其中当所述计算机可读程序在计算机上执行时，还导致所述计算机将所述缓存的数据版本从所述第一高速缓存单元离台到所述第一数据存储单元以及将所述缓存的数据版本从所述第二高速缓存单元离台到所述第二数据存储单元。

28. 如权利要求 25 中所述的计算机程序产品，其中当所述计算机可读程序在计算机上执行时，还导致所述计算机从所述第一存储单元提供离台的数据版本以及从所述第二存储单元提供旧的数据版本。

29. 如权利要求 25 中所述的计算机程序产品，其中当所述计算机可读程序在计算机上执行时，还导致所述计算机将缓存的数据版本从所述第一高速缓存单元离台到所述第一存储单元和所述第二存储单元。

30. 如权利要求 25 中所述的计算机程序产品，其中当所述计算机可读程序在计算机上执行时，还导致所述计算机将不同的数据版本离台到单个存储单元。

31. 一种包括第一高速缓存单元、第二高速缓存单元、第一存储单元以及第二存储单元的系统：

其中所述系统适于接收将新的数据版本写入所述第一高速缓存单元的特定高速缓存数据分配单元的请求；

将所述新的数据版本缓存在所述特定高速缓存数据分配单元处，以及响应于数据存储策略，将当前存储在所述特定高速缓存数据分配单元中的缓存的数据版本离台到所述第一和第二存储单元。

32. 如权利要求 31 中所述的系统，其中至少一个高速缓存单元是回写式高速缓存单元；

其中所述系统适于接收将新的数据版本写入特定高速缓存数据分配单元的请求，以及响应于数据存储策略，判定是覆盖缓存在所述特定高速缓存数据分配单元中的缓存的数据版本，还是在将所述新的数据版本写入所述特定高速缓存数据分配单元之前执行将所述缓存的数据版本离台到第一存储单元；

接收读取与特定时刻对应的数据版本的请求以及扫描代表写入操作的第一数据结构和代表回复操作的第二数据结构以确定所请求的数据版本的位置。

33. 如权利要求 31 或 32 中所述的系统，还适于将不同的数据版本离台到所述第一和第二存储单元以及维护指示要被登台到所述高速缓存的数据版本的位置的第四数据结构。

34. 如权利要求 31、32 或 33 中所述的系统，其中所述第一高速缓存单元为多版本高速缓存单元。

35. 如权利要求 31 中所述的系统，其中所述系统适于从所述第一存储单元提供离台的数据版本以及从所述第二存储单元提供旧的数据版本。

36. 如权利要求 32 中所述的系统，还包括将不同的数据版本离台到第

一和第二存储单元的装置以及维护指示将哪个数据版本从所述第一存储单元发送到所述第二存储单元的第三数据结构的装置。

37. 如权利要求 32 或 36 中所述的系统，还包括维护指示要被登台到所述高速缓存的数据版本的位置的第四数据结构的装置。

38. 如权利要求 32 或 36 中所述的系统，还包括使用截听来提供所请求的数据版本的位置的装置。

39. 如权利要求 32 中所述的系统，还包括将缓存的数据版本离台到第一存储单元以提供离台的数据版本的装置以及将先前离台到所述第一存储单元的旧的数据版本发送到所述第二存储单元的装置。

40. 如权利要求 39 中所述的系统，其中发送所述旧的数据版本的装置包括生成包含发送所述旧的数据版本的计时的伪入口。

41. 如权利要求 39 或 40 中所述的系统，其中发送所述旧的数据版本的装置包括将所述旧的数据版本登台到回写式高速缓存单元的装置以及将所述旧的数据版本从所述回写式高速缓存单元离台到所述第二存储单元的装置。

42. 如权利要求 39 中所述的系统，还包括响应于回复操作，将旧的数据版本从所述第二存储单元物理地复制到所述第一存储单元，同时更新指示要被登台到所述高速缓存的数据版本的位置的第四数据结构的装置。

43. 如权利要求 32 中所述的系统，其中响应于数据存储策略，判定是否覆盖所述缓存的数据版本的装置包括判定多版本高速缓存单元是否可以存储额外的数据版本的装置。

44. 如权利要求 32 中所述的系统，还包括将不同的数据版本离台到单个存储单元的装置。

45. 如权利要求 31 中所述的系统，还包括维护指示要被登台到所述高速缓存的数据版本的位置的第四数据结构的装置。

46. 如权利要求 31 或 45 中所述的系统，其中用于缓存和离台的装置包括将缓存的数据版本从所述第一高速缓存单元离台到所述第一存储单元和所述第二存储单元的装置。

47. 如权利要求 31 或 45 中所述的系统，其中用于缓存和离台的装置包括判定将所述缓存的数据版本离台到所述第一和第二存储单元中的一个还是两个数据存储单元的装置以及响应于所述判定而离台所述缓存的数据版本的装置。

48. 如权利要求 31 或 45 中所述的系统，其中所述系统还包括从所述第一存储单元提供离台的数据版本以及从所述第二存储单元提供旧的数据版本的装置。

49. 如权利要求 31 中所述的系统，其中所述系统包括接收读取与特定时刻对应的数据版本的请求的装置以及扫描代表写入操作的第一数据结构和代表回复操作的第二数据结构以确定所请求的数据版本的位置的装置。

50. 如权利要求 45 中所述的系统，其中响应于数据存储策略而判定是否覆盖缓存的数据版本的装置包括判定多版本高速缓存单元是否可以存储额外的数据版本的装置。

51. 如权利要求 45 中所述的系统，还包括将不同的数据版本离台到单个存储单元的装置。

52. 一种包括程序代码装置的计算机程序，当所述程序在计算机上运行时，所述程序代码装置适于执行如权利要求 1 至 18 中的任一权利要求所述的方法。

使用回写式高速缓存单元管理数据的系统、方法和计算机程序产品

技术领域

本发明涉及使用回写式高速缓存单元管理数据的方法、系统和计算机程序产品。

背景技术

数据会随着时间的推移而发展。在许多应用中，需要检索以前的数据版本。这些应用之一是连续数据保护（CDP）应用。在此引入作为参考的Perry等人的美国专利申请公开第2005/0066118号以及美国专利申请公开第2005/0193272号描述了用于连续数据保护的现有技术设备和方法。

已经提出了各种数据结构来跟踪数据随时间的发展。以下两篇在此引入作为参考的论文示出了两种称为B树和BTR树的数据结构，这两篇论文分别为：Jiang、Salzberg、Lomet、Barrena在2000年的第26届VLDB研讨会上所发表的“The BT-Tree: A Branched and Temporal Access Method”以及Jiang、Salzberg、Lomet、Barrena在2003年的时空数据库进展研讨会上发表的“The BT-Tree: Path-Defined Version-Range Splitting in a Branched and Temporal Structure”。

这些分支时态索引可以在满足特定假设的情况下进行维护。根据第一假设，按增大的时间戳的顺序插入分支时态索引的表项。

此外，更新这些结构相对复杂并且涉及复制数据，甚至复制元数据。此外，维护这些数据结构可能需要基准计数器。

以下在此引入作为参考的美国专利申请和美国专利也描述了各种管理数据的方法：

Rowan等人的美国专利申请公开第2005/0066222号，

Rowan等人的美国专利申请公开第2005/0076262号，

Rowan 等人的美国专利申请公开第 2005/0065962 号，

Rowan 等人的美国专利申请公开第 2005/0063374 号，

Rowan 等人的美国专利申请公开第 2005/0076264 号，

Rowan 等人的美国专利申请公开第 2005/0066225 号，

Rowan 等人的美国专利申请公开第 2005/0076261 号，

Perry 等人的美国专利申请公开第 2005/0066118 号，以及

Kekre 等人的美国专利申请公开第 2005/0071379 号，以及

标题为“Persistent Snapshot Methods”的美国专利申请公开第 2004/0117572 号。

存在提供能够管理数据，尤其是使用回写式高速缓存单元管理数据的设备、计算机程序产品和方法的不断增加的需要。

发明内容

根据第一方面，提供了一种管理数据的方法。所述方法包括：提供与至少一个存储单元相连的回写式高速缓存单元；接收将新的数据版本写入特定高速缓存数据分配单元的请求；响应于数据存储策略，判定是覆盖缓存在所述特定高速缓存数据分配单元中的缓存的数据版本，还是在将所述新的数据版本写入所述特定高速缓存数据分配单元之前执行将所述缓存的数据版本离台到第一存储单元；接收读取与特定时刻对应的数据版本的请求以及扫描代表写入操作的第一数据结构和代表回复操作的第二数据结构以确定所请求的数据版本的位置。

根据一个实施例，所述方法包括将不同的数据版本离台到第一和第二存储单元以及维护指示将哪个数据版本从所述第一存储单元发送到所述第二存储单元的第三数据结构。根据一个实施例，所述方法包括维护指示要被登台到所述高速缓存的数据版本的位置的第四数据结构。

根据一个实施例，所述方法包括使用截听来提供所请求的数据版本的位置。

根据一个实施例，所述方法包括将缓存的数据版本离台到第一存储单

元以提供离台的数据版本以及将先前离台到所述第一存储单元的旧的数据版本发送到第二存储单元。

根据一个实施例，发送所述旧的数据版本包括生成包含发送所述旧的数据版本的计时的伪入口（dummy entry）。

根据一个实施例，发送所述旧的数据版本包括将所述旧的数据版本登台到回写式高速缓存单元以及将所述旧的数据版本从所述回写式高速缓存单元离台到所述第二存储单元。

根据一个实施例，所述方法包括响应于回复操作，将旧的数据版本从所述第二存储单元物理地复制到所述第一存储单元，同时更新所述第四数据结构。

根据一个实施例，响应于数据存储策略，判定是否覆盖缓存的数据版本包括判定多版本高速缓存单元是否可以存储额外的数据版本。

根据一个实施例，所述方法包括将不同的数据版本离台到单个存储单元。

根据第二方面，提供了一种管理数据的方法，所述方法包括：提供第一级存储器单元和第二级存储器单元；其中所述第一级存储器单元包括第一高速缓存单元和第二高速缓存单元并且其中所述第二级存储器单元包括第一存储单元和第二存储单元；接收将新的数据版本写入所述第一高速缓存单元的特定高速缓存数据分配单元的请求；将所述新的数据版本缓存在所述特定高速缓存数据分配单元处，以及响应于数据存储策略，将当前存储在所述特定高速缓存数据分配单元中的缓存的数据版本离台到所述第一存储单元和所述第二存储单元。

根据第三方面，提供了一种包括计算机可用介质的计算机程序产品，所述计算机可用介质包含计算机可读程序，其中当所述计算机可读程序在计算机上执行时，可使所述计算机执行以下操作：接收将新的数据版本写入特定高速缓存数据分配单元的请求；响应于数据存储策略，判定是覆盖缓存在所述特定高速缓存数据分配单元中的缓存的数据版本，还是在将所述新的数据版本写入所述特定高速缓存数据分配单元之前执行将所述缓存

的数据版本离台到第一存储单元；接收读取与特定时刻对应的数据版本的请求以及扫描代表写入操作的第一数据结构和代表回复操作的第二数据结构以确定所请求的数据版本的位置。

根据第四方面，提供了一种包括计算机可用介质的计算机程序产品，所述计算机可用介质包含计算机可读程序，其中当所述计算机可读程序在计算机上执行时，可使所述计算机执行以下操作：接收将新的数据版本写入第一高速缓存单元的特定高速缓存数据分配单元的请求；将所述新的数据版本缓存在所述特定高速缓存数据分配单元处，以及响应于数据存储策略，将当前存储在所述特定高速缓存数据分配单元中的缓存的数据版本离台到第二级存储器单元；其中所述新的数据版本和所述缓存的数据版本以外的数据版本存储在属于同一存储器级别的两个不同存储器单元处。

根据第五方面，提供了一种包括第一高速缓存单元、第二高速缓存单元、第一存储单元以及第二存储单元的系统；其中所述系统适于接收将新的数据版本写入所述第一高速缓存单元的特定高速缓存数据分配单元的请求；将所述新的数据版本缓存在所述特定高速缓存数据分配单元处，以及响应于数据存储策略，将当前存储在所述特定高速缓存数据分配单元中的缓存的数据版本离台到所述第一和第二存储单元。

附图说明

现在将仅通过实例的方式参考以下附图描述本发明的优选实施例，这些附图是:

图 1A 示出了根据本发明的进一步实施例的管理数据的系统;

图 1B 示出了根据本发明的另一实施例的写入、读取登台和离台操作的示例性序列;

图 1C 示出了根据本发明的另一实施例的写入、读取登台和离台操作的示例性序列;

图 1D 示出了根据本发明的进一步实施例的写入、读取登台和离台操作的示例性序列;

图 1E 示出了根据本发明的进一步实施例的写入、读取登台和离台操作的示例性序列;

图 1F 示出了根据本发明的实施例的第一数据结构和第二数据结构的两个部分;

图 2 示出了根据本发明的实施例的写入操作和回复操作的示例性序列;

图 3 示出了根据本发明的实施例的写入操作和回复操作的示例性序列;

图 4 示出了根据本发明的另一实施例的写入操作和回复操作的示例性序列;

图 5 示出了根据本发明的实施例的检索数据版本的方法;

图 6 示出了根据本发明的实施例的粗略分析;

图 7a 和 7b 示出了根据本发明的实施例的全局精细分析;

图 8 示出了根据本发明的实施例的管理数据的方法;

图 9A 示出了根据本发明的另一实施例的管理数据的方法; 以及

图 9B 示出了根据本发明的进一步实施例的管理数据的方法。

具体实施方式

根据优选实施例，本发明提供了使用回写式高速缓存单元存储和检索多个数据版本的方法、系统和计算机程序产品。

根据本发明的实施例，数据被写入回写式高速缓存单元以提供缓存的数据版本，并且所述缓存的数据版本以后可以被离台到第一存储单元以提供离台的数据版本。所述离台的数据版本还可以称为当前数据版本。根据本发明的实施例，当前数据版本和旧的数据版本存储在同一逻辑存储单元中。要指出的是，本申请通篇中提到的术语“存储单元”可以包括逻辑存储单元或物理存储单元。逻辑存储单元可以是 LUN 或卷。要指出的是，多个逻辑存储单元可以位于单个物理存储单元中。

根据本发明的实施例,可以提供支持多个数据版本的高速缓存单元(多

版本高速缓存单元）。因此，可以延迟特定数据版本的离台，直到预定数量的版本已被存储在所述高速缓存单元中为止。

根据本发明的另一实施例，处理当前数据版本的方式与处理旧的数据版本的方式不同。所述不同的处理通常包括将旧的数据版本存储在逻辑上分离的存储单元中。该实施例还称为分布式实施例并且支持该实施例的系统具有分布式体系结构。

根据本发明的另一实施例，当前数据版本和旧的数据版本存储在同一数据存储单元中。该实施例称为集中式实施例并且支持该实施例的系统具有集中式体系结构。

根据本发明的进一步实施例，提供了分布式实施例和集中式实施例的组合。

根据本发明的进一步实施例，缓存的数据版本还存储在非易失性存储器单元中。

根据本发明的各种实施例，提供了多个高速缓存单元和多个数据存储单元。可以在这些高速缓存单元之间定义各种关系并且提供多个数据存储单元。

便利地，集中式体系结构包括高速缓存和第一存储单元。所述第一存储单元便利地为永久性存储器单元并且它存储当前数据版本和旧的数据版本。分布式体系结构包括高速缓存、第一存储单元和第二存储单元，其中当前数据版本存储在所述第一存储单元中并且旧的数据版本存储在所述第二存储单元中。

便利地，分布式体系结构维护两个指示写入操作和回复操作的数据结构以及诸如前向数据结构和后向数据结构之类的第三和第四数据结构。所述前向数据结构可以是前向位图并且所述后向数据结构可以是反向位图。便利地，反向位图和前向位图按照诸如盘之类的逻辑单元号（LUN）进行分配，其中每个位图中的表项都与诸如轨道之类的数据分配单元相关联。

要指出的是，可以在不偏离本发明的精神的情况下提供位图与逻辑（或物理）存储单元以及数据分配单元之间的其他关联。例如，数据分配单元

可以具有固定大小、可变大小，可以大于高速缓存页，小于或等于高速缓存页等。

便利地，数据分配单元大于高速缓存页并且高速缓存知道页与数据分配单元之间的关联。便利地，元数据（例如，时间戳）按照数据分配单元进行存储。登台和离台操作可以在不同于数据分配单元的存储内存（storage memory）部分上执行。所述高速缓存支持部分数据分配单元登台和离台操作。

每个位图对于每个数据分配单元都包含表项。所述前向位图协助判定是否应将数据版本从所述第一存储单元复制到所述第二存储单元。所述反向位图指示是将最近更新的数据版本存储在所述第一存储单元还是所述第二存储单元中。所述反向位图通常在回复操作期间使用。回复操作可使旧的数据版本被视为最近更新的数据版本。

便利地，所述高速缓存只能存储一个数据版本。当指向特定数据分配单元的新的写入请求到达时，系统可以强制所述高速缓存将当前缓存在所述数据分配单元中的缓存的数据版本离台到存储单元之一。此操作称为强制离台。根据本发明的另一实施例，只有缓存的数据版本的子集应被发送到存储设备，因此，仅当需要保存该缓存的数据版本时，系统才执行强制离台。

根据本发明的另一实施例，高速缓存可以存储一个以上的数据版本。可以离台旧的版本以释放高速缓存空间。

根据本发明的实施例，定义了数据存储策略。该策略判定是否将缓存在数据分配单元中的缓存的数据版本发送到存储单元，尤其是在将新的数据版本写入该数据分配单元之前。便利地，所述数据存储策略定义数据存储粒度。所述数据存储粒度设置由一个（或无）离台的数据版本表示的离台时段的长度。如果在离台时段期间生成多个缓存的数据版本，则只离台这些多个缓存的数据版本中的一个版本，以便提供表示离台时段的离台的数据版本。通常，离台的数据版本是在离台时段期间缓存的最后一个缓存的数据版本。

根据本发明的实施例，定义了多个数据存储粒度。通常，所述数据存储粒度随着数据版本变旧而变得粗略。所述数据存储粒度可以对诸如写入操作、登台操作、离台操作之类的事件做出响应，并且还会对应用生成的事件做出响应。

所述数据存储策略可以定义一个或多个相关性窗口。对于每个相关性窗口，数据存储粒度可以是固定的，但是并非必须如此。

每个写入操作与离台操作之间的延迟以及数据存储策略的应用减少了离台操作的次数。

便利地，所提出的解决方案不需要索引扫描并且不会增加索引的大小，也不会引入其他用户数据传输（也称为数据输入/输出）。

便利地，可以通过执行强制离台操作或通过扫描高速缓存的数据分配（不一定响应于写入新的数据版本的请求）并定位不同高速缓存数据分配单元中存储的未离台的缓存的数据版本，来离台缓存的数据版本。

根据本发明的实施例，可以直接将数据从所述第一存储单元复制到所述第二存储单元。

根据本发明的另一实施例，可以通过所述高速缓存在所述第一存储单元与所述第二存储单元之间传输数据。因此，离台的数据版本被从所述第一存储单元登台到所述高速缓存，然后被从所述高速缓存离台到所述第二存储单元以提供旧的数据版本。

便利地，对于分布式体系结构和集中式体系结构，回复存储设备的LUN不需要索引扫描，避免了数据和元数据的复制，同时不会对系统引入额外的用户数据输入/输出。此外，旧的数据版本也不会丢失。

要指出的是，第一回写式高速缓存单元131可以与多个接口、适配器等相连。

图1A示出了根据本发明的进一步实施例的管理数据的系统100''。

系统100''包括第一回写式高速缓存单元131、第二回写式高速缓存单元132、第一非易失性存储器141、第二非易失性存储器142、管理单元151、适配器143、第一存储单元121以及第二存储单元122。

系统 100''包括两个基本上相互独立的部分。一个部分作为另一部分的后备。便利地，每个部分都具有自己的供电单元、自己的管理单元等。为了简化说明，示出了单个管理单元 151 和单个适配器 143，但是它们可以被加倍。

系统 100''的第一部分包括第一非易失性存储器 141、第一回写式高速缓存单元 131 和第一存储单元 121。系统 100''的第二部分包括第二非易失性存储器 142、第二回写式高速缓存单元 132 和第二存储单元 122。

管理单元 151 与第一回写式高速缓存单元 131、第二回写式高速缓存单元 132、第一非易失性存储器 141 以及第二非易失性存储器 142 相连。适配器 143 与第一回写式高速缓存单元 131、第二回写式高速缓存单元 132、第一非易失性存储器 141 和第二非易失性存储器 142 相连。

第一存储单元 121 与第一回写式高速缓存单元 131 和第二回写式高速缓存单元 132 相连。第二存储单元 122 与第一回写式高速缓存单元 131 和第二回写式高速缓存单元 132 相连。

被发送到第一回写式高速缓存单元 131 的数据也被发送到第二非易失性存储器 142。被发送到第二回写式高速缓存单元 132 的数据也被发送到第一非易失性存储器 141。

此外，第一存储单元 121 和第二存储单元 122 中的每个单元都能够存储来自第一回写式高速缓存单元 131 和第二回写式高速缓存单元 132 的数据。

如上所述，系统 100''具有两个部分。为了简化说明，图 1A-1E 示出了针对第一部分（尤其是针对第一回写式高速缓存单元 131）的读写操作。本领域的技术人员将理解，可以使用对称的方式对第二部分执行读写操作。

系统 100''维护四个数据结构：（i）代表写入操作的第一数据结构 200，（ii）代表回复操作的第二数据结构 250，（iii）指示将哪些数据从第一存储单元发送到第二存储单元的第三数据结构（也称为前向数据结构）260，以及（iv）指示最近更新的数据版本是位于第一存储单元 121 还是位于第二存储单元 122 中的第四数据结构（也称为后向数据结构）270。

这些数据结构中的每个数据结构都可以存储在系统 100 内的各个位置，例如，所有数据结构 200、250、260、270 都可以存储在第二存储单元 122 中，但是并非必须如此。

在离台时段开始时和数据版本被从第一存储单元复制到第二存储单元时更新第三数据结构 260。

便利地，仅复制一个数据版本（对于每个数据分配单元）。为了知道是否应复制当前缓存的数据版本，可以对第一数据结构进行扫描。为了避免此扫描，第三数据结构 260 指示是否应复制当前缓存的数据。

便利地，第三数据结构 260 对于每个数据分配单元都包含一个位。所述位指示当前缓存的数据版本和新的数据版本（两者都与同一高速缓存数据分配单元关联）是否属于同一离台时段。如果回答是肯定的，则不复制当前缓存的数据版本。当第一缓存操作在当前离台时段中发生时，将设置第三数据结构中的位。在该离台时段的结束时会重置所述位。

将更新第四数据结构 270 作为对回复操作以及所述回复操作之后的登台和离台操作的响应。

回复操作需要将数据从第二存储单元发送到第一存储单元。此过程非常耗时，并且为了允许系统在此过程中响应数据请求，将使用第四数据结构。

要指出的是，在回复操作（回复到特定时刻）之后，最近更新的数据版本与该时刻对应。相应地，这些数据版本可以是回复操作之前的离台的数据版本和旧的数据版本。

第四数据结构 270 对于每个数据分配单元都可以包含一个位。可以在请求回复操作时设置所述位。一旦存储在第二存储单元的数据分配单元中的数据被存储在第一存储单元中，就重置所述位。

数据可以被缓存在第一回写式高速缓存单元 131 中，然后被离台到第一存储单元 121。来自第一存储单元 121 的数据可以经由第一回写式高速缓存单元 131 被发送到第二存储单元 122。第一存储单元 121 存储登台的（当前）数据版本，而第二存储单元 122 存储旧的数据版本。回复可以要

求从第一和/或第二存储单元 121 和 122 登台数据。

管理单元 151 适于：（i）接收将新的数据版本写入特定高速缓存数据分配单元的请求，（ii）响应于数据存储策略，判定是覆盖当前缓存在特定高速缓存数据分配单元中的缓存的数据版本，还是在将新的数据版本写入特定高速缓存数据分配单元之前执行将缓存的数据版本离台到第一存储单元；（iii）接收读取与特定时刻对应的数据版本的请求，以及（iv）扫描代表写入操作的第一数据结构和代表回复操作的第二数据结构以确定所请求的数据版本的位置。图 1K 中示出了第一和第二数据结构。

管理单元 151 控制系统 100 的各个组件以提供所请求的数据版本以及选择性地（响应于管理单元的判定）执行强制离台。

便利地，第一回写式高速缓存单元 131 不适于存储相同数据的多个版本。因此，在缓存新的数据版本之前，管理单元 151 必须判定是否应将缓存的数据版保存在数据存储单元中。所述判定响应于所述数据存储策略。

根据另一实施例，第一回写式高速缓存单元 131 适于存储相同数据的多个版本。

系统 100''，尤其是管理单元 151，可以通过使用截听来控制第一回写式高速缓存单元 131 的操作。

系统 100''可以定义，部分地定义，部分地接收或接收定义至少一个数据相关性窗口的数据存储策略。所述相关性窗口影响第一回写式高速缓存单元 131 与第一和第二存储单元 121 和 122 之间的登台和离台。

一种机制允许第一回写式高速缓存单元在登台和离台时调用截听来扫描代表写入操作的第一数据结构以及代表回复操作的第二数据结构以确定所请求的数据版本的位置。

图 1B 示出了根据本发明的另一实施例的写入、读取登台和离台操作的示例性序列。

图 1B 示出了集中式体系结构作为第一数据存储单元 121 存储离台的数据版本和旧的数据版本。

图 1D 示出了根据本发明的实施例的登台和离台操作的示例性序列。

所述序列始于从适配器 143 发送（由字母 A 指示）将新的数据版本写入第一回写式高速缓存单元 131 的特定高速缓存数据分配单元的请求的步骤。如果当前缓存在该特定缓存分配单元中的缓存的数据版本应被离台(根据数据存储策略），则将其离台到第一存储单元 121（由字母 B 指示）以提供离台的数据版本。完成此离台步骤后，新的数据版本被缓存在该特定高速缓存数据分配单元处。

离台的数据版本可以变为旧的数据版本，尤其是如果更加新的数据版本被缓存并且然后被离台。旧的数据版本也可以存储在第一数据存储单元 121 中。

当收到读取与特定时刻对应的数据版本的请求时，从第一存储单元 121 登台（由字母 C 指示）所请求的数据版本。所请求的数据版本然后被从第一回写式高速缓存单元 131 发送到适配器 143。

图 1C 示出了根据本发明的另一实施例的写入、读取登台和离台操作的示例性序列。

图 1C 示出了分布式体系结构作为第一数据存储单元 121 存储离台的数据版本以及作为第二存储单元 122 存储旧的数据版本。

图 1C 示出了一系列操作，始于从适配器 143 发送（由字母 A 指示）将新的数据版本写入第一回写式高速缓存单元 131 的特定高速缓存数据分配单元的请求的步骤。如果当前缓存在该特定缓存分配单元中的缓存的数据版本应被离台（根据数据存储策略），则将其离台到第一存储单元 121（由字母 B 指示）以提供离台的数据版本。在此离台步骤之后，新的数据版本被缓存在该特定高速缓存数据分配单元处。

离台的数据版本可以变为旧的数据版本，尤其是如果更加新的数据版本被缓存然后被离台。在其被覆盖之前，旧的数据版本可以被从第一存储单元 121 登台（由字母 C 指示）到第二回写式高速缓存 132，然后被离台（由字母 D 指示）到第二存储单元 122。

当收到读取与特定时刻对应的数据版本的请求时，可以从第一存储单元 121（由字母 E 指示）或从第二存储单元 122（由字母 E'指示）登台所

请求的数据版本。所请求的数据版本然后被从第一回写式高速缓存单元131发送到适配器143。

图 1D 示出了根据本发明的进一步实施例的写入、读取登台和离台操作的示例性序列。

图 1D 示出了高速缓存级别的分离体系结构。高速缓存级别的分离体系结构系统将新的数据版本缓存在第一和第二回写式高速缓存单元 131 和 132 中。在高速缓存级别的分离体系结构中，缓存的数据版本不会从第一存储单元 121 被发送到（通过高速缓存单元）第二数据存储单元 122。缓存的数据版本从第一回写式高速缓存单元 131 被发送到第一存储单元 121（由字母 B 指示）以及从第二回写式高速缓存单元 132 被发送到第二存储单元 122（由字母 B'指示）。旧的数据版本不存储在第一存储单元 121 中，而是存储在第二存储单元 122 中。第二存储单元 122 接收和存储离台的数据版本和旧的数据版本。

便利地，从第一存储单元 121 来提供（由字母 C 指示）离台的数据版本。从第二存储单元来提供（由字母 C'指示）旧的数据版本。要指出的是，还可以从第二存储单元 122 来提供离台的数据版本。

在高速缓存级别的分离体系结构中，不会将数据从第一存储单元 121 发送到第二存储单元。便利地，如果数据存储粒度非常精细，则高速缓存级别的分离体系结构可以节省许多登台和离台操作。另一方面，如果数据存储粒度较低，则该体系结构执行的离台操作多于非分离体系结构所需的操作。

图 1E 示出了根据本发明的进一步实施例的写入、读取登台和离台操作的示例性序列。

图 1E 示出了存储单元级别的分离体系结构。存储单元级别的分离体系结构系统将新的数据版本缓存在单个回写式高速缓存单元中，但是将缓存的数据版本发送到第一存储单元 121（由字母 B 指示）和第二存储单元 122（由字母 B'指示）。

第一存储单元 121 不存储旧的数据版本。将离台的数据版本从第一存

储单元 121（由字母 C 指示）提供给第一回写式高速缓存单元。旧的数据版本存储在第二存储单元 122 中并且可在以后从第一存储单元 121 被提供给第一回写式高速缓存单元 131（由字母 C'指示）。要指出的是，还可以从第二存储单元 122 来提供离台的数据版本。

在存储单元级别的分离体系结构中，不会将数据从第一存储单元 121 发送到第二存储单元。便利地，如果数据存储粒度非常精细，则存储单元级别的分离体系结构可以节省许多登台和离台操作。

根据本发明的另一实施例，系统 100''还可以应用选择性的存储单元级别的分离体系结构。因此，系统 100''可以在离台缓存的数据版本之前决定是将该数据版本发送到第一存储单元，还是将该数据版本发送到第一和第二存储单元 121 和 122 两者。

便利地，如果系统可以确定缓存的数据版本是在已经过的离台时段期间最后要缓存的缓存的数据版本，则该缓存的数据版本可以被发送到第一存储单元并且发送到第二存储单元。如果缓存的数据版本应在离台时段的中间或离台时段的开始被离台，则可以将其离台到第一存储单元 121。

要指出的是，响应于经过预定的离台时段部分，缓存的数据版本可以被发送到两个数据存储单元。

所述选择性的存储单元分离体系结构在粗略数据存储粒度处实质上作为分布式体系结构，而在精细数据存储粒度处实质上作为存储单元级别的分离体系结构。

要指出的是，尽管图 1A-1E 示出了第二回写式高速缓存和第二非易失性存储器并且假设当前数据版本和旧的数据版本存储在不同的高速缓存和非易失性存储器中，但这并非一定如此。因此，即使当前数据版本和旧的数据版本位于同一高速缓存和非易失性存储器中，系统 100''也可以基本上以相同的方式运行。

图 1F 分别示出了根据本发明的实施例的第一数据结构 200 和第二数据结构 250 的两个部分 201 和 251。在此实例中，关键字为 LBA 以及时间戳，数据为物理地址。但是也可以使用其他实例。为了简化说明，这些图

示出了到同一逻辑块地址（例如，LBA=12）的读取和回复操作。

要指出的是，第一数据结构 200 包括与到其他 LBA 的写入操作相关的元数据。

还要指出的是，部分 201 被示为表示到特定 LBA 的更新操作，但是它也可以表示与其他 LBA 有关的操作。备选地，其他图形可以表示与其他 LBA 有关的操作。

第一数据结构 200 包括三列 200（1）、200（2）和 200（3）。每个表项包括有关写入操作的信息。第一列 200（1）包括写入操作的逻辑块地址，第二列 200（2）包括写入时间戳，并且第三列 200（3）包括写入操作的物理地址。逻辑块地址和写入时间字段可用作第一数据结构 200（1）的关键字（索引）。

要指出的是，如果在特定时刻，出现到多个 LBA 的写入操作，则第一表将包括多个反映此写入操作的表项。

第二数据结构 250 包括四列 250（1）-250（4）。第一列 250（1）包括分支标识符，第二列 250（2）包括分支开始时间，第三列 250（3）包括分支结束时间，并且第四列 250（4）包括每个分支的回复时间。

数据结构 200 和 250 适于控制一系列写入和回复操作，其中在每个给定时刻有一个分支处于活动状态。为了支持多个并发的活动分支，这些表应被修改为包括分支标识信息。

图 2 示出了根据本发明的实施例的写入操作和回复操作的示例性序列 101。序列 101 包括在时间 10、30、40、60、90 和 110 的到诸如逻辑单元（LUN）之类的虚拟地址空间的写入操作，以及相应地回复时间 35 和 70 处的 LUN 的内容的请求（在时间 80 和 100 处接收）。

虚线表示回复操作。要指出的是，在任意给定的时刻，只有一个分支处于活动状态。

假设写入操作与逻辑块地址 12 关联并且与这些写入操作关联的物理地址相应地为 a、b、c、d、e 和 f。

第一数据结构 200 的第一列 200(1)指示写入操作指向逻辑块地址 12。

第一数据结构 200 的第二列 200（2）指示写入操作在时间 10、30、40、60、90 和 110 处发生。第一数据结构 200 的第三列 200（3）指示与这些写入操作关联的物理地址为 a、b、c、d、e 和 f。

第二数据结构 250 的第一表项指示第一分支在时间 0 处开始并在时间 80 处结束。当接受回复 LUN 的内容的第一请求时，第一分支结束。

第二数据结构 250 的第二表项指示第二分支在时间 80 处开始并在时间 100 处结束。当接受回复 LUN 的内容的第二请求时，第二分支结束。

第二数据结构 250 的第三表项指示第三分支在时间 100 处开始并没有结束。

根据本发明的另一实施例，第二数据结构包括诸如分支统计之类的其他元数据。分支统计可以例如包括属于该分支的第一数据结构表项的数目，在该分支存在期间被写入的不同逻辑块地址的数目等。便利地，分支统计可以协助判定要删除哪些分支，尤其是在存在频繁的读取和写入操作时。

图 3 示出了根据本发明的实施例的写入操作和回复操作的示例性序列 300。

通过三角形示出了对第一 LBA 的写入操作（W1、W3、W5 和 W6）。通过圆形示出了对第二 LBA 的写入操作（W4 和 W10）。通过正方形示出了对第三 LBA 的写入操作（W2、W7 和 W9）。

序列 300 包括四个分支 B1-B4 301-304 并且定义了跨 T14 和当前时刻（T_CURRENT）的相关性窗口 310。

第一分支（B1）在 T0 处开始（S1）并在 T8 处结束（E1）。第一分支 B1 包括以下写入操作: T1 处的 W1（写入第一 LBA）、T2 处的 W2（写入第三 LBA）、T3 处的 W3（写入第一 LBA）、T4 处的 W4（写入第二 LBA）以及 T6 处的 W5（写入第一 LBA）。B1 在 T8 处结束（E1）。

第二分支（B2）302 是 B1 301 的子分支并通过到时间 T5 的回复操作（RV1）在时间 T8 处开始（S2）。第二分支 B2 302 包括以下写入操作: T9 处的 W6（写入第一 LBA）以及 T11 处的 W7（写入第三 LBA）。B2 在 T12 处结束（E2）。

第三分支（B3）303 是 B2 302 的子分支并通过到时间 T10 的回复操作（RV2）在时间 T12 处开始（S3）。第三分支 B3 303 包括是 T13 处的 W8（写入第三 LBA）的单个写入操作。B3 在 T15 处结束（E3）。

第四分支（B4）304 是 B3 303 的子分支并通过到时间 T14 的回复操作（RV3）在时间 T15 处开始（S4）。B4 304 包括 T16 处的写入操作 W9（写入第三 LBA）以及 T17 处的另一写入操作 W10（写入第二 LBA）。第四分支 B4 304 在 T18 处结束（E4）。

图 4 示出了根据本发明的另一实施例的写入操作和回复操作的示例性序列 300'。

序列 300'与序列 300 的不同之处在于第二回复操作（RV2'）的回复时间。回复时间为 T7（属于第一分支 B1 301）而不是 T10（属于第二分支 B2 302）。因此，在序列 300 中具有子分支（B3 303）的第二分支 B2 302 在序列 300'中没有子分支。

图 5 示出了根据本发明的实施例的检索数据版本的方法 400。

方法 400 始于选择或获取当前分支的阶段 410。阶段 410 之后是阶段 412，在阶段 412，检查对于分支上始于该分支的开始处并在与数据版本的检索请求关联的所请求时间戳处结束的部分，是否存在写入操作。如果回答为否定的，则在阶段 412 后执行阶段 414，否则在阶段 412 后执行阶段 420。

阶段 420 包括返回最新的写入操作作为方法 400 的结果。

阶段 414 包括检查所述分支是否具有父分支。如果回答是否定的，则方法 400 在返回否定回答（Null）的阶段 416 处结束。如果回答是肯定的，则在阶段 414 后执行获取父分支的阶段 422 和检查对于父分支上始于该分支的开始处并在与所检查父分支的子分支关联的回复时间处结束的部分，是否存在写入操作的阶段 424。如果回答是否定的，则在阶段 424 后执行阶段 414，否则在阶段 424 后执行阶段 420。

图 6 示出了根据本发明的实施例的粗略分析。

粗略分析 500 始于将分支索引（J）设置为 1 的阶段 510。

在阶段 510 后执行确定第 J 个分支与相关性窗口之间的关系的阶段 520。

如果第 J 个分支和相关性窗口至少部分地重叠，则不删除第 J 个分支，并在阶段 520 后执行增大 J 并跳转到阶段 520 的阶段 530。要指出的是，此迭代在扫描整个第二数据结构后结束。

如果第 J 个分支和相关性窗口不重叠（甚至不部分地重叠），则在阶段 520 后执行检查第 J 个分支是否包括一个或多个子分支的阶段 540。

如果第 J 个分支具有单个子分支，则在阶段 540 后执行结合第 J 个分支与该子分支的阶段 550。根据本发明的实施例，阶段 550 包括将该分支标记为应被结合的分支（例如，通过将该子分支的回复时间之前的最后表项移动到该子分支的开始时间）。

如果第 J 个分支位于相关性窗口以外并且没有任何子分支，则在阶段 550 后执行删除与该分支关联的元数据的阶段 560。根据本发明的实施例，阶段 560 包括将该分支标记为应被删除的分支。所述删除可以在执行精细分析期间完成。

如果存在多个子分支，则在阶段 540 后执行阶段 530。

在阶段 560 和 550 后执行阶段 530。

例如，参考图 3，第四分支 B4 304 位于相关性窗口内并且第一至第三分支是第四分支的上级分支。相应地，不删除任何分支。

对于另一实例，参考图 4，第二分支 B2 302 不是第四分支 B4 304 的上级分支并且位于相关性窗口以外，因此其可以被删除。

要指出的是，所述删除可以包括将该分支标记为要删除的候选分支、响应于此类标记执行精细分析，然后立即删除该分支。

图 7a-7b 示出了根据本发明的实施例的全局精细分析 600。

图 7a 示出了分析 600 的第一部分 600（a），而图 7b 示出了分析 600 的部分 600（b）。

为特定分支执行所述全局精细分析。然后可以为另一分支执行此分析。

全局精细分析 600 始于定位至少在该特定分支的一个部分期间处于活

动状态的每个分支的阶段610。每个此类分支被称为相关分支。

在阶段610后执行将位于相关性窗口以外的相关分支分割为通过子分支的回复时间分隔的部分的阶段620。每个部分都与一个子分支相关联。

在阶段620后执行在假设分支的最后部分（该部分始于子分支的最后的回复时间并结束于分支的结束时间）位于相关性窗口以外的情况下，删除该部分中的任何写入操作的阶段625。

在阶段625后执行在某个部分（位于相关性窗口以外）具有多个写入的情况下，删除该部分的所有写入操作（除了最后一次写入）的阶段630。

在阶段630后执行针对多个部分中的特定部分，判定所述部分以及随后部分是否包括写入操作的阶段635。如果回答是肯定的，则在阶段635后执行将写入部分移动到子分支的开始时间的阶段642。如果答案是否定的，则在阶段635后执行判定当前部分没有写入操作还是应用了局部精细分析的阶段640。如果答案是肯定的，则在阶段640后执行阶段650，否则在之后执行在多个部分中选择下一部分并跳转到阶段635的阶段645。

阶段650包括接收特定部分的特定写入操作U以及子分支C1...Cn。在阶段650后执行将计数器X设为0，将分支b设为null以及将工作列表WL定义（或设置）为等于由写入操作U和子分支组成的对：WL={（U, C1）,...,（U, Cn）}的阶段652。

在阶段652后执行判定WL是否为空的阶段654。如果回答是肯定的，则在阶段654后执行阶段656，如果答案是否定的，则在阶段654后执行阶段664。

阶段656包括提供<x,b>。在阶段656后执行检查x的值的阶段658。如果x=0，则在阶段658后执行阶段660或在该阶段中删除写入操作。在阶段660后执行阶段645。如果x=1，则在阶段658后执行将写入操作移动到子分支的阶段662。如果x=2，则在阶段658后执行阶段645。在阶段662后执行阶段645。

阶段664包括从WL删除第一项（称为（w,d））以及检查分支d中的第一写入操作是否在相关性窗口开始之后发生。如果答案是否定的，则

在阶段 664 后执行阶段 674，否则在之后执行阶段 666。阶段 666 包括增大 x 并且在之后执行检查 x 是否大于 1 的阶段 668。如果 x 是 0 或 1，则在阶段 668 之后执行将<x,b>设置为<x,null>的阶段 670 并跳转到阶段 656。如果 x>1，则在阶段 668 后执行将 b 设置为 2 的阶段 672 并跳转到阶段 654。

阶段 674 包括向工作列表添加一组由分支 d 的开始时间以及分支 d 中回复时间小于时间 w 的子分支所组成的对。WL={（start time（d）..d1），（start time（d）..dn）}。在阶段 674 后执行阶段 654。

要指出的是，上述多数实例所指的情况都是在任意时刻只有一个活动的分支。如上所述，情况并不一定如此。为了支持多个共存的分支，需要额外的元数据以便在多个共存的分支之间进行选择。此额外的元数据可以包括分支标识符、父分支标识符等。

多个分支可以在各种情况下共存，所述情况包括但不限于测试环境、支持大量快照的主机可寻址逻辑单元的实施方式等。

要指出的是，如果有多个分支共存，则添加新的分支（例如，通过执行回复操作）不一定会终止另一分支。根据本发明的实施例，定义了第三数据结构。

所述第三数据结构包括针对每个 LBA 的第一写入操作。此类表可以简化各种扫描序列。对于另一实例，还可以定义包括每个分支的最后写入操作的另一数据结构。根据本发明的另一实施例，B 树由前缀 B 树所取代。

备选的解决方案是使 B 树表项仅为逻辑轨道号而没有时间分量，并且在树叶中存储具有所有 LBA 表项及其时间信息的可变长度结构。

图 8 示出了根据本发明的实施例的管理数据的方法 700。

方法 700 始于阶段 710，在阶段 710，提供代表针对由关键字（包括但不限于逻辑块地址）标识的信息的写入操作的第一数据结构，以及提供代表分支创建操作（包括但不限于回复操作、分支克隆操作等）的第二数据结构。

便利地，所述第一数据结构包括写入时间戳以及逻辑块地址与关联的物理地址之间的映射。便利地，所述第二数据结构包括分支标识符、分支

开始时间、分支结束时间以及分支回复时间。便利地，所述第一数据结构为 B 树。

在阶段 710 后执行阶段 720、730、740 和 750。

阶段 720 包括接收写入或更新数据的请求，以及相应地更新所述第一数据结构。

阶段 730 包括接收创建始于所请求时间戳处的数据版本的分支的请求，以及相应地更新所述第二数据结构。例如，创建分支的请求可以是将数据回复至特定回复时间（所述回复时间被认为是所请求的时间）的请求的结果。对于另一实施例，创建分支的请求可以是逻辑地复制（克隆）数据的请求的结果，并且在此情况下，该请求的时间就是所请求的时间。

阶段 740 包括接收读取在读取时间戳处的数据版本的请求，以及扫描所述第一和第二数据结构以定位该版本。阶段 740 可以包括方法 400 的一个或多个阶段。

阶段 750 包括响应于数据存储策略而更新所述第一和第二数据结构。阶段 750 可以包括执行粗略分析，执行精细分析等。阶段 750 可以包括图 6 和 7 中所示的一个或多个阶段。

图 9A 示出了根据本发明的另一实施例的管理数据的方法 800。

方法 800 从提供与至少一个存储单元相连的回写式高速缓存单元的步骤 810 开始。

在步骤 810 后执行定义数据存储策略的步骤 820。所述数据存储策略可以定义一个或多个相关性窗口，从而定义一个或多个数据粒度时段。此定义影响登台和离台决策。

在步骤 820 后执行接收将新的数据版本写入特定数据分配单元的请求的步骤 830。

在步骤 830 后执行步骤 840，在步骤 840 中，响应于数据存储策略，判定是覆盖缓存在特定高速缓存数据分配单元处的缓存的数据版本、还是在将新的数据版本写入特定高速缓存分配单元之前将缓存的数据版本离台到第一存储单元，还是维护多个缓存的数据版本。便利地，所述数据存储

策略（针对每个高速缓存数据分配单元）定义了由单个离台的数据版本表示的离台时段。如果该时段过期，则可以离台最近缓存的数据版本。如果新的数据版本和当前缓存的数据版本在同一离台时段内被发送到高速缓存，则可以在不离台当前缓存的数据版本的情况下覆盖该当前缓存的数据版本。

在步骤 840 后执行根据判定选择性地离台缓存的数据版本的步骤 850。

便利地，步骤 850 可以应用于各种体系结构，包括分布式体系结构、集中式体系结构、存储单元级别的分离体系结构以及选择性存储单元级别的分离体系结构。

便利地，步骤 850 包括将不同的数据版本离台到第一和第二存储单元处的步骤 852 以及维护指示将哪个数据版本从所述第一存储单元发送到所述第二存储单元的第三数据结构的步骤 854。

便利地，步骤 850 包括维护指示要登台到高速缓存的数据版本的位置的第四数据结构的步骤 856。

便利地，步骤 850 包括使用截听来提供所请求的数据版本。

便利地，步骤 850 包括将缓存的数据版本离台到第一存储单元以提供离台的数据版本。

根据本发明的实施例，在步骤 856 后执行将先前离台到第一存储单元的旧的数据版本发送到第二存储单元的步骤 865。

便利地，步骤 865 包括生成包含发送旧的数据版本的计时的伪入口。所述伪入口是第一数据结构的一部分。便利地，步骤 865 包括将旧的数据版本登台到回写式高速缓存单元以及将旧的数据版本从回写式高速缓存单元离台到第二存储单元。

在步骤 850 后执行接收读取与特定时刻对应的数据版本的请求以及扫描代表写入操作的第一数据结构和代表回复操作的第二数据结构以确定所请求数据版本的位置的步骤 860。

便利地，步骤 860 包括维护指示所请求的数据版本是存储在所述第一存储单元还是存储在所述第二存储单元中的后向数据结构。

在步骤 860 后执行提供所请求的数据版本的步骤 870。便利地，步骤 870 包括在数据存储单元出现故障的情况下从非易失性存储器单元提供所请求的数据版本。

根据本发明的实施例，步骤 840 包括在非易失性存储器单元处存储新的数据版本。

图 9B 示出了根据本发明的进一步实施例的管理数据的方法 900。

方法 900 始于提供第一高速缓存单元、第二高速缓存单元、第一存储单元以及第二存储单元的步骤 910。

在步骤 910 后执行定义数据存储策略的步骤 920。所述数据存储策略可以定义一个或多个相关性窗口，从而定义一个或多个数据粒度时段。此定义影响登台和离台决策。所述体系结构定义数据版本如何被缓存（到一个或两个高速缓存单元）以及如何离台（到一个或多个数据存储单元）。离台到一个或多个数据存储单元可以是固定的，也可以动态地改变。固定的离台策略由存储单元级别的分离体系结构和高速缓存级别的分离体系结构表示。动态的离台策略由选择性的存储单元分离策略表示。

要指出的是，图 9B 示出了一系列包括接收写入新的数据版本的请求，然后接收读取数据版本的请求的步骤。要指出的是，这些请求的顺序可以更改，相应地，方法 900 的各个阶段的顺序也可以更改。

在步骤 920 后执行接收将新的数据版本写入所述第一高速缓存单元的特定高速缓存数据分配单元的请求的步骤 930。

在步骤 930 后执行步骤 940，其中将新的数据版本缓存在特定高速缓存数据分配单元处以及响应于数据存储策略，将当前存储在特定高速缓存数据分配单元中的缓存的数据版本离台到所述第一和第二存储单元。

例如，如果新的数据版本缓存在所述第一和第二高速缓存单元中，则离台包括将缓存的数据版本从所述第一高速缓存单元离台到所述第一数据存储单元以及将缓存的数据版本从所述第二高速缓存单元离台到所述第二数据存储单元。

在步骤 940 后执行接收读取与特定时刻对应的数据版本的请求以及提

供所请求的数据版本的步骤950。

便利地，步骤950包括从所述第一存储单元提供离台的数据版本以及从所述第二数据单元提供旧的数据版本。

此外，本发明可以采取可从计算机可用或计算机可读介质访问的计算机程序产品的形式，所述计算机可用或计算机可读介质提供了可以被计算机或任何指令执行系统使用或与计算机或任何指令执行系统结合的程序代码。出于此描述的目的，计算机可用或计算机可读介质可以是任何能够包含、存储、传送、传播或传输由指令执行系统、装置或设备使用或与所述指令执行系统、装置或设备结合的程序的装置

所述介质可以是电、磁、光、电磁、红外线或半导体系统（或装置或设备）或传播介质。计算机可读介质的实例包括半导体或固态存储器、磁带、可移动计算机盘、随机存取存储器(RAM)、只读存储器(ROM)、硬磁盘和光盘。光盘的当前实例包括光盘-只读存储器(CD-ROM)、光盘-读/写(CR-R/W)和DVD。

适合于存储和/或执行程序代码的数据处理系统将包括至少一个通过系统总线直接或间接连接到存储器元件的处理器。所述存储器元件可以包括在程序代码的实际执行期间采用的本地存储器、大容量存储装置以及提供至少某些程序代码的临时存储以减少必须在执行期间从大容量存储装置检索代码的次数的高速缓冲存储器。

输入/输出或I/O设备（包括但不限于键盘、显示器、指点设备等）可以直接或通过中间I/O控制器与系统相连。

网络适配器也可以被连接到系统以使所述数据处理系统能够通过中间专用或公共网络变得与其他数据处理系统或远程打印机或存储设备相连。调制解调器、电缆调制解调器和以太网卡只是几种当前可用的网络适配器类型。

根据本发明的实施例，数据被写入回写式高速缓存单元，并且当前数据版本以及先前的数据版本被发送到诸如盘、盘阵列、磁带之类的一个或多个存储单元。数据存储策略帮助刷新数据和元数据，并且还协助判定是

否将特定数据版本发送到盘。

在不偏离所要求保护的本发明的精神和范围的情况下，本领域的技术人员将想到此处所述的内容的变型、修改和其他实施方式。

相应地，本发明不是由上述示例性说明来限定，而是由以下权利要求的精神和范围来限定。

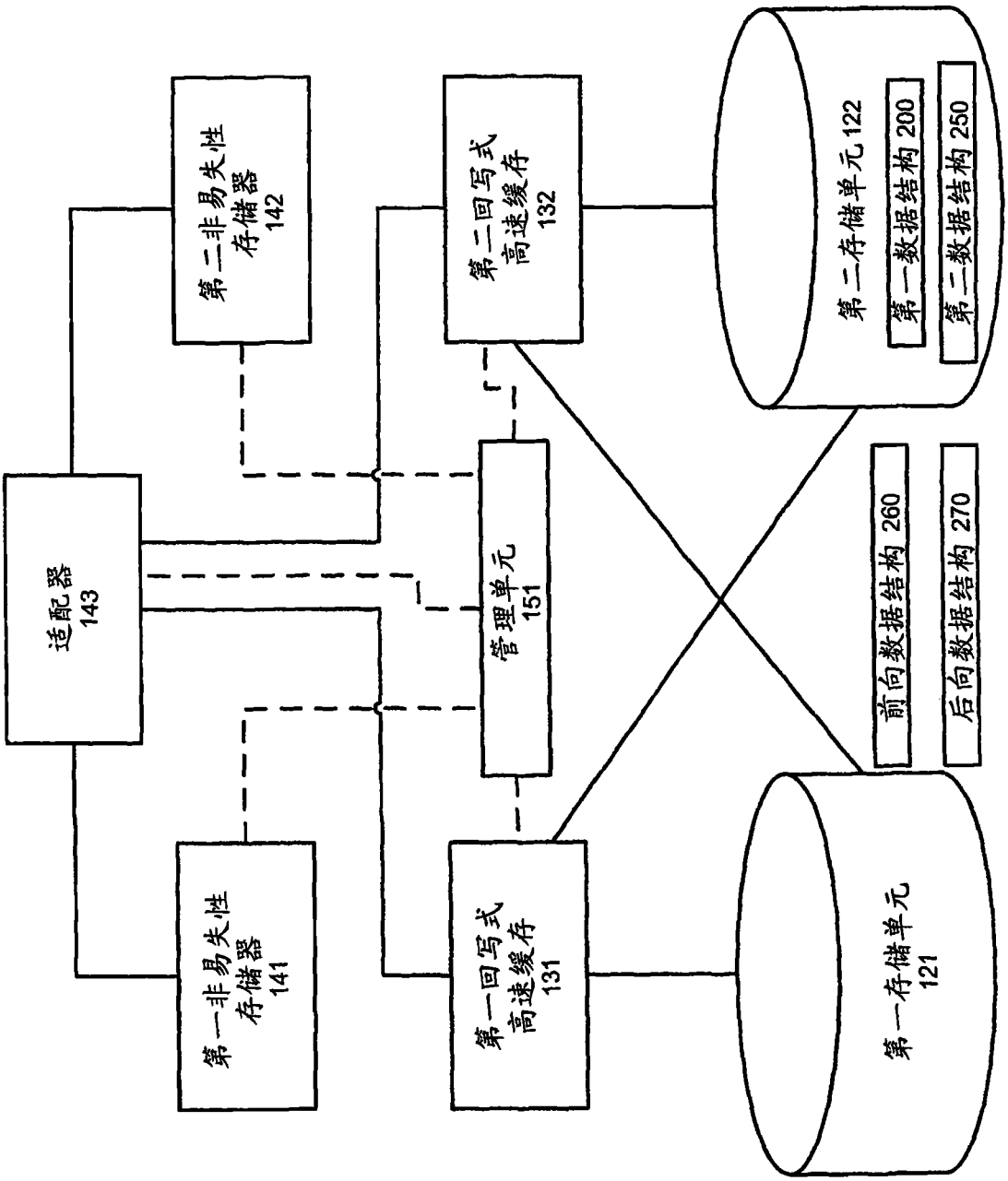

适配器
143
第一非易失性存储器
141
第二非易失性存储器
142
第一回写式高速缓存
131
第二回写式高速缓存
132
管理单元
151
第一存储单元
121
第二存储单元 122
第一数据结构 200
第二数据结构 250
前向数据结构 260
后向数据结构 270
"100

图 1A

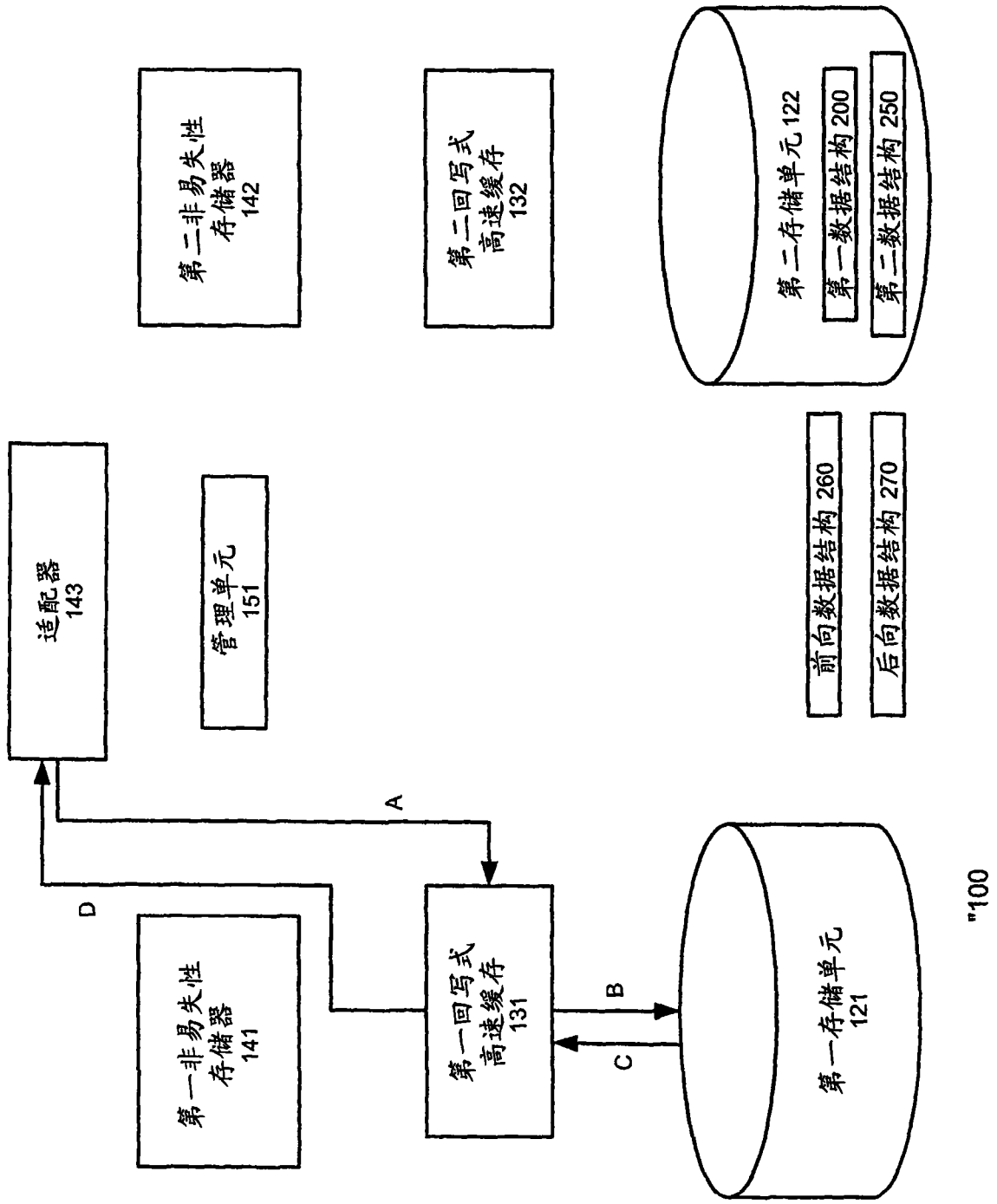

适配器
143
管理单元
151
第二非易失性
存储器
142
第二回写式
高速缓存
132
第二存储单元 122
第一数据结构 200
第二数据结构 250
前向数据结构 260
后向数据结构 270
第一非易失性
存储器
141
第一回写式
高速缓存
131
第一存储单元
121
A
B
C
D
"100

图 1 B

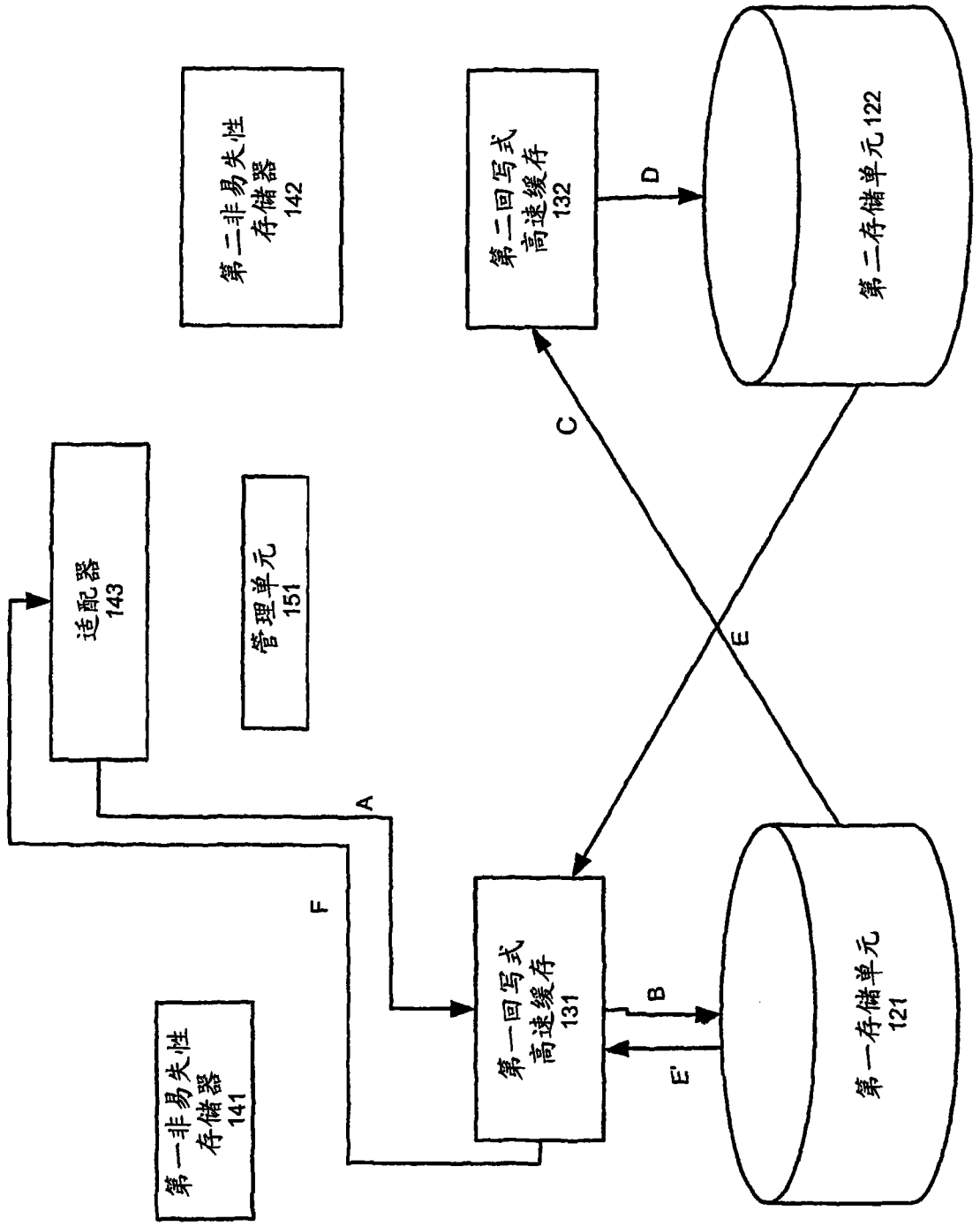

第二非易失性存储器 142
第二回写式高速缓存 132
第二存储单元 122
适配器 143
管理单元 151
第一非易失性存储器 141
第一回写式高速缓存 131
第一存储单元 121
A
B
C
D
E
E'
F
100

图 1C

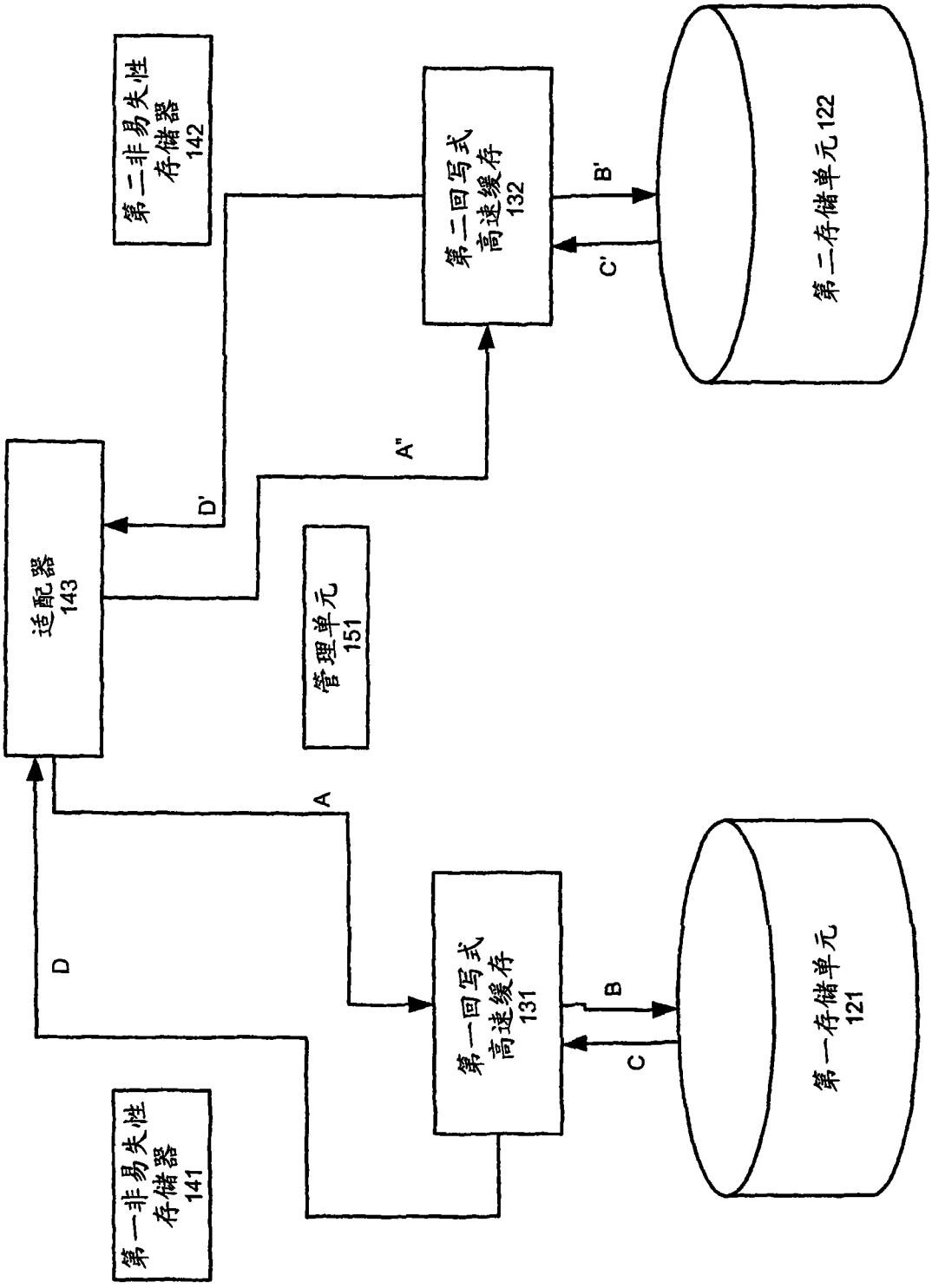

第二非易失性存储器 142
第二回写式高速缓存 132
第二存储单元 122
适配器 143
管理单元 151
第一非易失性存储器 141
第一回写式高速缓存 131
第一存储单元 121
A
A"
B
B'
C
C'
D
D'
"100

图 1D

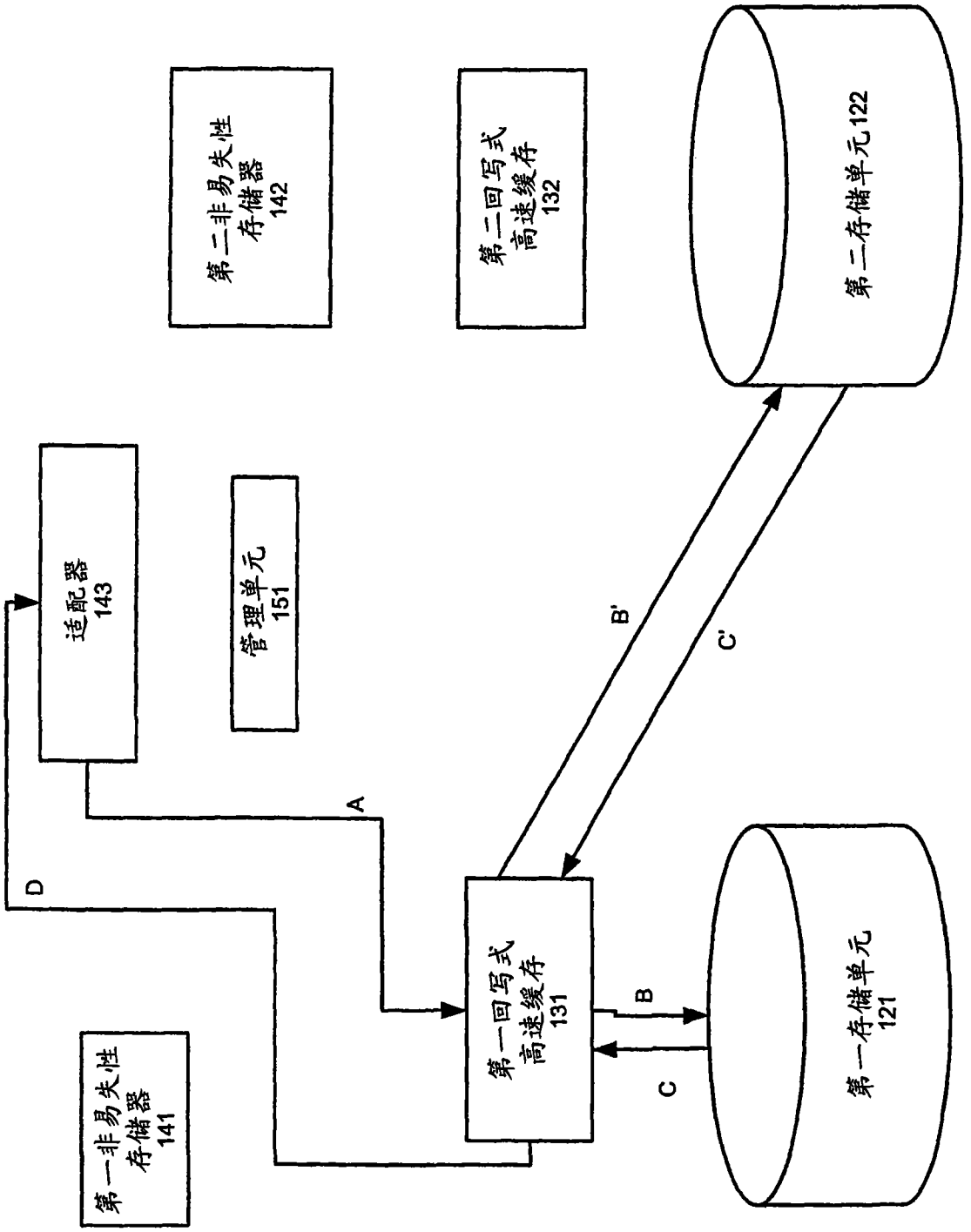

第二非易失性存储器 142
第二回写式高速缓存 132
第二存储单元122
适配器 143
管理单元 151
B'
C'
A
D
第一回写式高速缓存 131
B
C
第一存储单元 121
第一非易失性存储器 141
100

图 1E

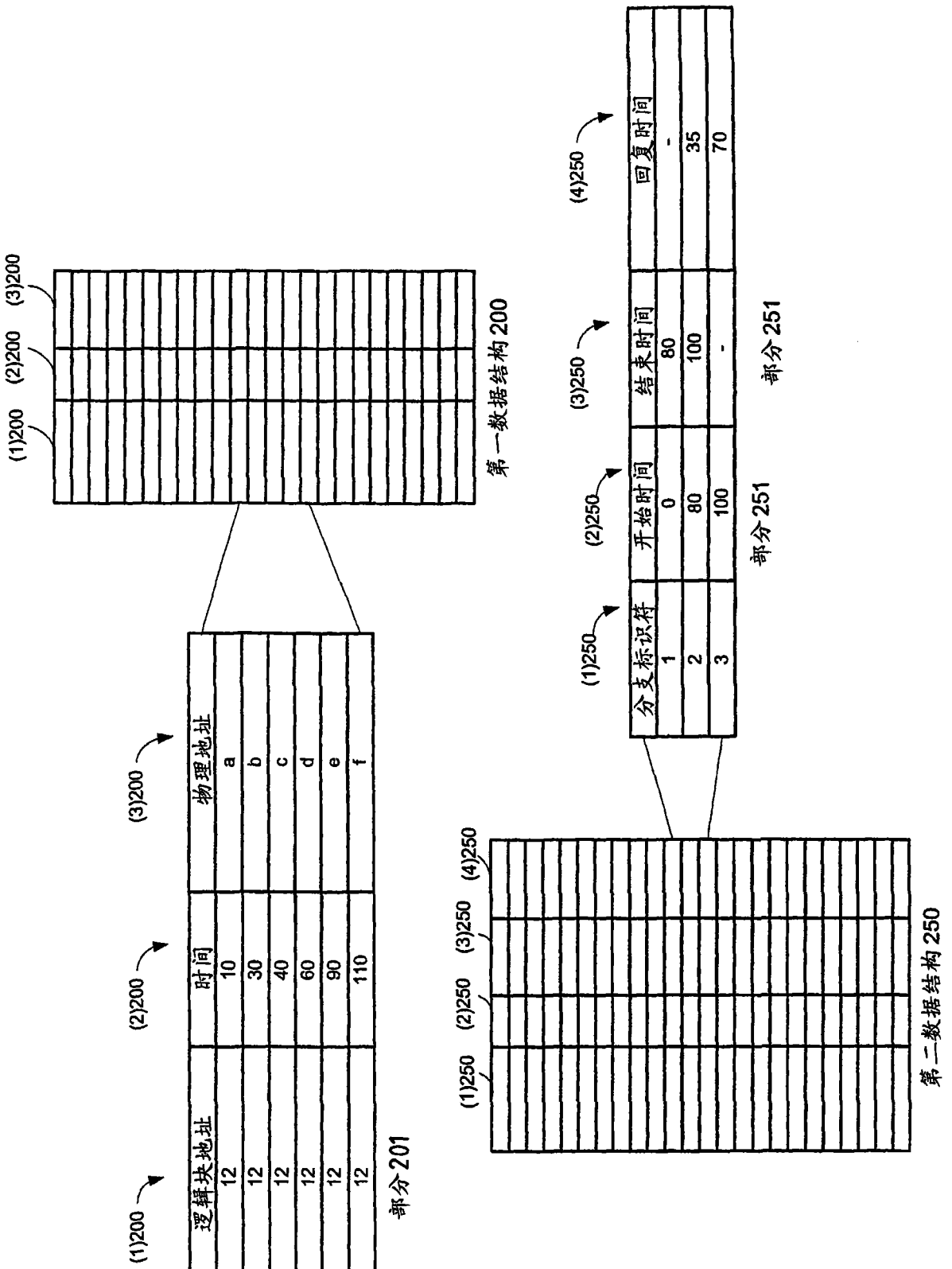

(1)200
(2)200
(3)200
第一数据结构 200
逻辑块地址
时间
物理地址
12 10 a
12 30 b
12 40 c
12 60 d
12 90 e
12 110 f
部分 201
(1)250
(2)250
(3)250
(4)250
第二数据结构 250
分支标识符
开始时间
结束时间
回复时间
1 0 80 -
2 80 100 35
3 100 - 70
部分 251

图 1F

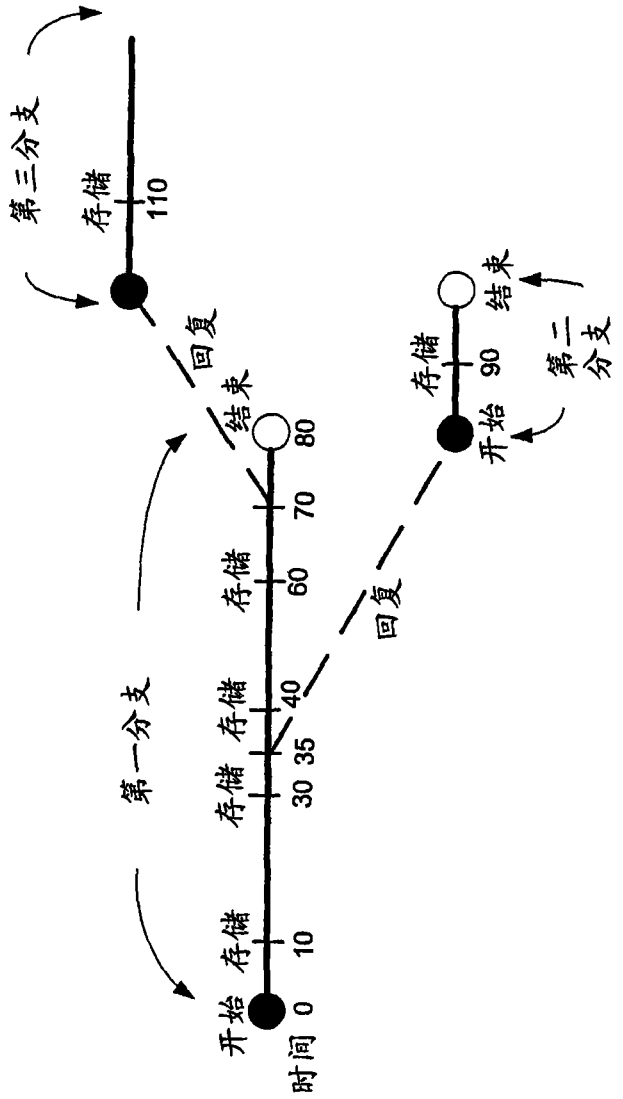

第三分支
存储
110
回复
结束
存储
90
结束
第二分支
开始
回复
第一分支
存储
80
存储
70
存储
60
存储
40
存储
35
存储
30
存储
10
开始
0
时间
101

图 2

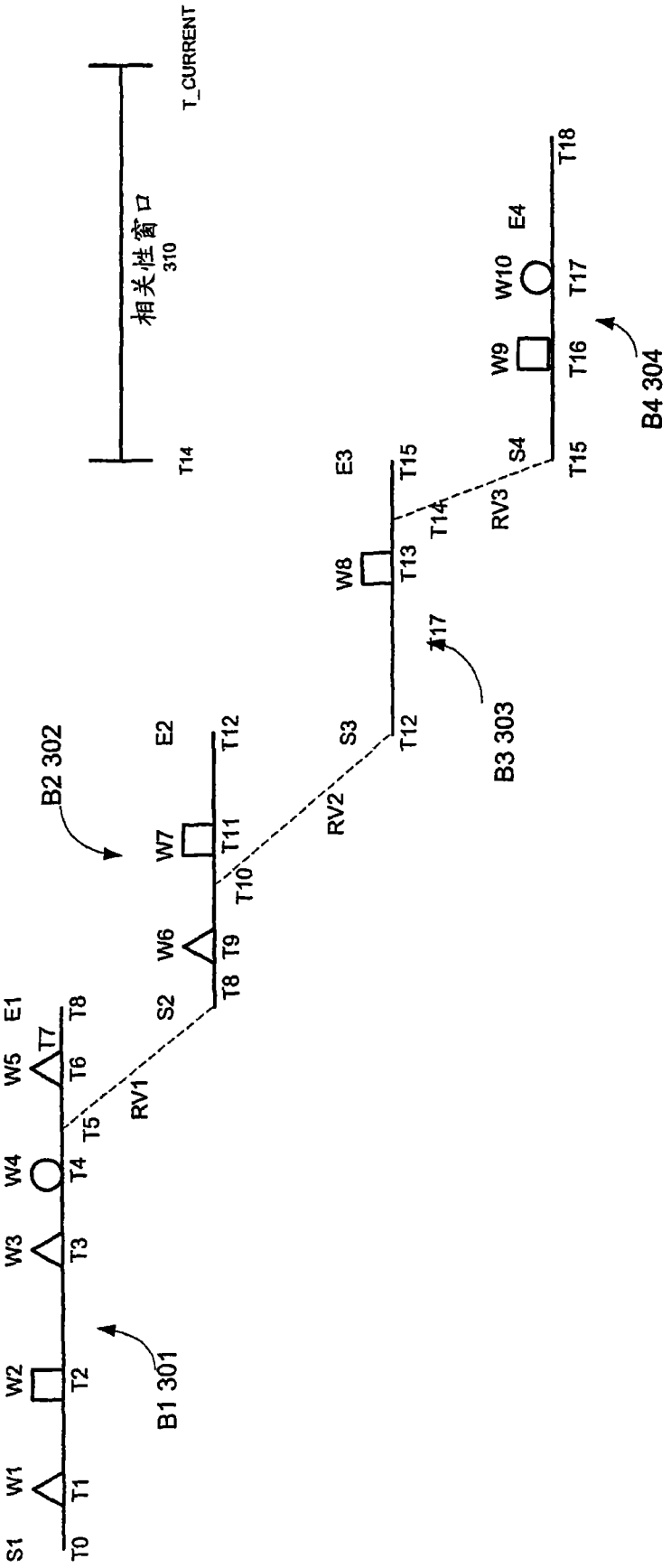

T_CURRENT
相关性窗口
310
T14
T15
T16
T17
T18
W9
W10
E4
S4
RV3
B4 304
T12
T13
T14
W8
E3
S3
RV2
B3 303
T11
T12
W7
E2
T10
B2 302
T8
T9
W6
S2
RV1
T5
T6
T7
W5
E1
T3
T4
W4
W3
T2
W2
B1 301
T0
T1
W1
S1
300

图 3

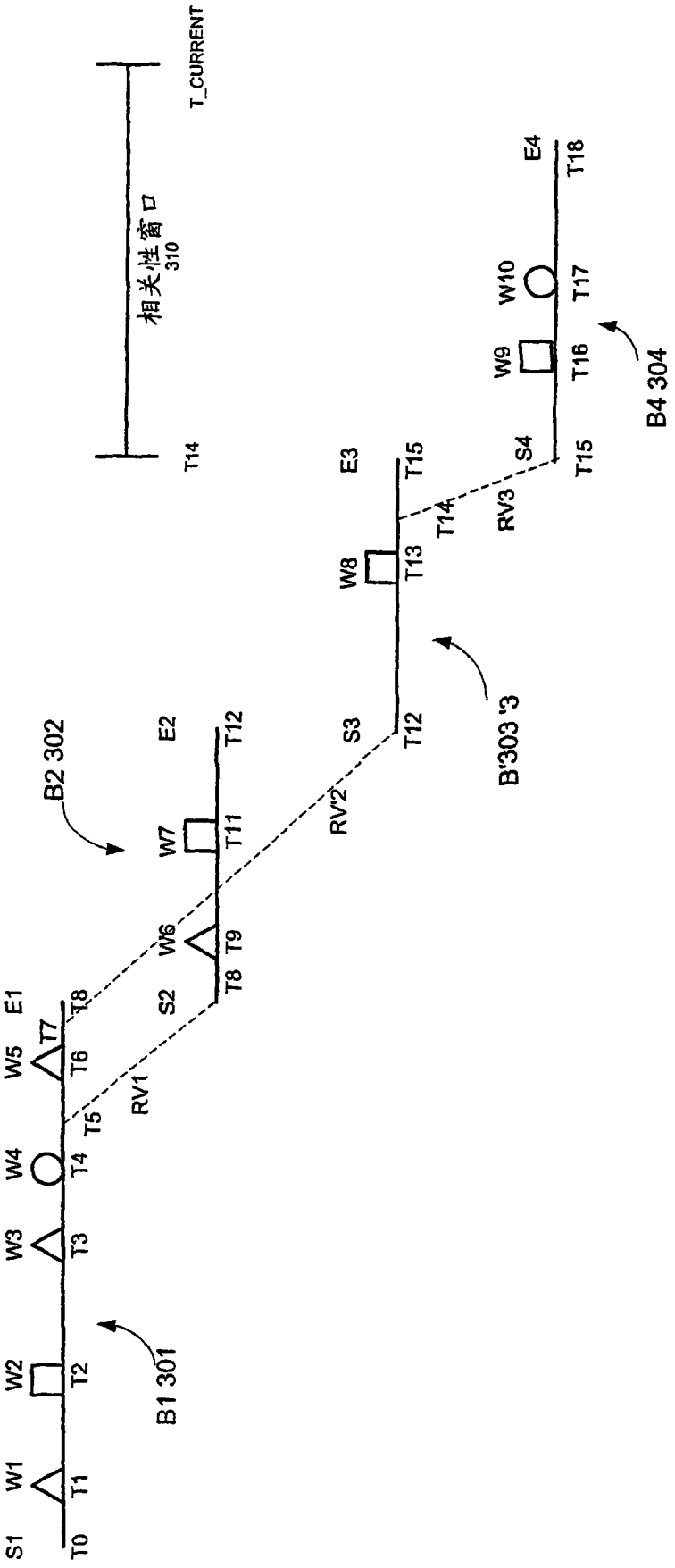
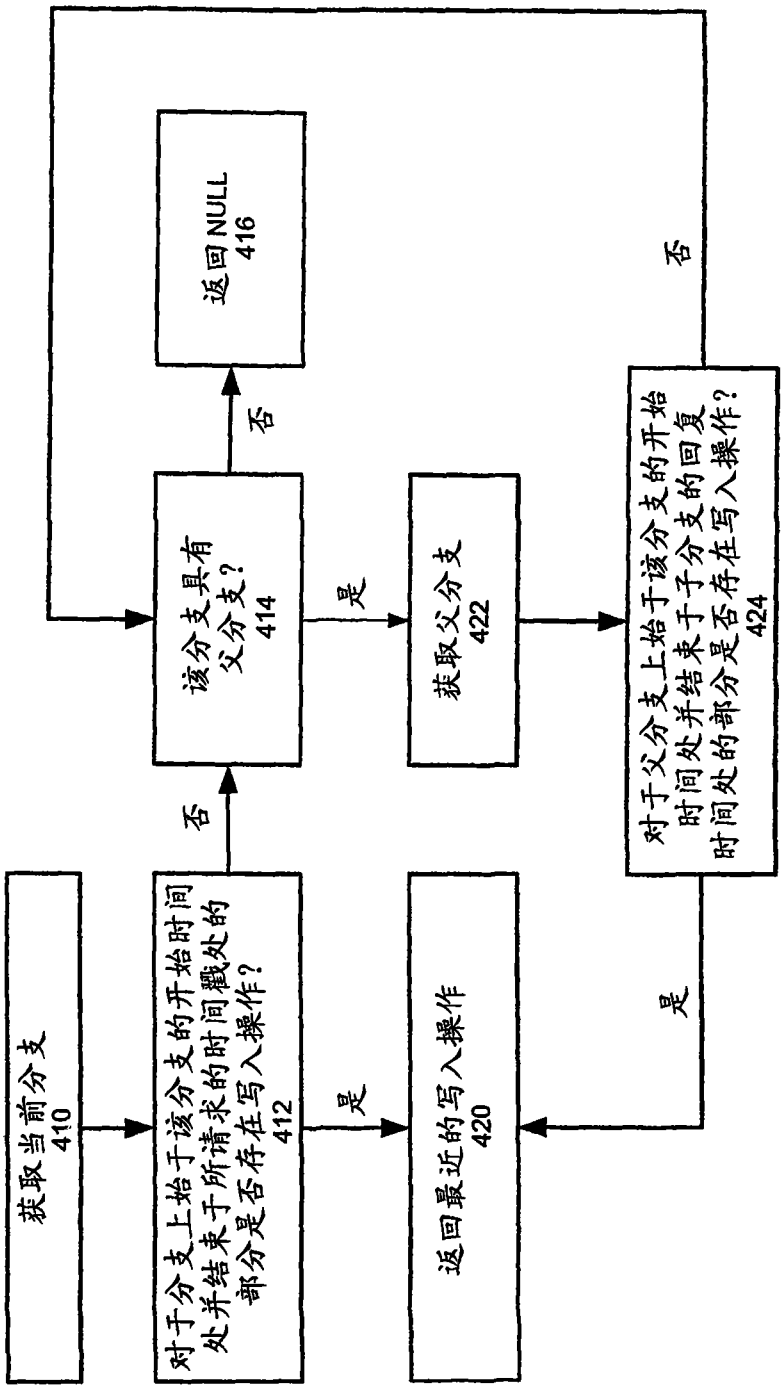

T_CURRENT
相关性窗口
310
T14
B2 302
B1 301
B'303 '3
B4 304
'300
S1
W1
T0
T1
W2
T2
W3
T3
W4
T4
W5
T5
E1
T6
T7
T8
RV1
S2
W6
T8
T9
W7
E2
T11
T12
RV'2
S3
T12
W8
E3
T13
T15
T14
RV3
S4
T15
W9
T16
W10
T17
E4
T18

图 4


获取当前分支
410
对于分支上始于该分支的开始时间处并结束于所请求的时间戳处的部分是否存在写入操作?
412
否
是
该分支具有父分支?
414
否
返回NULL
416
是
获取父分支
422
返回最近的写入操作
420
对于父分支上始于该分支的开始时间处并结束于子分支的回复时间处的部分是否存在写入操作?
424
否
是

400

图 5

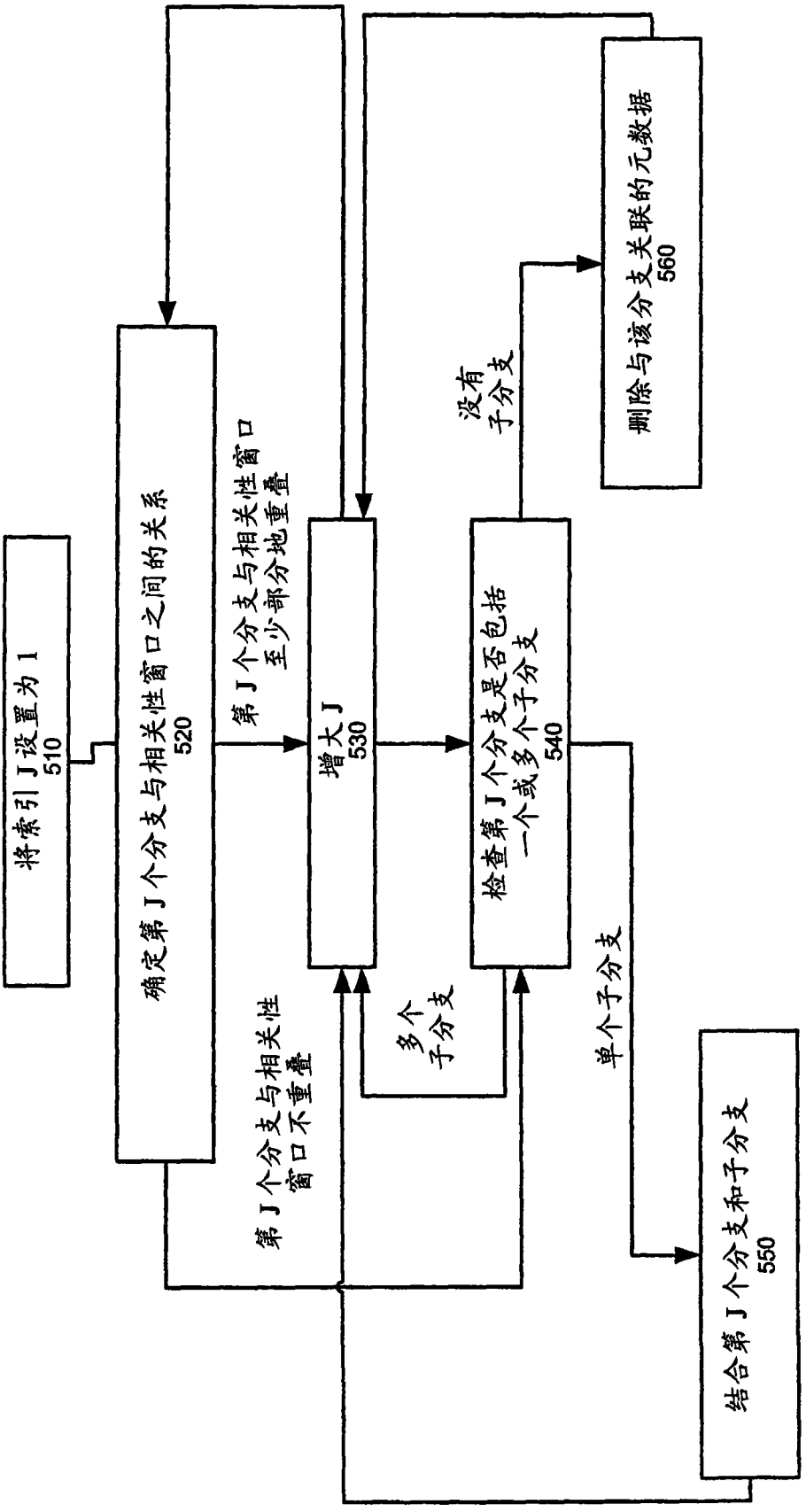

将索引 J 设置为 1
510
确定第 J 个分支与相关性窗口之间的关系
520
第 J 个分支与相关性窗口至少部分地重叠
第 J 个分支与相关性窗口不重叠
增大 J
530
检查第 J 个分支是否包括一个或多个子分支
540
多个子分支
没有子分支
单个子分支
删除与该分支关联的元数据
560
结合第 J 个分支和子分支
550
500

图 6

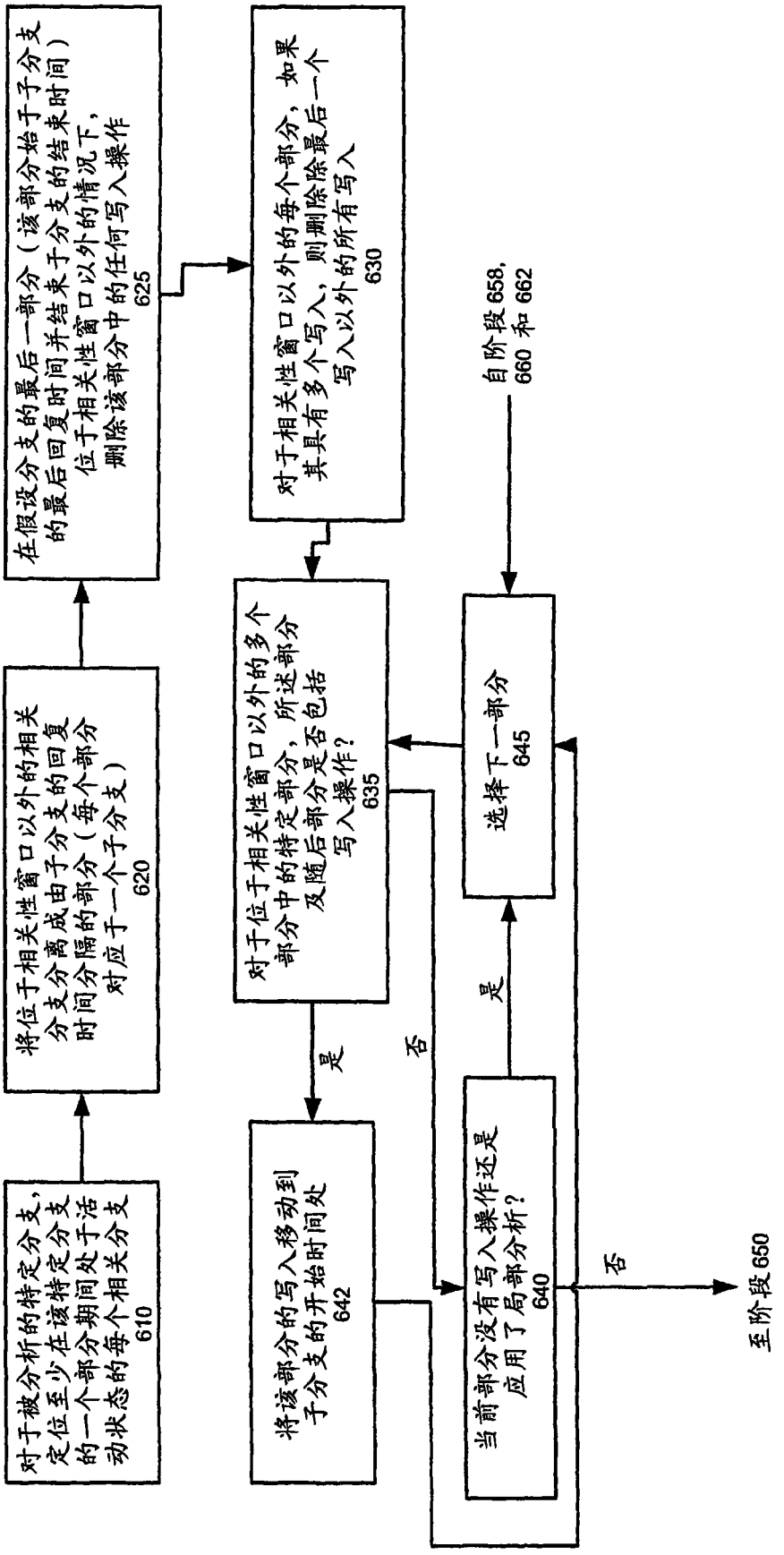
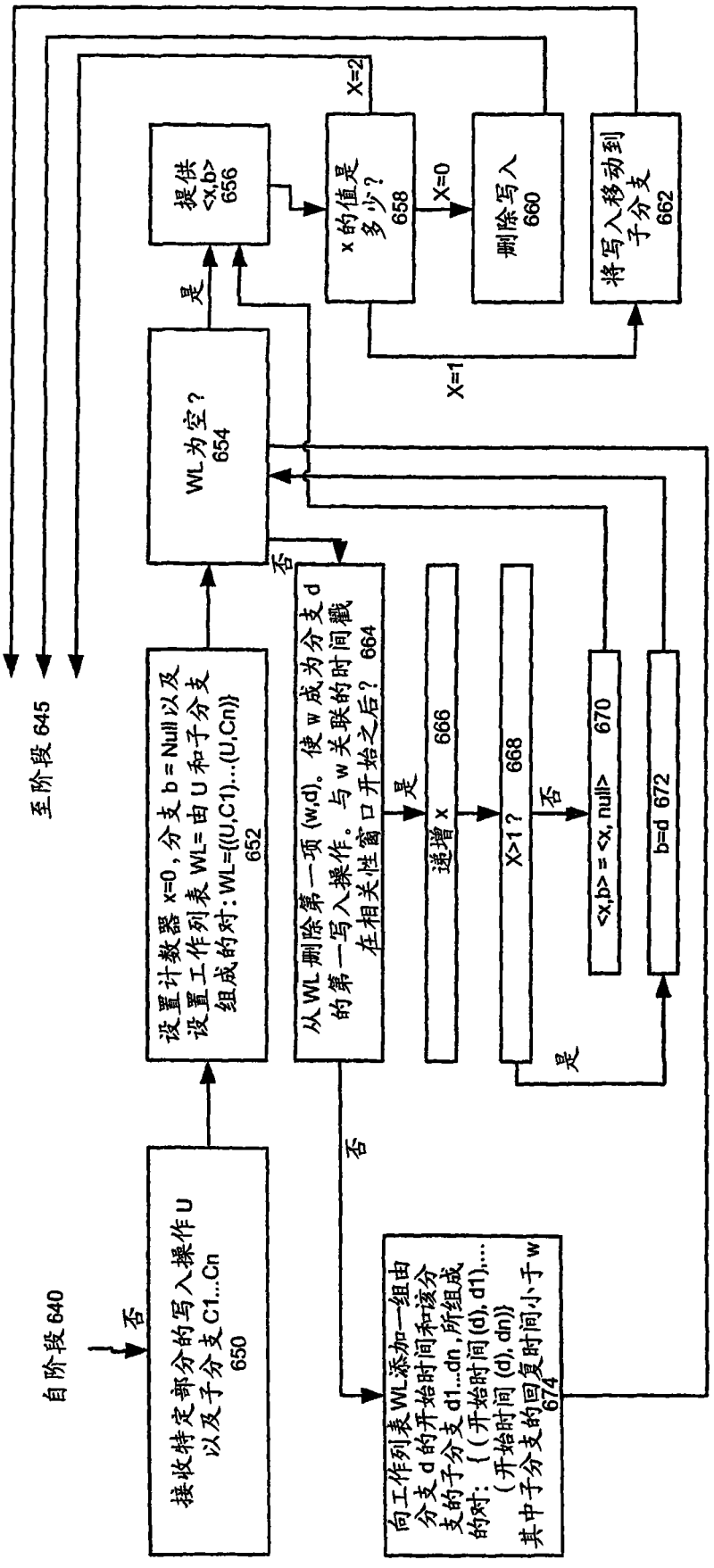

对于被分析的特定分支，定位至少在该特定分支的一个部分期间处于活动状态的每个相关分支
610
将位于相关性窗口以外的相关分支分离成由子分支的回复时间分隔的部分（每个部分对应于一个子分支）
620
在假设分支的最后一部分（该部分始于子分支的最后回复时间并结束于分支的结束时间）位于相关性窗口以外的情况下，删除该部分中的任何写入操作
625
对于相关性窗口以外的每个部分，如果其具有多个写入，则删除除最后一个写入以外的所有写入
630
自阶段 658, 660 和 662
对于位于相关性窗口以外的多个部分中的特定部分，所述部分及随后部分是否包括写入操作？
635
是
否
将该部分的写入移动到子分支的开始时间处
642
选择下一部分
645
是
当前部分没有写入操作还是应用了局部分析？
640
否
至阶段 650
600(a)

图 7a


自阶段 640
否
接收特定部分的写入操作 U
以及子分支C1...Cn
650
设置计数器 x=0，分支 b = Null 以及
设置工作列表 WL= 由 U 和子分支
组成的对：WL={(U,C1)...(U,Cn)}
652
WL 为空？
654
是
提供
<x,b>
656
x 的值是
多少？
658
X=2
X=0
X=1
删除写入
660
将写入移动到
子分支
662
至阶段 645
否
从 WL 删除第一项 (w,d)。使 w 成为分支 d
的第一写入操作。与 w 关联的时间戳
在相关性窗口开始之后？ 664
是
递增 x 666
X>1？ 668
否
是
<x,b> = <x, null> 670
b=d 672
向工作列表 WL 添加一组由
分支 d 的开始时间和该分
支的子分支 d1...dn，所组成
的对：{（开始时间 (d), d1),...
（开始时间 (d), dn)}
其中子分支的回复时间小于 w
674

600(b)

图 7b

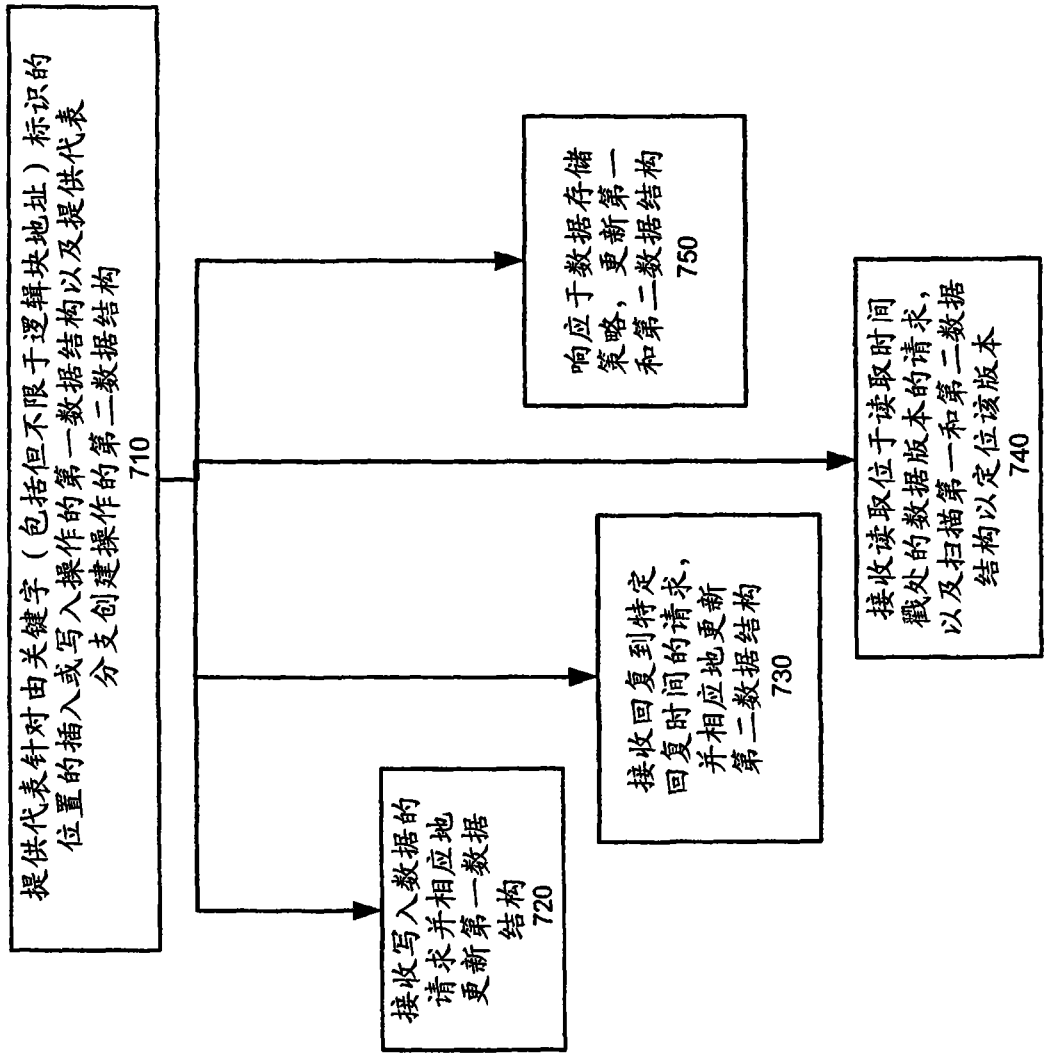

提供代表针对由关键字（包括但不限于逻辑块地址）标识的位置的插入或写入操作的第一数据结构以及提供代表分支创建操作的第二数据结构
710
接收写入数据的请求并相应地更新第一数据结构
720
接收回复到特定回复时间的请求，并相应地更新第二数据结构
730
响应于数据存储策略，更新第一和第二数据结构
750
接收读取位于读取时间戳处的数据版本的请求，以及扫描第一和第二数据结构以定位该版本
740
700

图 8

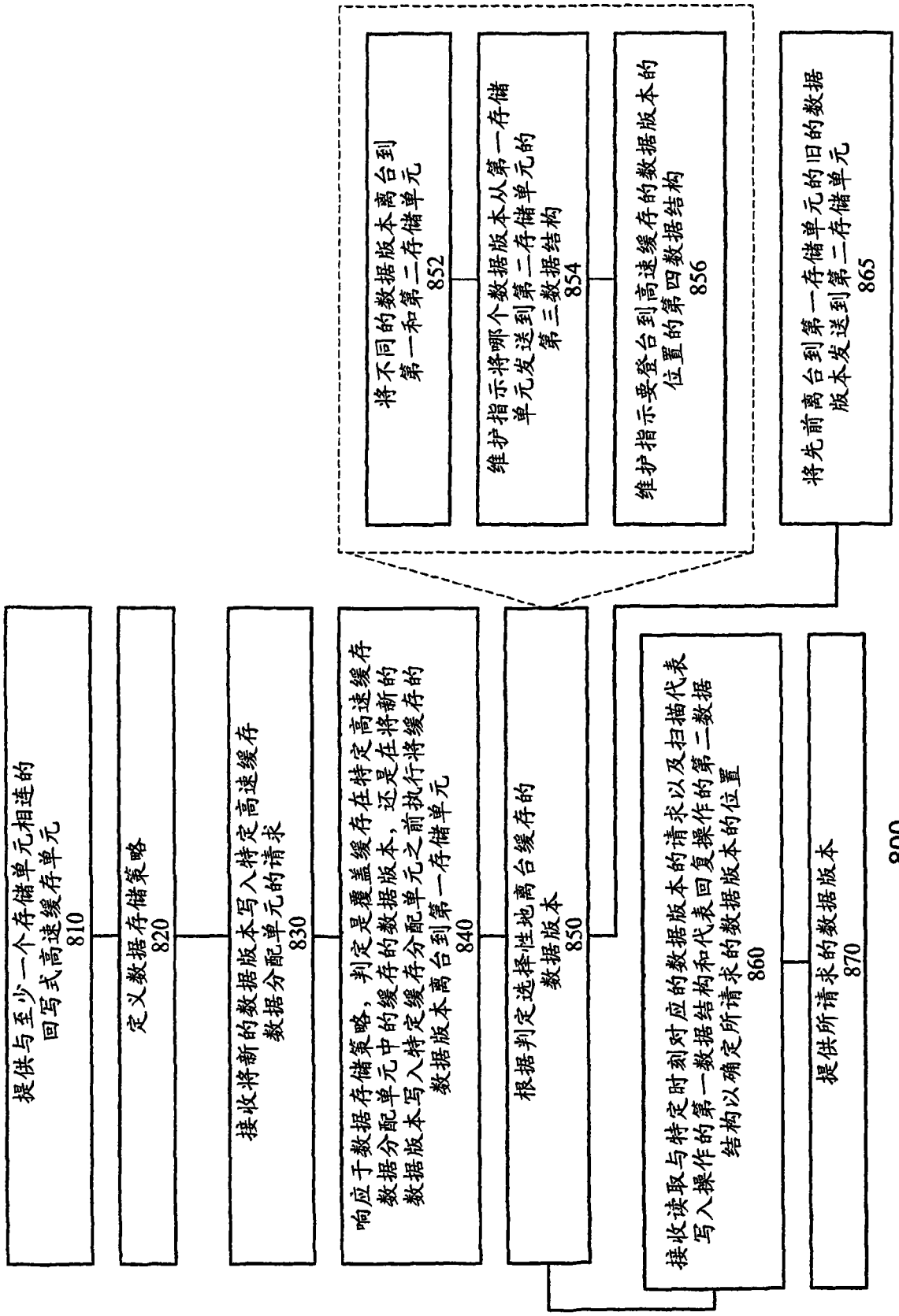

提供与至少一个存储单元相连的回写式高速缓存单元
810
定义数据存储策略
820
接收将新的数据版本写入特定高速缓存数据分配单元的请求
830
响应于数据存储策略，判定是覆盖缓存在特定高速缓存数据分配单元中的缓存的数据版本，还是在将新的数据版本写入特定缓存分配单元之前执行将缓存的数据版本离台到第一存储单元
840
根据判定选择性地离台缓存的数据版本
850
将不同的数据版本离台到第一和第二存储单元
852
维护指示将哪个数据版本从第一存储单元发送到第二存储单元的第三数据结构
854
维护指示要登台到高速缓存的数据版本的位置的第四数据结构
856
将先前离台到第一存储单元的旧的数据版本发送到第二存储单元
865
接收读取与特定时刻对应的数据版本的请求以及扫描代表写入操作的第一数据结构和代表回复操作的第二数据结构以确定所请求的数据版本的位置
860
提供所请求的数据版本
870
800

图 9A

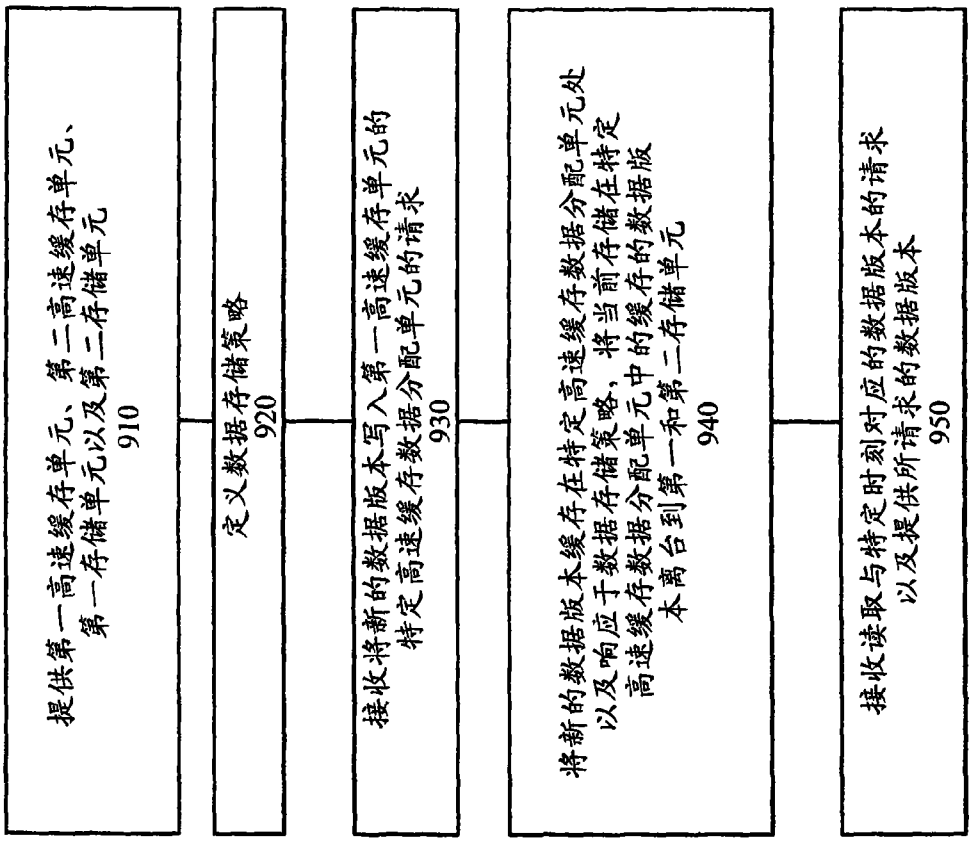

提供第一高速缓存单元、第二高速缓存单元、
第一存储单元以及第二存储单元
910
定义数据存储策略
920
接收将新的数据版本写入第一高速缓存单元的
特定高速缓存数据分配单元的请求
930
将新的数据版本缓存在特定高速缓存数据分配单元处
以及响应于数据存储策略，将当前存储在特定
高速缓存数据分配单元中的缓存的数据版
本离台到第一和第二存储单元
940
接收读取与特定时刻对应的数据版本的请求
以及提供所请求的数据版本
950
900

图 9 B