(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110033799 A
(43)申请公布日 2019.07.19

(21)申请号 201811590628.6

(22)申请日 2018.12.20

(30)优先权数据
10-2018-0068127 2018.06.14 KR
62/616,718 2018.01.12 US

(71)申请人 三星电子株式会社
地址 韩国京畿道

(72)发明人 黄珠荣

(74)专利代理机构 北京市柳沈律师事务所 11105
代理人 邵亚丽

(51)Int.Cl.
G11C 7/10(2006.01)

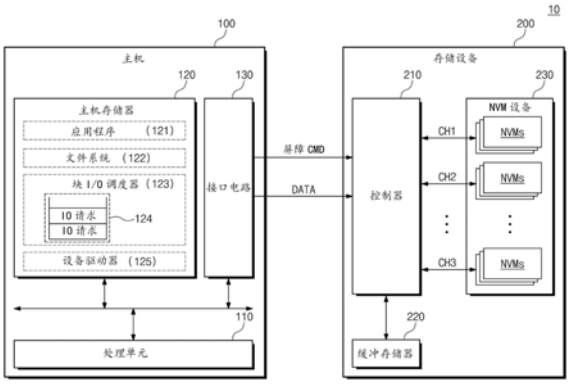
权利要求书2页 说明书17页 附图18页

(54)发明名称

基于屏障命令按顺序存储数据的存储设备

(57)摘要

一种将数据编程到包括非易失性存储设备的存储设备的方法，包括：从主机接收第一屏障命令至第三屏障命令；从主机接收与第一屏障命令至第三屏障命令相对应的第一数据至第三数据；合并第一屏障命令和第二屏障命令并基于第一屏障命令和第二屏障命令的顺序将第一数据和第二数据顺序地编程到非易失性存储设备；验证第一数据和第二数据两者的编程完成，当第一数据和第二数据的编程完成时映射入第一数据和第二数据的映射信息；以及当第一数据和第二数据中的至少一个的编程未完成时映射出第一数据和第二数据两者的信息；以及在映射入或映射出之后将第三数据编程到非易失性存储设备。



CN 110033799 A

1.一种将数据编程到包括非易失性存储器设备的存储设备的方法,该方法包括:

从主机接收第一屏障命令、第二屏障命令和第三屏障命令;

从所述主机接收与所述第一屏障命令相对应的第一数据、与所述第二屏障命令相对应的第二数据和与所述第三屏障命令相对应的第三数据;

合并所述第一屏障命令和所述第二屏障命令,并基于所述第一屏障命令和所述第二屏障命令的顺序将所述第一数据和所述第二数据顺序地编程到所述非易失性存储器设备;

验证所述第一数据和所述第二数据两者的编程完成;

当所述第一数据和所述第二数据的编程完成时映射入所述第一数据和所述第二数据,或者当所述第一数据和所述第二数据中的至少一个的编程未完成时映射出第一数据和第二数据两者;和

在所述映射入或所述映射出之后,将所述第三数据编程到所述非易失性存储器设备。

2.根据权利要求1所述的方法,其中,所述第一屏障命令至所述第三屏障命令中的每一个包括编程命令。

3.根据权利要求1所述的方法,其中,所述第一屏障命令至所述第三屏障命令中的每一个被标记为所述非易失性存储器设备的备用区域中的标志信息。

4.根据权利要求3所述的方法,其中,所述标志信息包括时期号,并且合并的第一屏障命令和第二屏障命令具有相同的时期号。

5.根据权利要求4所述的方法,其中,提交记录位被另外写入所述第二屏障命令的时期号中。

6.根据权利要求4所述的方法,其中,参考提交记录位来确定所述第一数据和所述第二数据两者的编程是否完成。

7.根据权利要求1所述的方法,其中,在接收到所述第一屏障命令之后,从所述主机顺序地提供所述第一数据。

8.根据权利要求1所述的方法,其中,在接收到所述第二屏障命令之后,从所述主机提供跟随所述第二屏障命令的所述第二数据。

9.根据权利要求1所述的方法,其中,在确定所述第一数据和所述第二数据两者的编程是否完成之后,编程所述第三数据。

10.根据权利要求9所述的方法,其中,所述第一数据和所述第二数据被编程到所述非易失性存储器设备的第一块,并且所述第三数据被编程到所述非易失性存储器设备的第二块。

11.根据权利要求1所述的方法,其中,所述映射入包括将所述第一数据和所述第二数据分类为有效数据。

12.根据权利要求1所述的方法,其中,所述映射出包括将所述第一数据和所述第二数据分类为无效数据。

13.根据权利要求12所述的方法,进一步包括:

对分类为所述无效数据的所述第一数据和所述第二数据执行垃圾收集。

14.根据权利要求1所述的方法,其中,所述第一数据和所述第二数据的大小等于或小于所述非易失性存储器设备的编程单元。

15.根据权利要求1所述的方法,进一步包括当所述第一数据和所述第二数据的大小小

于所述非易失性存储器设备的编程单元时，由所述存储设备将所述第一数据和所述第二数据与第四数据一起编程到所述非易失性存储器设备。

16.一种控制非易失性存储器设备的存储器控制器的操作方法，该操作方法包括：

由所述存储器控制器从主机接收第一屏障命令和第一编程命令、第二屏障命令和第二编程命令以及第三屏障命令和第三编程命令；

由所述存储器控制器从所述主机接收与所述第一屏障命令相对应的第一数据、与所述第二屏障命令相对应的第二数据和与所述第三屏障命令相对应的第三数据；

由所述存储器控制器合并所述第一屏障命令至所述第三屏障命令，并将所述第一数据至所述第三数据顺序地编程到所述非易失性存储器设备；和

由所述存储器控制器验证所述第一数据至所述第三数据是否被编程，当所有所述第一数据至所述第三数据被编程时，将所述第一数据至所述第三数据分类为有效数据，或者当所述第一数据至所述第三数据中的至少一个未被编程时，将所述第一数据至所述第三数据分类为无效数据。

17.根据权利要求16所述的操作方法，其中，所述第一数据至所述第三数据的大小之和等于或小于所述非易失性存储器设备的编程单元。

18.根据权利要求16所述的操作方法，其中，参考提交记录位来确定所述第一数据至所述第三数据是否被编程。

19.一种用于编程数据的计算机系统，包括：

主机；和

存储设备，被配置为从所述主机接收第一屏障命令、第二屏障命令和第三屏障命令以及分别与所述第一屏障命令、所述第二屏障命令和所述第三屏障命令相对应的要被编程的第一数据、第二数据和第三数据，

其中，所述存储设备包括：

多个非易失性存储器设备，被配置为存储所述第一数据、所述第二数据和所述第三数据，和

存储器控制器，被配置为：

控制所述多个非易失性存储器设备，

合并所述第一屏障命令、所述第二屏障命令和所述第三屏障命令，

顺序地编程所述第一数据、所述第二数据和所述第三数据，

确定所述第一数据、所述第二数据和所述第三数据的编程是否完成，

当所有所述第一数据、所述第二数据和所述第三数据的编程完成时，映射入所述第一数据、所述第二数据和所述第三数据作为有效数据，并且

当所述第一数据、所述第二数据和所述第三数据中的至少一个的编程未完成时，映射出所述第一数据、所述第二数据和所述第三数据作为无效数据。

20.根据权利要求19所述的计算机系统，其中，所述主机被配置为当向所述存储设备提供所述第一屏障命令、所述第二屏障命令和所述第三屏障命令时，不刷新分别与所述第一屏障命令、所述第二屏障命令和所述第三屏障命令相对应的所述第一数据、所述第二数据和所述第三数据。

基于屏障命令按顺序存储数据的存储设备

[0001] 相关申请的交叉引用

[0002] 要求于2018年1月12日向美国专利商标局提交的美国临时专利申请第62/616,718号和于2018年6月14日向韩国知识产权局提交的韩国专利申请第10-2018-0068127号的优先权,其全部内容通过引用结合于此。

技术领域

[0003] 本文的公开涉及一种存储设备,更具体地,涉及一种基于屏障命令按顺序存储数据的存储设备。

背景技术

[0004] 存储在作为非易失性存储介质的存储设备中的数据可以永久或半永久保留,而不管存储设备是否通电。通常,这种存储设备可以首先存储从主机提供给缓冲存储器的数据,然后可以将数据从缓冲存储器存储到非易失性存储器。然而,由于上述编程操作,在将数据存储到非易失性存储器之前,首先将数据存储到缓冲存储器,因此主机的写入请求的顺序可能无法得到保证。

[0005] 为了确保写入顺序,主机可以将数据传送到存储设备,可以等待直到传送的数据被存储(或刷新)到存储设备的非易失性存储器,然后可以将下一个数据传送到存储设备。然而,此操作可能会降低主机的性能。因此,需要一种可以保证写入请求的顺序而不会降低主机性能的存储设备。

发明内容

[0006] 本发明构思的实施例提供了一种基于屏障命令按顺序存储数据的存储设备。

[0007] 本发明构思的实施例提供了一种用于将数据编程到包括非易失性存储设备的存储设备的方法,该方法包括:由存储设备从主机接收第一屏障命令、第二屏障命令和第三屏障命令;由存储设备从主机接收与第一屏障命令相对应的第一数据、与第二屏障命令相对应的第二数据和与第三屏障命令相对应的第三数据;由存储设备合并第一屏障命令和第二屏障命令,并基于第一屏障命令和第二屏障命令的顺序将第一数据和第二数据顺序地编程到非易失性存储设备;由存储设备验证第一数据和第二数据两者的编程完成;当第一数据和第二数据的编程完成时,由存储设备将第一数据和第二数据的映射信息映射入(mapping in)存储设备的映射表中,并且当第一数据和第二数据中的至少一个的编程未完成时,映射出(mapping out)第一数据和第二数据两者的映射信息;以及在映射入和映射出之后,由存储设备将第三数据编程到非易失性存储设备。

[0008] 本发明构思的实施例还提供了一种控制非易失性存储设备的存储器控制器的操作方法,包括由存储器控制器从主机接收第一屏障命令和第一编程命令、第二屏障命令和第二编程命令、以及第三屏障命令和第三编程命令;由存储器控制器从主机接收与第一屏障命令相对应的第一数据、与第二屏障命令相对应的第二数据、和与第三屏障命令相对应

的第三数据;由存储器控制器将第一至第三屏障命令合并,并将第一至第三数据顺序地编程到非易失性存储设备;以及由存储器控制器验证第一至第三数据是否被编程,当所有第一至第三数据被编程时,将第一至第三数据分类为有效数据,并且当第一数据至第三数据中的至少一个未被编程时,将第一数据至第三数据分类为无效数据。

[0009] 本发明构思的实施例还提供了一种计算机系统,其包括:主机;以及存储设备,被配置为从主机接收第一屏障命令、第二屏障命令和第三屏障命令、以及分别与第一屏障命令、第二屏障命令和第三屏障命令相对应的第一数据、第二数据和第三数据。该存储设备包括:多个非易失性存储设备,被配置为存储第一数据、第二数据和第三数据;以及存储器控制器,被配置为控制多个非易失性存储设备,合并第一屏障命令、第二屏障命令和第三屏障命令,顺序地将第一数据、第二数据和第三数据编程到非易失性存储设备,确定第一数据、第二数据和第三数据的编程是否完成,当所有第一数据、第二数据和第三数据的编程完成时,将第一数据、第二数据和第三数据映射入存储设备的映射表作为有效数据,并且当第一数据、第二数据和第三数据中的至少一个的编程未完成时,将第一数据、第二数据和第三数据映射出到映射表作为无效数据。

附图说明

[0010] 从下面结合附图的详细描述中将更清楚地理解本发明构思的实施例。

[0011] 图1示出了根据本发明构思的实施例的计算机系统的框图。

[0012] 图2示出了图1的主机执行日志记录的操作的时序图。

[0013] 图3示出了图1的控制器的框图。

[0014] 图4示出了根据本发明构思的实施例的存储设备不支持屏障命令的操作的图。

[0015] 图5示出了根据本发明构思的实施例的存储设备处理屏障命令的操作的图。

[0016] 图6示出了根据本发明构思的另一实施例的存储设备处理屏障命令的操作的图。

[0017] 图7示出了根据本发明构思的另一实施例的存储设备处理屏障命令的操作的图。

[0018] 图8示出了根据本发明构思的实施例的在其中存储设备按顺序处理接收到的屏障命令,并且所有数据被正常编程的操作的图。

[0019] 图9示出了根据本发明构思的实施例的图8中的在其中存储设备按顺序处理接收到的屏障命令,并且所有数据未被编程操作的图。

[0020] 图10示出了根据本发明构思的另一实施例的在其中存储设备按顺序处理接收到的屏障命令,并且数据被正常编程的操作的图。

[0021] 图11示出了根据本发明构思的另一实施例的图10中的在其中存储设备按顺序处理接收到的屏障命令,并且所有数据未被正常编程的操作的另一图。

[0022] 图12示出了根据本发明构思的另一实施例的在其中存储设备按顺序处理接收到的屏障命令的操作的图。

[0023] 图13示出了根据本发明构思的另一实施例的在其中存储设备按顺序处理接收到的屏障命令的操作的图。

[0024] 图14示出了图1的非易失性存储设备中的非易失性存储器的框图。

[0025] 图15示出了包括在图14的存储器单元阵列中的三维存储器块的第一块的电路图。

[0026] 图16示出了图1的存储设备的操作方法的流程图。

[0027] 图17示出了根据本发明构思的实施例的支持对写入顺序的系统调用的主机的软件堆栈的图。

[0028] 图18示出了图17的主机在存储设备上执行日志记录的操作的时序图。

[0029] 图19示出了根据本发明构思的实施例的存储设备的IOPS(Input/Output Operations Per Second,每秒的读取/写入操作)和命令队列深度的图。

具体实施方式

[0030] 以下将详细且清楚地描述本发明构思的实施例,以达到本领域普通技术人员可以容易地实施本发明构思的程度。

[0031] 如本发明构思领域中的传统,可以按照执行所描述的一个或多个功能的块来描述和示出实施例。这些块在本文可以被称为单元或模块等,它们由模拟和/或数字电路物理得实施,诸如逻辑门、集成电路、微处理器、微控制器、存储器电路、无源电子组件、有源电子组件、光学组件、硬连线电路等,并且可以可选地由固件和/或软件驱动。电路可以例如体现在一个或多个半导体芯片中,或者体现在诸如印刷电路板等的衬底支撑件上。构成块的电路可以由专用硬件实施,或者由处理器(例如,一个或多个编程的微处理器和相关电路系统)实施,或者由执行块的一些功能的专用硬件和执行块的其他功能的处理器的组合实施。在不脱离本发明构思的范围的情况下,实施例的每个块可以物理上分离成两个或更多个相互作用和离散的块。同样,在不脱离本发明构思的范围的情况下,实施例的块可以物理地组合成更复杂的块。

[0032] 图1示出了根据本发明构思的实施例的计算机系统的框图。计算机系统10包括主机100和存储设备200。计算机系统10可以被应用于电子设备或实施在电子设备内,诸如例如个人计算机、服务器、工作站、笔记本、平板电脑、移动设备、智能手机等。主机100可以将屏障命令和数据传送到存储设备200(例如,固态驱动器(solid state drive,SSD))。存储设备200可以基于屏障命令存储数据。主机100可以发出屏障命令来通知存储设备200的数据的写入顺序,并且存储设备200可以遵循或保持由主机100基于屏障命令请求的数据写入顺序。主机100包括处理单元110、主机存储器120、和接口(interface,I/F)电路130。

[0033] 处理单元110可以运行加载到主存储器120上的软件。例如,处理单元110可以运行应用程序、文件系统、块输入/输出调度器、和设备驱动器。处理单元110可以包括同构多核处理器或异构多核处理器。例如,处理单元110可以包括中央处理单元(central processing unit,CPU)、图像信号处理单元(image signal processing,ISP)、数字信号处理单元(digital signal processing,DSP)、图形处理单元(graphics processing unit,GPU)、视觉处理单元(vision processing unit,VPU)和神经处理单元(neural processing unit,NPU)中的至少一个。

[0034] 管理计算机系统10中的所有硬件和软件的操作系统OS(operating system,操作系统)可以被加载到主机存储器120上。详细地,用户空间中包括的应用程序121、内核空间中包括的文件系统122、块输入/输出(input/output,I/O)调度器123、和设备驱动器125可以被加载到主机存储器120上。加载到主机存储器120上的软件层121至123和125可以被包括在用于访问存储设备200的软件堆栈中。主机存储器120可以包括存储介质,诸如例如动态随机存取存储器(dynamic random access memory,DRAM)设备或静态随机存取存储器

(static random access memory,SRAM)设备。

[0035] 应用程序121可以作为基本(或默认)服务来运行,或者可以由用户的请求(或响应于用户的请求)来运行。存储应用程序121的用户空间和存储包括文件系统122、块I/O调度器123、设备驱动程序125等的内核的内核空间可以彼此分离。应用程序121不能直接访问诸如存储设备200的资源。相反,应用程序121可以调用在包含系统调用函数的库(未示出)中定义的函数,并且可以向内核请求必要的任务。在调用系统调用功能的情况下,可以从用户模式切换到内核模式。

[0036] 文件系统122可以管理存储到存储设备200的文件或数据。例如,文件系统122可以包括FAT(file allocation table,文件分配表)、NTFS™(new technology file system,新技术文件系统)、HFS(hierarchical file system,分层文件系统)、HPFS((high performance file system,高性能文件系统)、UFS(unix file system,Unix文件系统)、ext2(secondary extended file system,二级扩展文件系统)、ext3、ext4、JFS(journaling file system,日志记录文件系统)、ISO 9660、File-11、VxFS(veritas file system,Veritas文件系统)、ZFS™、ReiserFS、UDF(universal disk format,通用磁盘格式)等。特别地,文件系统122可以执行日志记录,以防止数据库、文件或数据的一致性由于突然断电(sudden power off,SPO)或系统崩溃而无法维护。

[0037] 块I/O调度器123可以存在于块层中。块I/O调度器123可以从文件系统122接收IO请求,并且可以将接收到的IO请求存储到调度器队列124。块I/O调度器123可以管理调度器队列124。块I/O调度器123可以合并IO请求或者可以调整IO请求的顺序(重新排序)。例如,调度器队列124可以包括无操作(Noop)调度器、截止期(Deadline)调度器、预期(Anticipatory)调度器、完全公平队列(completely fair queuing,CFQ)调度器等。

[0038] 设备驱动器125是作为内核的一部分操作的程序,用于控制硬件设备,诸如存储设备200。设备驱动器125可以从调度器队列124中移除IO请求,并且可以生成用于控制存储设备200的命令。设备驱动器125可以处理调度器队列124的IO请求。例如,设备驱动器125可以是按块在存储设备200上执行数据输入/输出的块设备。在其他实施例中,加载到主机存储器120上的程序和软件层不限于图1的示例。

[0039] 接口电路130可以提供主机100和存储设备200之间的物理连接。例如,接口电路130可以根据与存储设备200通信的方案来转换(或翻译)命令、地址和数据,这些命令、地址和数据与从主机100生成的各种IO请求相对应。

[0040] 主机100可以将屏障命令和数据传送到存储设备200。主机100可以通过发出屏障命令来请求存储设备200按顺序写入数据。存储设备200可以从主机100接收屏障命令和与屏障命令相对应的数据。存储设备200包括控制器210、缓冲存储器220和非易失性存储器设备230。

[0041] 控制器210(或存储器控制器)可以处理从主机100接收的命令。控制器210可以控制缓冲存储器220和非易失性存储器设备230的操作。控制器210可以将从主机100接收的数据存储或缓冲到缓冲存储器220,缓冲存储器220的数据I/O速度比非易失性存储器设备230的数据I/O速度快,然后可以将存储在缓冲存储器220中的数据写入或编程到非易失性存储器设备230。

[0042] 在一个实施例中,控制器210和接口电路130可以基于各种接口协议中的一种或多

种相互通信,诸如例如通用串行总线(universal serial bus,USB)、小型计算机系统接口(small computer system interface,SCSI)、外围组件互连快速(peripheral component interconnect express,PCIe)、非易失性存储器快速(nonvolatile memory express,NVME)、移动PCIe(mobile PCIe,M-PCIe)、高级技术附件(advanced technology attachment,ATA)、并行ATA(parallel ATA,PATA)、串行ATA(serial ATA,SATA)、串行连接SCSI(serial attached SCSI,SAS)、集成驱动电子设备(integrated drive electronics,IDE)、通用闪存(universal flash storage,UFS)和Firewire™。

[0043] 缓冲存储器220可以临时存储从主机100接收的数据或者从非易失性存储器设备230接收的数据。缓冲存储器220可以存储映射表,该映射表指示主机100的逻辑地址LA(logical address,LA)(或逻辑块地址LBA(logical block address,LBA))和非易失性存储器设备230的物理地址PA(physical address)(或物理块地址PBA(physical block address,PBA))之间的关系。缓冲存储器220可以通过使用DRAM设备或SRAM设备来实施。

[0044] 非易失性存储器设备230可以包括通过第一通道CH1与控制器210通信的非易失性存储器、通过第二通道CH2与控制器210通信的非易失性存储器、以及通过第三通道CH3与控制器210通信的非易失性存储器。非易失性存储器设备230和控制器210之间的通道的数量不限于图1所示的示例。每个非易失性存储器可以包括例如非易失性存储器单元,诸如NAND闪存单元、NOR闪存单元、电阻随机存取存储器(resistive random access memory,RERAM)单元、铁电随机存取存储器(ferroelectric random access memory,FRAM)单元、相变随机存取存储器(phase change random access memory,PRAM)单元、或磁随机存取存储器(magnetic random access memory,MRAM)单元。在下文中,将在假设第一至第三非易失性存储器中的每一个包括NAND闪存单元的情况下给出描述。

[0045] 在一个实施例中,控制器210可以使用缓冲存储器220作为高速缓冲存储器。控制器210可以将与非易失性存储器设备230的编程单元相对应的数据存储和合并到缓冲存储器220,并且可以同时将合并的数据编程到非易失性存储器设备230。非易失性存储器设备230的寿命和性能可以通过上述操作来提高,但是数据集可以不按照从主机100接收的写入命令的顺序被编程到非易失性存储器设备230。这里,术语“数据集”可以被用于指示分别与写入命令相对应的多种形式的数据,该数据集可以被称为“多数据”或“多个数据”。主机100可以发出屏障命令,从而基于要存储到存储设备200或应用程序121的数据的类型,将数据集按顺序编程到非易失性存储器设备230。

[0046] 图2示出了图1的主机执行日志记录的操作的时序图。将参考图1描述图2。例如,加载到主机存储器120上的操作系统(例如,Android™ OS)可以频繁地生成高速缓存刷新命令。操作系统可以生成高速缓存刷新命令以保证数据实际被编程到非易失性存储器设备230,但是可以生成高速缓存刷新命令以保证数据集被写入到非易失性存储器设备230的顺序。如上所述,文件系统122可以执行日志记录。在主机100修改数据库文件的一部分的情况下,主机100可以将待修改的原始数据备份到日志,可以修改数据库文件,并且可以删除日志。文件系统122可以将高速缓存刷新命令传送到存储设备200,用于提交日志事务或维护数据库的一致性。高速缓存刷新命令可以用于保持写入顺序。

[0047] 参考图2,当备份日志时或修改数据库文件后,可以调用系统调用,诸如fsync()。在fsync()开始的情况下,文件系统122可以将写入请求插入(或排队)到块层的调度器队

列124中，用于将文件数据“D”传送到存储设备200。写入请求可以被分派到存储设备200。文件系统122等待直到对文件数据“D”的直接存储器访问(direct memory access，DMA)传送完成。DMA传送可以意味着主机100中独立于处理单元110的DMA控制器(未示出)直接与存储设备200交换数据。在对文件数据“D”的DMA传送完成的情况下，文件系统122可以触发日志块设备(以下称为“JBD(journal block device)”)来提交日志事务。JBD可以是在主机存储器120的一部分运行的线程，文件系统122为日志记录保护其。

[0048] JBD可以将写入请求插入到调度器队列124中，用于将日志数据JD传送到存储设备200。写入请求可以由设备驱动器125分派给存储设备200。JBD等待直到对日志数据JD的DMA传送完成。在对日志数据JD的DMA传送完成的情况下，文件系统122可以将刷新请求插入到调度器队列124中，使得日志数据JD从缓冲存储器220刷新到非易失性存储器设备230。刷新请求可以由设备驱动器125分派给存储设备200(即对高速缓存刷新命令的传送)。在对日志数据JD的刷新完成的情况下，插入到调度器队列124中的写入请求可以被分派到存储设备200。在根据写入请求完成对日志提交JC的DMA传送并且日志提交JC被完全刷新的情况下，可以返回fsync()。只有在日志提交JC被写入到非易失性存储器设备230之后，文件系统122才可以提交日志记录事务。在提交日志记录事务之后，文件系统122可以执行另一操作。

[0049] 参考图2，为了在日志记录的运行期间保持写入顺序，文件系统122必须等待直到对文件数据“D”的DMA传送、对日志数据JD的DMA传送和刷新、以及对日志提交JC的DMA传送和刷新全部完成。参考图2描述的日志记录可以中和存储设备200中的并行处理，或者可以减少控制器210的命令队列的深度。具体而言，主机100的操作可以被延迟存储设备200中的单元编程所需的时间。因此，主机100可以生成屏障命令而不是高速缓存刷新命令，以保持写入顺序并减少由于高速缓存刷新命令引起的延迟。主机100可以发出屏障命令而不是高速缓存刷新命令，并且可以执行另一操作，而无需等待与屏障命令相对应的存储设备200的操作完成。

[0050] 在一个实施例中，屏障命令可以在主机100和存储设备200之间的接口协议中定义。屏障命令可以占据上述调度器队列124的一个条目。在另一实施例中，主机100可以通过设置写入命令的标志(例如，REQ_BARGE)来将写入命令设置为屏障命令。存储设备200可以解码写入命令的标志，并且可以确定写入命令是否是屏障命令。存储设备200可以按顺序存储与屏障命令相对应的数据。屏障命令可以包括编程命令，即写入命令。

[0051] 调用fsync()的示例在图2中示出，但是可以调用fdatasync()。fdatasync()类似于fsync()。fsync()可以修改文件元数据。然而，当调用fdatasync()时，在没有额外修改用于读取新写入数据的文件元数据的情况下，文件元数据可能不会被修改。

[0052] 图3是示出图1的控制器的框图。将参考图1描述图3。控制器210包括处理单元211、工作存储器212、主机接口(I/F)电路214、缓冲存储器接口(I/F)电路215、和闪存接口电路216。控制器210可以例如通过使用片上系统(system on chip，SoC)、专用集成电路(application specific integrated circuit，ASIC)、或现场可编程门阵列(field programmable gate array，FPGA)来实施。缓冲存储器220可以被提供为独立于控制器210，如图1所示，或者可以被包括在控制器210中，如图2所示。

[0053] 处理单元211可以解码从主机100提供的命令。处理单元211可以基于该命令控制控制器210的其他组件212至216的操作。处理单元211可以运行用于执行用于管理非易失性

存储器设备230的垃圾收集(garbage collection)的闪存转换层(flash translation layer,FTL)、指示逻辑地址和物理地址之间的关系的映射表、损耗均衡等。处理单元211可以包括上述处理单元中的至少一个。

[0054] 工作存储器212可以作为高速缓冲存储器操作。工作存储器212可以存储处理单元211的解码结果。例如,按照从主机100传送的命令CMD1到命令CMD3的顺序存储命令CMD1到命令CMD3的命令队列213可以被分配给工作存储器212的区域。这里,要存储到命令队列213的命令的数量不限于图3所示的示例。与图3的图示不同,命令队列213可以被放置在处理单元211中或者缓冲存储器220的部分区域中。

[0055] 主机接口电路214可以按照上述通信协议与主机100通信。例如,主机接口电路214可以按照NVMe协议操作。处理单元211可以通过主机接口电路214接收命令,并且可以按顺序将接收到的命令插入命令队列213中。

[0056] 缓冲存储器接口电路215可以在处理单元211的控制下控制缓冲存储器220的读取操作和写入操作。缓冲存储器接口电路215可以向缓冲存储器220提供指示逻辑地址和物理地址之间的关系的映射表。缓冲存储器接口电路215可以将存储在缓冲存储器220中的数据提供给主机接口电路214或闪存接口电路216。缓冲存储器接口电路215可以向缓冲存储器220提供从主机接口电路214或闪存接口电路216提供的数据。

[0057] 缓冲存储器220可以包括回写高速缓存221被分配到的区域和写入缓冲器222被分配到的区域。例如,在从主机100提供并与命令相对应的数据的大小小于非易失性存储器设备230的编程单元的情况下,数据可以被存储到回写高速缓存221。在从主机100提供并与命令相对应的数据的大小不小于非易失性存储器设备230的编程单元的情况下,数据可以被存储到写入缓冲器222。

[0058] 闪存接口电路216可以与非易失性存储器设备230交换数据。闪存接口电路216可以通过图1的通道CH1至通道CH3将从缓冲存储器220提供的数据写入非易失性存储器设备230。闪存接口电路216可以通过通道CH1至通道CH3从非易失性存储器设备230接收数据,并且可以将接收到的数据提供给缓冲存储器220。在下文中,将描述控制器210处理屏障命令的各种范例。

[0059] 图4和图5示出了根据本发明构思的实施例的存储设备处理屏障命令的操作的图。将参考图1一起描述图4和图5。存储设备200的命令队列213、回写高速缓存221、写入缓冲器222和非易失性存储器设备230(参考图1)仅为了便于描述而示出。非易失性存储器设备230可以包括多个块。然而,在图4和图5中仅示出了一个块。该块可以包括至少一个或多个物理页面。物理页面可以包括与读取单元或写入(或编程)单元相对应的存储单元。块的存储单元可以与擦除单元相对应。

[0060] 存储设备200的控制器210可以按顺序接收第一写入命令WCMD1、第二屏障命令BCMD2和第三写入命令WCMD3,并且第一写入命令WCMD1、第二屏障命令BCMD2和第三写入命令WCMD3可以按顺序插入命令队列213中。假设图4的存储设备200不支持第二屏障命令BCMD2,并且图5的存储设备200支持第二屏障命令BCMD2并保持写入顺序。并且,假设与第一写入命令WCMD1相对应的第一数据DATA1的大小和与第二屏障命令BCMD2相对应的第二数据DATA2的大小中的每一个都是4KB,其小于非易失性存储器设备230的编程单元(例如,16KB),并且与第三写入命令WCMD3相对应的第三数据DATA3的大小为16KB,其与非易失性存

储器设备230的编程单元相同。在图4和图5中示出了一个示例，非易失性存储器设备230的编程单元与物理页面的大小相匹配。在其他实施例中，编程单元可以不同于物理页面的大小。

[0061] 例如，控制器210可以在交织方案(interleaving scheme)中通过多通道、多路和多平面同时编程多个物理页面，以减少非易失性存储器设备230的编程时间。也就是说，非易失性存储器设备230的编程单元可以根据连接非易失性存储器设备230和控制器210的通道的数量、连接到每个通道的通道的数量、非易失性存储器的平面的数量、物理页面的大小以及存储器单元存储的位数来确定。

[0062] 控制器210可以将大小小于编程单元的第一数据DATA1和第二数据DATA2存储到回写高速缓存221。例如，控制器210可以将小于编程单元的数据集合并到回写高速缓存221，并且可以将与编程单元相对应的合并数据编程到非易失性存储器设备230。回写高速缓存221可以被用于合并数据集，数据的大小小于编程单元。控制器210可以将与编程单元相对应的第三数据DATA3存储到写入缓冲器222。写入缓冲器222可以被用于存储数据，其大小与编程单元相同或大于编程单元。

[0063] 参考图4，在另一数据被合并到第一数据DATA1和第二数据DATA2之前，第一数据DATA1和第二数据DATA2中的每一个不与编程单元相对应，与编程单元相对应的第三数据DATA3可以在第二数据DATA2之前被编程到非易失性存储器设备230。在第三数据DATA3被编程然后发生SPO(即突然断电)或系统崩溃的情况下，已经存储在回写高速缓存221中的第一数据DATA1和第二数据DATA2可能丢失。图4的控制器210不能根据第二屏障命令BCMD2按顺序编程第一数据DATA1、第二数据DATA2和第三数据DATA3。

[0064] 相反，图5的控制器210可以解码第二屏障命令BCMD2，并且可以按顺序编程第一数据DATA1、第二数据DATA2和第三数据DATA3。控制器210可以合并第一数据DATA1、第二数据DATA2、和第三数据DATA3的一部分(例如，8KB)，并且可以将合并的数据编程到非易失性存储器设备230。这里，合并的数据的大小可以与编程单元相对应。即使与第二屏障命令BCMD2相对应的第二数据DATA2的大小小于编程单元，控制器210也可以从回写高速缓存221或写入缓冲器222借用任何其他数据，可以合并借用的数据和第二数据DATA2，并且可以将与编程单元相对应的合并的数据编程到非易失性存储器设备230。也就是说，控制器210可以基于第二屏障命令BCMD2保持第二数据DATA2的写入顺序。

[0065] 图6示出了根据本发明构思的另一实施例的在其中存储设备处理屏障命令的操作的图。将参考图1和图5一起描述图6。参考图4和图5描述的假设也可以应用于图6，并且假设图6的存储设备200支持第二屏障命令BCMD2并且保持写入顺序。

[0066] 再次返回图5，在控制器210将第三数据DATA3的一部分与第二数据DATA2合并的情况下，第三数据DATA3的其余部分可以被编程到另一物理页面。即使第三数据DATA3的大小与编程单元相对应，第三数据DATA3也可以被划分成不同的页面，并且可以被编程到不同的页面。在这种情况下，为了读取所有第三数据DATA3，非易失性存储器设备230必须激活第三数据DATA3存储在其中的至少两个物理页面。

[0067] 参考图6，控制器210可以将第一数据DATA1和第二数据DATA2与虚拟数据Dummy而不是第三数据DATA3合并。控制器210可以通过使用虚拟数据来调整与第二屏障命令BCMD2相对应的第二数据DATA2的大小到编程单元。控制器210可以合并第一数据DATA1、第二数据

DATA2和虚拟数据，可以将与编程单元相对应的合并的数据编程到非易失性存储器设备230的物理页面，然后可以将第三数据DATA3编程到非易失性存储器设备230的另一物理页面。这里，第二数据DATA2被编程在其中的物理页面的位置和第三数据DATA3被编程在其中的物理页面的位置可以彼此相邻或者可以彼此远离。控制器210可以基于第二屏障命令BCMD2保持第二数据DATA2的写入顺序。

[0068] 图7示出了根据本发明构思的另一实施例的在其中存储设备处理屏障命令的操作的图。将一起参考图1和图5描述图7。参考图4和图5描述的假设也可以被应用于图7，并且假设图7的存储设备200支持第二屏障命令BCMD2并且保持写入顺序。

[0069] 控制器210可以将存储在回写高速缓存221中的第一数据DATA1和第二数据DATA2存储到第一块BLK1，然后可以将存储在写入缓冲器222中的第三数据DATA3存储到第二块BLK2。尽管在图7中未示出，但是控制器210可以合并图6的第一数据DATA1、第二数据DATA2和虚拟数据，并且可以将合并的数据存储到第一块BLK1。

[0070] 在图7中，小于编程单元的第一数据DATA1和第二数据DATA2可以是相对频繁更新的热数据(hot data)，并且与编程单元相对应的第三数据DATA3可以是不相对频繁更新的冷数据(cold data)。控制器210可以分离存储热数据的第一块BLK1和存储冷数据的第二块BLK2，同时基于第二屏障命令BCMD2保持第二数据DATA2的写入顺序。例如，第一块BLK1的存储器单元中每一个可以是存储一位的单级单元(single level cell，SLC)，第二块BLK2的存储器单元中每一个可以是存储至少两位的多级单元(multi-level cell，MLC)。

[0071] 参考图5至图7，控制器210可以按以下顺序执行编程操作：1)将在与第二屏障命令BCMD2相对应的第二数据DATA2之前接收到的第一数据DATA1编程到非易失性存储器设备230，2)将第二数据DATA2编程到非易失性存储器设备230，以及3)将在第二数据DATA2之后接收到的第三数据DATA3编程到非易失性存储器设备230。尽管在图5至图7中未示出，但是第一数据DATA1可以被编程，另一数据可以被编程，并且第二数据DATA2可以被编程。另外，第二数据DATA2可以被编程，另一数据可以被编程，第三数据DATA3可以被编程。在所有情况下，控制器210可以保持第二数据DATA2的写入顺序，而不管是否编程另一数据。

[0072] 图8和图9示出了根据本发明构思的实施例的在其中存储设备按顺序处理接收到的屏障命令的操作的图。图8和图9将一起描述，并将参考图1描述。在图8和图9中，假设第一数据DATA1至第四数据DATA4中的每一个的大小与编程单元相对应。在计算机系统10中，主机100可以按顺序将第一屏障命令BCMD1至第三屏障命令BCMD3传送到存储设备200，并且可以在DMA方案中将第一数据DATA1至第三数据DATA3传送到存储设备200。主机100可以按顺序传送第一屏障命令BCMD1、第一数据DATA1、第二屏障命令BCMD2、第二数据DATA2、第三屏障命令BCMD3和第三数据DATA3。第一数据DATA1可以在接收到第一屏障命令BCMD1之后被顺序地提供。第二数据DATA2可以在接收到第二屏障命令BCMD2之后被顺序地提供。然而，主机100可以在传送第一数据DATA1的同时传送第二屏障命令BCMD2。

[0073] 存储设备200的控制器210可以按顺序从主机100接收第一屏障命令BCMD1至第三屏障命令BCMD3，并且可以按顺序将第一屏障命令BCMD1至第三屏障命令BCMD3插入命令队列213中。控制器210可以接收分别与第一屏障命令BCMD1至第三屏障命令BCMD3相对应的第一数据DATA1至第三数据DATA3，并且可以将第一数据DATA1至第三数据DATA3存储到缓冲存储器220(①)。

[0074] 控制器210可以解码第一屏障命令BCMD1至第三屏障命令BCMD3，并且可以基于接收第一屏障命令BCMD1至第三屏障命令BCMD3的顺序将第一数据DATA1至第三数据DATA3按顺序顺序地编程到非易失性存储器设备230(②)。控制器210可以原子地执行第一数据DATA1至第三数据DATA3的编程操作。在控制器210执行原子编程操作的情况下，所有第一数据DATA1至第三数据DATA3可以正常编程到非易失性存储器设备230(参考图8)，或者所有第一数据DATA1至第三数据DATA3可以不被编程到非易失性存储器设备230(参考图9)。根据原子编程操作，不发生第一数据DATA1至第三数据DATA3的仅一部分被编程到非易失性存储器设备230的情况。控制器210可以将第一数据DATA1至第三数据DATA3编程到非易失性存储器设备230，并且可以提交第一数据DATA1至第三数据DATA3的编程操作。可替换地，控制器210可以回退(roll back)，而不编程第一数据DATA1至第三数据DATA3。在完成对第一数据DATA1至第三数据DATA3的原子编程操作之后，控制器210可以将第四数据DATA4编程到非易失性存储器设备230。图8和图9中未示出用于第四数据DATA4的命令。

[0075] 参考图8，第一数据DATA1可以被编程到第一物理页面<P1>，第二数据DATA2可以被编程到第二物理页面<P2>，并且第三数据DATA3可以被编程到第三物理页面<P3>(原子编程成功)。在这种情况下，控制器210可以在映射表L2P中映射入或更新第一至第三物理页面<P1:P3>的映射信息(③)。这里，术语“映射入”可以表示控制器210在映射表L2P中更新第一至第三物理页面<P1:P3>的映射信息的操作。控制器210可以通过映射入操作将所有第一数据DATA1至第三数据DATA3分类为有效数据。控制器210可以映射入第一数据DATA1至第三数据DATA3。

[0076] 参考图9，第一数据DATA1可以被编程到第一物理页面<P1>；第三数据DATA3可以被编程到第三物理页面<P3>，但是由于SPO或系统崩溃(原子编程失败)，第二数据DATA2可以不被编程到第二物理页面<P2>。在这种情况下，控制器210可以在映射表L2P中映射出第一物理页面<P1>和第三物理页面<P3>的映射信息以及第二物理页面<P2>的映射信息(③)。这里，术语“映射出”可以表示控制器210在映射表L2P中不更新第一至第三物理页面<P1:P3>的映射信息的操作。控制器210可以通过映射出操作将所有第一数据DATA1至第三数据DATA3分类为无效数据。控制器210可以对第一数据DATA1至第三数据DATA3执行垃圾收集，并且可以擦除第一数据DATA1至第三数据DATA3(④)。

[0077] 图10和图11示出了根据本发明构思的另一实施例的在其中存储设备按顺序处理接收到的屏障命令的操作的图。图10和图11将一起描述，并将参考图1描述。与图8和图9的情况一样，控制器210可以按顺序接收第一屏障命令BCMD1至第三屏障命令BCMD3，并且可以按顺序将第一屏障命令BCMD1至第三屏障命令BCMD3插入命令队列213中。控制器210可以合并和处理多个屏障命令。例如，控制器210可以合并第一屏障命令BCMD1和第二屏障命令BCMD2。控制器210可以首先处理合并的第一屏障命令BCMD1和第二屏障命令BCMD2，然后可以处理第三屏障命令BCMD3。

[0078] 控制器210可以接收分别与第一屏障命令BCMD1至第三屏障命令BCMD3相对应的第一数据DATA1至第三数据DATA3，并且可以将第一数据DATA1至第三数据DATA3存储到缓冲存储器220(①)。控制器210可以基于第一屏障命令BCMD1和第二屏障命令BCMD2的顺序，将第一数据DATA1和第二数据DATA2顺序地编程到非易失性存储器设备230的第二页面<P2>和第三页面<P3>。接下来，控制器210可以将提交页面编程到非易失性存储器设备230的第四页

面<P4>,用于确定是否提交第一数据DATA1和第二数据DATA2的编程操作(②)。

[0079] 控制器210可以通过读取提交页面来确定是否提交第一数据DATA1和第二数据DATA2的编程操作。控制器210可以读取或扫描第一物理页面<P1>和第四物理页面<P4>的提交页面,并且可以确定是否提交在提交页面之间的第一数据DATA1和第二数据DATA2的编程操作。参考图10,控制器210可以将第一数据DATA1和第二数据DATA2分类为有效数据,并且可以在映射表L2P中映射入第一物理页面<P1>和第二物理页面<P2>的映射信息(③)。

[0080] 相反,参考图11,由于SPO或系统崩溃,第二数据DATA2可能没有被编程到第三物理页面<P3>,提交页面可能没有被编程到第四物理页面<P4>。控制器210可以将第一数据DATA1和第二数据DATA2分类为无效数据,并且可以在映射表L2P中映射出第一物理页面<P1>和第二物理页面<P2>的映射信息(③)。这里,即使在第二数据DATA2被编程到第三物理页面<P3>并且只有提交页面没有被编程到第四物理页面<P4>的情况下,控制器210也可以在映射表L2P中映射出第一物理页面<P1>和第二物理页面<P2>的映射信息(③)。之后,控制器210可以对第一数据DATA1和第二数据DATA2执行垃圾收集(④)。

[0081] 在映射表L2P中映射入或映射出第一物理页面<P1>和第二物理页面<P2>的映射信息之后,控制器210可以将第三数据DATA3编程到第五物理页面<P5>。接下来,控制器210可以将提交页面编程到第六物理页面<P6>,用于确定是否提交第三数据DATA3的编程操作(④)。控制器210可以读取或扫描提交页面,并且可以确定是否提交提交页面之间的第三数据DATA3的编程操作。

[0082] 参考图10,控制器210可以将编程到第四物理页面<P4>和第六物理页面<P6>之间的第五物理页面<P5>的第三数据DATA3分类为有效数据,并且可以在映射表L2P中映射入第五物理页面<P5>的映射信息(⑤)。参考图11,控制器210可以将编程到第一物理页面<P1>和第六物理页面<P6>之间的第五物理页面<P5>的第三数据DATA3分类为有效数据,并且可以在映射表L2P中映射入第五物理页面<P5>的映射信息(⑤)。例如,参考图11,对第一数据DATA1和第二数据DATA2执行垃圾收集的时间可以在第三数据DATA3被编程的时间(⑤)之后。

[0083] 图12示出了根据本发明构思的另一实施例的在其中存储设备按顺序处理接收到的屏障命令的操作的图。将参考图1描述图12。在图12中,假设数据的原子编程操作都是成功的。

[0084] 与图8至图11的情况一样,控制器210可以按顺序接收第一屏障命令BCMD1至第三屏障命令BCMD3,并且可以按顺序将第一屏障命令BCMD1至第三屏障命令BCMD3插入命令队列213中。例如,控制器210可以合并第二屏障命令BCMD2和第三屏障命令BCMD3。控制器210可以首先处理第一屏障命令BCMD1,然后可以处理合并的第二屏障命令BCMD2和第三屏障命令BCMD3。控制器210可以接收分别与第一屏障命令BCMD1至第三屏障命令BCMD3相对应的第一数据DATA1至第三数据DATA3,并且可以将第一数据DATA1至第三数据DATA3存储到缓冲存储器220(①)。

[0085] 与图8至图11的情况不同,控制器210可以将数据连同标志信息一起编程到非易失性存储器设备230,而不是仅将数据编程到非易失性存储器设备230。标志信息可以标记指示数据的写入顺序的屏障命令。

[0086] 可以根据主机100的屏障命令来确定数据的时期(epoch)。时期号表示数据的时

期,并用于区分相对于屏障命令的第一编程数据与随后编程的数据。控制器210可以在数据的标志信息中包括数据的时期号。控制器210可以将相同的时期号分配给与合并的屏障命令相对应的数据集。参考图12,控制器210可以将第一时期号EP<1>分配给第一数据DATA1,并且可以将第二时期号EP<2>分配给第二数据DATA2和第三数据DATA3。提交记录位“C”可以由控制器210包括在数据的标志信息中。如同在提交页面中一样,提交记录位“C”可以被用于确定是否提交数据的编程操作。

[0087] 控制器210可以将第一数据DATA1、第一时期号EP<1>和提交记录位“C”编程到非易失性存储器设备230的第一物理页面<P1>(②)。控制器210可以对第一物理页面<P1>的数据区域中的第一数据DATA1以及第一物理页面<P1>的备用区域中的第一时期号EP<1>和提交记录位“C”进行编程。然而,不同于图12所示,在其他实施例中,第一时期号EP<1>和提交记录位“C”可以被存储到非易失性存储器设备230的不同页面或不同块。

[0088] 控制器210可以读取第一物理页面<P1>的提交记录位“C”,并且可以确定是否提交第一数据DATA1的编程操作。控制器210可以将第一数据DATA1分类为有效数据,并且可以在映射表L2P中映射入第一物理页面<P1>的映射信息(③)。

[0089] 如同在第一数据DATA1中一样,控制器210可以将第二数据DATA2和第二时期号EP2编程到第二物理页面P2,并将第三数据DATA3、第二时期号EP2和提交记录位“C”编程到第三物理页面P3(④)。控制器210可以读取第三物理页面<P3>的提交记录位“C”,并且可以确定是否提交第二数据DATA2和第三数据DATA3的编程操作。控制器210可以将第二数据DATA2和第三数据DATA3分类为有效数据,并且可以在映射表L2P中映射入第二物理页面<P2>和第三物理页面<P3>的映射信息(⑤)。

[0090] 在本发明构思的实施例中,控制器210可以将提交记录位“C”和与合并的命令相对应并具有相同时期号的数据集的最终编程数据(参考图12,第三数据DATA3)一起编程。控制器210可以不将提交记录位“C”包括在除了最终编程数据之外的具有相同时期号的数据集的剩余数据的标志信息中。在其他实施例中,控制器210可以重新排列与合并的命令相对应并具有相同时期号的数据集中的写入顺序。例如,控制器210可以将第三数据DATA3编程到第二物理页面<P2>,并将第二数据DATA2和提交记录位“C”编程到第三物理页面<P3>。

[0091] 图13示出了根据本发明构思的另一实施例的在其中存储设备按顺序处理接收到的屏障命令的操作的图。将参考图1描述图13。在图13中,假设数据的原子编程操作都成功,并且第一数据DATA1至第三数据DATA3中的每一个的大小小于编程单元。

[0092] 如同图8至图12的情况一样,控制器210可以按顺序接收第一屏障命令BCMD1至第三屏障命令BCMD3,并且可以按顺序将第一屏障命令BCMD1至第三屏障命令BCMD3插入命令队列213中。例如,控制器210可以合并第一屏障命令BCMD1至第三屏障命令BCMD3。例如,控制器210可以基于屏障数据的大小合并屏障命令。与合并的屏障命令相对应的数据集的大小可以是编程单元。

[0093] 控制器210可以接收分别与第一屏障命令BCMD1至第三屏障命令BCMD3相对应的第一数据DATA1至第三数据DATA3,并且可以将第一数据DATA1至第三数据DATA3存储到缓冲存储器220(①)。例如,第一数据DATA1可以是8KB,第二数据DATA2可以是4KB,第三数据DATA3可以是4KB,并且合并的数据集的大小可以是16KB,并且可以是编程单元。

[0094] 控制器210可以将第一数据DATA1至第三数据DATA3编程到非易失性存储器设备

230的第一物理页面<P1>(②)。尽管在图13中未示出,但是控制器210可以在第一物理页面<P1>的备用区域中编程第一数据DATA1至第三数据DATA3的标志信息(即图12的时期号和提交记录位)。在本发明构思的其他实施例中,第一数据DATA1至第三数据DATA3的位置可以由控制器210重新排列而没有限制,并且可以不同于图13中所示。控制器210可以将第一数据DATA1至第三数据DATA3分类为有效数据,并且可以在映射表L2P中映射入第一物理页面<P1>的映射信息(③)。

[0095] 图14示出了图1的非易失性存储设备中的一个非易失性存储器的框图。非易失性存储器231包括存储器单元阵列231_1、地址解码器231_2、页面缓冲器231_3、输入/输出(input/output,I/O)电路231_4、以及控制逻辑和电压生成电路231_5。非易失性存储器231也可以被称为“非易失性存储器芯片”。

[0096] 存储器单元阵列231_1可以包括多个存储器块。存储器块中的每一个可以包括多个单元串。单元串中的每一个可以包括存储器单元。存储器单元可以与字线WL连接。每个存储器单元可以包括存储一位的单级单元(SLC)或存储至少两位的多级单元(MLC)。

[0097] 在一个实施例中,存储器单元阵列231_1可以包括三维存储器阵列。三维(three-dimensional,3D)存储器阵列可以单片地形成在存储器单元阵列的一个或多个物理层级中,该存储器单元阵列具有布置在硅衬底上的电路上的有源区,该电路与存储器单元的操作相关。与存储器单元的操作相关联的电路可以位于衬底中或衬底上。术语“单片”表示3D存储器阵列的每一层级的层直接沉积在阵列的每一底层级的层上。3D存储器阵列包括垂直定向的垂直NAND串,使得至少一个存储器单元位于另一存储器单元之上。该至少一个存储器单元可以包括电荷俘获层。每个垂直NAND串可以包括位于存储器单元之上的至少一个选择晶体管。至少一个选择晶体管可以具有与存储器单元相同的结构,并且与存储器单元一起单片地形成。以下专利文献(通过引用并入本文)描述了三维存储器阵列的合适配置,其中三维存储器阵列被配置为多个层级,其中字线和/或位线在层级之间共享:美国专利第7,679,133号;第8,553,466号;第8,654,587号;第8,559,235号;和美国专利公开第2011/0233648号。

[0098] 地址解码器231_2通过字线WL、串选择线SSL和接地选择线GSL与存储器单元阵列231_1连接。地址解码器231_2可以从控制器210接收和解码物理地址ADD,并且可以基于解码结果驱动字线WL。例如,地址解码器231_2可以选择字线WL中的至少一条。

[0099] 页面缓冲器231_3通过位线BL与存储器单元阵列231_1连接。在控制逻辑和电压生成电路231_5的控制下,页面缓冲器231_3可以驱动位线BL,使得由页面从输入/输出电路231_4接收的数据“DATA”被存储到存储器单元阵列231_1。可替换地,在控制逻辑和电压生成电路231_5的控制下,页面缓冲器231_3可以按页面读取存储在存储器单元阵列231_1中的数据,并且可以将读取的数据提供给输入/输出电路231_4。

[0100] 输入/输出电路231_4可以从控制器210接收数据“DATA”,并且可以将数据“DATA”提供给页面缓冲器231_3。可替换地,输入/输出电路231_4可以从页面缓冲器231_3接收数据“DATA”,并且可以将数据“DATA”提供给控制器210。输入/输出电路231_4可以基于控制信号CTRL与外部设备交换数据。

[0101] 控制逻辑和电压生成电路231_5可以响应于从控制器210接收的存储命令CMD和控制信号CTRL来控制地址解码器231_2、页面缓冲器231_3和输入/输出电路231_4。例如,控制

逻辑和电压生成电路231_5可以响应于信号CMD和CTRL来控制其他组件，使得数据“DATA”被存储到存储器单元阵列231_1。可替换地，控制逻辑和电压生成电路231_5可以响应于信号CMD和CTRL来控制其他组件，使得存储在存储器单元阵列231_1中的数据“DATA”被传送到外部设备。控制逻辑和电压生成电路231_5可以生成非易失性存储器231操作所需的各种电压。控制逻辑和电压生成电路231_5可以例如生成编程电压、通过电压(pass voltages)、选择读取电压、非选择读取电压、擦除电压和验证电压。控制逻辑和电压生成电路231_5可以将生成的电压提供给地址解码器231_2或存储器单元阵列231_1的衬底。

[0102] 图15示出了包括在图14的存储器单元阵列中的三维存储器块的第一块的电路图。在第一块BLK1中，单元串的数量、由单元串组成的行和列的数量、单元晶体管GST、MC、DMC、SST等的数量、与单元晶体管连接的线GSL、WL、DML、SSL等的数量、以及第一块BLK1的高度不受限制，如图15所示。非易失性存储器设备230中包括的剩余存储器块也可以具有与第一块BLK1的结构类似的结构。

[0103] 第一块BLK1可以包括单元串CS11至CS22。单元串CS11至CS22可以沿着行方向和列方向布置。单元串CS11和CS12可以与串选择线SSL1a和SSL1b(第一行)连接。单元串CS21和CS22可以与串选择线SSL2a和SSL2b(第二行)连接。单元串CS11和CS21可以与第一位线BL1(第一列)连接。单元串CS12和CS22可以与第二位线BL2(第二列)连接。

[0104] 单元串CS11至CS22中的每一个可以包括单元晶体管。单元串CS11至CS22中的每一个可以包括串选择晶体管SSTa和SSTb、存储器单元MC1至MC8、接地选择晶体管GSTa和GSTb、以及虚拟存储器单元DMC1和DMC2。存储器单元MC1至MC8中的每一个可以是电荷俘获闪存(charge trap flash,CTF)存储器单元。

[0105] 存储器单元MC1至MC8可以串联连接，并且可以在垂直于由行方向和列方向定义的平面的方向的高度方向上堆叠。在每个单元串中，串选择晶体管SSTa和SSTb可以彼此串联连接，并且可以布置在存储器单元MC1至MC8和位线BL之间。在每个单元串中，接地选择晶体管GSTa和GSTb可以彼此串联连接，并且可以布置在存储器单元MC1至MC8和公共源极线CSL之间。在每个单元串中，第一虚拟存储器单元DMC1可以设置在存储器单元MC1至MC8与接地选择晶体管GSTa和GSTb之间。在每个单元串中，第二虚拟存储器单元DMC2可以设置在存储器单元MC1至MC8与串选择晶体管SSTa和SSTb之间。单元串CS11至CS22的接地选择晶体管GSTa和GSTb可以共同连接到接地选择线GSL。

[0106] 第一行中的单元串CS11和CS12的第一接地选择晶体管GSTa可以连接到第一接地选择线，第二行中的单元串CS21和CS22的第一接地选择晶体管GSTa可以连接到第二接地选择线。从衬底(未示出)以相同高度提供的接地选择晶体管可以连接到相同的接地选择线，并且以不同高度提供的接地选择晶体管可以连接到不同的接地选择线。例如，单元串CS11至CS22的第一接地选择晶体管GSTa可以连接到第一接地选择线，并且其第二接地选择晶体管GSTb可以连接到第二接地选择线。

[0107] 距衬底相同高度的存储器单元(或接地选择晶体管GSTa和GSTb)被共同连接到相同的字线，并且距其不同高度的存储器单元被连接到不同的字线。单元串CS11至CS22中的第一存储器单元MC1至第八存储单元MC8可以分别共同连接到第一字线WL1至第八字线WL8。从相同高度的第一串选择晶体管SSTa当中属于相同行的串选择晶体管可以连接到相同的串选择线，并且从相同高度的第一串选择晶体管SSTa当中属于不同行的串选择晶体管可以

连接到不同的串选择线。例如，第一行中的单元串CS11和CS12的第一串选择晶体管SSTa可以共同连接到串选择线SSL1a，并且第二行中的单元串CS21和CS22的第一串选择晶体管SSTa可以共同连接到串选择线SSL2a。并且，第一行中的单元串CS11和CS12的第二串选择晶体管SSTb可以共同连接到串选择线SSL1b，并且第二行中的单元串CS21和CS22的第二串选择晶体管SSTb可以共同连接到串选择线SSL2b。

[0108] 相同行中的单元串的串选择晶体管可以共同连接到串选择线。例如，第一行中的单元串CS11和CS12的第一串选择晶体管SSTa和第二串选择晶体管SSTb可以共同连接到相同的串选择线。第二行中的单元串CS21和CS22的第一串选择晶体管SSTa和第二串选择晶体管SSTb可以共同连接到相同的串选择线。相同高度的虚拟存储器单元可以与相同的虚拟字线连接，并且不同高度的虚拟存储器单元可以与不同的虚拟字线连接。例如，第一虚拟存储器单元DMC1可以与第一虚拟字线DWL1连接，并且第二虚拟存储器单元DMC2可以与第二虚拟字线DWL2连接。

[0109] 在第一块BLK1中，读取和写入操作可以由该行执行。例如，第一块BLK1中的一行可以由串选择线SSL1a、SSL1b、SSL2a和SSL2b选择。当导通电压被供应给串选择线SSL1a和SSL1b并且截止电压被供应给串选择线SSL2a和SSL2b时，第一行中的单元串CS11和CS12可以连接到位线BL1和BL2。在相反的情况下，第二行中的单元串CS21和CS22可以连接到位线BL1和BL2。当驱动字线时，从所选择的行中的单元串的存储器单元当中选择属于相同高度的存储器单元。所选择的存储器单元可以与物理页面单元相对应。可以对所选择的存储器单元执行读取或写入操作。

[0110] 图16示出了图1的存储设备的操作方法的流程图。在操作S110中，存储设备200的控制器210接收屏障命令和分别与屏障命令相对应的数据(或数据集)。如上所述，数据可以以DMA方案从主机100传送。

[0111] 在操作S120中，控制器210合并屏障命令并基于屏障命令的顺序或根据屏障命令的顺序，将数据顺序地编程到非易失性存储器设备230。控制器210可以基于非易失性存储器设备230的编程单元、要同时访问的物理页面的数量、要存储到存储器单元的位数、一个物理页面的大小、多通道、多路和多平面来合并屏障命令。在图8至图13中示出了实施例，数据被存储在任何一个物理页面中。然而，数据可以被分发并存储到非易失性存储器设备230中的多个芯片、多个平面、多个块或多个页面。可以原子地执行操作S120的编程。

[0112] 在操作S130中，控制器210验证操作S120的编程完成。控制器210可以读取图10和图11的提交页面，或者可以读取图12的提交记录位。控制器210确定被分配到提交页面或提交记录位的数据为有效数据，并确定未被分配到提交页面或提交记录位的数据为无效数据。

[0113] 在编程成功的情况下(S130中的编程成功)，在操作S140中，控制器210在映射表L2P中映射入有效数据被编程的物理页面的映射信息。在编程失败的情况下(S130中的编程失败)，在操作S150中，控制器210在映射表L2P中映射出无效数据被编程的物理页面的映射信息。在以原子编程操作为目标的数据的一部分被分发并存储到非易失性存储器设备230中的多个芯片、多个平面、多个块或多个页面，而其余数据未被存储的情况下，控制器210映射出存储该部分数据的物理页面的映射信息。在本发明构思的实施例中，在映射出或映射入与屏障命令相对应的数据之后，或者在完全编程数据之后，控制器210可以将与不同屏障

命令相对应的不同数据编程到非易失性存储器设备230。

[0114] 图17示出了根据本发明构思的实施例的支持对写入顺序的系统调用的主机的软件堆栈的图。图18示出了图17的主机在存储设备上执行日志记录的操作的时序图。将参考图1和图2一起描述图17和图18。存储设备(即屏障兼容的存储设备)400可以是支持屏障命令的图1的存储设备200。

[0115] 图17的主机300可以包括图1的主机100的组件。就硬件而言,主机300可以实施为与图1的主机100基本相同。加载到图1的主机存储器120上的多个软件也可以在图17的主机300中运行。然而,与主机100不同,主机300可以支持系统调用,诸如fbarrier()或fdatababarrier()。屏障文件系统(双模日志记录)322、分派器(即顺序保持分派)323和输入/输出(I/O)调度器(即基于时期的调度器)324可以被加载到主机300的主机存储器(参考图1的主机存储器120)上。

[0116] 屏障文件系统322的操作可以类似于文件系统122的操作。在本发明构思的实施例中,在调用fsync()或fdatasync()的情况下,屏障文件系统322可以通过使用提示信息来确定是否调用fsync()或fdatasync()来按顺序写入数据。屏障文件系统322可以确定文件的扩展名是否是预先确定的单词,文件名是否是预先确定的单词,或者调用fsync()或fdatasync()的进程的名称是否是预先确定的单词。提示信息可以包括预先确定的文件的扩展名、预先确定的文件名或调用fsync()或fdatasync()的进程的名称。

[0117] 在本发明构思的另一实施例中,在调用fbarrier()或fdatababarrier()的情况下,屏障文件系统322可以确定调用fbarrier()或fdatababarrier()来按顺序写入数据。fbarrier()类似于fdatababarrier()。文件元数据可以由fbarrier()修改。然而,当调用fbarrier()时,在没有额外修改用于读取新写入数据的文件元数据的情况下,文件元数据可能不会被修改。屏障文件系统322可以将提交线程分配给主机存储器,以向存储设备400分派写入请求,并且将刷新线程分配给主机存储器,以刷新与写入请求相对应的数据。屏障文件系统322可以生成提交线程和刷新线程以执行双模日志记录。

[0118] 参考图18,屏障文件系统322可以执行日志记录。当备份日志时或修改数据库文件后,可以调用fbarrier()或fdatababarrier()。在fbarrier()开始的情况下,屏障文件系统322可以将写入请求插入(或入队)块层的调度器队列(未示出)(参考图1的调度器队列124),用于将文件数据“D”传送到存储设备400。写入请求可以被分派到存储设备400。与文件系统122不同,屏障文件系统322可以触发提交线程,而无需等待文件数据“D”的DMA传送完成。

[0119] 提交线程可以将写入请求插入到调度器队列中,用于将日志数据JD和日志提交JC传送到存储设备400。写入请求可以被分派到存储设备400。提交线程可以等到日志数据JD的DMA传送和日志提交JC的DMA传送完成。提交线程可以在日志数据JD的DMA传送和日志提交JC的DMA传送完成的情况下触发刷新线程。刷新线程可以将刷新请求插入调度器队列中,以便刷新日志数据JD和日志提交JC。刷新请求可以被分派到存储设备400。在日志数据JD和日志提交JC被完全刷新的情况下,可以返回fbarrier()。

[0120] 参考图18,在调用fsync()的情况下,图1的文件系统122必须等待,直到文件数据“D”的DMA传送、日志数据JD的DMA传送和刷新以及日志提交JC的DMA传送和刷新全部完成。相反,在如图18中调用fbarrier()的情况下,屏障文件系统322可以将IO请求插入调度器

队列,而无需等到文件数据“D”、日志数据JD和日志提交JC的DMA传送和刷新完成。

[0121] 再次返回图17,块层的分派器323可以将输入到调度器队列的IO请求分派到存储设备400的命令队列(参考图3至图13的命令队列213)。例如,分派器323可以向命令队列分派屏障命令。分派器323可以保证以下顺序:1)处理命令队列的现有命令,2)处理屏障命令,以及3)处理屏障命令之后的命令。分派器323可以允许从主机300向存储设备400分派命令的顺序与在存储设备400的命令队列处处理命令的顺序一致。分派器323可以被称为“顺序保持分派器”。

[0122] 在实施例中,从分派器323分派的屏障命令可以是具有屏障标志的写入命令。在另一实施例中,从分派器323分派的屏障命令可以从独立于写入请求的输入请求生成并占据调度器队列的一个条目。

[0123] 在时期的基础上,输入/输出调度器324可以允许IO请求被插入调度器队列中的顺序与命令从主机300被分派到存储设备400的顺序一致。输入/输出调度器324可以保持各时期之间的顺序。输入/输出调度器324可以确定插入到调度器队列中的IO请求是否是屏障写入请求。在插入的IO请求是屏障写入请求的情况下,输入/输出调度器324可以不再接收IO请求。因此,在屏障写入请求之前输入到调度器队列并且在屏障写入请求之后出现在调度器队列中的所有IO请求可以属于一个时期。输入/输出调度器324可以重新排列或合并属于该时期的IO请求。输入/输出调度器324可以将调度器队列中存在的IO请求发送到设备驱动器(参考图1的设备驱动器125)。输入/输出调度器324可以将最终输出IO请求指定为新的屏障写入请求。在调度器队列中存在的所有IO请求出列或被发送的情况下,输入/输出调度器324可以接收新的IO请求。

[0124] 图19示出了根据本发明构思的实施例的存储设备的IOPS和命令队列深度的图。根据本发明构思的实施例的存储设备可以包括参考图1至图18描述的存储设备200或400。在图19中,“XnF”表示传送和刷新方案(参考图2),“X”表示等待传送方案,“OP”表示根据本发明构思的实施例的顺序保持方案(参考图18)。在与三星Galaxy 6中通用闪存2.0的实施方式相对应的UFS 2.0(GS6)、与三星SSD 850 Pro相对应的SSD(850 PRO)和与三星SSD 843TN相对应的SSD(843TN)中的每一个都可以从三星电子有限公司获得的情况下,以XnF方案、“X”方案和OP方案处理IO请求,每秒输入/输出操作(IOPS)和命令队列深度如图19所示。在根据本发明构思的实施例的OP方案的情况下,主机100/300可以根据IO请求向存储设备100/200传送命令,而无需等到数据的DMA传送和数据的刷新完成。因此,与XnF方案或“X”方案相比,存储设备100/200的IOPS和队列深度可以通过OP方案增加。

[0125] 根据本发明构思的实施例,存储设备可以支持用于保持写入顺序的屏障命令。与存储设备通信的主机可以按顺序向存储设备提供写入请求,而无需等到存储设备完成按顺序生成的写入请求中的每一个。

[0126] 虽然已经参考本发明的示例性实施例描述了本发明的构思,但是对于本领域的普通技术人员显而易见的是,在不脱离如以下权利要求所述的本发明构思的精神和范围的情况下,可以对其进行各种改变和修改。

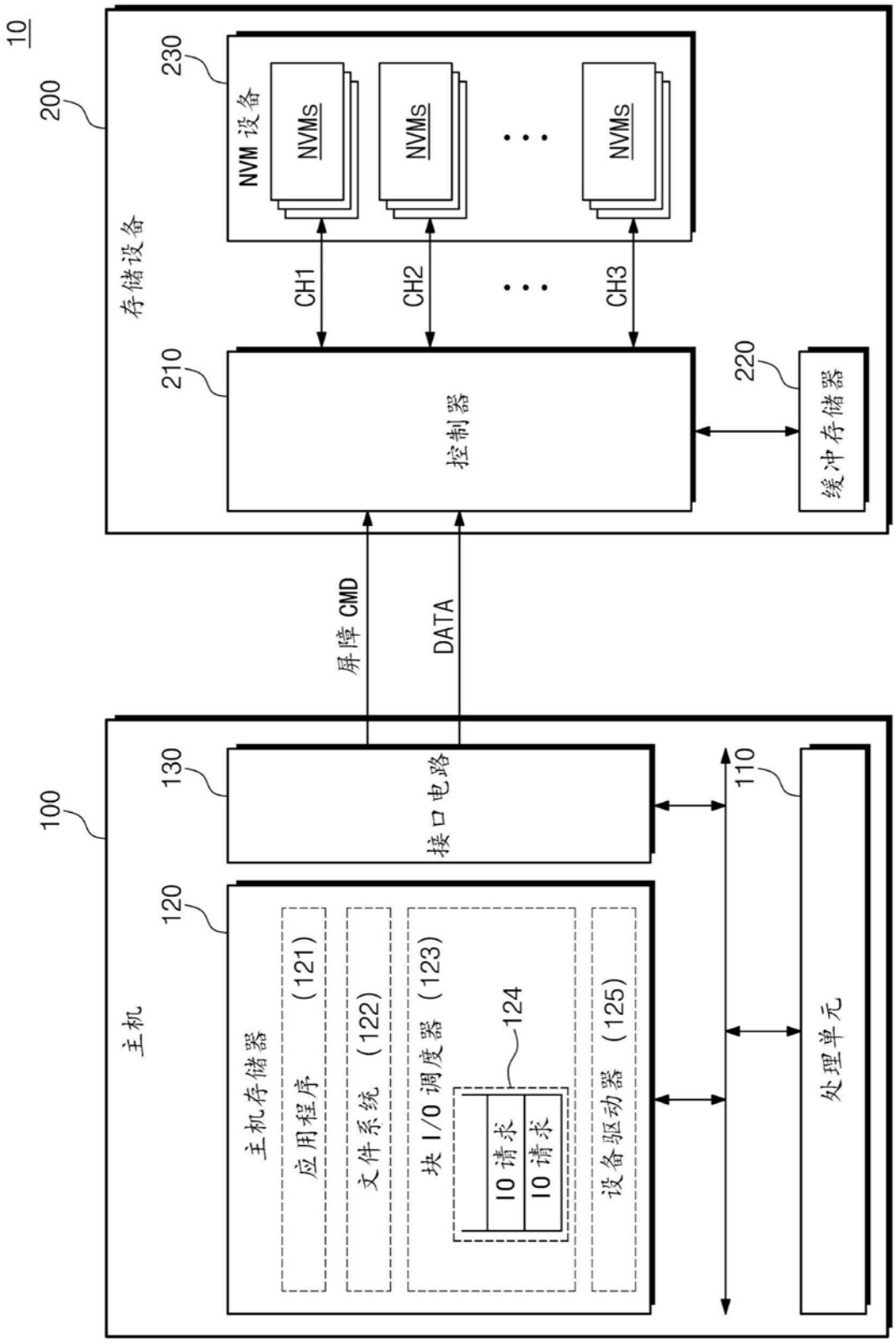

10
200
存储设备
230
NVM 设备
NVMs
NVMs
NVMs
CH1
CH2
CH3
210
控制器
220
缓冲存储器
屏障 CMD
DATA
100
主机
130
接口电路
110
120
主机存储器
应用程序 (121)
文件系统 (122)
块 I/O 调度器 (123)
124
IO 请求
IO 请求
设备驱动器 (125)
处理单元

图1

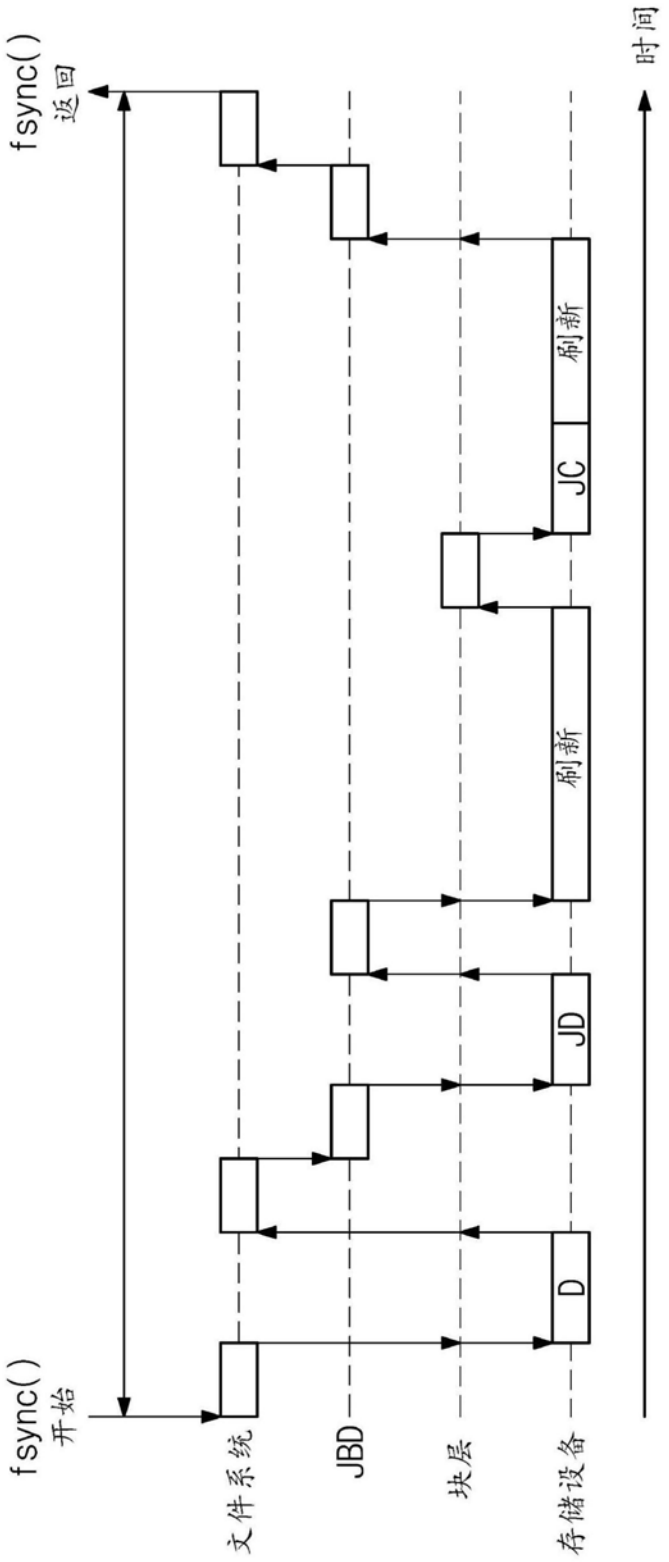


图2

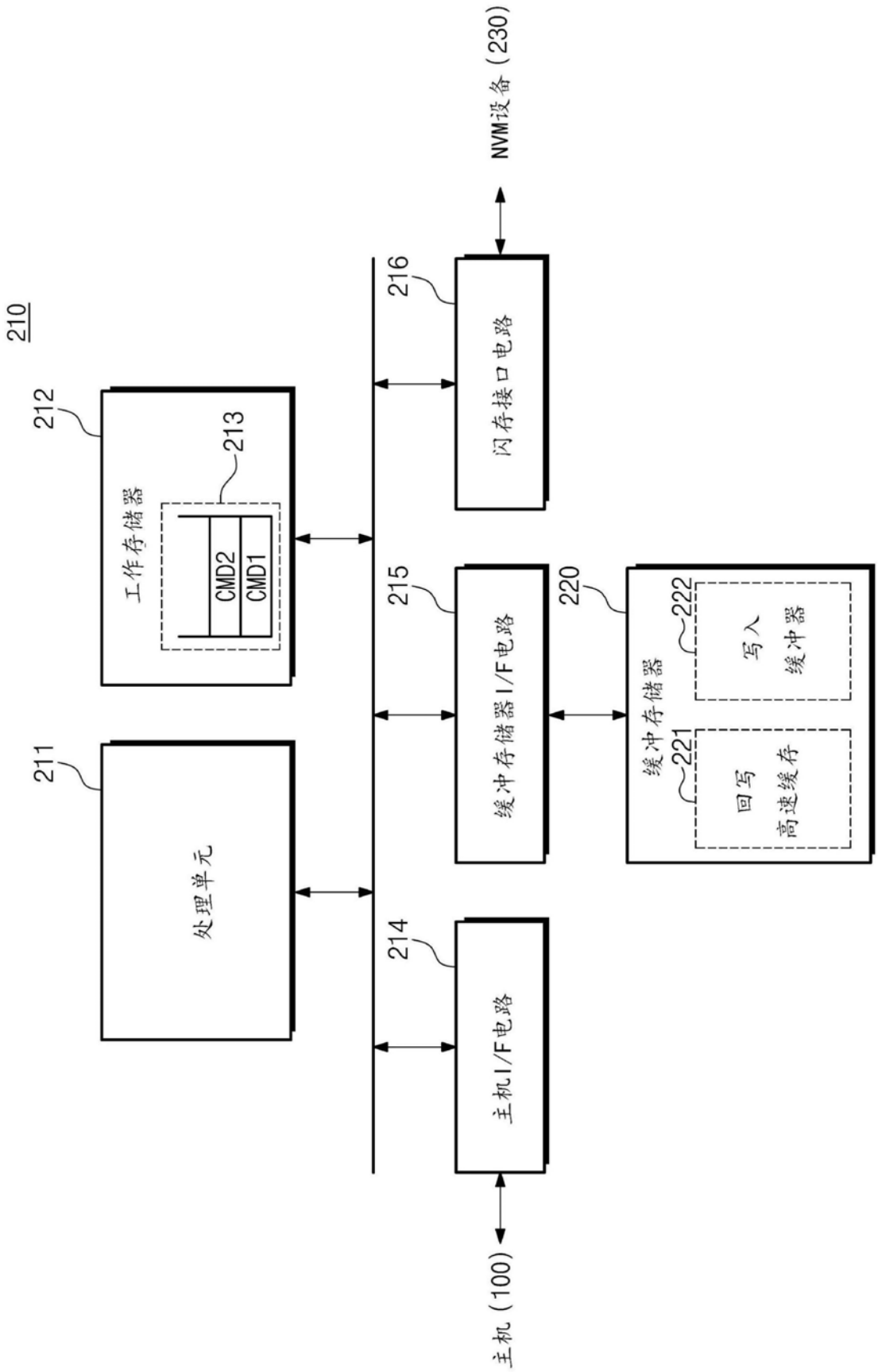

NVM设备（230）
210
216
闪存接口电路
212
213
工作存储器
CMD2
CMD1
215
缓冲存储器I/F电路
220
缓冲存储器
222
写入
缓冲器
221
回写
高速缓存
211
处理单元
214
主机I/F电路
主机（100）

图3

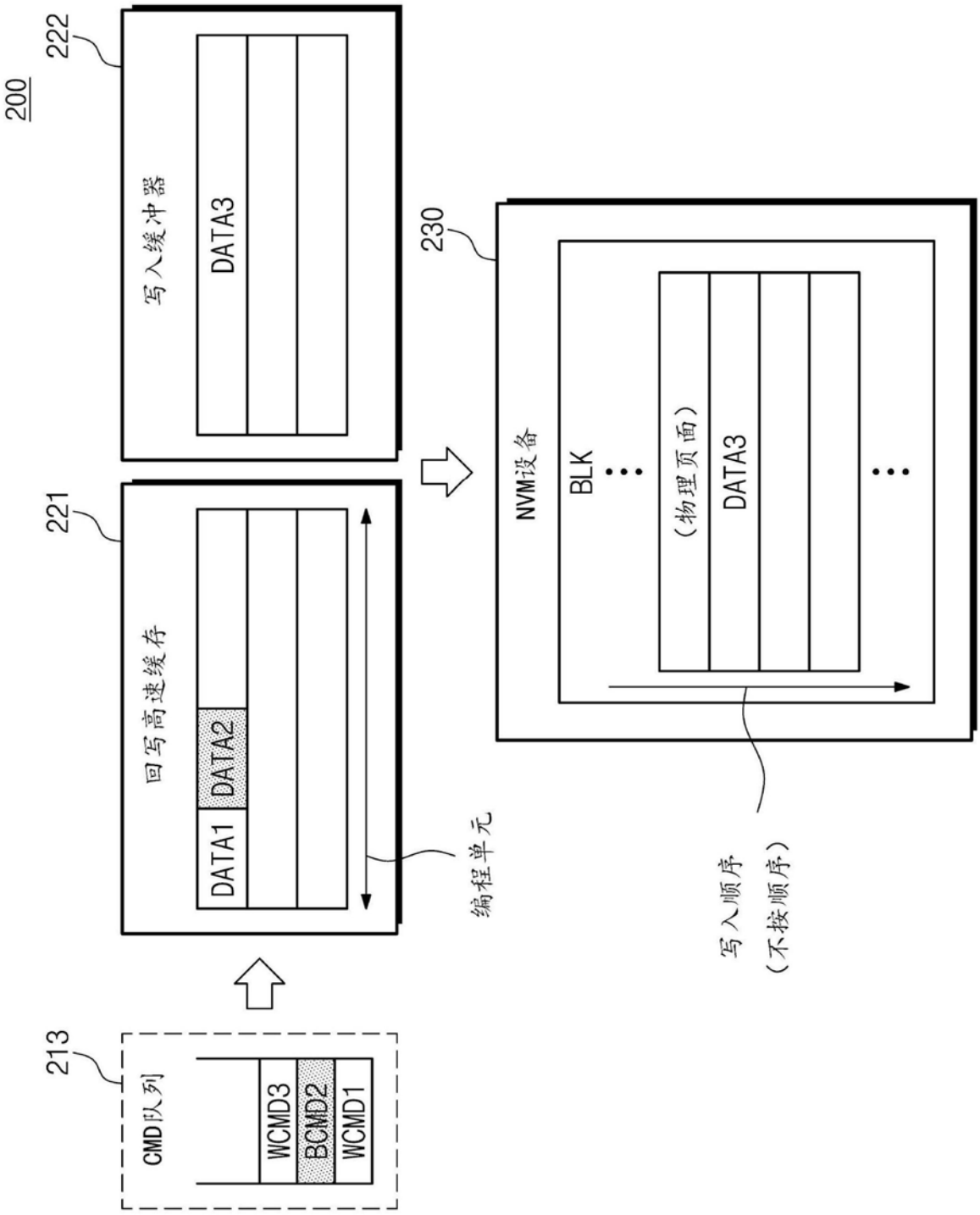

200
222
写入缓冲器
DATA3
221
回写高速缓存
DATA1
DATA2
213
CMD队列
WCMD3
BCMD2
WCMD1
编程单元
230
NVM设备
BLK
(物理页面)
DATA3
写入顺序
(不按顺序)

图4

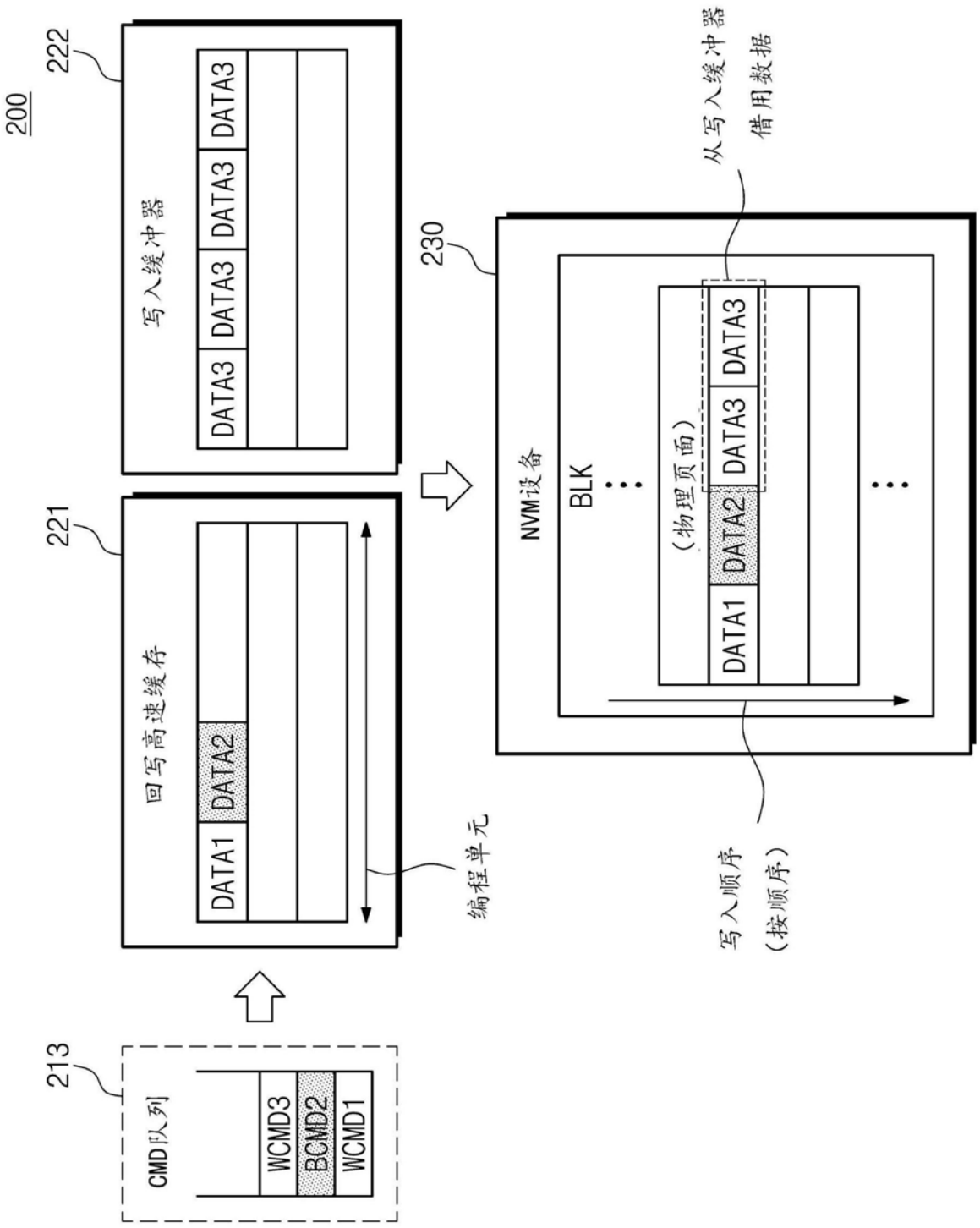

200
222
写入缓冲器
DATA3
DATA3
DATA3
DATA3
221
回写高速缓存
DATA1
DATA2
编程单元
213
CMD队列
WCMD3
BCMD2
WCMD1
230
NVM设备
BLK
(物理页面)
DATA1
DATA2
DATA3
DATA3
从写入缓冲器
借用数据
写入顺序
(按顺序)

图5

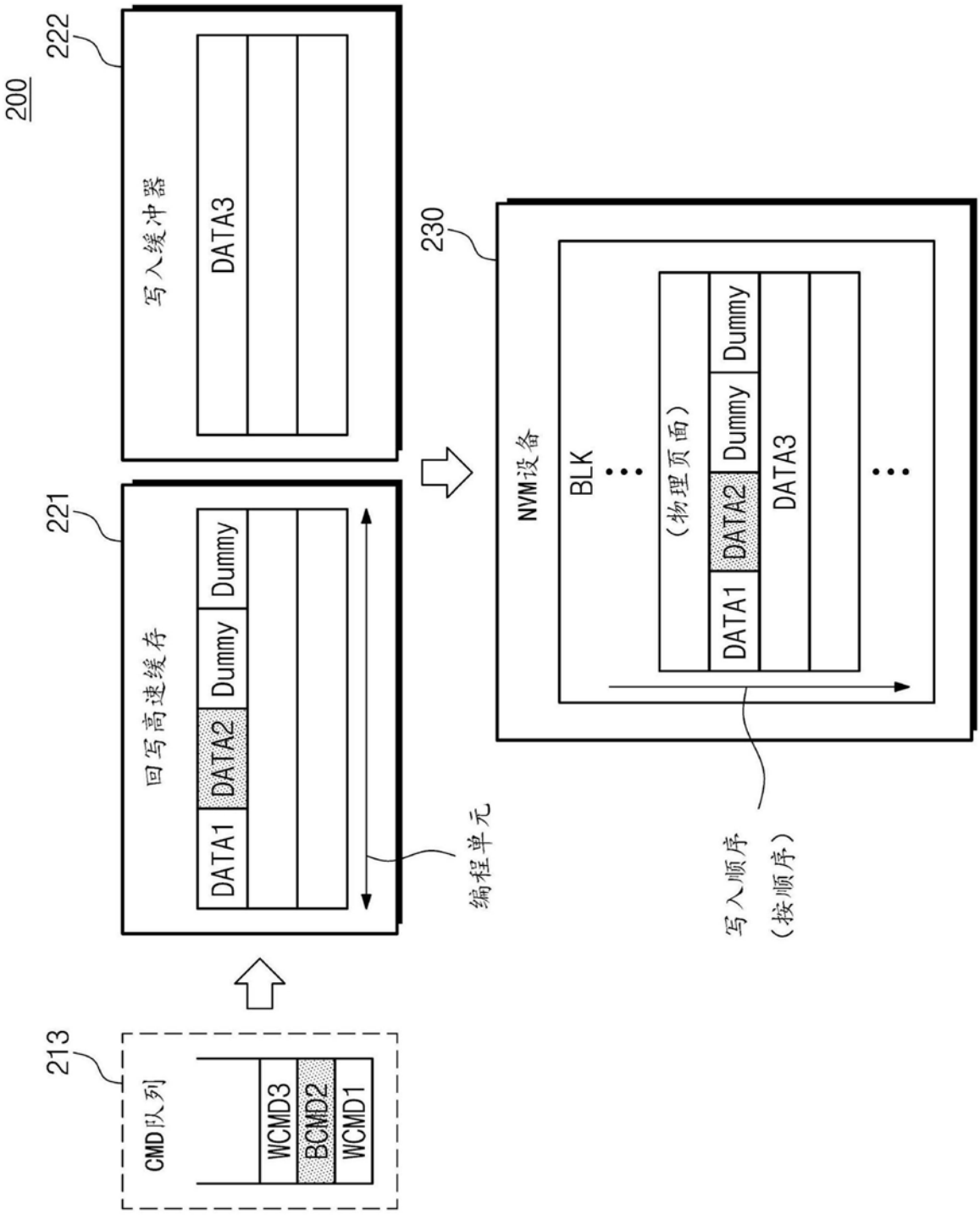

200
222
写入缓冲器
DATA3
230
NVM设备
BLK
(物理页面)
DATA1
DATA2
Dummy
Dummy
DATA3
221
回写高速缓存
DATA1
DATA2
Dummy
Dummy
编程单元
写入顺序
(按顺序)
213
CMD队列
WCMD3
BCMD2
WCMD1

图6

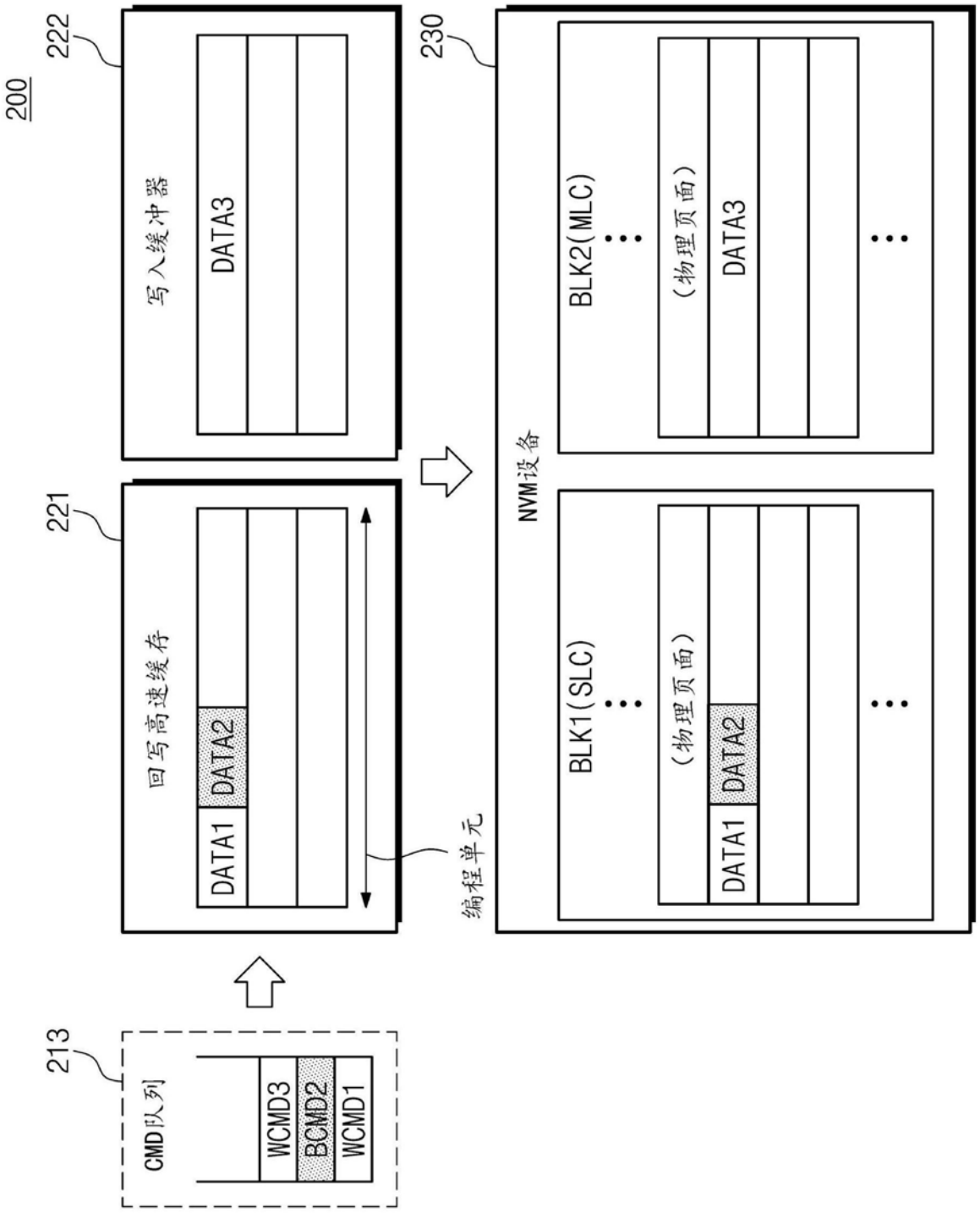

200
222
写入缓冲器
DATA3
221
回写高速缓存
DATA1
DATA2
编程单元
230
NVM设备
BLK1(SLC)
(物理页面)
DATA1
DATA2
BLK2(MLC)
(物理页面)
DATA3
213
CMD队列
WCMD3
BCMD2
WCMD1

图7

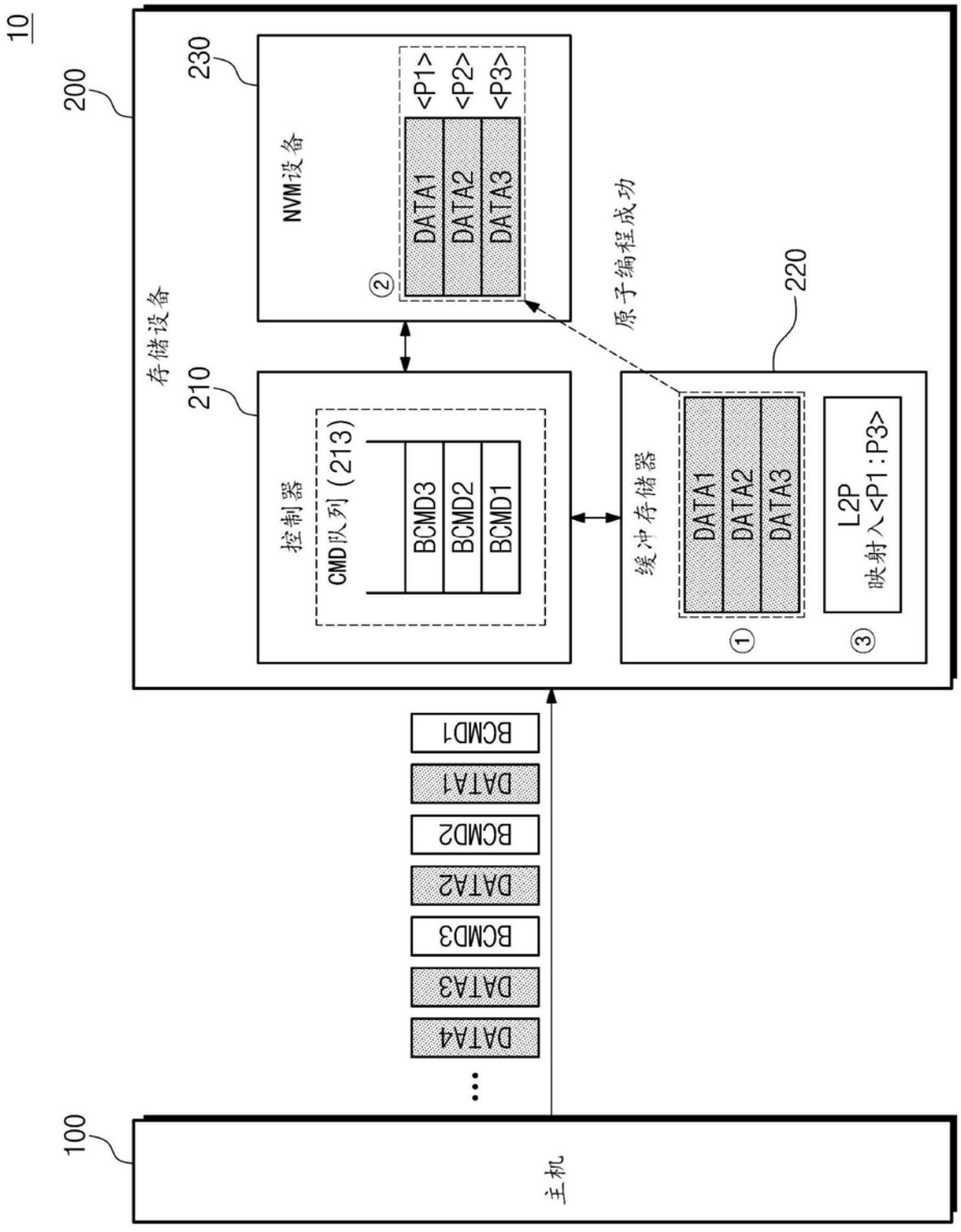

10
200
230
NVM设备
<P1>
<P2>
<P3>
DATA1
DATA2
DATA3
②
原子编程成功
220
存储设备
210
控制器
CMD队列 (213)
BCMD3
BCMD2
BCMD1
缓冲存储器
DATA1
DATA2
DATA3
①
L2P
映射入<P1:P3>
③
BCMD1
DATA1
BCMD2
DATA2
BCMD3
DATA3
DATA4
100
主机

图8

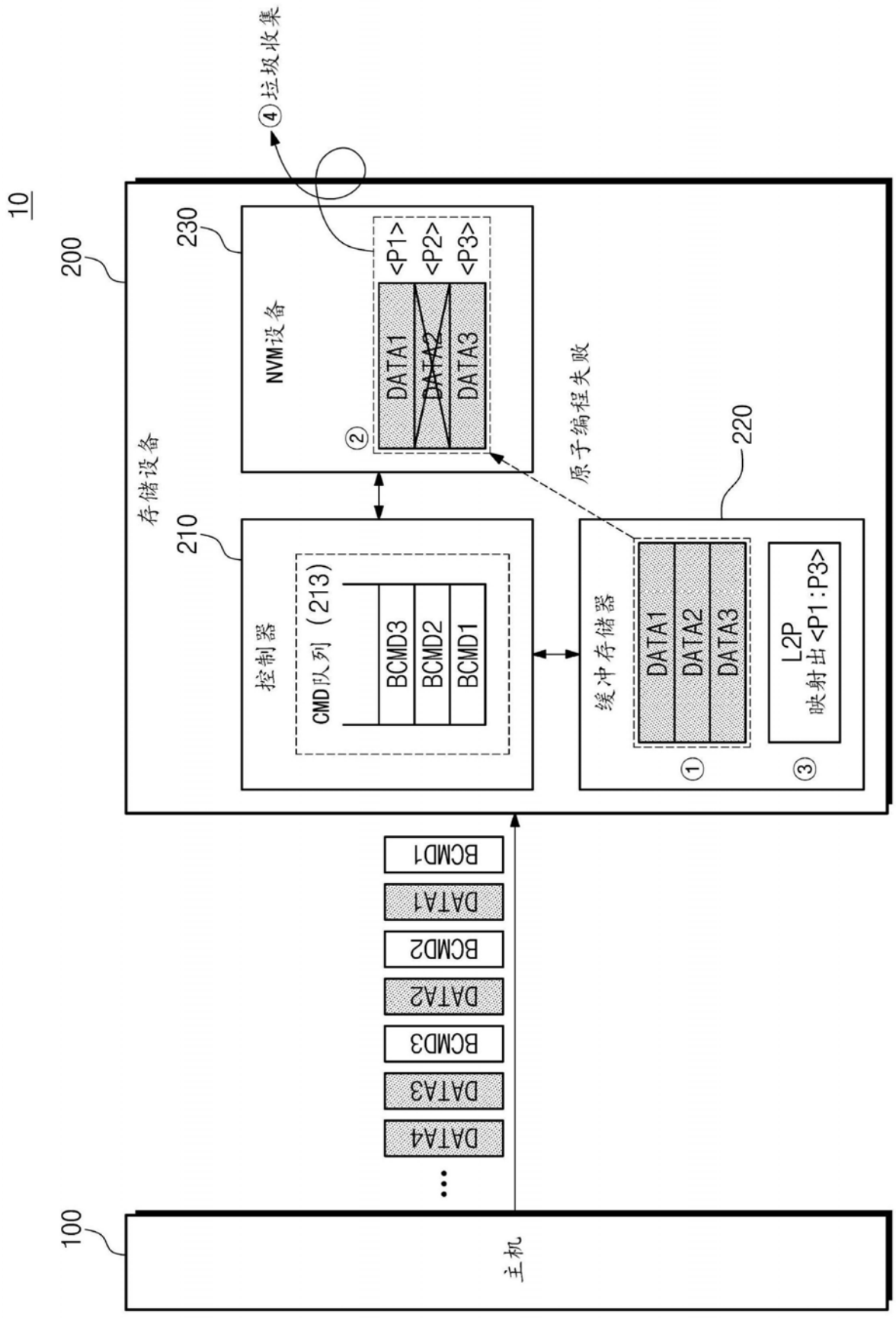

10
200
230
NVM设备
<P1>
<P2>
<P3>
DATA1
DATA2
DATA3
②
④垃圾收集
存储设备
210
控制器
CMD队列（213）
BCMD3
BCMD2
BCMD1
原子编程失败
220
缓冲存储器
DATA1
DATA2
DATA3
①
L2P
映射出<P1:P3>
③
BCMD1
DATA1
BCMD2
DATA2
BCMD3
DATA3
DATA4
100
主机

图9

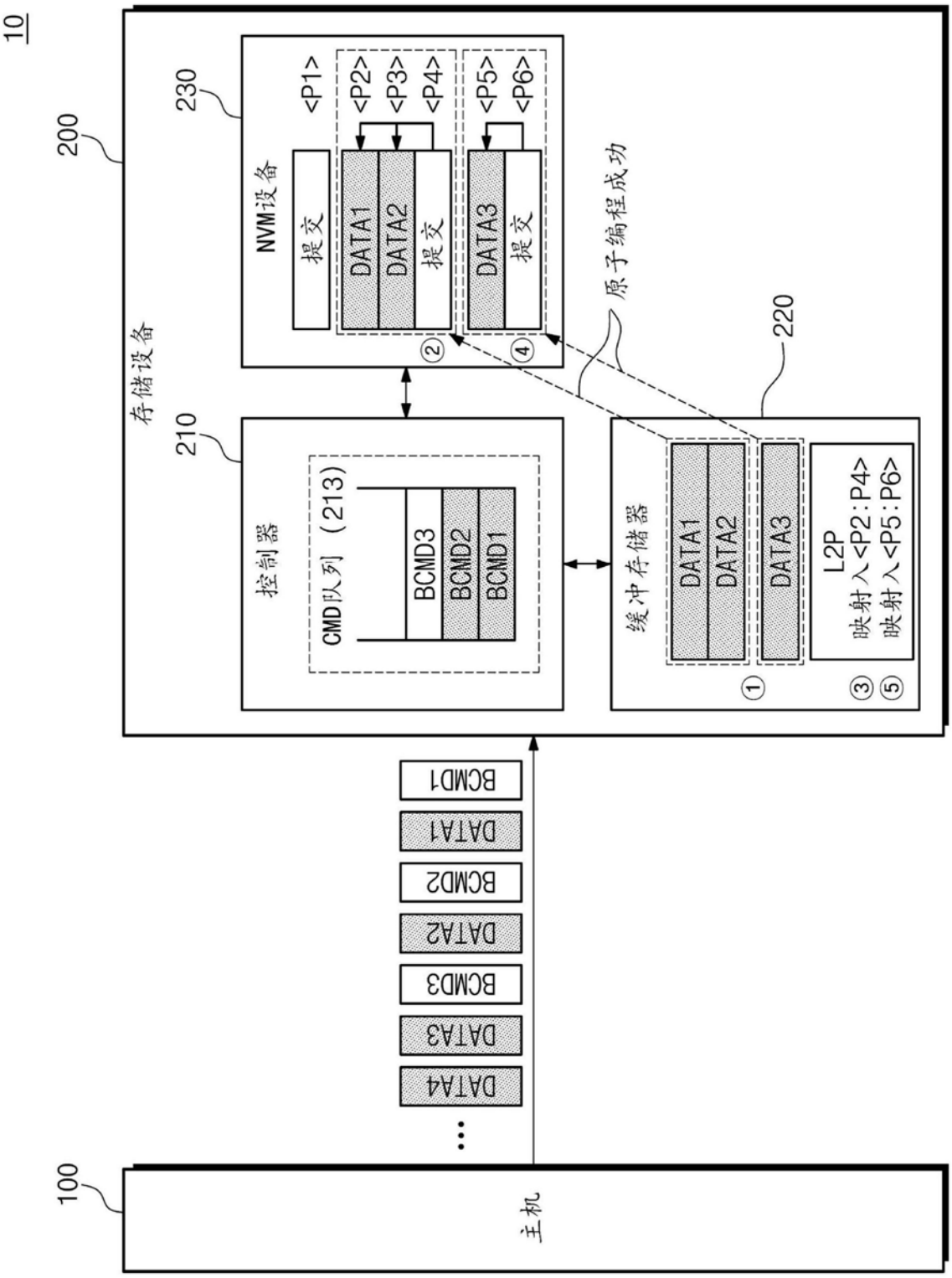

10
200
230
NVM设备
<P1>
<P2>
<P3>
<P4>
<P5>
<P6>
提交
DATA1
DATA2
提交
DATA3
提交
原子编程成功
220
存储设备
210
控制器
CMD队列 (213)
BCMD3
BCMD2
BCMD1
缓冲存储器
DATA1
DATA2
DATA3
L2P
映射入<P2:P4>
映射入<P5:P6>
①
②
③
④
⑤
BCMD1
DATA1
BCMD2
DATA2
BCMD3
DATA3
DATA4
100
主机

图10

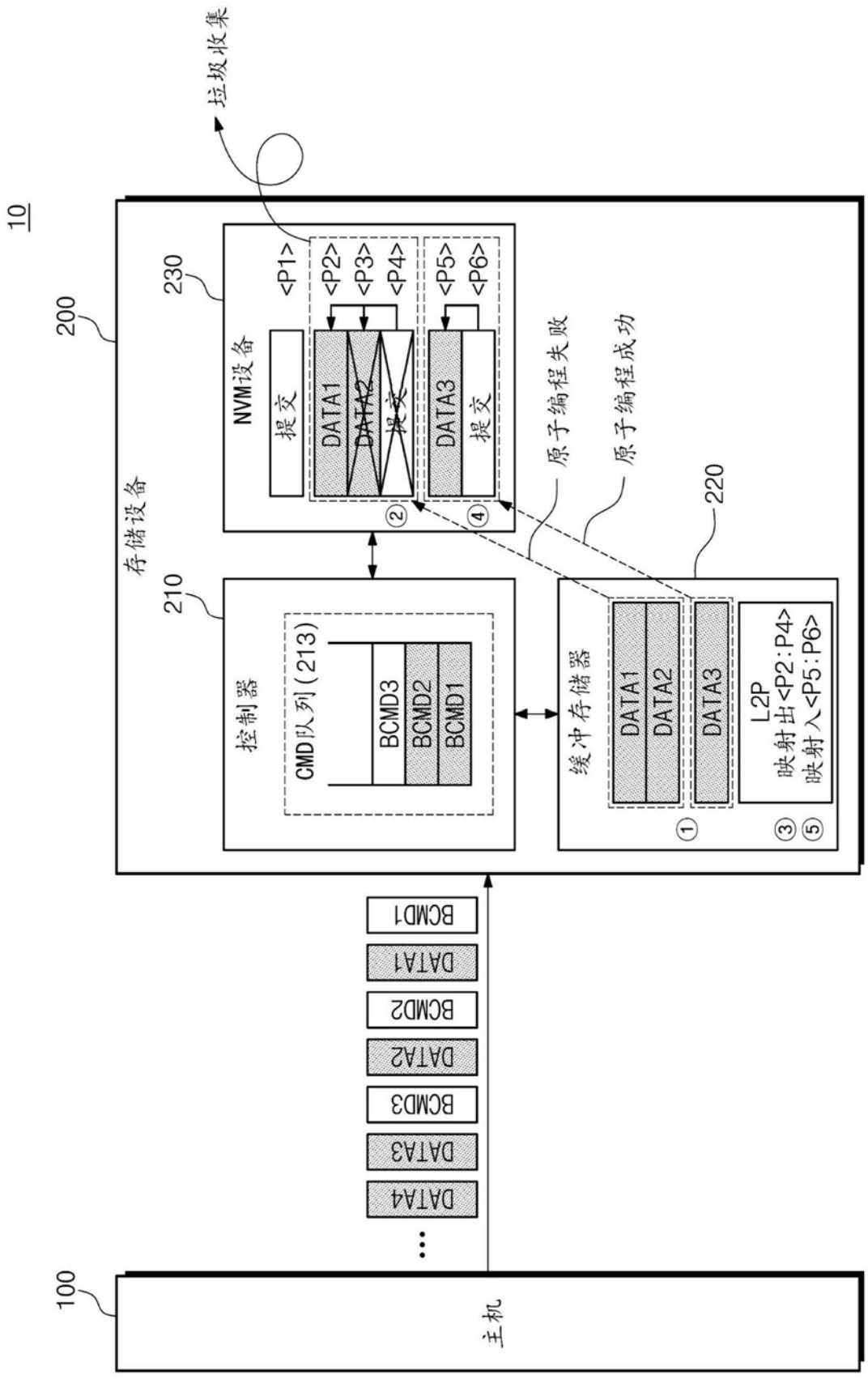

10
垃圾收集
200
230
NVM设备
<P1>
<P2>
<P3>
<P4>
<P5>
<P6>
提交
DATA1
DATA2
提交
DATA3
提交
②
④
原子编程失败
原子编程成功
220
存储设备
210
控制器
CMD队列(213)
BCMD3
BCMD2
BCMD1
缓冲存储器
DATA1
DATA2
DATA3
L2P
映射出<P2:P4>
映射入<P5:P6>
①
③
⑤
BCMD1
DATA1
BCMD2
DATA2
BCMD3
DATA3
DATA4
...
100
主机

图11

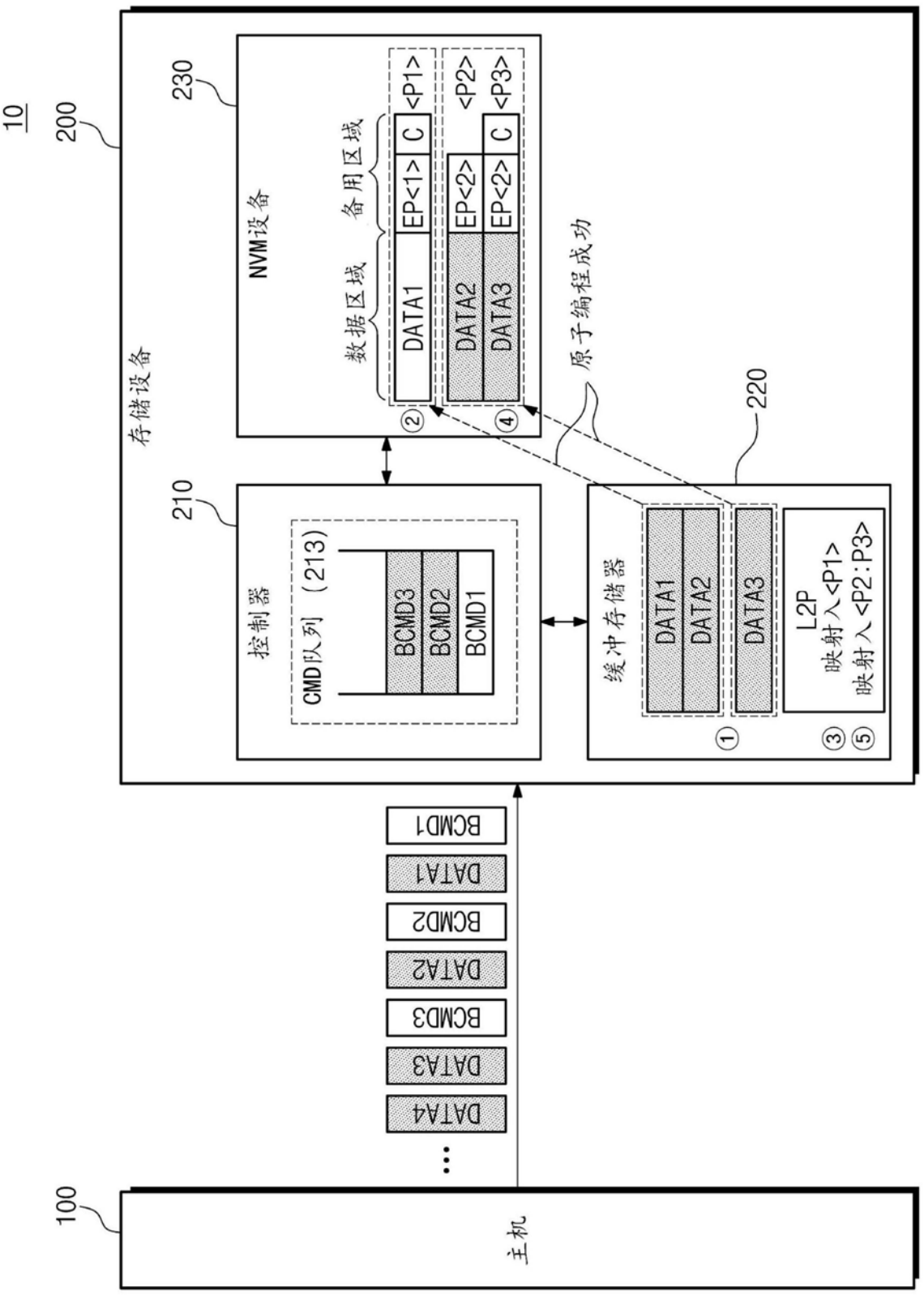

10
200
230
NVM设备
数据区域
备用区域
DATA1
EP<1>
C
<P1>
DATA2
EP<2>
<P2>
DATA3
EP<2>
C
<P3>
②
④
原子编程成功
存储设备
210
220
控制器
CMD队列 (213)
BCMD3
BCMD2
BCMD1
缓冲存储器
DATA1
DATA2
DATA3
L2P
映射入<P1>
映射入<P2:P3>
①
③
⑤
BCMD1
DATA1
BCMD2
DATA2
BCMD3
DATA3
DATA4
100
主机

图12

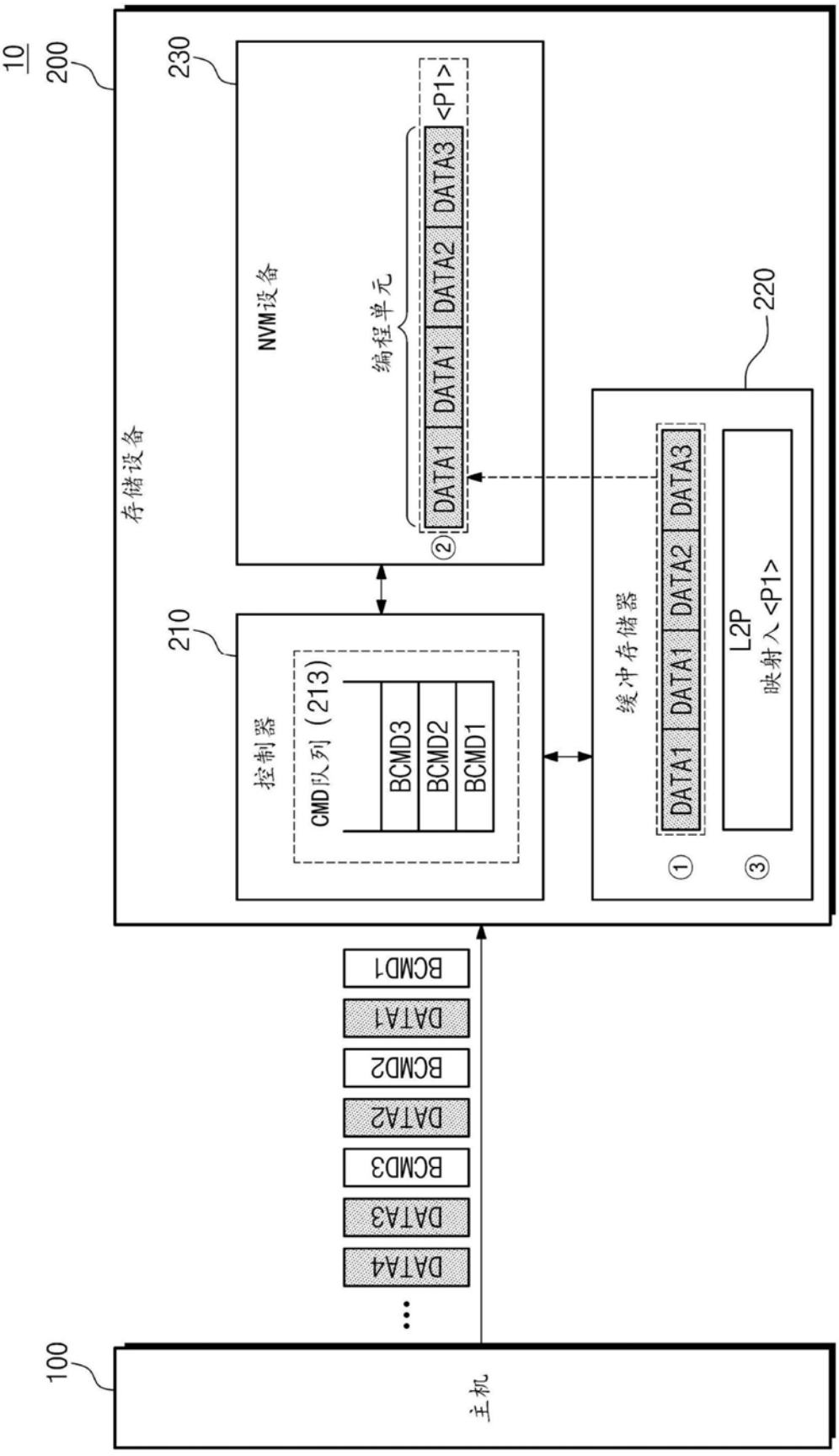

10
200
230
NVM设备
编程单元
DATA1
DATA1
DATA2
DATA3
<P1>
②
220
存储设备
210
控制器
CMD队列 (213)
BCMD3
BCMD2
BCMD1
缓冲存储器
DATA1
DATA1
DATA2
DATA3
①
③
L2P
映射入 <P1>
BCMD1
DATA1
BCMD2
DATA2
BCMD3
DATA3
DATA4
100
主机

图13

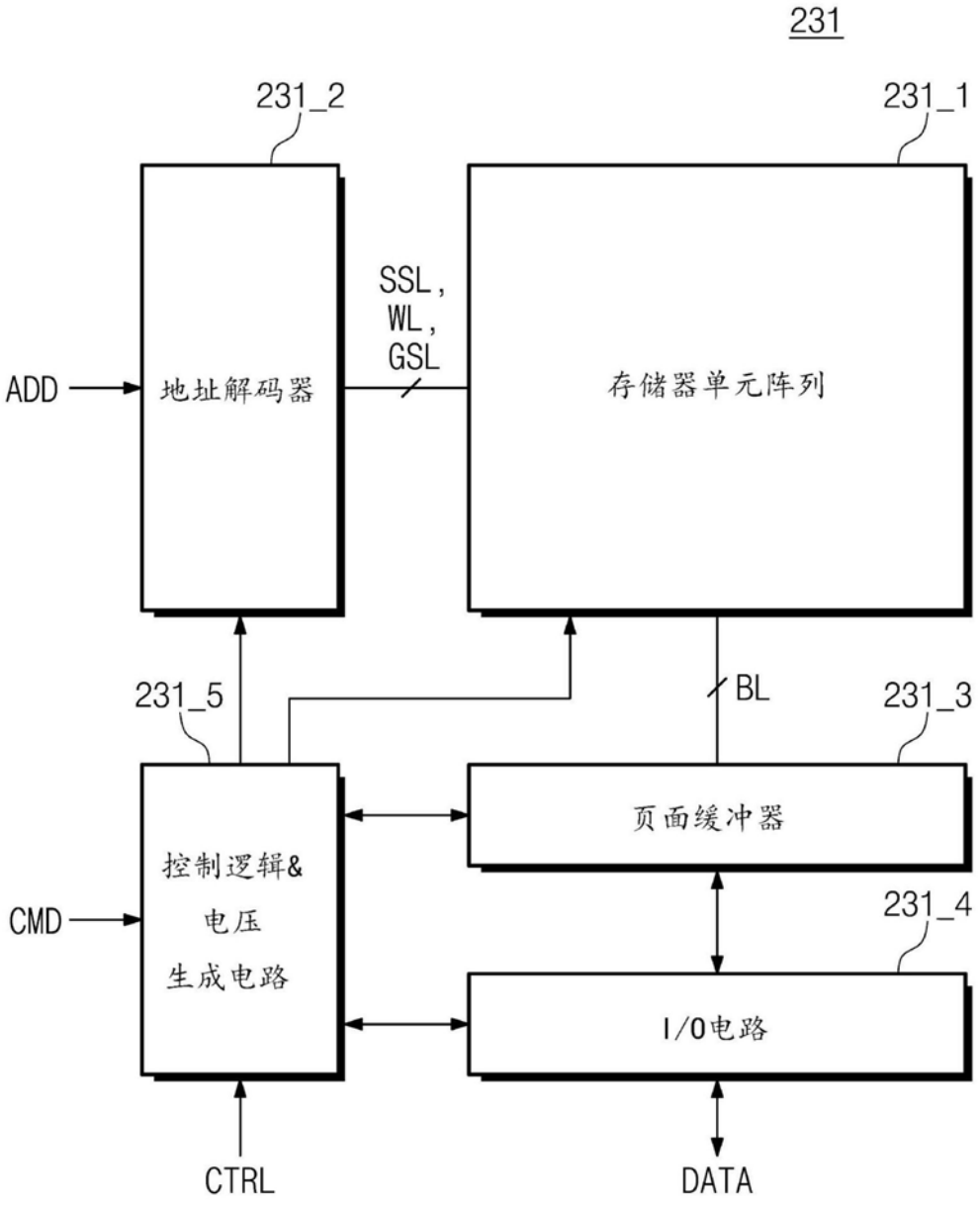

231
231_2
231_1
SSL,
WL,
GSL
ADD
地址解码器
存储器单元阵列
231_5
BL
231_3
页面缓冲器
控制逻辑&
电压
生成电路
CMD
231_4
I/O电路
CTRL
DATA

图14

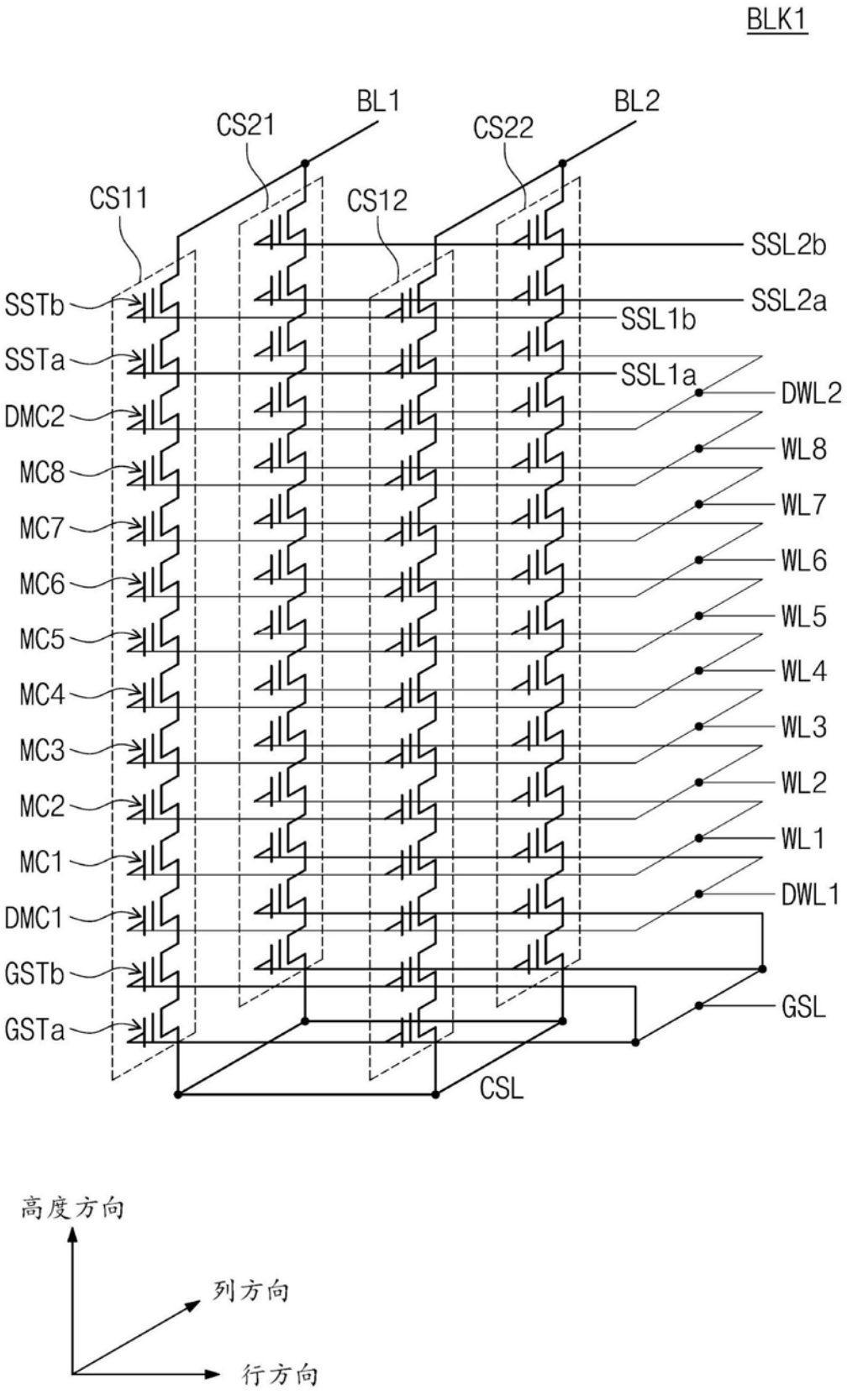

BLK1
BL1
BL2
CS21
CS22
CS11
CS12
SSL2b
SSL2a
SSL1b
SSL1a
DWL2
WL8
WL7
WL6
WL5
WL4
WL3
WL2
WL1
DWL1
GSL
SSTb
SSTa
DMC2
MC8
MC7
MC6
MC5
MC4
MC3
MC2
MC1
DMC1
GSTb
GSTa
CSL
高度方向
列方向
行方向

图15

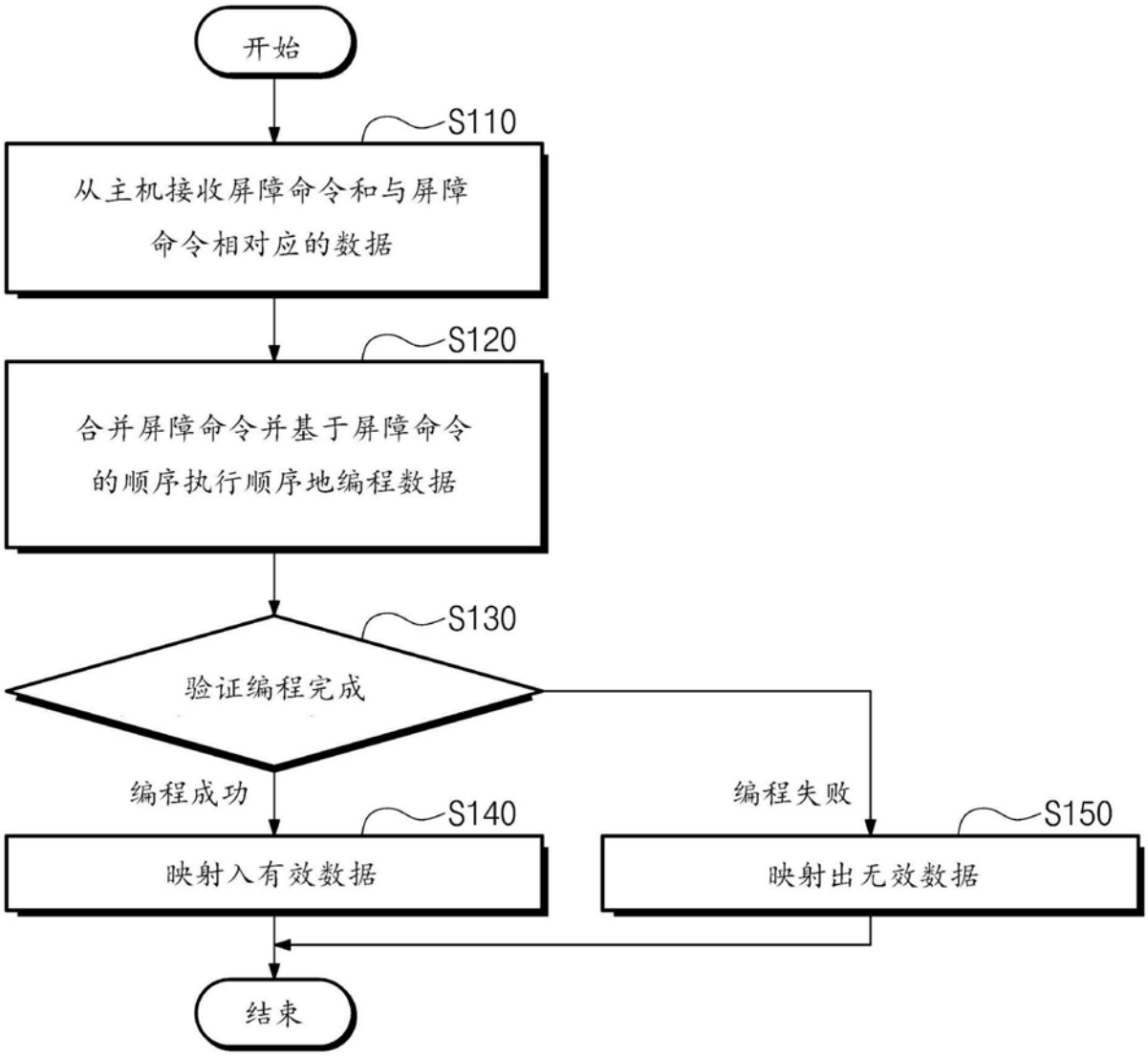

开始
S110
从主机接收屏障命令和与屏障命令相对应的数据
S120
合并屏障命令并基于屏障命令的顺序执行顺序地编程数据
S130
验证编程完成
编程成功
编程失败
S140
映射入有效数据
S150
映射出无效数据
结束

图16

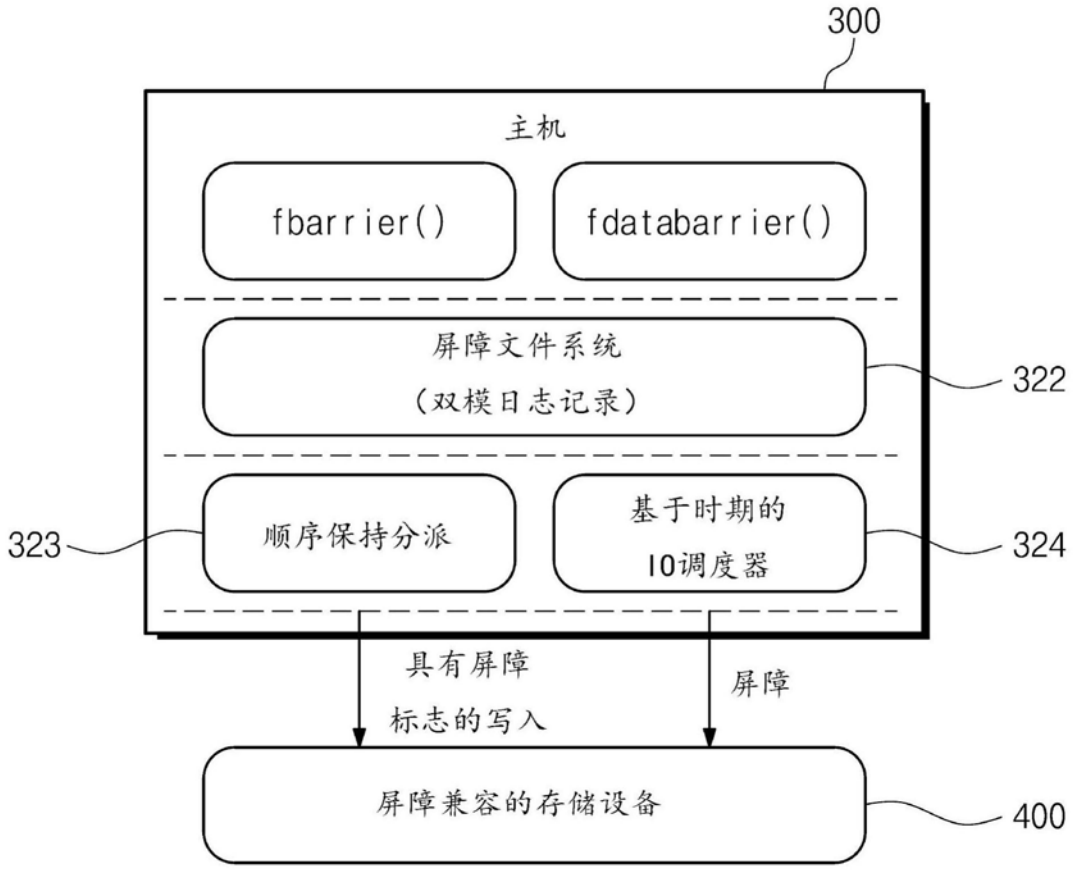

300
主机
fbarrier()
fdatabarrier()
屏障文件系统
（双模日志记录）
322
323
顺序保持分派
基于时期的
IO调度器
324
具有屏障
标志的写入
屏障
屏障兼容的存储设备
400

图17

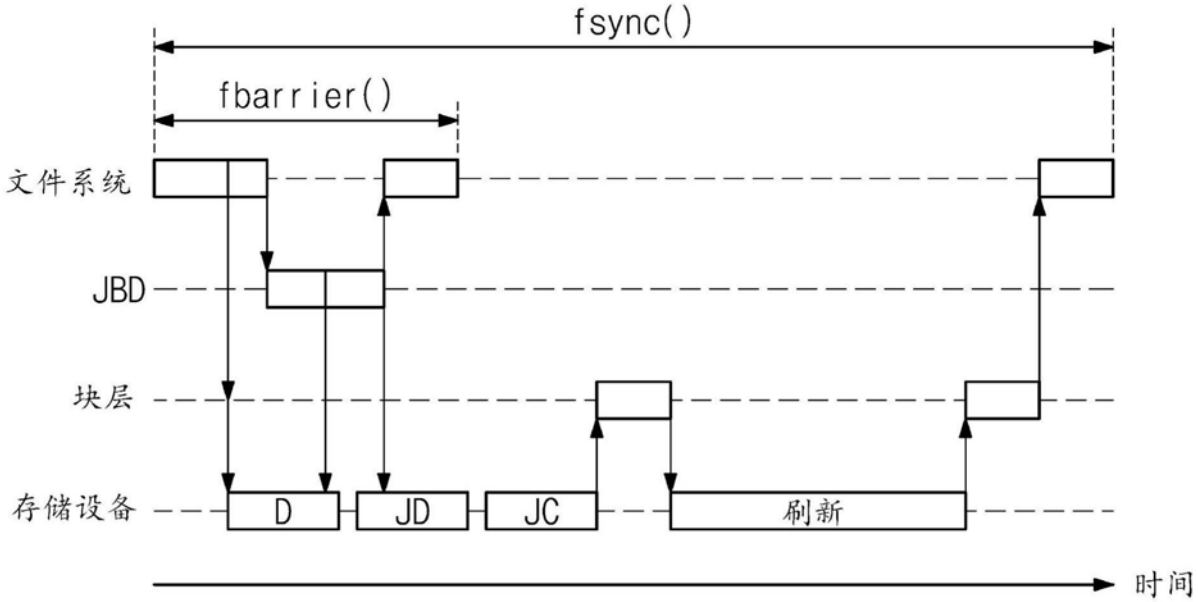

fsync()
fbarrier()
文件系统
JBD
块层
存储设备
D
JD
JC
刷新
时间

图18

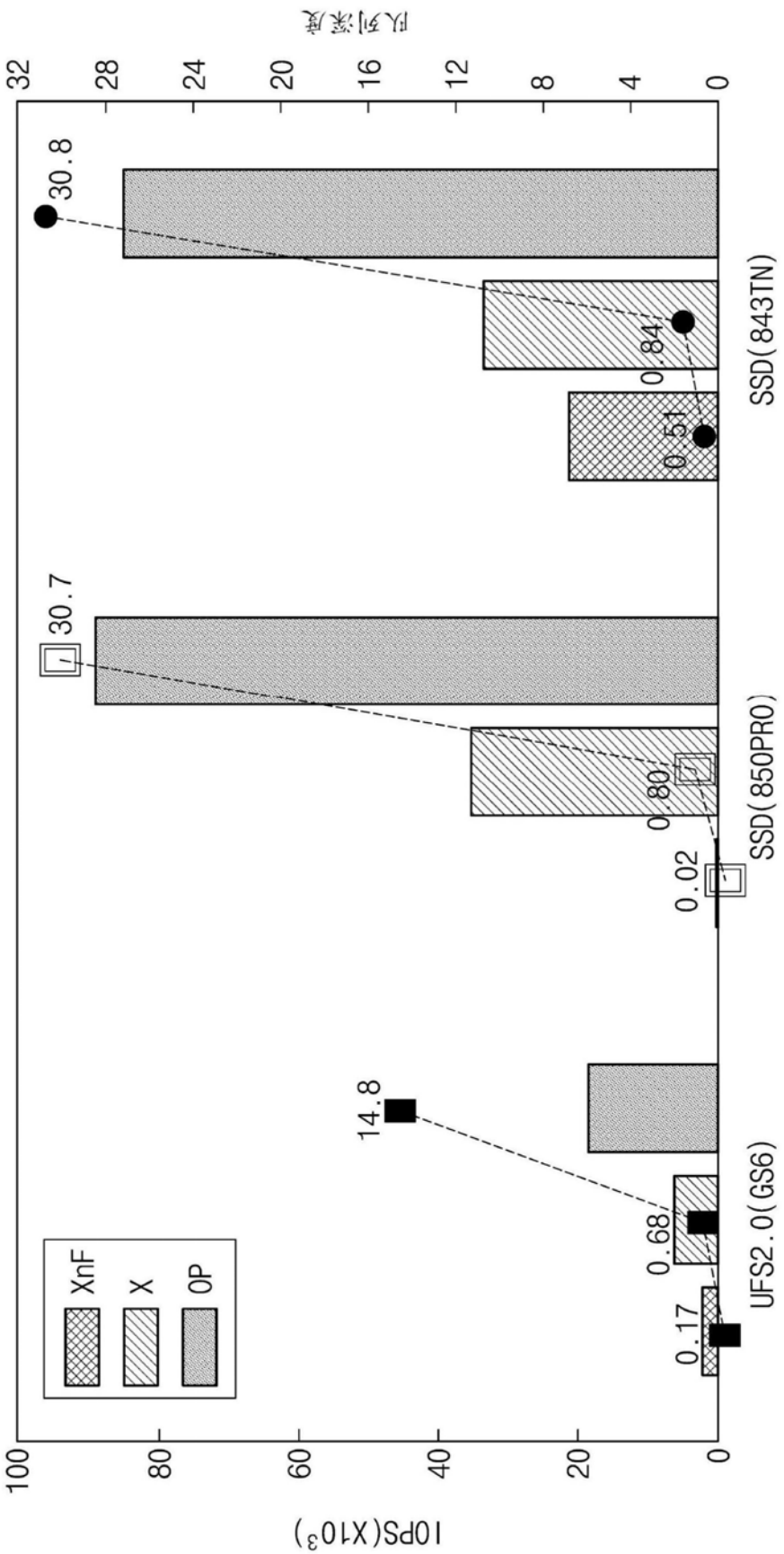

队列深度
32
28
24
20
16
12
8
4
0
30.8
0.84
0.51
SSD(843TN)
30.7
0.80
0.02
SSD(850PRO)
14.8
0.68
0.17
UFS2.0(GS6)
XnF
X
0P
100
80
60
40
20
0
IOPS(X10³)

图19