



(12)发明专利

(10)授权公告号 CN 104471523 B

(45)授权公告日 2017.04.05

(21)申请号 201280072911.4

(22)申请日 2012.05.18

(65)同一申请的已公布的文献号
申请公布号 CN 104471523 A

(43)申请公布日 2015.03.25

(85)PCT国际申请进入国家阶段日
2014.10.31

(86)PCT国际申请的申请数据
PCT/JP2012/003289 2012.05.18

(87)PCT国际申请的公布数据
W02013/171809 EN 2013.11.21

(73)专利权人 株式会社日立制作所
地址 日本东京都

(72)发明人 工藤晋太郎 野中裕介

(74)专利代理机构 北京市金杜律师事务所
11256

代理人 王茂华 辛鸣

(51)Int.Cl.
G06F 3/06(2006.01)

(56)对比文件
US 6182177 B1,2001.01.30,
US 8090832 B1,2012.01.03,
CN 1285036 C,2006.11.15,
审查员 周静奇

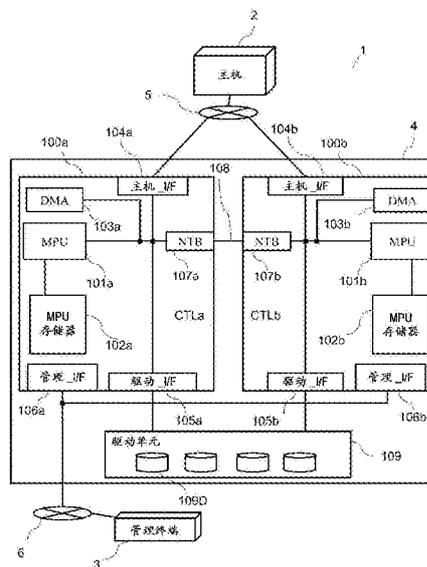
权利要求书3页 说明书17页 附图27页

(54)发明名称

计算机系统及其控制方法

(57)摘要

本发明是一种群集式存储系统,利用该群集式存储系统,即使在从一个控制器的处理器发送对另一控制器的处理器的访问时,第二控制器的处理器也能够使对这一访问的处理优先,从而使得防止I/O处理被延迟。利用本发明的存储系统,第一控制器的第一处理器通过在针对其的处理将被第二控制器的第二处理器优先的请求信息和针对其的处理将不被优先的请求信息之间区分来向第二处理器传输将由第二处理器处理的请求信息,并且第二处理器通过在针对其的处理将被优先的请求信息与针对其的信息将不被优先的请求信息之间区分来获取请求信息。



1. 一种控制在主机与存储区域之间的数据传送的存储系统,包括:

第一控制器;

第二控制器;

所述第一控制器的第一处理器和第一存储器;

所述第二控制器的第二处理器和第二存储器;以及

连接所述第一处理器和所述第二处理器的总线,

其中被存储在所述第一存储器中的共享区域和所述第二存储器中的共享区域中的数据将由所述第一处理器和所述第二处理器使用,

其中所述第一处理器向所述第二处理器传输将由所述第二处理器处理的请求信息,

其中所述第二处理器处理所述请求信息并且向所述第一处理器发送回响应,

其中所述第一处理器确定用于访问所述第二存储器中的所述共享区域的所述请求信息将被优先并且向所述第二处理器传输指示所述请求信息的优先级的所述请求信息,并且在处理下一过程之前等待来自所述第二处理器的响应,

其中所述第一处理器确定用于请求数据传输的所述请求信息将不被优先并且向所述第二处理器发送不指示所述请求信息的优先级的所述请求信息,并且处理所述下一过程而不等待来自所述第二处理器的所述响应,并且

其中所述第二处理器通过在针对其的处理将被优先的请求信息与针对其的处理将不被优先的请求信息之间区分,来获取所述请求信息,并且使具有优先级的所述请求信息的处理优先于不具有所述优先级的所述请求信息。

2. 根据权利要求1所述的存储系统,

其中,在确定所述请求信息是针对其的处理将被优先的信息时,所述第一处理器向所述第二存储器的第一区域提供所述请求信息,

其中,在确定所述请求信息是针对其的处理无需被优先的信息时,所述第一处理器向所述第二存储器的第二区域提供所述请求信息,并且

其中所述第二处理器用比所述第二区域更高的优先级处理所述第一区域。

3. 根据权利要求2所述的存储系统,

其中所述第一区域是与用于对来自所述主机的I/O的处理的非中断处理有关的信息被记录在其中的第一记录单元,

其中所述第二区域是与用于对来自所述主机的I/O的处理的非中断处理有关的信息被记录在其中的第二记录单元,

其中,在从所述第一处理器接收到针对其的处理将被优先的请求信息已经被传输到所述第一记录单元的通知时,所述第二处理器使对所述请求信息的所述处理优先于对来自所述主机的I/O的所述处理。

4. 根据权利要求1所述的存储系统,

其中针对其的处理将不被优先的所述请求信息是用于出于数据传送的目的而请求启动所述第二控制器的硬件资源的信息。

5. 根据权利要求1所述的存储系统,

其中,在从所述主机接收到I/O时,所述第一处理器确定被存储在所述第二存储器的所述共享区域中的所述数据已经被更新,并且在确定所述第二存储器中的所述共享区域中的

所述数据已经被更新时,所述第一处理器向所述第二处理器传输用于访问所述第二存储器中的所述共享区域的所述请求信息。

6. 根据权利要求3所述的存储系统,

其中,在预定数目或者更多的请求信息项被记录于所述第一区域中时,所述第一处理器向所述第二处理器传输中断请求,并且其中,在接收到所述中断请求时,所述第二处理器集体地处理在所述第一区域中记录的所述预定数目或者更多的请求信息。

7. 根据权利要求2所述的存储系统,

其中所述第二处理器由多个核配置,

其中所述第一区域是所述多个核之中的一个或者多个专用核,

其中所述第二区域是所述多个核之中的、除了所述专用核之外的普通核,

其中所述专用核使对所述请求信息的所述处理优先于对来自所述主机的I/O的所述处理,

其中所述普通核与对来自所述主机的I/O的处理一起执行对所述请求信息的处理,并且

其中所述第一处理器通过使所述专用核优先于所述普通核来向所述专用核传输针对其的处理将被优先的所述请求信息。

8. 根据权利要求7所述的存储系统,

其中所述第二处理器根据所述多个核的负荷状态来改变所述多个核之中的所述专用核和所述普通核的比率。

9. 根据权利要求7所述的存储系统,

其中,如果所述专用核上的负荷高,则所述第一处理器使得所述普通核将对所述请求信息的所述处理优先于对来自所述主机的I/O的所述处理。

10. 根据权利要求7所述的存储系统,

其中多个虚拟核能被配置于所述第二处理器中,并且所述多个虚拟核中的一个或者多个虚拟核借助来自所述第一处理器的请求而被从所述普通核改变成所述专用核。

11. 根据权利要求1所述的存储系统,

其中所述第一处理器根据所述第二控制器的访问目的地硬件和访问类型来确定所述请求信息被传输到的所述第二处理器的所需响应时间,并且

其中所述第一处理器选择用于向所述第二处理器传输所述请求信息的传输格式以便满足该响应时间。

12. 根据权利要求11所述的存储系统,

其中所述第一处理器能够校正所述第二处理器的所述所需响应时间。

13. 一种使其中第一控制器的第一处理器和第二控制器的第二处理器通过总线被连接、并且其中被存储在所述第一控制器的第一存储器中的共享区域和所述第二控制器的第二存储器中的共享区域中的数据将由所述第一处理器和所述第二处理器使用的存储系统能够控制在主机与存储区域之间的数据传送的控制方法,包括:

所述第一处理器向所述第二处理器传输将由所述第二处理器处理的请求信息,

所述第二处理器处理所述请求信息并且向所述第一处理器发送回响应,

所述第一处理器确定用于访问所述第二存储器中的所述共享区域的所述请求信息将

被优先并且向所述第二处理器传输指示所述请求信息的优先级的所述请求信息,并且在处理下一过程之前等待来自所述第二处理器的响应,

其中所述第一处理器确定用于请求数据传输的所述请求信息将不被优先并且向所述第二处理器发送不指示所述请求信息的优先级的所述请求信息,并且处理所述下一过程而不等待来自所述第二处理器的所述响应,并且

所述第二处理器通过在针对其的处理将被优先的请求信息与针对其的处理将被优先的请求信息之间区分,来获取所述请求信息,并且使具有优先级的所述请求信息的处理优先于不具有所述优先级的所述请求信息。

计算机系统及其控制方法

技术领域

[0001] 本发明涉及一种存储系统及其控制方法,利用该存储系统和该控制方法在主机与存储区域之间的数据I/O处理被控制,并且更特别地涉及一种群集式存储系统,在该群集式存储系统中用于执行在主机与存储设备之间的数据传输和接收的控制器被复用。

背景技术

[0002] 如下存储系统被称为群集式存储系统,在该存储系统中用于控制在主机与存储区域之间的访问的控制器被复用。利用这一类型的存储系统,为了有效地操作多个控制器,某个控制器的处理器必须访问硬件资源,比如另一控制器的主存储器 and 接口。因此,为了使这一访问迅速和可靠,处理器不直接地访问硬件,相反地,处理器经由用于访问的专用LSI获得访问。作为包括群集式控制器的存储系统,已知在第3719976号日本公开专利和第2008-269424号日本公开未审申请中公开的存储系统。

[0003] 引用列表

[0004] 专利文献

[0005] PTL 1:第3719976号日本公开专利

[0006] PTL 2:第2008-269424号日本公开未审申请

发明内容

[0007] 技术问题

[0008] 近年来,部分地由于已经促成通用LSI性能改进的条件,已经产生了对于通过使用通用产品而未使用专用LSI来构建用于存储系统的群集控制器的需要。在这一类型的存储系统中,可以通过经由PCI-快速总线或者相似总线连接两个控制器的处理器来访问多个控制器。

[0009] 在这一类型的存储系统中,由于未使用专用LSI,所以第一控制器的处理器不能直接地访问硬件资源(比如第二控制器的数据传送硬件和存储器等),但是第一控制器的处理器可以向第二控制器的处理器发布对于访问资源的请求并且可以从第二控制器的处理器接收这一访问的结果。

[0010] 然而,问题在于第二控制器的处理器频繁地执行主机I/O处理,并且在这样的处理期间,尽管从第一控制器的处理器接收请求,但是不可能直接地对请求做出响应,因此与使用专用LSI的存储系统比较减少响应性能。

[0011] 因此,必须使第二控制器的处理器对来自第一控制器的处理器的请求的处理优先。存在一种方法,在该方法中,在处理器之间使用中断请求以用于这一类型的优先的处理。然而,在生成中断时,有必要将处理从OS迄今一直工作于的进程切换到普通中断处置器,并且问题在于开销由于为了切换进程所需要的时间而为大并且在于在广泛地使用中断请求时性能代之以下降。

[0012] 另外,也可以考虑一种用于通过使用多核处理器作为处理器、选择多个核之一作

为用于从另一控制器接收请求的专用核并且保证这一个核不执行普通处理(比如主机I/O)处理来缩短对从另一控制器接收的请求的响应时间的方法。然而,在专用核上有大量负荷时,对来自另一控制器的请求的处理被延迟,而在增加专用核的数目时,普通核的比例小,并且也很可能的是主机I/O处理的性能将受影响。

[0013] 因此,本发明的目的是提供一种群集式存储系统,利用该群集式存储系统,即使在从一个控制器的处理器发送对另一控制器的处理器的访问时,另一控制器的处理器也能够使对这一访问的处理优先,从而使得也防止I/O处理被延迟。

[0014] 对问题的解决方案

[0015] 为了实现前述目的,本发明是一种存储系统及其控制方法,其中第一控制器的第一处理器通过在针对其的处理将被第二控制器的第二处理器优先的请求信息与针对其的处理将不被优先的请求信息之间区分来向第二处理器传输将由第二处理器处理的请求信息,并且第二处理器通过在针对其的处理将被优先的请求信息与针对其的处理将不被优先的请求信息之间区分来获取请求信息。

[0016] 根据本发明,第二处理器不使对来自第一处理器的所有请求信息的处理优先,相反地,处理请求信息,因为需要其优先以便操作存储系统,因此也防止第二处理器对来自主机的I/O的处理被延迟。本发明的有利效果

[0017] 利用本发明,在群集式存储系统中,即使在从一个控制器的处理器发送对另一控制器的处理器的访问时,第二控制器的处理器也能够使对这一访问的处理优先,从而使得也防止I/O处理被延迟。

附图说明

[0018] [图1]图1是包括根据第一实施例的存储系统的计算机系统的硬件框图。

[0019] [图2]图2是根据控制器的MPU存储器的配置示例的框图。

[0020] [图3]图3是示出了在MPU存储器的文本区域中的软件资源配置的框图。

[0021] [图4]图4是请求传输程序的过程流程的示例。

[0022] [图5]图5是示出了请求传输程序的另一方面的流程图。

[0023] [图6]图6是中断接收程序的过程流程的示例。

[0024] [图7]图7是普通接收程序的过程流程的示例。

[0025] [图8]图8是代理访问程序的过程流程的示例。

[0026] [图9]图9是定时器同步程序的过程流程的示例。

[0027] [图10]图10是请求处理标识表的配置示例。

[0028] [图11]图11是访问目的地属性表。

[0029] [图12]图12是内部/外部系统确定表的配置示例。

[0030] [图13]图13是主机2的读取I/O处理的流程图。

[0031] [图14]图14具体地描述图13的流程图的部分。

[0032] [图15]图15具体地描述图13的流程图的部分。

[0033] [图16]图16是根据第二实施例的MPU存储器的配置示例。

[0034] [图17]图17是专用核确定表的配置示例。

[0035] [图18]图18是根据第二实施例的MPU存储器的软件配置图。

- [0036] [图19]图19是根据第二实施例的请求传输程序的过程流程的示例。
- [0037] [图20]图20是专用核计数调整程序的流程图。
- [0038] [图21]图21是请求系统要求表的示例。
- [0039] [图22A]图22A是请求系统可用性表的第一示例。
- [0040] [图22B]图22B是请求系统可用性表的第二示例。
- [0041] [图23]图23是根据第四实施例的请求传输程序的流程图。
- [0042] [图24]图24是为每个卷确定前述性能要求的卷性能要求表的配置示例。
- [0043] [图25]图25是使用户能够设置卷I/O性能要求的输入屏幕的示例。
- [0044] [图26]图26是用于校正预计通信响应时间的流程图。

具体实施方式

[0045] 实施例1

[0046] 接着将基于附图描述根据本发明的存储系统的实施例。存储系统包括双控制器配置。图1是包括根据第一实施例的存储系统的计算机系统的硬件框图。计算机系统1包括主机2、管理终端3、存储系统4以及网络5和网络6。网络5服务于将主机2连接到存储系统4并且特别地为SAN。网络6将存储系统4连接到管理终端3并且特别地为LAN。主机2例如是大型通用计算机、服务器或者客户端终端。主机2也可以连接到存储系统4而未经由SAN 5通过。另外,管理终端3也可以连接到存储系统而未经由LAN 6通过。

[0047] 存储系统4包括第一控制器100a、第二控制器100b和包括多个存储驱动109D的驱动单元109。第一控制器100a可以被称为“CTLa”,并且第二控制器100b可以被称为“CTLb”。另外,在引用控制器中的部件时,如果区分部件CTLa与部件CTLb,则在前一个部件的标号之后追加“a”,并且在后一个部件的标号之后追加“b”。向相同部件指派相同标号。如果在部件CTLa与CTLb二者之间区分是不必需的,则不追加“a”和“b”。在有在两个控制器之间进行的区分时,一个控制器被称为内部系统控制器并且另一控制器被称为外部系统控制器。

[0048] CTLa包括MPU 101a、MPU存储器102a、DMA 103a、主机_I/F(I/F:用于接口的缩写词,下文相同)104a、驱动_I/F 105a和管理_I/F 106a。这对于CTLb同样如此。CTLa的主机_I/F 104a和CTLb的主机_I/F 104b各自经由SAN 5连接到主机2。CTLa的驱动_I/F 105和CTLb的DRIVE_I/F 105b各自连接到驱动单元109。CTLa的管理_I/F 106a和CTLb的管理_I/F 106b各自经由LAN 6连接到管理终端3。

[0049] CTLa的NTB 107a和CTLa的NTB 107b各自是非透明网桥。MPU 101b和MPU 101a使用具有至少5 Gbps的传送速度的全双工系统高速总线108经由NTB 107a和107b来连接并且能够交换用户数据和控制信息等。可以有多个NTB 107和连接路径108。驱动单元109包括作为多个逻辑存储区域的LU(逻辑单元)。驱动单元109的存储驱动109d各自由存储介质(比如磁盘或者SSD)配置。

[0050] 接着将基于图2描述CTLa的MPU存储器102a的配置。CTLb的MPU存储器102b相同。MPU 102a包括文本区域10a、用于本地存储器11a的区域、用于通信区域14a的区域、共享存储器12a(下文被称为“SM”)的区域、高速缓存存储器13a的区域、每个区域被配置为地址区域。区域各自存储程序、控制信息或者各种数据。

[0051] 文本区域10a存储使存储系统能够实施它的功能的各种程序。本地存储器11a存储

被文本区域10a的程序引用的表。这些表不被CTLb的文本区域的程序引用。MPU程序102b的文本区域的程序引用MPU存储器102b的本地存储器11b。

[0052] 作为本地存储器11a的表,例如,内部/外部系统确定表110、代理访问标识表111和硬件属性表112。随后将提供对每个表的描述。

[0053] 通信区域14a包括中断通信区域141a和普通通信区域142a。控制信息和数据被MPU 101b写入到这一通信区域。控制信息和数据被MPU 101写入到MPU存储器102b的通信区域14b的中断通信区域141b和普通通信区域142b。

[0054] SM 12a存储必须由CTLa和CTLb共享的存储系统配置信息和控制信息。这一信息包括内部系统控制器的信息和外部系统控制器的信息。外部系统控制器的SM 12b相同。这一信息将在下文被称为共享信息。

[0055] 共享信息是为了控制存储系统而需要的信息。更具体而言,共享信息是用于在控制器之间建立用于在存储器中存储并且与主机计算机交换的数据的高速缓存管理信息、用于数据传送硬件资源(比如主机接口、盘接口或者DMA)的启动寄存器、独占管理信息(比如用于实施用于独占访问这些硬件资源的锁定机制的锁定位)、用于必须也可由另一控制器识别的硬件资源配置的设置或者更新信息和存储应用(例如,用于创建卷的重复映像的卷复制功能、用于虚拟化和如果必需则指派物理卷容量的虚拟卷功能以及用于通过执行向远程地点的数据复制来实施灾难恢复的远程复制功能)的一致性的控制信息。

[0056] 如果这一共享信息在一个控制器的SM 12a中存在,则共享信息经由一个控制器的MPU 101a被外部系统控制器的MPU 101b引用。例如,由于高速缓存存储器13a的索引信息在包括高速缓存存储器的控制器100a中存在,所以,如果外部系统控制器100b引用索引信息,则外部系统控制器100b经由控制器100a的包括高速缓存存储器13a的MPU 101a访问索引信息。

[0057] 高速缓存存储器13a暂时地保持在盘上存储的用户数据而未保持主机2的用户数据。外部系统控制器的高速缓存存储器13b也是这样。

[0058] 图3是文本区域10a的软件资源的配置图。软件资源包括配置管理程序201、基本I/O程序202、定时器同步程序203a、代理访问程序204a、中断接收程序205a、普通接收程序206a、请求传输程序207a、SM访问程序208a、DMA访问程序209a、主机_IF访问程序210a和驱动_IF访问程序211a。这些程序由MPU 101a执行。CTLb的文本区域也是这样。

[0059] 配置管理程序210a由来自管理终端3的命令启动并且能够引用和更新配置管理信息。配置管理信息是用于管理硬件(比如驱动单元109、主机_IF 104和驱动_IF 105)以及逻辑部件(比如从一个或者多个存储驱动配置的逻辑卷)的信息等。CTLa的配置信息和CTLb的配置信息被记录在SMa(12a)和SMb(12b)中。

[0060] 为了多个控制器由主机集中地操作,如果一个控制器的配置信息被更新,则希望同步这一控制器与另一控制器。在MPU 102a更新内部系统控制器的SM 12a的配置信息时实现这一同步以例如作为MPU 102a将这一更新写入到外部系统控制器的SM 12b的结果。

[0061] 控制器100a(该控制器100a包括由管理终端3访问的管理_I/F106a)的MPU 101a更新相同SM 12a的配置管理信息并且将用于外部系统控制器100b的MPU存储器102b中的通信区域14b的配置信息更新确定标志设置成ON。外部系统控制器的MPU 101b引用更新标志并且如果标志为ON则从控制器100a的MPU 101a获取SM 12a的最新配置信息并且更新它的自

有控制器100b的SM 12b的更新信息并将标志设置成OFF。

[0062] 基本I/O程序202a从主机2接收I/O请求命令并且执行数据读取I/O或者写入I/O。在执行读取I/O或者写入I/O时,基本I/O程序202a在必需时调用请求传输程序207a、SM访问程序208a、DMA访问程序209a、主机_IF访问程序210a和驱动_IF访问程序211a等并且执行每个程序。

[0063] 定时器同步程序203a服务于同步内部系统控制器100a的定时器与外部系统控制器100b的定时器。在多个控制器之中,一个控制器的定时器被视为主控(master),并且另一控制器的定时器被视为从属(slave)。被视为主控的控制器的定时器同步程序被置于非操作模式中。

[0064] 代理访问程序204a是使内部系统控制器100a的MPU 101a而不是MPU 101b能够基于从外部系统控制器的MPU 101b传输的请求信息(命令等)访问存储器102a、接口104a和105a以及DMA 103a中的任何一项或者多项的程序。

[0065] 请求传输程序207a向或者经由外部系统控制器101a的MPU101b发布用于访问的请求以使配置管理程序201a或者基本I/O程序202a能够读取或者向外部系统控制器100b的MPU存储器102b的SM 12b和高速缓存存储器13b写入并且启动接口104b和105b以及DMA 103b。这一请求被实现作为处理器101a经由网桥108在外部系统控制器的MPU存储器102b的通信区域14b中写入请求信息(命令等)并且外部系统控制器的MPU 101b执行这一命令的结果。如更早描述的那样,通信区域14a(14b)包括用于中断处理的请求信息被记录在其中的中断通信区域141a(141b)和用于除了中断处理之外的普通处理的命令信息被记录在其中的普通通信区域142a(142b)。

[0066] 中断通信区域141a记录请求信息,该请求信息请求外部系统控制器100b的优先的处理,并且普通通信区域142a记录命令,这些命令请求外部系统控制器的非优先的处理。CTLa的配置管理程序201a和基本I/O程序202a在中断通信区域141a中写入希望被CTLb的MPU 101b迅速地处置的处理,并且向CTLb发出中断信号以便补偿在处理器之间的通信中不使用专用LSI的缺点。CTLb的MPU 101b通过执行代理访问程序204b来执行在中断通信区域141b和普通通信区域142b中存储的命令并且用存储器(12b,13b)的读取的信息和从命令的执行而产生的数据传送完成信息等向发布命令请求的MPU 101a做出响应。

[0067] MPU 101a的请求传输程序207a使将在中断通信区域141b中存储的请求信息(命令)的类别和属性等限于针对优先的处理所需要的范围。假设广泛多种请求信息都被存储在中断通信区域141b中,MPU 101b必须使对用于外部系统控制器的大量命令的处理优先,并且MPU 101必须实质上执行处理而在内部系统控制器中的对来自主机的I/O的处理等被延迟。

[0068] 因此,将描述针对其的处理被外部系统控制器优先的请求信息的性质。如果第一控制器的处理器要求第二控制器的处理器引用前述共享信息,则第一控制器的处理器应当直至从第二控制器的处理器获得共享信息才继续下一处理。例如,如果第一控制器的处理器不能从第二控制器获取共享信息,则第一控制器的处理器不能访问驱动单元109的正确卷并且不能正确的执行I/O处理。其中第一控制器将它的自有共享信息的更新信息复制到外部系统控制器的情况也被希望迅速地完成。也就是说,必须在很短时间内完成对共享信息的访问,并且必须有来自请求目的地处理器的向请求源处理器的响应。因此,本发明假设

针对其的处理被优先的请求信息的优选示例,其中对外部系统控制器的共享存储器进行访问。

[0069] 接着将描述针对其的处理无需被优先的请求信息。例如,内部系统控制器请求向外部系统控制器的数据传送或者审查数据传送请求。即使在处理将很可能由于传送数据传送大小为大的数据而需要长时间时,也必须通过访问共享信息来使外部系统控制器的数据传送硬件的启动优先。

[0070] 换言之,无需外部系统控制器的处理器的短响应时间。这是因为除了共享信息不同并且不使处理优先的影响小之外,控制器还能够经由并行复用来处理主机I/O,并且因此在处理器已经由于处理某个主机I/O命令而向外部系统控制器的处理器发布了对启动用于数据传送的硬件资源的请求之后,处理器继续处理另一主机I/O命令而未等待来自外部系统控制器的处理器的响应。无需缩短用于这一种类的数据传送启动的响应时间,并且这一处理被视为普通处理(未优先的处理)而不是优先的处理。

[0071] 因此,进行第一控制器的处理器是否应当使对从第二控制器的处理器请求的信息的处理优先或者不使这一处理优先的区分,这意味着仅在前一种情况下向第二控制器的优先处理单元传输请求信息,并且在后一种情况下向第二控制器的普通处理单元传输请求信息。第二控制器的处理器使在优先处理单元中的对请求信息的信息处理优先并且在普通处理单元中普通地处理请求信息。

[0072] 第一实施例采用中断处理作为用于实现优先的处理的手段。第二控制器100b的处理器101b根据从第一控制器100a的处理器101a发送的请求信息是否被写入到中断通信区域141b或者写入到普通通信区域142b来在使对来自第一控制器100a的请求信息的信息处理优先或者普通地处理请求信息而未优先之间区分。对于向普通通信区域142b写入的命令,第二控制器的处理器101b以规律间隔轮询普通通信区域142b,并且当在普通通信区域142b中发现请求信息(该请求信息是普通处理目标)时,处理器101b通过检测比如在主机I/O处理完成之后或者在处理期间的某个间隔来处理命令。

[0073] 图3中的中断接收程序205a是如下程序,该程序在从外部系统控制器100b接收中断请求时被调用和启动并且在中断通信区域141a中接收命令并且向代理访问程序204a转交命令,并且普通接收程序206a是如下程序,该程序在前述轮询期间被启动并且如果有在普通通信区域142a中的命令则接收并且向代理访问程序204a转交命令。

[0074] 程序在必需时调用SM访问程序208、DMA访问程序209、主机_IF访问程序210和驱动_IF访问程序211等并且启动这些程序中的一些或者所有程序。

[0075] SM访问程序208a是如下程序,该程序在配置管理程序201a和基本I/O程序202a需要SM 12a的共享信息时被调用并且支持用于引用和更新SM 12a的控制信息的处理的执行。

[0076] DMA访问程序209a是如下程序,该程序启动DMA 103a并且执行在高速缓存存储器13a中存储的用户数据的数据传送。

[0077] 主机_IF访问程序210a是用于访问主机_IF 105a并且向和从主机2传输和接收命令和数据的程序。

[0078] 驱动_IF访问程序211a是用于访问驱动_IF 105a并且向和从驱动109D传输和接收命令和数据的程序。

[0079] 这些访问程序与访问请求信息一起由基本I/O程序等调用,该访问请求信息包括

信息,比如标识访问目标硬件的编号、访问类型(读取/写入/数据传送等)、目标数据存在于的存储器地址和在硬件处理完成时存储结果信息(比如成功/失败和读取的值等)的存储器地址。访问程序各自根据指明的访问类型执行对由基本I/O程序指定的硬件的访问。

[0080] 驱动_IF访问程序211a与访问请求信息一起由基本I/O程序202a调用,该访问请求信息伴随有管理信息,比如用于标识驱动_IF 105a的驱动_IF编号、访问类型(比如读取/写入)、传送数据存储区域(高速缓存存储器、驱动单元)的地址和响应存储目的地地址(SM地址)。被这样调用的驱动_IF访问程序211a基于信息(比如访问类型以及数据传送目的地地址)启动由驱动_IF编号指明的驱动_IF105并且结束。启动的驱动_IF 105向驱动单元109传送在指明的高速缓存存储器13a中的数据或者向指明的数据存储高速缓存区域传送驱动单元109中的指明的数据。此外,驱动_IF 105a向在MPU存储器101a中的指明的响应存储目的地区域(通信区域14a)写入指示是否已经普通地完成了传送的完成状态信息。

[0081] 主机_IF访问程序210a与访问请求信息一起由基本I/O程序202a调用,该访问请求信息伴随有管理信息,比如用于标识主机_IF 104a的主机_IF编号、访问类型(比如读取/写入)、传送数据存储区域(高速缓存存储器、驱动单元)的地址和响应存储目的地地址(SM地址)。被调用的主机_IF访问程序基于信息(比如访问类型以及数据传送目的地地址)启动由主机_IF编号指明的主机_IF并且结束。启动的主机_IF向主机传送在高速缓存存储器13的指明的区域中的数据或者从主机向高速缓存存储器中的指明的区域传送数据。此外,驱动_IF 105a向在MPU存储器101a中的指明的响应存储目的地区域(通信区域14a)写入指示是否已经普通地完成了传送的完成状态信息。

[0082] DMA访问程序正在接收的对启动的请求包含信息,比如正被启动的DMA的标识编号、关于数据传送源/传送目的地高速缓存存储器属于的控制器信息、存储器地址和用于存储涉及DMA数据传送完成的状态信息的响应存储目的地地址(SM通信区域)。

[0083] SM访问程序208a根据由调用器程序指定的请求类型向MPU存储器101a的SM 12a执行读取/写入/原子访问等。另外,SM访问程序208a向通信区域的结果返回区域写入访问请求。

[0084] 图4是请求传输程序207的过程流程。请求传输程序207与请求信息(该请求信息包括标识信息,该标识信息示出针对其的代理访问将被请求的外部系统控制器的硬件)、请求内容(读取、写入、数据传送等的启动)和如果必要则包括读取目标存储器的地址等、内部系统控制器的MPU存储器101a(该MPU存储器用于存储来自外部系统控制器的硬件或者访问程序的返回值)的地址和伴随参数一起由调用器程序(比如基本I/O程序202或者配置信息管理程序201)调用。

[0085] 首先,请求传输程序207a从供应自调用器程序的请求信息获取用于标识请求处理信息的标识代码(步骤S1201)。请求传输程序207a然后引用请求处理标识表111a(图10:随后提供的细节)并且根据标识代码获取关于外部系统控制器的硬件资源的信息,该外部系统控制器是访问目的地(步骤S1202)。请求传输程序207a继续步骤SP1203并且确定访问目的地硬件是否包括前述共享信息。更具体而言,请求传输程序207a引用地址目的地属性表112(图12,随后提供的细节),并且如果在与访问目的地硬件对应的共享信息字段中存储了共享信息,则请求传输程序207a确定对外部系统控制器100b的访问目的地硬件的访问是对共享信息的访问。在获得肯定结果时,请求传输程序207a向外部系统控制器100b的MPU存储

器102b的中断通信区域141b传输请求信息(S1204)。

[0086] 在访问请求时,请求传输程序207a向外部系统控制器的MPU存储器102b的通信区域14b的中断通信区域141b写入请求信息并且然后经由控制器间连接总线108向外部系统控制器的MPU 101b传输中断信号(例如,PCI快速MSI分组)

[0087] 同时,在步骤S1203产生否定结果并且确定对外部系统控制器的访问不是对共享信息的访问时,请求传输程序207a无需对外部系统控制器100b的请求目的地硬件执行中断处理并且向外部系统控制器100b的MPU存储器102b的普通通信区域142b传输请求信息。

[0088] 图5是服务于图示请求传输程序207a的另一实施例的流程图。这一流程图与图4的流程图不同在于图4的步骤S1204被步骤S1210至S1214替换。在请求传输程序207a确定调用器程序的访问目的地是在外部系统控制器101b的访问目的地(SM 12b等)中记录的共享信息时,请求传输程序207a递增计数器,该计数器更新和记录在向外部系统控制器100b的中断处理请求之中的未完成中断处理数目。请求传输程序207a在将传输中断信号的阶段将计数器递增1并且在外部系统控制器发送回响应信号时的时刻将计数器递减1。控制器在内部系统控制器的MPU存储器102的本地存储器11中包括计数器。

[0089] 请求传输程序207a在步骤S1210中确定当前计数器的值是否小于阈值(S1210),并且如果该值等于或者大于阈值(S1210:否),则请求传输程序207a不向外部系统控制器100b的MPU 102b产生中断处理请求并且在内部系统控制器100a的MPU存储器102a的中断待命区域中记录请求,因此结束流程图[的处理](步骤S1211)。

[0090] 在确定计数器小于阈值(S1212:是)时,请求传输程序207a检查是否有在中断待命区域中记录的请求(S1212),并且如果有在中断待命区域中的请求信息,则请求传输程序207a组合待命请求信息与由S1203确定的当前请求信息并且向外部系统控制器的中断通信区域141b传输组合的请求信息(S1214)并且向外部系统控制器100b传输用于多个请求信息的中断信号(S1214)。请求传输程序207a递增请求信息计数计数器(步骤S1215)、同时,在请求传输程序确定无待命请求信息时,请求传输程序207a传输当前请求信息(S1213)、传输用于当前请求信息的中断信号(S1214)并且将计数器递增1(S1215)。

[0091] 利用这一流程图,由于请求传输程序207a收集中断请求,所以有可能避免向请求目的地控制器100b连续地发布大量请求信息传输(这些请求信息传输是中断处理目标)并且减少以由请求目的地控制器100b基于请求信息而调用的过程的频繁切换为基础的开销。

[0092] 图6示出了用于中断接收程序205的过程流程。在MPU 101b从外部系统控制器101a接收中断信号的时刻中止正被执行的处理时执行中断接收程序205。在中止正被执行的处理时,记录程序计数器和局部变量等,并且中断处理结束,其中重启中止的程序。这一布置通常是OS(操作系统)的功能。

[0093] 首先,中断接收程序205a检查是否在中断通信区域141a中记录了请求信息(请求)(S1802),并且如果有请求,则中断接收程序205a从中断通信区域141a获取请求信息(S1800)并且相应地启动代理访问程序204a(S1801)并且返回到步骤S1802。如果无请求(S1802:否),则结束流程图[的处理]。

[0094] 图7示出了普通接收程序206的过程流程。以规律间隔向MPU101调用但是在中止主机命令处理(比如I/O处理)时不执行普通接收程序206。因此,虽然普通接收程序在正被执行的处理中不生成开销,但是不同于中断处理,处理可能在接收请求之后需要长时间。普通

接收程序206a首先检查是否在普通通信区域142a中存储请求信息(S1900)并且如果不是则结束处理。如果存储了请求信息,则普通接收程序206a从普通通信区域142a获取请求信息(步骤S1901)并且启动代理访问程序204a(1902)。在代理访问程序204a的启动结束时,处理返回到步骤S1900。

[0095] 图8示出了用于代理访问程序204的流程图。代理访问程序204a由中断接收程序205a或者普通通信程序206a启动。代理访问程序204a经由中断通信程序205a或者普通通信程序206a获取请求信息(步骤S1700)。代理访问程序204a然后引用请求处理标识表111(图10)以基于在请求信息中包含的标识代码标识外部系统控制器的访问目的地硬件(步骤S1701)。另外,在步骤S1702至S1709中,代理访问程序204a鉴别访问目的地硬件并且启动用于每个访问目的地硬件的访问程序。

[0096] 图9是用于定时器同步程序203的过程流程。定时器信息是将由多个控制器共享的共享信息之一。定时器同步处理是用于同步多个控制器的MPU内部定时器的处理。定时器同步程序203a以固定间隔由内部系统控制器100a调用并且获取外部系统控制器100b的MPU内部定时器的值并且向内部系统控制器100a的MPU内部定时器复制读取的值。在这一情况下,外部系统控制器的定时器被视为主控。

[0097] 定时器同步程序203a启动请求传输程序207a并且在外部系统控制器的中断通信区域141b中记录定时器读取请求信息(S1101)。外部系统控制器的MPU 101b借助中断接收程序205b获取请求信息并且借助代理访问程序204a读取MPU 101b的内部定时器值。MPU101b执行请求传输程序207a并且在控制器100a的中断通信区域141a中存储读取的信息作为对请求信息的响应。MPU 101a执行中断接收程序205a并且接收中断通信区域141a的定时器值,并且代理访问程序204a访问MPU 101a的内部定时器并且复制读取的定时器值。因此,定时器同步程序在忙循环中等待直至有来自外部系统控制器的响应(S1102)并且如果接收响应则从响应信息获取定时器值(S1103)并且向内部系统控制器的MPU内部定时器复制定时器值(S1104),因此结束流程图[的处理]。

[0098] 接着将描述由代理访问程序、中断接收程序、普通接收程序和请求传输程序中的一个或者两个或者更多程序使用的控制和管理表。控制和管理表包括前述内部/外部系统确定表110、请求处理标识表111和地址目的地属性表112(见图2)。

[0099] 图10示出了请求处理标识表111的配置示例。请求标识表111是用于根据标识代码标识通信类型和访问目的地的表。每个条目包括标识代码字段1110、通信类型字段1111和访问目的地字段1112。请求传输程序207基于来自配置管理程序201或者基本I/O程序202的命令确定通信类型1111并且确定访问目的地1112。

[0100] 请求传输程序207基于这些确定项目确定标识代码。请求传输程序207a根据访问目的地进行关于请求传输程序207a将向外部系统控制器100b的通信区域14b的中断通信区域141b和普通通信区域142b中的哪个通信区域写入标识代码的判决。如果访问目的地是共享信息或者定时器信息,则向中断通信区域141b写入标识代码,否则向普通通信区域142b写入标识代码。

[0101] 图11示出了访问目的地属性表112的配置示例。访问目的地属性表112是指示每个访问目的地是否存储共享信息的表。每个条目包括访问目的地字段1120和共享信息字段1121。共享存储器(SM)12包括共享信息并且定时器也包括共享信息,但是其它访问目的地

不包括共享信息。

[0102] 为了确定向外部系统控制器100b的通信区域14b的中断通信区域141b和普通通信区域142b中的哪个通信区域写入标识代码,请求传输程序207引用访问目的地属性表112以确定访问目的地包括共享信息(是),并且请求传输程序207a在中断通信区域141b中存储标识代码。在确定访问目的地不包括共享信息时,请求传输程序207a在普通通信区域142b中存储标识代码。

[0103] 图12示出了内部/外部系统确定表110的配置示例。内部/外部系统确定表110基于前述标识代码(ID)示出了访问目的地类型和访问目的地是否为CTLa或者CTLb。每个条目包括ID字段1100、类型字段1101和CTL字段1102。如果在CTL字段中的注册信息是“b”,则请求传输程序207a在外部系统控制器的通信区域14b中存储标识代码,并且在检测到在CTL字段中的注册信息是“a”时,由于配置管理程序201a或者基本程序202a的访问目的地是在它的自有控制器中的资源,程序无需在外部系统控制器100b的通信区域14b中存储标识信息而对请求处理标识表111的访问目的地硬件资源执行与通信类型对应的处理。

[0104] 接着,在描述对来自主机2的读取I/O的处理之时,将基于图13至图16描述代理访问程序204、中断接收程序205、普通接收程序206和请求传输程序207的操作。图13是读取I/O处理流程图。在MPU 101从主机IF 104接收读取I/O命令时启动这一流程。

[0105] 首先,基本I/O命令202向配置管理程序201发布关于存在配置信息更新的查询。配置管理程序201a访问通信区域14a并且检查是否已经设置了配置信息已更新标志。在检测到已经设置了用于外部系统控制器100b的配置信息已更新标志时,由于必须从外部系统控制器读取最新共享信息(S1015:是),所以配置管理程序201a启动请求传输程序207a(S1016)。

[0106] 在启动请求传输程序207a时,向外部系统控制器100b的通信区域14b的中断通信区域141b写入标识代码(0x00),并且向外部系统控制器100b传输中断信号。标识代码(0x00)是如下命令,该命令请求外部系统控制器100b的MPU 101b访问共享存储器12b以从共享存储器12b读取更新的共享信息。请求传输程序207a在忙循环中等待来自外部系统控制器100b的响应(S1017)。注意,可以如更早提到的那样使内部系统控制器100a对来自外部系统控制器100b的响应的处理优先。

[0107] 在从内部系统控制器100a接收中断信号时,外部系统控制器100b的MPU 101b启动中断接收程序205b,并且中断接收程序205b读取在中断通信区域141b中的标识代码并且向代理访问程序204b转交标识代码。代理访问程序204b基于标识代码引用请求处理标识表111b并且在确定访问目的地是SM 12b时访问SM 12b。

[0108] 另外,代理访问程序204a读取更新的共享信息,并且内部系统控制器的经由请求传输程序207b向内部系统控制器100a传输包含共享信息的响应的基本I/O程序202a从响应信息获取更新的共享信息(S1018),并且经由配置管理程序201a向共享存储器12a复制更新的信息(S1019)。因此,在基本I/O程序202a执行对来自主机2的读取I/O的处理时,由于将对存在对将由在具有双控制器配置的存储系统中的两个控制器共享的配置信息的更新进行检查,所以即使当在一个控制器中实现配置信息更新时,也可以向外部系统控制器迅速地复制这一更新。注意,在配置管理程序201a在S1015中检查通信区域14a时,如果尚未设置配置信息更新标志,则可以跳过步骤S1016在S1019。

[0109] 基本I/O程序202a引用SM 12a的共享信息(内部系统控制器100a的高速缓存存储器13a的高速缓存索引和外部系统控制器100b的高速缓存存储器13b的高速缓存索引)并且执行高速缓存命中/未命中确定(S1002),并且在高速缓存命中的情况下前进到S1008。

[0110] 在高速缓存未命中的情况下,基本I/O程序202a前进到步骤S1004。在S1004中,为了在内部系统控制器100a的高速缓存存储器13a或者外部系统控制器100b的高速缓存存储器13b中保护用于从驱动单元109倒盘(stage)读取的数据的新高速缓存区域,基本I/O程序202a更新相应的高速缓存存储器配置信息。在向内部系统控制器100a的SMA复制涉及保护高速缓存区域的信息时,配置管理程序201a将标志设置成更新信息已经借助请求传输程序207a在外部系统控制器100b的通信区域14b的中断通信区域141b中被更新的效果。外部系统控制器100b的MPU 101b使对标志的处理优先并且在内部系统控制器100a的中断通信区域141a中写入用于读取SM12a的高速缓存配置信息的请求。

[0111] 随后,在确定已经在内部系统高速缓存存储器中保护了高速缓存区域时,基本I/O程序202a启动内部系统控制器100a的驱动_IF105a,也就是启动驱动_IF访问程序211a(S1006)。如果保护的高速缓存区域在外部系统控制器100b的高速缓存存储器13a中,则基本I/O程序202a启动请求传输程序207a并且在外部系统控制器的普通通信区域142b中写入对启动驱动_IF 105b的请求。外部系统控制器100b的MPU 101b执行代理访问程序204a、启动驱动_IF访问程序211b并且然后启动驱动_IF 105b。

[0112] 同时,如果基本I/O程序202a确定高速缓存命中(S1002:是),则基本I/O程序202a在步骤S1008中引用在SM 12a中的共享信息并且确定包括读取的数据的高速缓存存储器13是否在与从主机接收读取I/O的主机_IF 104相同的控制器中(S1008)。在进行否定确定时,基本I/O程序202a确定包括读取的数据的高速缓存存储器是否在将处理读取I/O的控制器(这在图13的情况下是内部系统存储器100a)中(S1009),并且在确定高速缓存存储器13属于内部系统控制器100a时,基本I/O程序202a启动内部系统控制器100a的DMA访问程序209a并且启动DMA 103a(S1010)。DMA 209a向主机_IF104a传送高速缓存存储器的读取的数据。

[0113] 然而,在确定高速缓存存储器在外部系统控制器100b中时,基本I/O程序202a启动请求传输程序207a以便向外部系统控制器发布对于启动DMA 109b的请求(S1011)。

[0114] 在步骤S1008中,基本I/O程序202a引用内部系统控制器100a的SM 12a中的配置信息,并且在确定包括读取的数据的高速缓存存储器在与从主机2接收读取I/O的主机_IF 104相同的控制器中时,基本I/O程序202a确定主机_IF是否在内部系统控制器100a中。在确定主机_IF在内部系统控制器100a中时,基本I/O程序启动主机_IF访问程序210a并且启动主机_IF 104a(步骤S1013)。

[0115] 在确定主机_IF在外部系统控制器100b中(S1008:是)时,基本I/O程序启动请求传输程序207a以便请求启动外部系统控制器100b的主机_IF访问程序210b(步骤S1014)。

[0116] 基本I/O程序经由步骤S1013、S1014、S1010、S1011、S1006和S1007前进到S1015。在S1015中,基本I/O程序202a从前述启动的程序设置响应待命状态并且暂时地结束图13中的流程图[的处理]。在识别来自启动目标程序的响应时,基本I/O程序202a重启图13中的流程图[的处理]以便继续读取I/O命令处理。

[0117] 当在图13的S1006或者S1007中识别对启动请求的响应时,基本I/O程序202a重启图13中的流程图[的处理]。在重启之后的流程如在图14的流程图中所示。在有来自启动目

标程序的响应的状态(S1006或者S1007)中,在内部系统控制器100a或者外部系统控制器100b中保护新高速缓存区域。另外,读取的数据从驱动单元109被倒盘到新保护的高速缓存区域中。因此,基本I/O程序202a在S1002中确定肯定结果并且执行S1008和后续步骤。对图14中的每个步骤的描述与对于图13的描述相同。

[0118] 在基本I/O程序202a由于S1013或者S1014而从程序(该程序是启动目标)接收响应时,假设如下状态,在该状态中,读取的目标数据可以由在内部系统控制器100a和外部系统控制器100b中的DMA 103从高速缓存存储器13传送到主机_IF 104。因此,为了支持向主机的读取数据传送,基本I/O程序必须启动主机_IF访问程序210。出于这一原因,在由于S1013或者S1014而从启动目标程序接收响应时,基本I/O程序202a如在图15中所示执行图13的S1012和后续步骤的处理。图15的S1012和后续步骤的处理与图13的处理相同。

[0119] 当在S1013中从内部系统控制器100a的主机_IF访问程序210a(该主机_IF访问程序210a是启动目标)接收响应或者在S1014中从外部系统控制器100b的主机_IF访问程序210b(该主机_IF访问程序210b是启动目标)接收响应(图15:S1015)时,读取的数据经由主机_IF 104从高速缓存存储器13被传送到发布读取I/O的主机2,并且因此基本I/O程序202a结束读取I/O处理。

[0120] 实施例2

[0121] 虽然在第一实施例中通过中断处理实现了前述优先的处理,但是在第二实施例的情况下,采用多核型处理器作为处理器,从而使得通过为优先的处理独占地配置核中的一些核来实现优先的处理,从而使得防止这些核执行主机I/O处理。图16是第二实施例的MPU存储器102a的配置。第二实施例与第一实施例不同在于向本地存储器11a添加用于在多个核处理器之中确定专用核处理器的表113a以及向通信区域14a添加个别核中断通信区域143a和个别核普通通信区域144a。这对于外部系统控制器100b的MPU存储器102b也成立。

[0122] 利用第二实施例,进行任何给定的核是否为专用核的确定作为每个核在启动期间引用专用核确定表113a的结果,并且仅如果确定核为普通核而不是专用核才执行基本I/O程序202a和配置管理程序201a。

[0123] 图17示出了专用核确定表113a的示例。专用核确定表113a包含如下条目,在这些条目中,数目对应于核数目,并且每个条目包括核字段113、核类型字段1131和中断可接收/不可接收字段132。核字段1130存储用于标识核的信息,核类型字段1131存储用于在专用核与普通核之间区分的标识信息,并且中断可接收/不可接收字段132存储指示普通核是否可以接收中断处理的信息。

[0124] 如更早提到的那样,专用核专用于优先的处理,而普通核能够执行除了优先的处理之外的处理,即,主机I/O处理和对来自管理终端3的管理请求的处理。请求信息向普通核和专用核的传输使用个别核通信区域。专用核无需来自外部系统控制器的对于处理请求的中断请求,并且因此个别核普通通信区域142a用于向专用核的请求。如在图17中所示,为普通核进行设置以指定是否可以接收中断。被设置为不接收中断的普通核不从外部系统控制器接收中断请求,并且因此中断通信区域不用于这一个核。因此,普通核与前述普通处理而不是前述优先的处理兼容。已经被设置为能够接收中断的普通核从外部系统控制器接收中断请求,并且因此与优先的处理和普通处理二者兼容。

[0125] 可以取决于核类型、是否可以接收中断和在每个核上的负荷来动态地改变专用核

确定表113。另外,示出了如下示例,在该示例中,两个处理器各自包括四个核,并且也在该示例中,每个处理器具有一个专用核,但是这仅为示例。专用核和普通核的平衡不限于在该表中描述的平衡。由重写专用核确定表113的配置管理程序使在专用核与普通核之间切换有可能。在专用核确定表中,CTL0是第一控制器的MPU 101a,并且CTL1是第二控制器的MPU 101b。

[0126] 图18是根据第二实施例的MPU存储器102的软件配置图。与在第一实施例中不同,已经添加了专用核计数调整程序220。随后将提供关于这一程序的细节。

[0127] 图19是用于根据第二实施例的请求传输程序207a的过程流程。这一请求传输程序207a与根据第一实施例的请求传输程序(图4)不同在于已经在S1203与S1204之间添加步骤S1253至S1256。接收主机I/O的普通核(专用核不接收主机I/O)启动请求传输程序207。请求传输程序207a在S1201至S1203之后的S1253中引用专用核确定表113并且确定是否有在外部系统控制器100b的MPU 101b中的至少一个专用核(S1254)。

[0128] 在确定没有在外系统控制器中的专用核时,执行请求传输程序207a的普通核在外系统控制器中的MPU存储器101b的普通核的中断通信区域143b中写入请求信息并且传输中断请求(S1204)。

[0129] 在另一方面,如果确定存在专用核,则请求传输程序207a前进到S1257并且选择外部系统控制器的已经被设置为专用核的一个核(S1257)。如果有已经被设置为专用核的多个核,则请求传输程序207a例如经由轮循等选择这些核中的任何核。请求传输程序207a然后在外系统控制器100b中的选择的核的普通通信区域144b中写入请求信息(步骤S1256)。专用核不执行非优先的处理(比如主机I/O处理),并且因此能够在短时间内对请求信息传输源做出响应,这是有利的,因为与借助中断处理来执行优先的处理的情况比较,未基于正在运行的必须被切换到中断处理的程序来生成开销。

[0130] 图20是专用核计数调整程序220的流程图。以规律间隔(例如,每分钟)执行专用核计数调整程序200。专用核计数调整程序连续地监视每个处理核的负荷(操作速率),并且在普通核的处理器负荷低时设置专用核(或者产生大的专用核比率),并且请求传输程序207能够防止由向普通核发送的中断请求所引起的任何开销。同时,调整程序220在普通核的处理器负荷高时取消具有专用核设置的核(或者产生小的专用核比率)。

[0131] 专用核计数调整程序220在S2600中获取用于多个核中的每个核的负荷信息并且在S2601中确定通过从用于普通核负荷的总计值减去用于专用核负荷的总计值而获得的差值是否等于或者大于阈值。如果总计值等于或者大于阈值,则专用核计数调整程序220减少专用核[的数目](S2602)。

[0132] 用于减少专用核数目的过程如下。专用核计数调整程序220a能够通过改变这一程序属于的内部系统控制器100a的专用核确定表113a来减少内部系统控制器100a或者外部系统控制器100b的专用核数目。由此,专用核计数调整程序220a优选地在更新内部系统控制器100a的专用核确定表之前更新外部系统控制器100b的专用核确定表113b。

[0133] 专用核确定表113包括关于内部系统控制器的MPU 101a和外部系统控制器的MPU 101b的信息,并且因此是由内部系统控制器和外部系统控制器共享的共享信息,并且获得访问以便更新这一信息是优先的处理的目标。因此,对外部系统控制器100b的专用核进行或者对普通核的中断通信区域143b进行内部系统控制器100a的访问。从内部系统控制器

100a对外部系统控制器100b的访问可以由专用核或者普通核进行。内部系统控制器的专用核计数调整程序220a等待来自外部系统控制器100b的响应并且修改内部系统控制器的专用核确定表113a。

[0134] 使对外部系统控制器100b的专用核确定表113b的更新优先于对内部系统控制器100a的专用核确定表113a的更新,因为外部系统控制器100b不能在内部系统控制器100a的专用核之前进行请求。注意,虽然内部系统控制器100a在等待来自外部系统控制器100b的响应之后更新专用核确定表113a,但是也可以更新内部系统控制器100a的表113a而未等待来自外部系统控制器100b的响应。

[0135] 在另一方面,如果在S2601中进行否定确定,则专用核计数调整程序220a确定通过从专用核的总负荷减去普通核的总负荷而获得的负荷差值是否等于或者大于阈值(S2603)。如果负荷等于或者大于阈值,则专用核计数调整程序220a增加专用核(S2604)。首先,专用核计数调整程序220a更新内部系统控制器100a的专用核确定表113a。由此,专用核计数调整程序220a假设其属性将被改变成专用核的普通核正在执行主机I/O处理或者其它处理并且在等待处理完成之后固定时间已经流逝之后将普通核改变成专用核。随后,内部系统控制器100a向外部系统控制器100b发布对于将专用核确定表113b更新作为优先的处理的请求。注意,专用核计数调整程序220a也可以中止正被处理的所有请求处理并且将普通核立即地切换成专用核。

[0136] 根据第二实施例,由于通过比较专用核负荷与普通核负荷来动态地改变专用核的比率,所以如果I/O处理负荷与优先的处理比较高,则提高普通核的比率以加速I/O处理,并且如果优先的处理与I/O负荷比较高,则提高专用核的比率以加速优先的处理。作为结果,即使当在多核处理器中配置专用于优先的处理的专用核时,也总是有在I/O处理与优先的处理之间产生的高平衡水平。根据第二实施例,如果专用核负荷为高,则已经被设置为接收中断的普通核可以请求对共享信息的访问

[0137] 实施例3

[0138] 第三实施例是第二实施例的修改并且其特征在于使用同时多线程化(SMT)。通过使用SMT,单个物理核可以被视为被称为线程的多个虚拟核。这些线程被称为虚拟核。根据这一实施例,在前述专用核确定表113中预备数目与虚拟核数目对应的条目。在核字段中记录虚拟核标识编号。这一实施例与第二实施例相同在于虚拟核中的一个或者多个虚拟核可以被设置为专用核。

[0139] 这一实施例使用SMT来使得有可能总是将虚拟核置于暂停状态(停止的状态),并且作为控制器在外部系统控制器的虚拟核中生成中断信号的结果来将虚拟核启动作为专用核。该实施例因此有利在于通过使具有专用核设置的虚拟核空闲来避免压缩其它虚拟核(普通核)的处理性能。另外,根据这一实施例,在多个虚拟核设置为专用核之后,停止这些核中的一个或者多个核并且将该一个或者多个核置于待命,并且如果由于故障等而停止尚未被停止的具有专用核设置的虚拟核,则可以通过向被置于待命的具有专用核设置的虚拟核传输核间中断并且从待命状态恢复虚拟核来高速地对虚拟核进行故障转移(failover)。

[0140] 实施例4

[0141] 接着将作为前述第二实施例的修改描述第四实施例。这一实施例的特征在于关于针对确定控制器的请求传输程序207a请求来自外部系统控制器100b的处理完成的要求进

行判决并且在于选择用于实现这一要求的请求系统。由于这些特性,根据第二实施例向MPU存储器的本地存储器11(图16)添加用于确定要求的表和用于选择用于实现这些要求的请求系统的表。前者的示例被称为要求系统要求表,并且后者的示例称为请求系统可用性表。

[0142] 图21是请求系统要求表114的示例。请求系统要求表114由请求传输程序207a指明。该表包括由请求传输程序207a指明的条目,这些条目用于记录外部系统控制器100b的访问目的地1140、用于访问目的地的访问类型1141和在外系统控制器的MPU 101b对访问目的地执行与访问类型有关的处理并且向请求传输源控制器100a发送回关于完成该处理的响应之前的预计时间(预计响应时间)1142。

[0143] 图22A是请求系统可用性表115的第一示例。请求系统可用性表115包括用于定义请求传输程序207a将向其传输请求信息的外部系统控制器100b的核的请求目的地字段1153、用于定义来自请求传输程序的通信类型的请求系统字段1150、用于定义请求目的地核是否能够处理请求的可用性字段1151和用于定义在请求源控制器向请求目的地控制器传输请求信息并且请求目的地核处理请求信息之后向请求源控制器发送回响应之前的时间的预计响应时间字段1152。预计响应时间根据在每个核上的负荷波动。配置管理程序连续地监视在每个核上的负荷并且根据要求来更新预计响应时间。

[0144] 图22B是请求系统可用性表115的第二示例。核0和核1是专用核。因此,向核0或者核1传输的请求无需是中断请求,而是可以是对于普通处理的请求。在专用核负荷低时,核0或者核1的预计响应时间低。因此,向专用核(比如核0或者核1)传输对于普通处理的请求一般比向普通核(比如核2)传输中断请求更有利。然而,在专用核(比如核0或者核1)的负荷变成高时,预计响应时间可以是向普通核(比如核2)传输中断请求比向专用核(比如核0或者核1)传输对于普通处理的请求更短。

[0145] 图23是根据这一实施例的请求传输程序207的流程图。与图5和图19的前述流程图的不同在于步骤S1203和后续步骤被步骤S1280至S1283替换。在跟随步骤S1202的步骤S1280中,请求传输程序207引用请求系统要求表114以获取硬件1140(该硬件是访问目的地)和预计响应时间1142作为与访问类型对应的要求。

[0146] 要求传输程序207然后引用请求系统可用性表115(步骤S1281)并且选择满足要求1142(预计响应时间)的、在请求目的地1153、请求系统1150和可用性1151的属性中的每个属性的组合(步骤S1282)。为了满足要求,在要求系统可用性表115中的预计响应时间字段1152中包含的预计响应时间等于或者少于在步骤S1280中获取的用于访问类型的预计响应时间1142。请求传输程序207a根据基于请求系统可用性表115选择的请求系统(也就是具有请求目的地核1153的通信系统(中断通信或者普通通信)1150)、由于向外部系统的请求而执行访问(步骤S1283)。

[0147] 专用核计数调整程序220通过确定用于每个核的负荷状态来更新请求系统可用性表115。如果普通核由于主机I/O处理而具有高负荷,则用于普通通信系统的长预计响应时间被设置为长。同时,如果负荷为低,则用于普通通信系统的预计响应时间被设置为短。在后一种情况下,中断通信系统的预计响应时间比普通通信系统的预计响应时间更短,而在后一种情况下相反。例如,根据排队模型等确定普通通信系统的预计响应时间。

[0148] 在图21中,由于第一控制器100a对第二控制器100b的共享存储器12b的访问(也就是读取、写入(非标示(non-posted)写入、原子更新))的完成响应时间为10微秒,所以在下

文中,支持向第一控制器100a的响应的系统根据图22A、图22B被称为用于第二控制器100b的处理器101b的非核选择普通通信系统或者核1普通通信系统。专用核的预计响应时间无需总是比普通核更短,并且中断系统的通信无需总是比普通系统的通信更短。使用哪个核以及哪个通信系统能够满足预计响应时间根据核负荷和I/O处理负荷而变化。在非核选择普通通信系统中,未指明请求目的地核,并且多个核中的任何一个或者多个核可以接收请求。将通信区域划分成用于个别核的通信区域,并且通信区域由核共享。在向用于个别核的通信区域写入请求时,设置核1普通通信系统。在向由核共享的通信区域写入请求时,设置非核选择普通通信系统。在设置非核选择普通通信系统时,响应时间变成一般比在指明请求目的地核时更短,因为已经完成I/O处理的核接收请求。在这样的情况下,需要独占控制以用于核从通信区域读取请求。

[0149] 实施例5

[0150] 现在将描述第五实施例,该第五实施例是第四实施例的修改。在第五实施例中,示出了如下示例,在该示例中,用户判决对于I/O处理的性能要求并且相应地校正用于通信的预计响应时间。示出了如下情况,其中对于I/O处理的性能要求是响应时间(在主机发布I/O命令之后直至从存储系统发送回响应的时间)。

[0151] 图24是卷性能要求表116的配置示例,在该卷性能要求表116中为每个卷判决前述性能要求。向MPU存储器102的本地存储器11添加卷性能要求表。卷性能要求表116包括多个条目,这些条目包括卷编号字段1161和预计通信响应时间校正值字段1162。卷编号字段1161存储用于标识在存储系统中的卷的编号,并且预计通信响应时间校正值字段1162支持由于对卷的I/O请求而在通信中校正在选择通信系统时使用的要求。

[0152] 图25示出了让用户能够选择对于卷的I/O性能要求的输入屏幕。在管理终端3上显示这一屏幕,并且用户经由这一屏幕输入对于每个卷的性能要求。屏幕显示用于定义性能要求的性能要求输入表300。性能要求输入表300显示如下条目,这些条目包括两个输入字段,即卷,编号字段3001和预计I/O响应时间字段3002。

[0153] 用户输入用于卷(确定在每个字段中对于这些卷的要求)的卷编号3001并且为用于卷的I/O响应输入预计响应时间作为预计I/O响应时间3002。在接收输入时,管理终端3向存储系统4输出信息。存储配置管理程序201向卷性能要求表116添加新条目、输入向添加的条目的卷编号字段输入的卷编号并且在预计通信响应时间校正值字段中存储从这样输入的预计I/O响应时间确定的校正值。

[0154] 发现校正值的方式例如是通过确定标准I/O响应时间,并且用于标准I/O响应时间的预计I/O响应时间的比值可以是预计通信响应时间校正值。因此,对于具有短预计I/O响应时间的卷,选择与具有长预计I/O响应时间的卷比较具有短通信响应时间的请求系统。在第四实施例中描述了用于请求短通信响应时间的系统。因而,如果有对具有不同性能要求的卷的I/O请求的混合,则使对具有严格性能要求的卷的I/O处理优先,从而使得可以在总体上满足每个卷的性能要求。

[0155] 在图26中示出了用于校正预计通信响应时间的具体过程。图26是图23的修改。不同首先在于在调用请求传输程序207时,从基本I/O程序202(该基本I/O程序202是调用器程序)向请求传输程序207传递如下信息,该信息指示调用与对特定卷的I/O处理关联。这一信息包括卷编号。这被称为目标卷编号。另外,除了图23的流程之外还在步骤1280之前添加步

骤S1290。

[0156] 在步骤S1290中,[请求传输程序207]引用卷性能要求表116以获取在用于与目标卷编号对应的条目的预计通信响应时间校正值字段1162中存储的值。这一个值然后通过与在后继步骤S1280中获取的预计响应时间值相乘被校正。通过取校正的预计响应时间值作为通信系统要求并且执行后继步骤1281的处理和后续处理,可以根据每个卷的性能要求来选择通信系统。作为结果,可以选择用于通信响应时间的预计值为小的通信系统以用于与被请求I/O响应性能的卷的I/O处理关联的通信,并且可以满足对于该卷的I/O性能要求。

[0157] 注意,在这一实施例中,示出了如下示例,在该示例中借助一种对根据对于每个卷的I/O性能要求而确定的系数进行积分的方法来校正预计响应时间,但是也可以根据预定预计通信响应时间是否降至预定阈值以下来将例如用户输入的对于每个卷的I/O性能要求校正成该时间。另外,虽然在这一实施例中为每个卷判决性能要求,但是也可以例如为发布I/O的每个主机确定性能要求。此外,取代使用响应时间作为I/O性能要求,性能要求可以由吞吐量性能(对于每个时间单位的I/O命令数目或者向主机发送回响应的数据量)确定。在这一情况下,请求的吞吐量性能的值越高,设置的用于与I/O处理关联的通信的预计响应时间就越短。

[0158] 注意,本发明不限于前述实施例,而是包括各种修改。例如,具体描述前述实施例以便易于理解的方式描述本发明,但是本发明未必限于包括描述的所有配置。另外,某个实施例的配置中的一些配置也可以被来自其它实施例的配置替换,并且也可以向某个实施例的配置添加来自其它实施例的配置。另外,其它配置也可以被添加到每个实施例的配置、从每个实施例的配置被删除或者替换每个实施例的配置中的一些配置。另外,也可以例如借助LSI设计等通过硬件实施前述配置、功能、处理单元和处理装置等中的一些或者所有配置、功能、处理单元和处理装置等。另外,也可以通过软件实现配置和功能等中的每个配置和功能等作为处理器解译和执行程序的结果,这些程序实现相应功能。此外,可以示出视为说明书必需的控制接线和连接接线,但是出于制造目的而未必要示出所有接线。在现实中,也可以考虑人工地连接几乎所有配置。

[0159] 在第一、第二、第三、第四和第五实施例中将来自控制器的处理器的请求被写入到的通信区域划分成中断通信区域和普通通信区域。然而,不是必须划分并且因此可以共享通信区域。因此,基于用于优先的处理的中断信号是否存在,可以在优先的处理与非优先的处理(普通处理)之间切换处理。

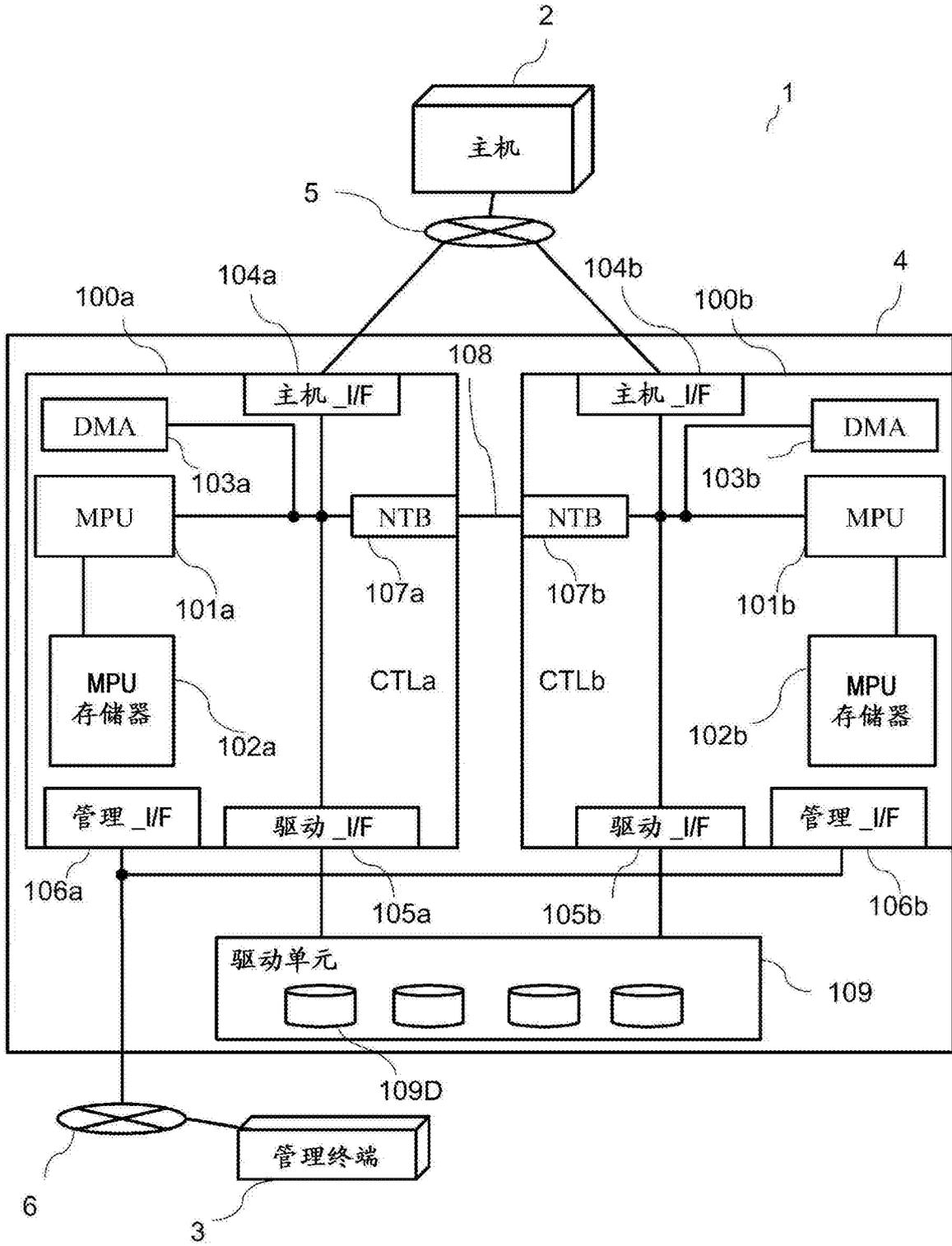


图1

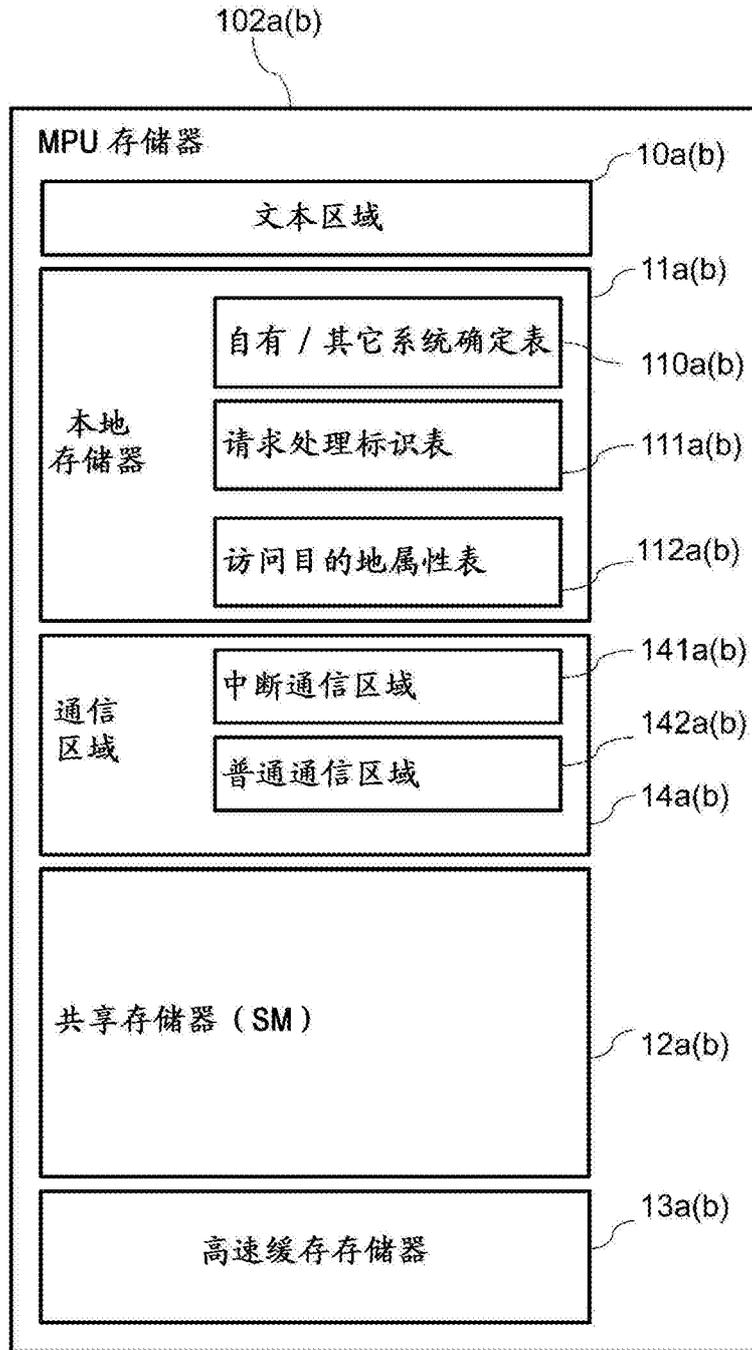


图2



图3

请求传输程序

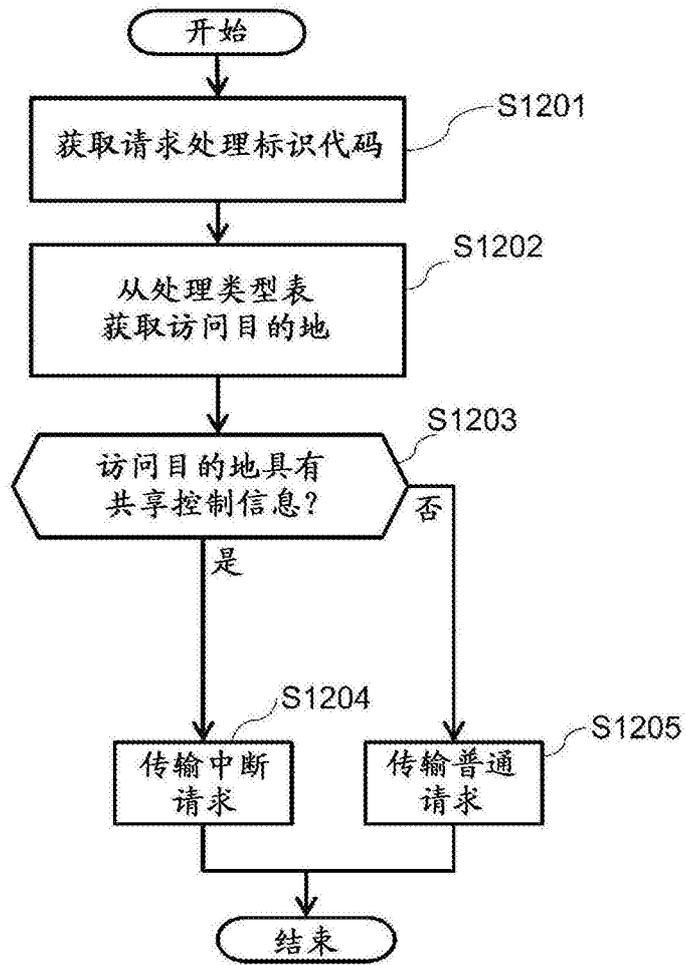


图4

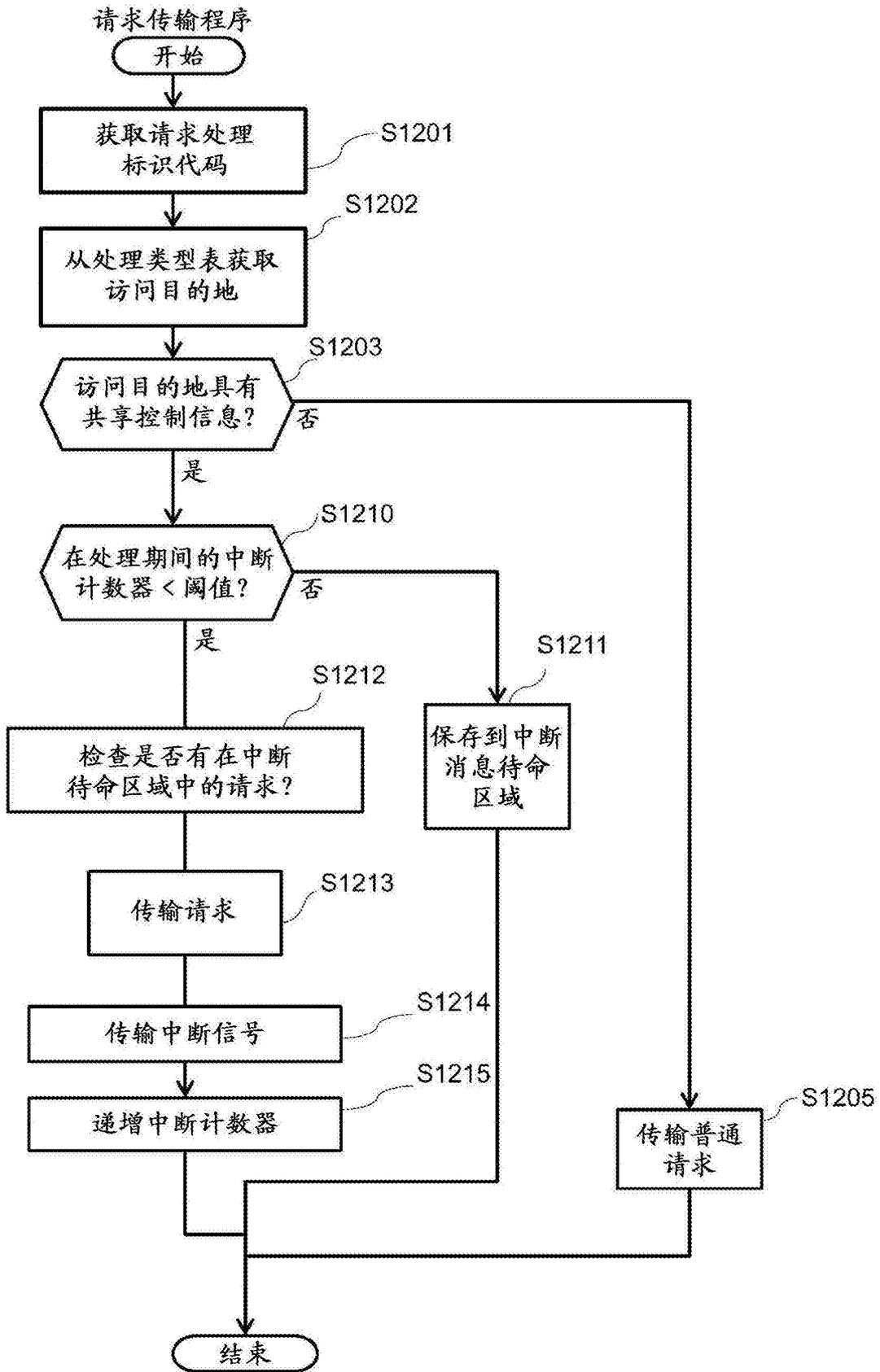


图5

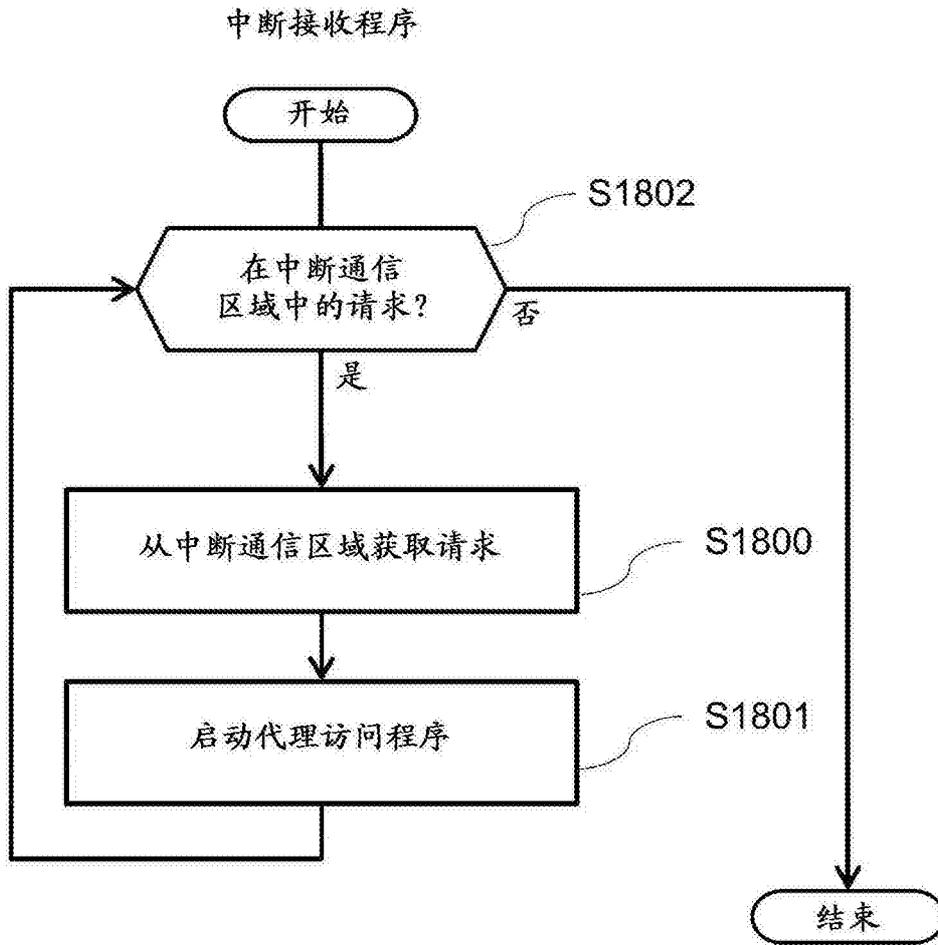


图6

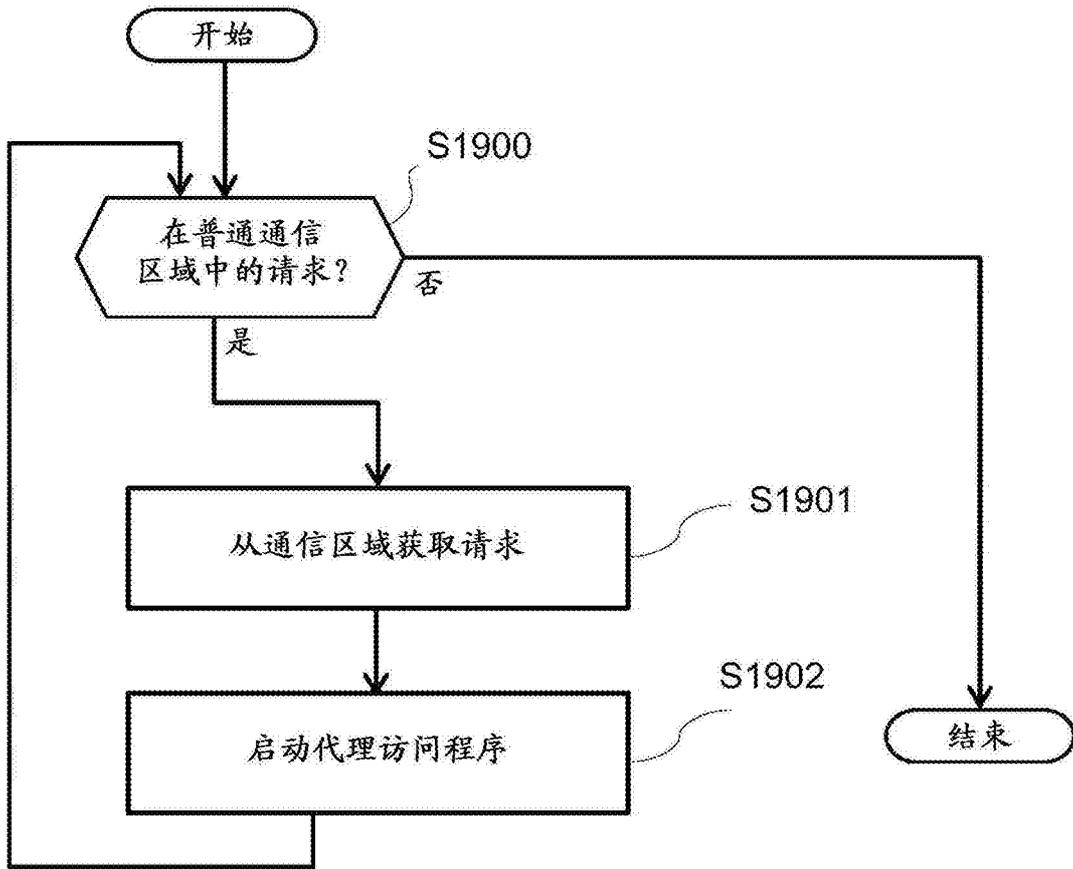


图7

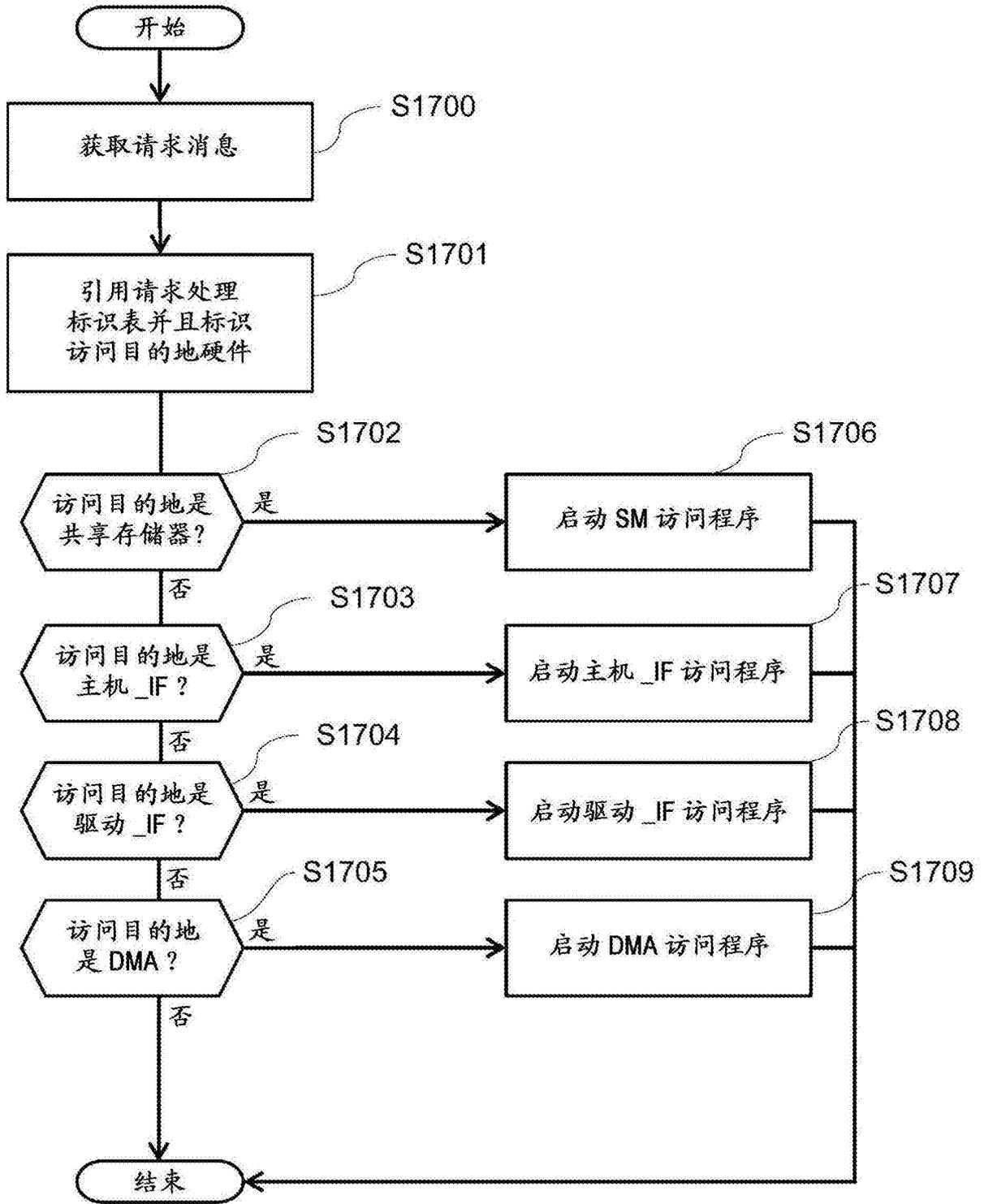


图8

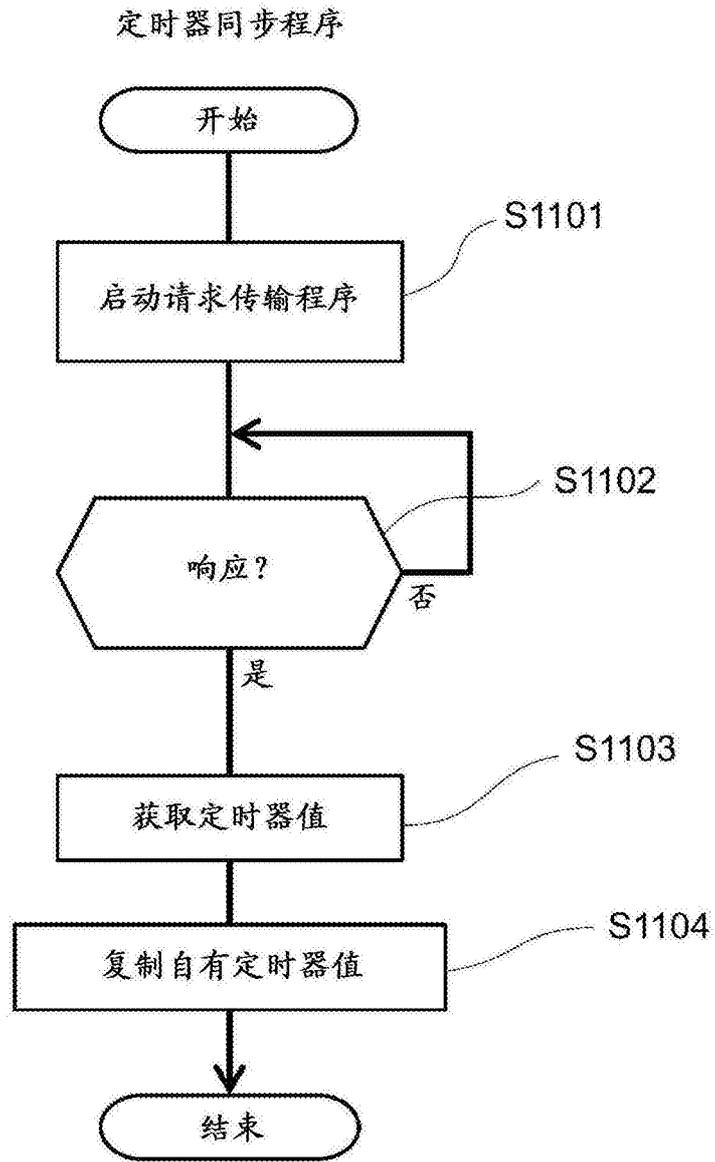


图9

请求处理标识表 111

标识代码	通信类型	访问目的地
0x00	读取	共享存储器
0x01	写入	共享存储器
0x02	原子	共享存储器
0x10	激活传送	DMA
0x11	零数据文件	DMA
0x20	读取传送	主机_IF
0x21	写入传送	主机_IF
0x30	读取传送	驱动_IF
0x31	写入传送	驱动_IF
0x40	读取	定时器
0x41	写入	定时器
0x50	奇偶校验生成	奇偶校验生成
...	...	

图10

访问目的地属性表 112

1120

1121

访问目的地	共享信息
共享存储器	是
DMA	否
主机_IF	否
驱动_IF	否
定时器	是
奇偶校验生成器	否
...	...

图11

自有 / 其它系统确定表 110

ID	类型	CTL
0x00	共享存储器	a
0x01	共享存储器	b
0x10	主机_IF	a
0x11	主机_IF	a
0x12	主机_IF	b
0x13	主机_IF	b
...

图12

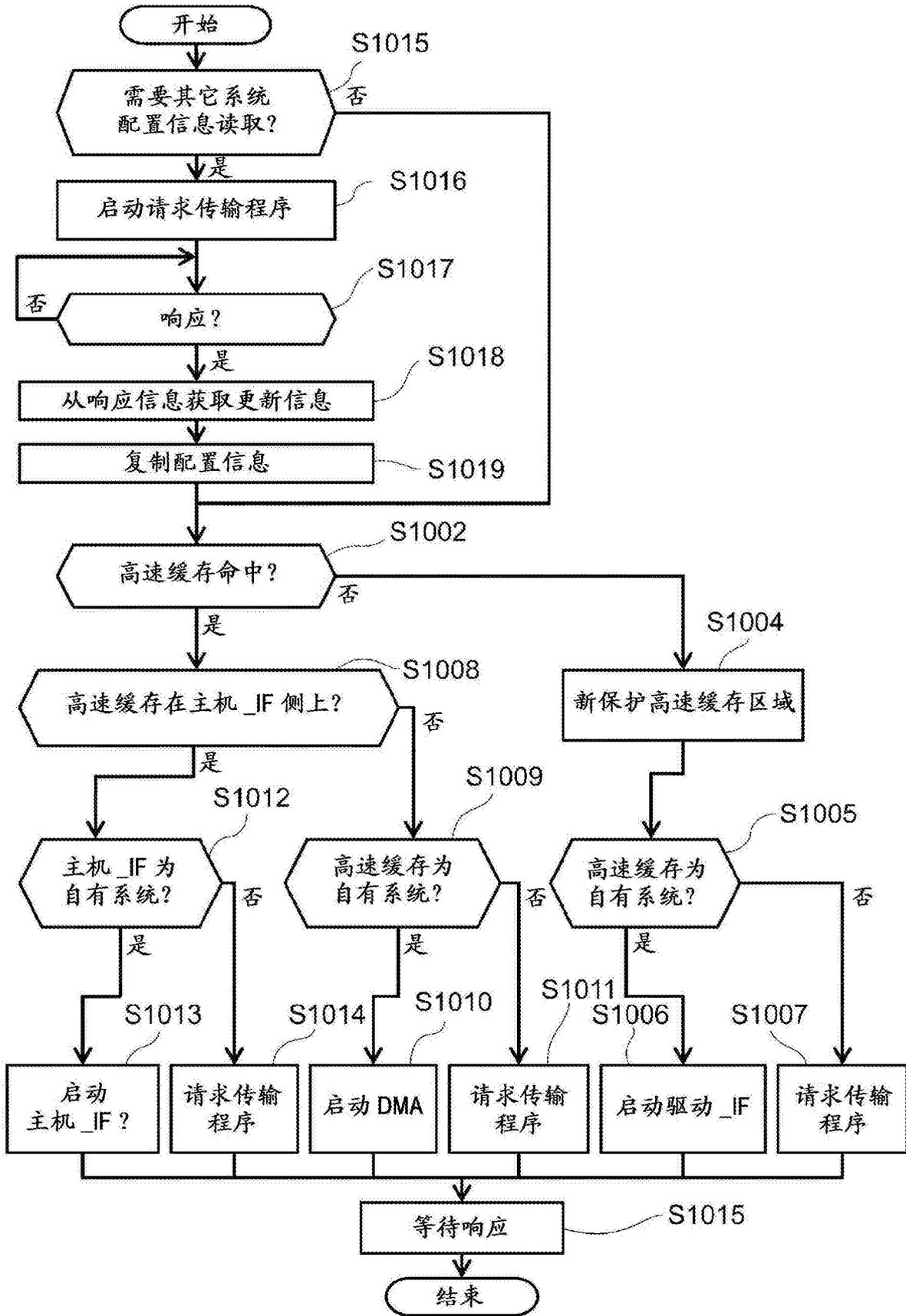


图13

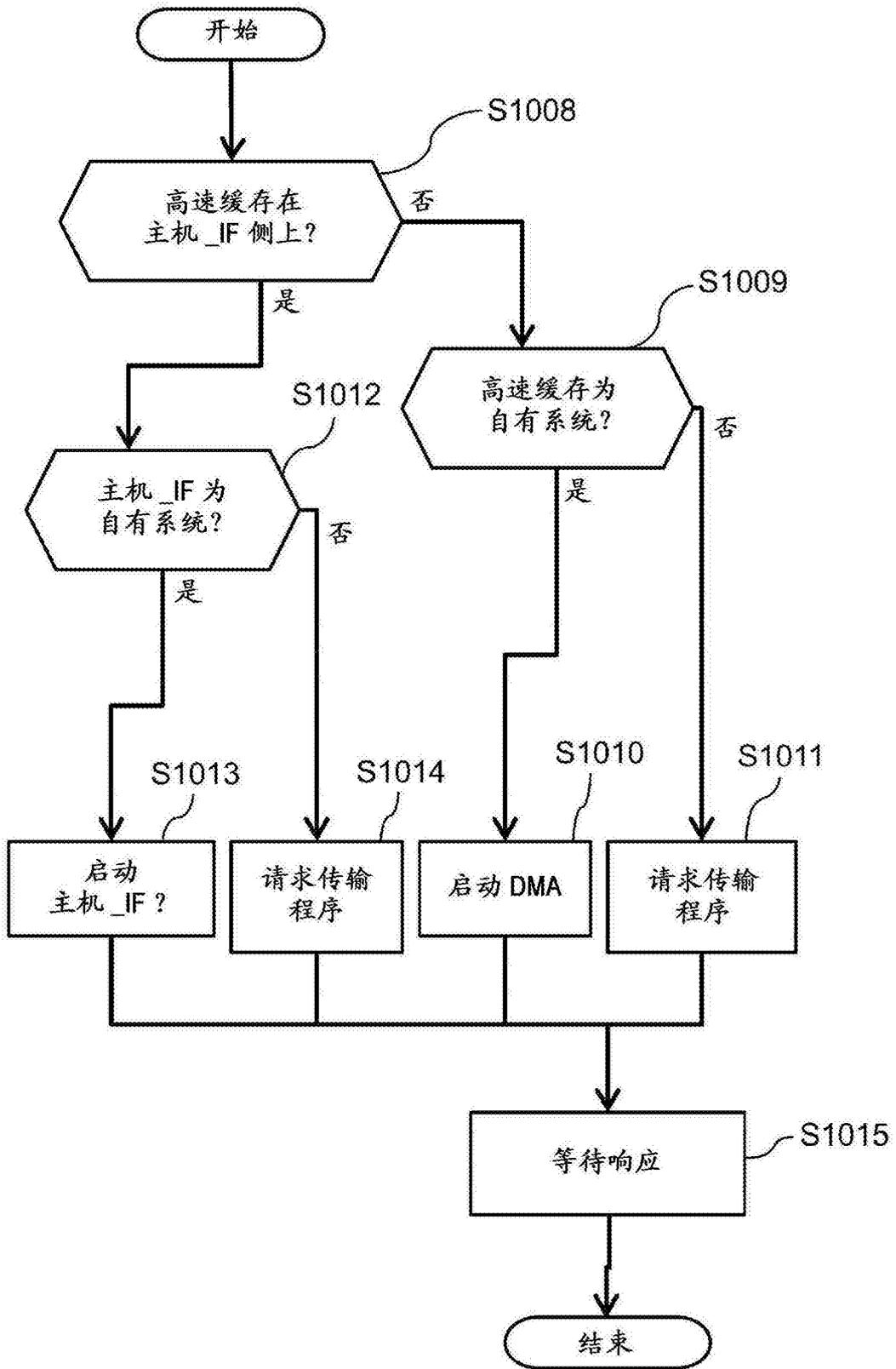


图14

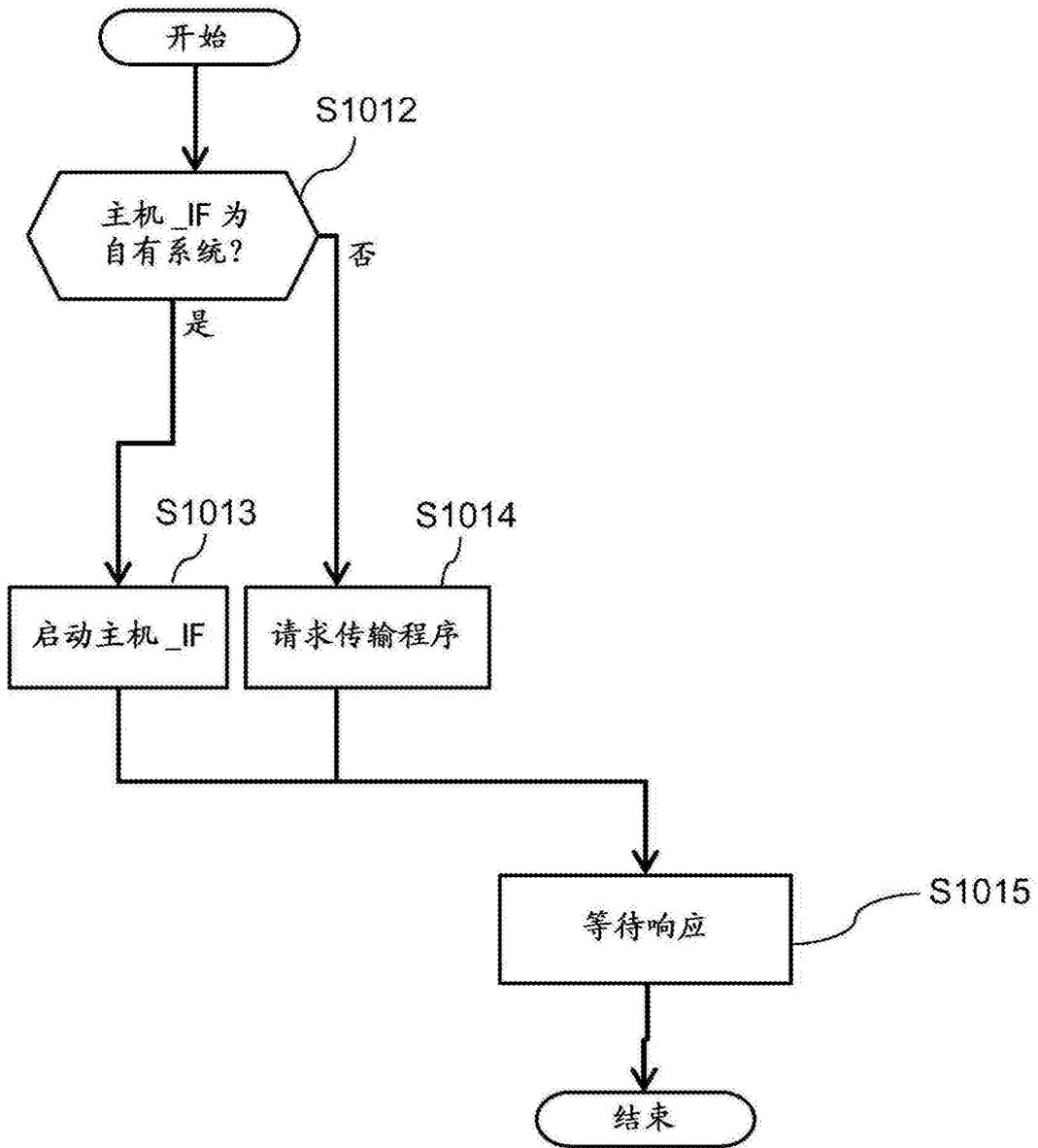


图15

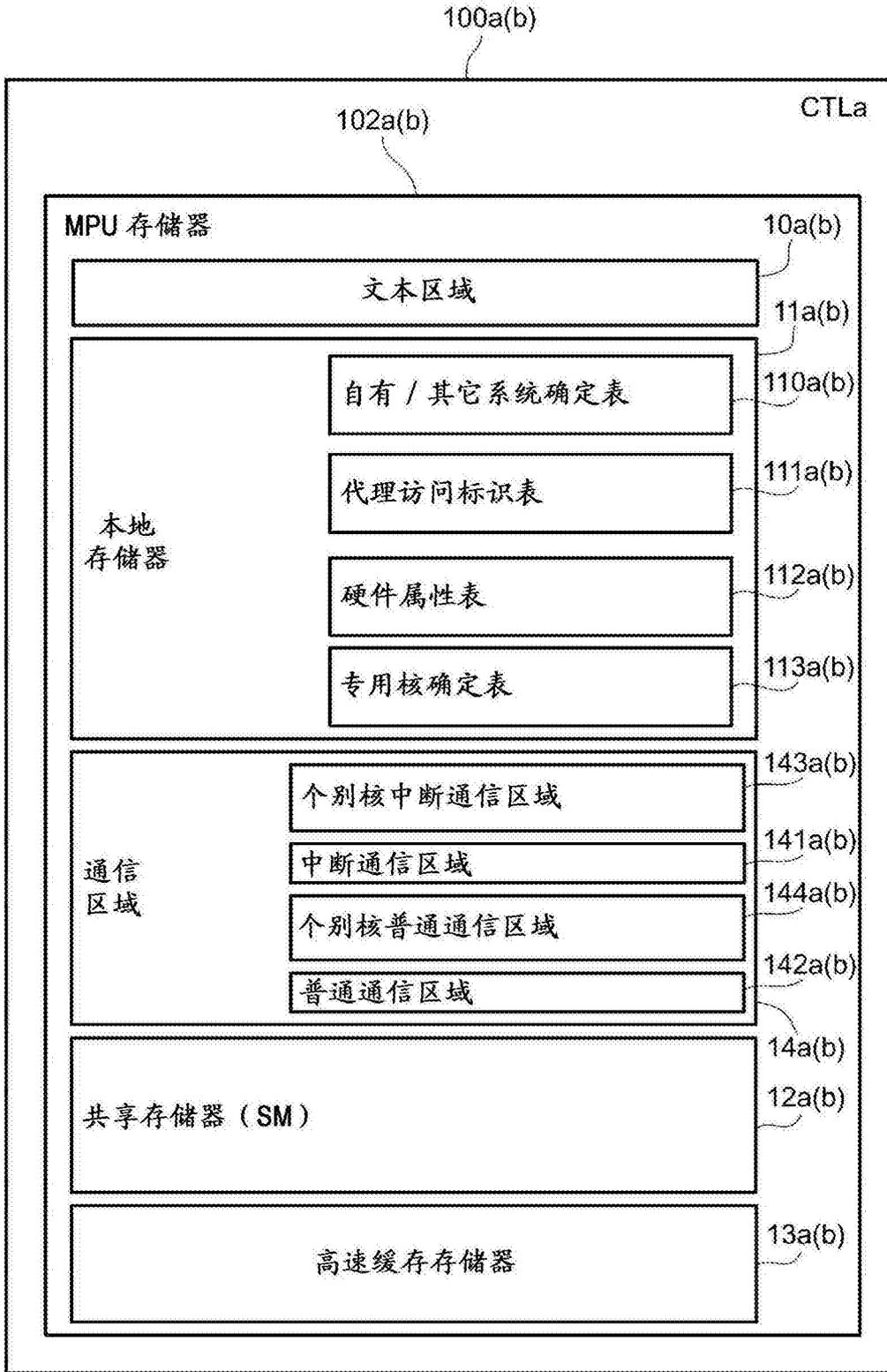


图16

专用核确定表 113

核	核类型	中断可接收性
CTL0-0	专用核	-
CTL0-1	普通核	可能
CTL0-2	普通核	可能
CTL0-3	普通核	不可能
CTL1-0	专用核	-
CTL1-1	普通核	可能
CTL1-2	普通核	可能
CTL1-3	普通核	不可能
...

图17

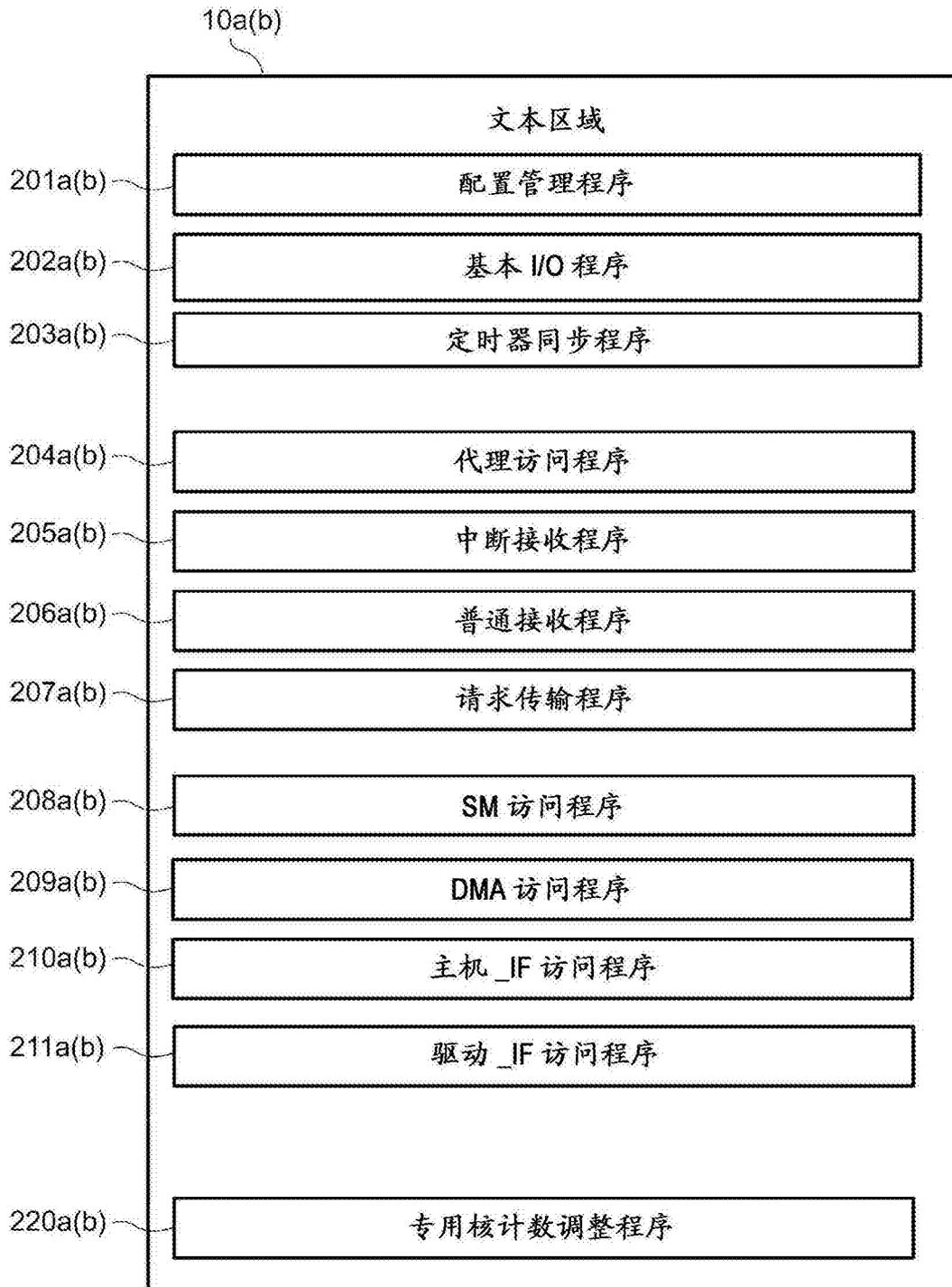


图18

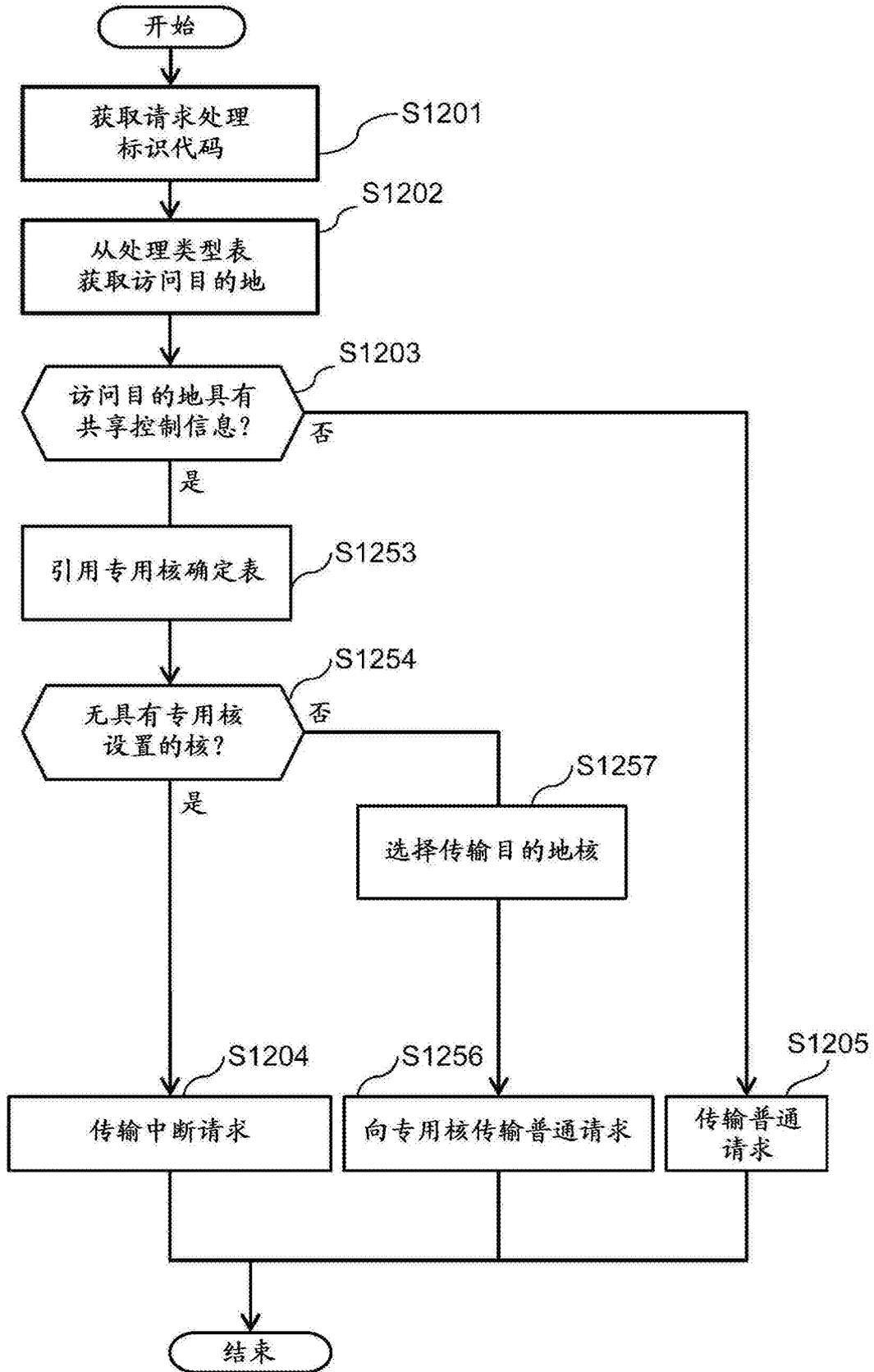


图19

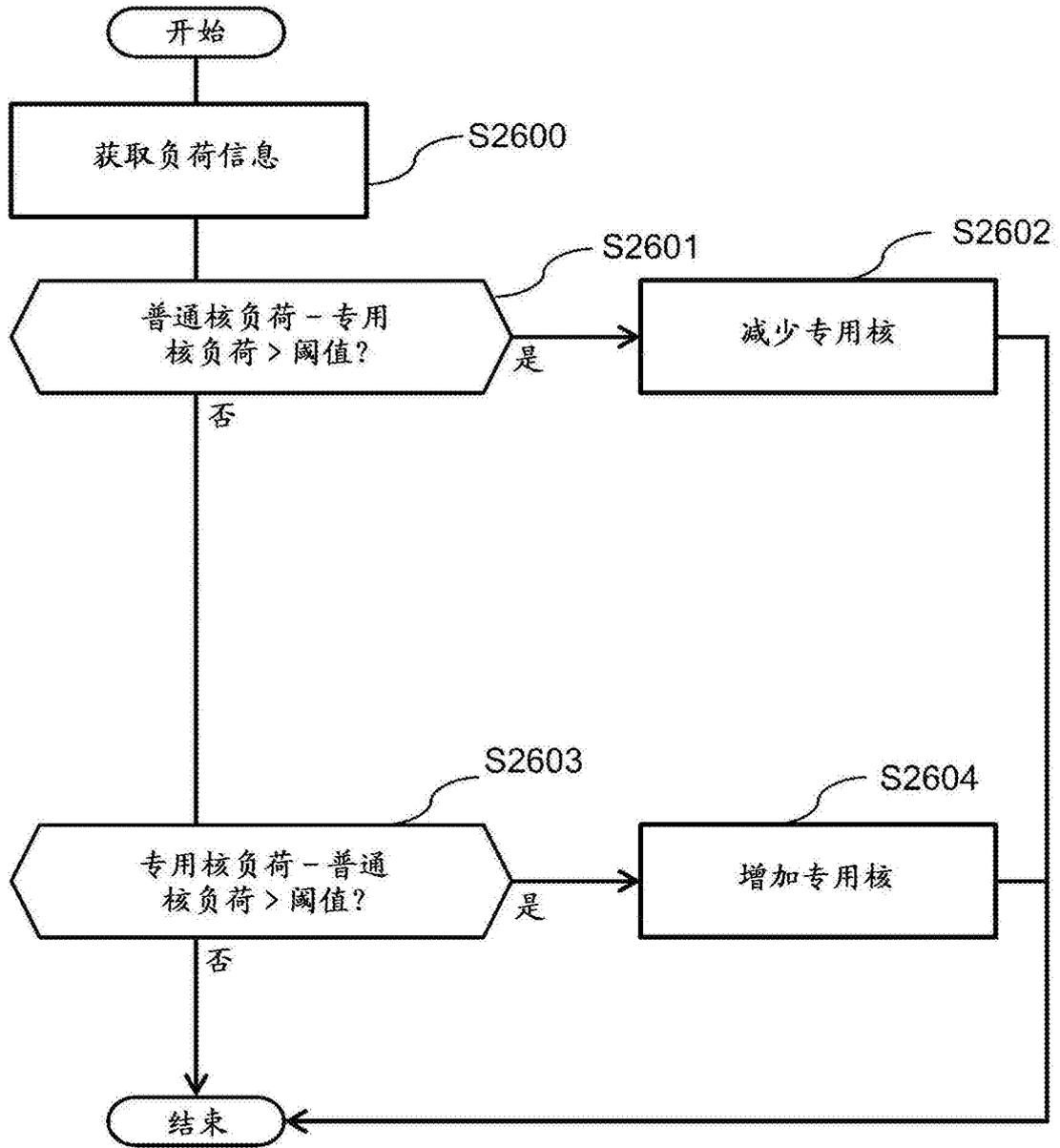


图20

请求系统要求表 114

访问目的地	访问类型	预计响应时间
共享存储器	读取	10
	非标示写入	10
	原子更新	10
DMA	启动传送	100
主机_IF	读取传送	100
	写入传送	100
驱动_IF	读取传送	100
	写入传送	100
定时器	读取	10
	写入	10
...

图21

请求系统可用性表 115

1153 请求目的地	1150 请求系统	1151 可用性	1152 预计响应时间
核 0	中断	○	10
	普通	×	-
核 1	中断	○	10
	普通	○	5
核 2	中断	○	10
	普通	○	30
核 3	中断	×	-
	普通	×	-
未选择的核	中断	○	10
	普通	○	2
***		***	***

图22A

请求系统可用性表 115

请求目的地	请求系统	可用性	预计响应时间
核 0	中断	×	-
	普通	○	5
核 1	中断	×	-
	普通	○	30
核 2	中断	○	10
	普通	○	30
核 3	中断	×	-
	普通	×	-
未选择的核	中断	○	10
	普通	○	15
...	

图 22B

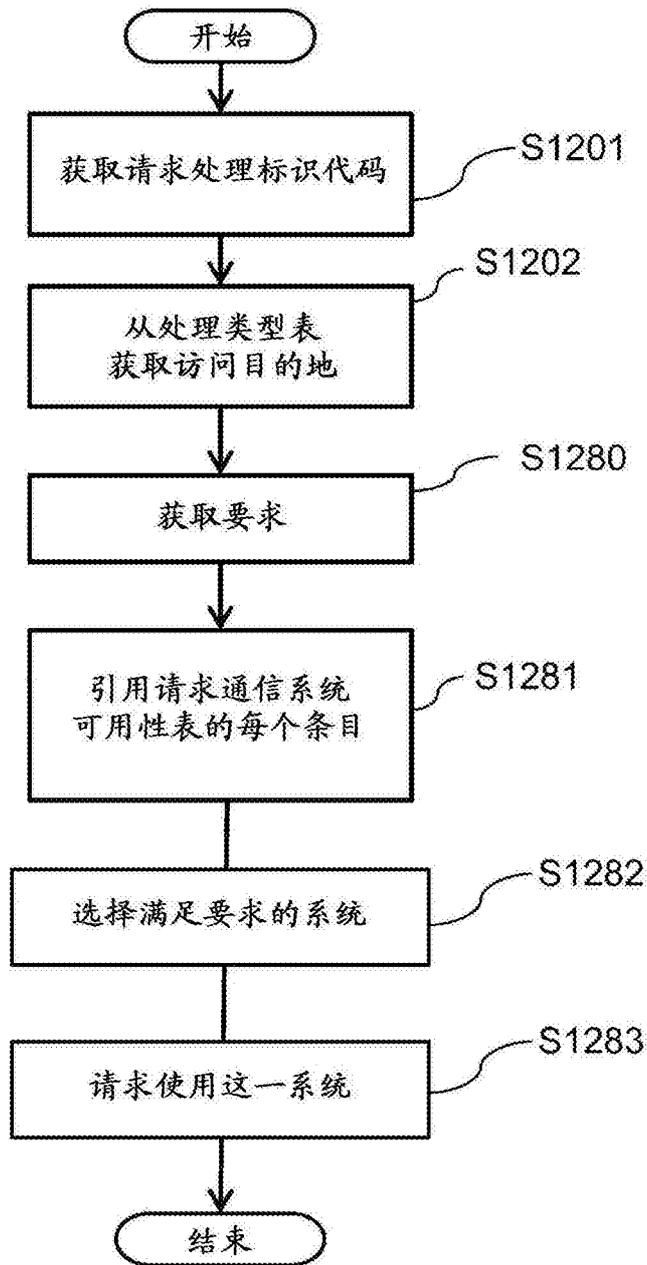


图23

卷性能要求表 116

卷编号	预计通信响应时间校正值
0	1.0
1	0.8
2	0.5
3	1.5
...	...

图24

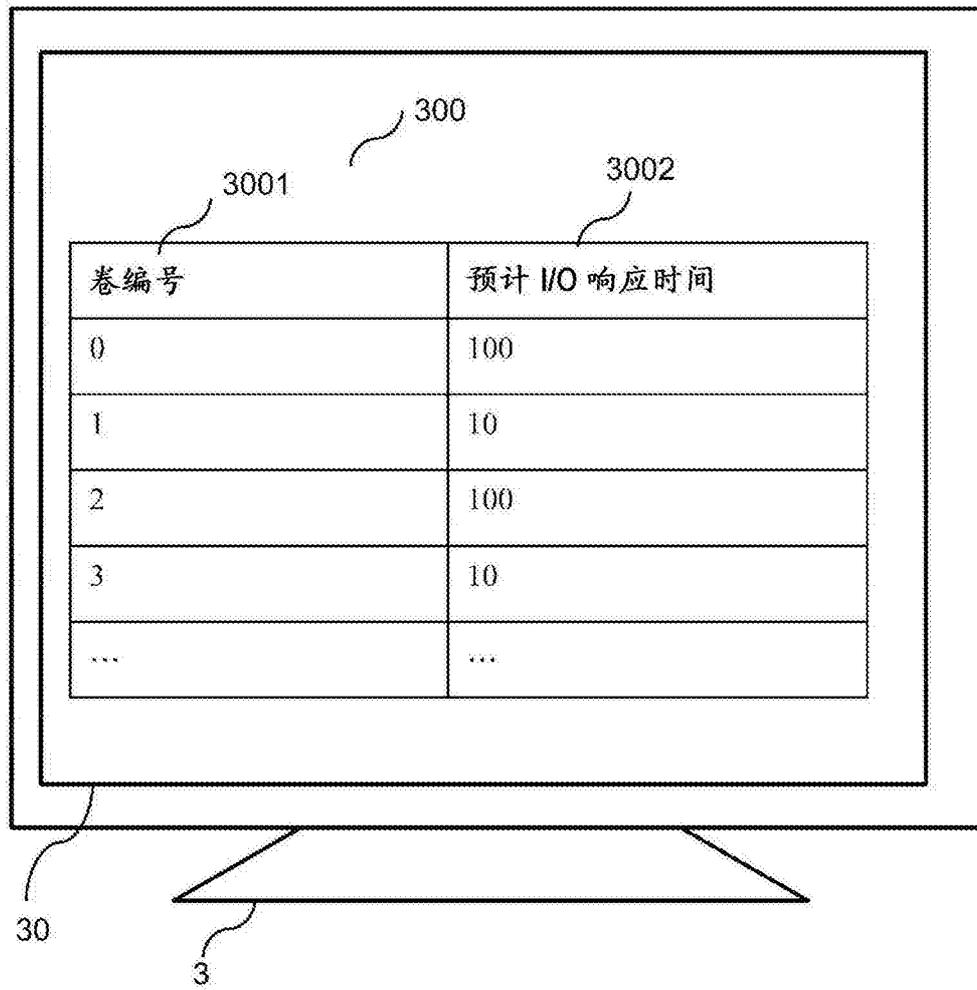


图25

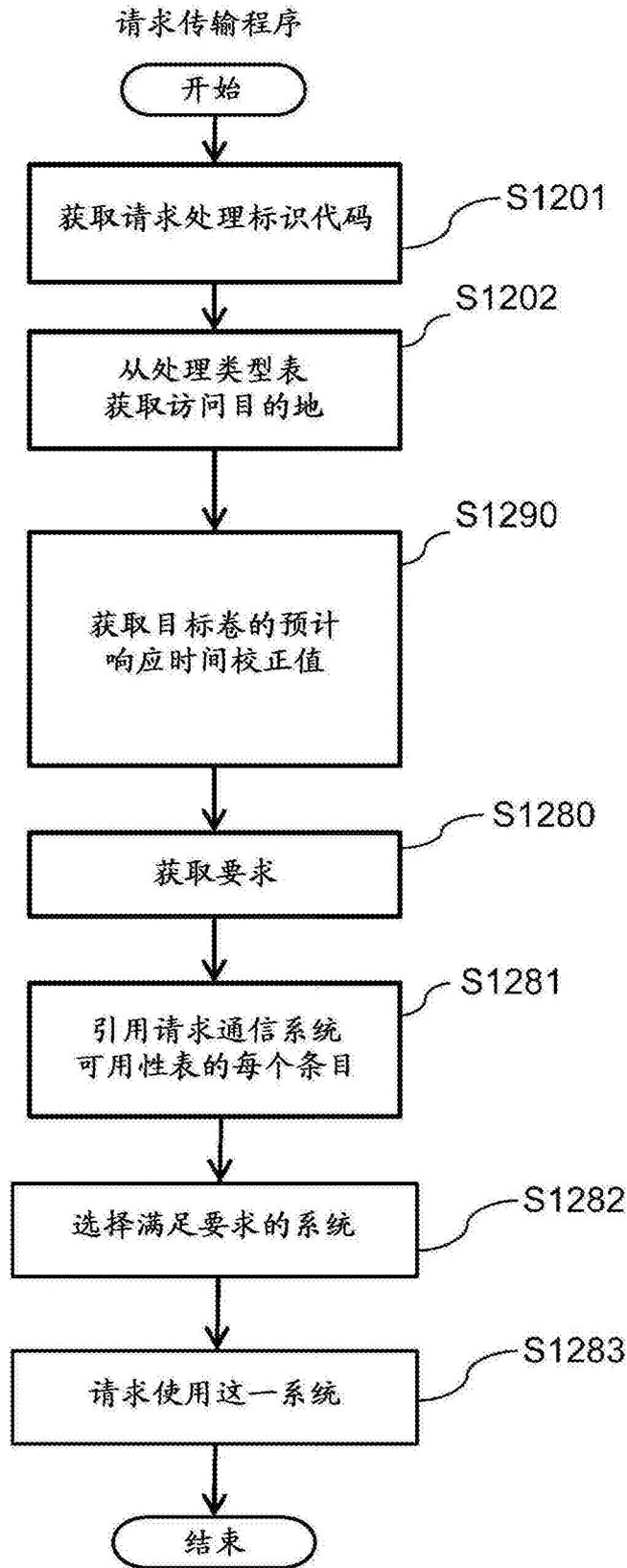


图26