(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103294616 B
(45)授权公告日 2016.01.20

(21)申请号 201210390225.3
(22)申请日 2012.10.15
(30)优先权数据
2012-044848 2012.02.29 JP
(73)专利权人 富士通株式会社
地址 日本神奈川县
(72)发明人 鲤沼秀之 出井裕之
(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 朱胜 陈炜
(51)Int.Cl.
G06F 12/16(2006.01)
(56)对比文件
US 2012036412 A1,2012.02.09,
JP H11175409 A,1999.07.02,
US 2007220313 A1,2007.09.20,
审查员 赵鹏翔

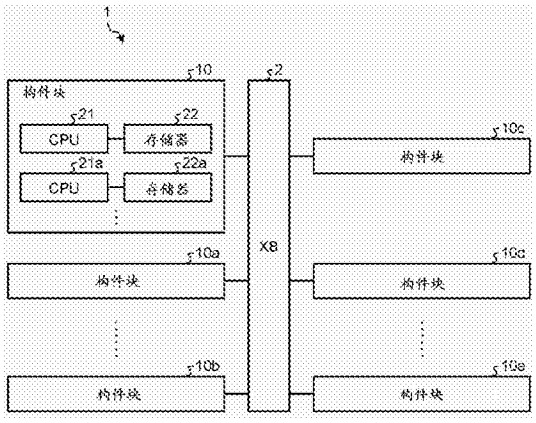
权利要求书2页 说明书23页 附图26页

(54)发明名称

信息处理设备和控制方法

(57)摘要

本发明公开了一种信息处理设备和控制方法。信息处理设备中的多个节点中的至少一个节点执行用于包括在一个节点或其它节点的存储器中且存储在一个节点和其它节点访问的共享存储器区域中的数据的以下处理。即,该节点检测在预定时间内发生超过预定次数的ICE或在共享存储器区域中的单一位置处发生的PCE。当检测到错误时,节点执行控制以阻止该节点和其它节点访问共享存储器。节点在不同于共享存储器区域的存储器区域中恢复数据。节点向其它节点通知关于不同存储器区域的信息。节点执行控制以重新开始从节点和其它节点对数据的访问。



CN 103294616 B

1. 一种信息处理设备，包括：

多个节点，每个节点均包括存储装置；和

互连装置，连接在所述多个节点之间，

其中，所述多个节点中的至少一个节点包括：

检测单元，检测包括在所述一个节点或其它节点中的存储装置中的共享存储器区域中所存储的数据的可校正错误，所述共享存储器区域是所述一个节点和所述其它节点访问的区域，并且所述可校正错误是 (i) 在预定的时间段内发生多于预定次数的错误或者 (ii) 在所述共享存储器区域中的单一位置处发生的错误；

阻止控制单元，当所述检测单元检测到所述可校正错误时，执行控制以通过从第一地址转换信息中删除将在其中检测到所述可校正错误的所述共享存储器区域中的虚拟地址与物理地址相关联的条目、并且通过将用以从第二地址转换信息中删除所述条目的指令发送至所述其它节点来阻止所述一个节点和所述其它节点访问所述共享存储器区域，其中所述第一地址转换信息将用于所述一个节点和所述其它节点的存储器访问的虚拟地址与表示所述一个节点的存储装置中的数据存储区域的物理地址相关联，所述第二地址转换信息将用于所述一个节点和所述其它节点的存储器访问的虚拟地址与表示所述其它节点的存储装置中的数据存储区域的物理地址相关联；

恢复单元，在与所述共享存储器区域不同的存储器区域中恢复存储在所述共享存储器区域中的所述数据；

通知单元，向所述其它节点通知关于该不同的存储器区域的信息；以及

重新开始控制单元，执行控制以重新开始从所述一个节点和所述其它节点对所恢复的数据的访问。

2. 根据权利要求 1 所述的信息处理设备，其中，所述阻止控制单元将用于停止从所述其它节点执行的应用程序对所述共享存储器的访问的指令发送至所述其它节点，并且停止从所述一个节点执行的应用程序对所述共享存储器的访问。

3. 根据权利要求 1 或权利要求 2 所述的信息处理设备，其中，所述恢复单元获取其它存储器区域的页面以复制包括所述共享存储器区域的错误的页面，并且将存储在包括所述共享存储器区域的错误的页面的区域中的数据复制到所述其它存储器区域的页面的区域。

4. 根据权利要求 3 所述的信息处理设备，其中，所述恢复单元将所述共享存储器区域的页面分割成多个页面，并且将存储在包括所述错误的分割后的页面的区域中的数据复制到所述其它存储器区域的页面的区域。

5. 根据权利要求 3 所述的信息处理设备，其中，当获取所述其它存储器区域的页面时，如果可获取页面容量是预定容量或更小，或者所述其它存储器区域的页面的获取失败，则所述恢复单元分割所述共享存储器区域的页面。

6. 根据权利要求 3 所述的信息处理设备，其中，当在本地节点中启动操作系统 OS 或者启动使用本地节点或其它节点的共享存储器的应用程序时，所述恢复单元预先确保能够用作所述其它存储器区域的预定大小的区域。

7. 一种信息处理设备，包括：

多个节点，每个节点均包括存储装置；以及

互连装置，连接在所述多个节点之间，

其中，所述多个节点中的至少一个节点包括：

访问控制单元，控制对所述一个节点的所述存储装置的访问，并且包括检测从所述存储装置读取的数据的错误的错误检测单元；以及

处理单元，执行处理，所述处理包括：

当所述错误检测单元检测到存储在共享存储器区域中的数据的可校正错误时，通过从第一地址转换信息中删除将在其中检测到所述可校正错误的所述共享存储器区域中的虚拟地址与物理地址相关联的条目、并且通过将用以从第二地址转换信息中删除所述条目的指令发送至其它节点来阻止从所述一个节点和所述其它节点对所述共享存储器区域的访问，其中所述第一地址转换信息将用于所述一个节点和所述其它节点的存储器访问的虚拟地址与表示所述一个节点的存储装置中的数据存储区域的物理地址相关联，所述第二地址转换信息将用于所述一个节点和所述其它节点的存储器访问的虚拟地址与表示所述其它节点的存储装置中的数据存储区域的物理地址相关联，所述共享存储器区域包括在所述一个节点的存储装置中且通过所述一个节点和所述其它节点访问，所述可校正错误是 (i) 在预定的时间段内发生多于预定次数的错误或者 (ii) 在所述共享存储器区域中的单一位置处发生的错误；

在不同于所述共享存储器区域且包括在所述一个节点的所述存储装置中的存储器区域中恢复存储在所述共享存储器区域中的所述数据；

向所述其它节点通知关于该不同的存储器区域的信息；以及

重新开始从所述一个节点和所述其它节点对所恢复的数据的访问。

8. 一种用于信息处理设备的控制方法，所述控制方法由所述信息处理设备中的多个节点中的至少一个节点执行，所述信息处理设备包括每一个均包括存储装置的多个节点以及连接在所述多个节点之间的互连装置，所述控制方法包括：

检测包括在所述一个节点或其它节点的存储装置中的共享存储器区域中所存储的数据的可校正错误，所述共享存储器区域是所述一个节点和所述其它节点访问的区域，并且所述可校正错误是 (i) 在预定的时间段内发生多于预定次数的错误或者 (ii) 在所述共享存储器区域中的单一位置处发生的错误；

当检测到所述可校正错误时，执行控制以通过从第一地址转换信息中删除将在其中检测到所述可校正错误的所述共享存储器区域中的虚拟地址与物理地址相关联的条目、并且通过将用以从第二地址转换信息中删除所述条目的指令发送至所述其它节点来阻止所述一个节点和所述其它节点访问所述共享存储器区域，其中所述第一地址转换信息将用于所述一个节点和所述其它节点的存储器访问的虚拟地址与表示所述一个节点的存储装置中的数据存储区域的物理地址相关联，所述第二地址转换信息将用于所述一个节点和所述其它节点的存储器访问的虚拟地址与表示所述其它节点的存储装置中的数据存储区域的物理地址相关联；

在不同于所述共享存储器区域的存储器区域中恢复存储在所述共享存储器区域中的所述数据；

向所述其它节点通知关于该不同的存储器区域的信息；以及

执行控制以重新开始从所述一个节点和所述其它节点对所恢复的数据的访问。

信息处理设备和控制方法

技术领域

[0001] 在本文中讨论的实施例针对一种信息处理设备、控制方法以及控制程序。

背景技术

[0002] 传统上，多个计算处理装置共享存储装置的 SMP（对称多处理器）技术是公知的。应用这样的 SMP 技术的信息处理系统的示例是如下信息处理系统（信息处理设备）：其通过单条总线连接每一个均具有计算处理装置和存储装置的多个节点，并且在其中，每个计算处理装置共享每个存储装置。即，该信息处理系统具有在多个节点之间共享的存储装置（共享存储器）。

[0003] ECC ICE（错误检查和校正间歇的可校正错误）或 ECC PCE（永久的可校正错误）发生在这样的共享存储器的数据中。同时，“ECC ICE”是间歇的可校正错误。即，“ECC ICE”是在预定时间内发生超过预定次数的可校正错误。另外，“ECC PCE”是固定的可校正错误。即，“ECCPCE”是在存储器区域中的单一位置处发生的可校正错误。

[0004] 另外，存在这样的一种技术：当在给定页内频繁地执行错误校正时，将该页的内容从包括错误被校正的位置的第一存储器区域复制到第二存储器区域，并且将在 TLB 中的物理页从第一存储器区域的地址写入到第二存储器区域的地址。

[0005] 此外，存在这样的一种技术：当在对共享存储器进行访问时发生了可校正的一位错误时，在给定装置写回数据的同时阻止另一装置访问给定装置正在访问的存储器。

[0006] 专利文献 1：第 11-175409 号日本早期公开专利公布

[0007] 专利文献 1：第 09-128303 号日本早期公开专利公布

[0008] 专利文献 1：第 08-077078 号日本早期公开专利公布

[0009] 然而，上述技术具有如下问题：如果 CE（诸如 ICE 或 PCE）留在存储器上，则在一些情况下（诸如，在已存在 ICE 或 PCE 的存储器空间发生软错误），信息处理设备发生故障（go down）。

[0010] 因此，本发明的实施例的一个方面的目的是抑制信息处理设备发生故障的可能性。

发明内容

[0011] 根据实施例的一方面，一种信息处理设备包括每一个均包括存储装置的多个节点和连接在多个节点之间的互连装置，其中多个节点中的至少一个节点包括：检测单元，检测包括在这一个节点或其它节点的存储装置中的共享存储器区域中所存储的数据的可校正错误，该共享存储器区域是这一个节点和其它节点访问的区域，并且该可校正错误是（i）在预定时间段内发生多于预定次数的错误或者（ii）在共享存储器区域中的单一位置处发生的错误；阻止控制单元，当检测单元检测到可校正错误时，执行控制以阻止这一个节点和其它节点访问共享存储器区域；恢复单元，在与共享存储器区域不同的存储器区域中恢复存储在共享存储器区域中的数据；通知单元，向其它节点通知关于不同的存储器区域的信息；

以及重新开始(resumption)控制单元,执行控制以重新开始从这一个节点和其它节点对所恢复的数据的访问。

附图说明

[0012] 图 1 是用于说明根据实施例的信息处理系统的示例的图;

[0013] 图 2 是用于说明根据实施例的构件块(building block)的功能配置的图;

[0014] 图 3 是示出当另一节点附接到被分配了共享存储器的节点时存储器映射的示例的图;

[0015] 图 4 是用于说明根据实施例的 CPU 的功能配置的图;

[0016] 图 5 是用于说明根据实施例的节点映射的数据配置的示例的图;

[0017] 图 6 是用于说明根据实施例从 CPU 发送的分组(packet)的图;

[0018] 图 7 是用于说明根据实施例从 CPU 发送请求的处理的示例的图;

[0019] 图 8 是用于说明根据实施例当 CPU 接收分组时所执行的处理的示例的图;

[0020] 图 9 是用于说明根据实施例从 I/O 装置发送请求的处理的示例的图;

[0021] 图 10 是用于说明根据实施例在 I/O 装置处接收响应的处理的示例的图;

[0022] 图 11 是用于说明控制共享区域的处理的流程的流程图;

[0023] 图 12 是用于说明分配共享存储器的处理的流程图;

[0024] 图 13 是用于说明共享存储器附接处理的流程图;

[0025] 图 14 是用于说明在应用程序中使用共享存储器的处理的流程图;

[0026] 图 15 是用于说明拆卸在节点之间的共享存储器的处理的流程图;

[0027] 图 16 是用于说明释放节点间共享存储器的处理的流程图;

[0028] 图 17 是用于说明发出请求的处理的流程的流程图;

[0029] 图 18 是用于说明当接收到请求时所执行的处理的流程的流程图;

[0030] 图 19 是用于说明 ECC 检查单元执行的处理的流程的流程图;

[0031] 图 20 是用于说明 ECC 检查单元执行的处理的流程的流程图;

[0032] 图 21 是用于说明 ECC 检查单元执行的检测 ICE 的发生的处理的示例的图;

[0033] 图 22 是用于说明 ECC 检查单元执行的检测 ICE 的发生的处理的示例的图;

[0034] 图 23 是用于说明 ECC 检查单元执行的检测 ICE 的发生的处理的示例的图;

[0035] 图 24 是用于说明 ECC 检查单元执行的检测 ICE 的发生的处理的示例的图;

[0036] 图 25 是用于说明 ECC 检查单元执行的检测 ICE 的发生的处理的示例的图;

[0037] 图 26 是用于说明 ECC 检查单元执行的检测 ICE 的发生的处理的示例的图;

[0038] 图 27 是用于说明 ECC 检查单元执行的检测 ICE 的发生的处理的示例的图;

[0039] 图 28A 是示出根据实施例由 CPU 执行的 OS 的功能配置的示例的图;

[0040] 图 28B 是示出在 OS 的处理中所参考的表的数据配置的示例的图;

[0041] 图 28C 是示出在 OS 的处理中所参考的表的数据配置的示例的图;

[0042] 图 29 是用于说明 OS 执行的处理的流程的流程图;

[0043] 图 30 是用于说明 OS 执行的处理的流程的流程图;

[0044] 图 31 是用于说明 OS 执行的处理的流程的流程图;

[0045] 图 32 是用于说明 OS 执行的处理的流程的流程图;

[0046] 图 33 是用于说明 OS 执行的处理的流程的流程图；

[0047] 图 34 是示出存储器管理表的数据配置的示例的图；

[0048] 图 35 是示出地址转换表的数据配置的示例的图；

[0049] 图 36 是用于说明 OS 执行的处理的流程的流程图；

[0050] 图 37 是示出访问重新开始等待列表的数据配置的示例的图；

[0051] 图 38 是示出调度等待列表的数据配置的示例的图；

[0052] 图 39 是用于说明 OS 执行的处理的流程的流程图；以及

[0053] 图 40 是示出根据实施例的 CPU 执行的处理和 OS 的功能配置的示例的图。

具体实施方式

[0054] 将参照附图说明本发明的优选实施例。

[0055] 将使用图 1 以下述实施例描述具有多个节点的信息处理系统的示例。图 1 是用于说明根据实施例的信息处理系统的示例的图。根据图 1 所示的示例，信息处理系统 1 具有 XB 2(交叉开关)和多个构件块 10 至 10e。XB 2 是将各构件块 10 至 10e 相互连接的交叉开关。另外，XB 2 具有未示出的服务处理器，该服务处理器用作以下描述的构件块 10 至 10e 中的每个构件块的各自服务处理器的母机(master)。另外，在连接有少量节点的小规模配置的情况下，构件块可在没有 XB2 的情况下直接相连。

[0056] 另外，构件块 10 具有多个 CPU(中央处理单元)21 至 21c 和多个存储器 22 至 22c。此外，其它构件块 10a 至 10e 还使用与构件块 10 相同的配置，因此，以下将不进行描述。另外，根据图 1 所示的示例，没有绘出 CPU 21b 和 21c 以及存储器 22b 和 22c。另外，在每个构件块中设置有未示出的 I/O (输入 / 输出)装置。同时，根据本示例，CPU 之间的高速缓存一致性控制通过目录系统来实现，并且以下描述的、具有存储器上的数据的主 CPU 管理该目录。

[0057] 构件块 10 至 10e 中的每一个均独立地操作 OS。即，CPU 21 至 21c 中的每一个均独立地执行 OS。由构件块 10 至 10e 中的每一个执行的 OS 在每个构件块不同的分区中运行。同时，分区指的是同一 OS 运行且从运行的 OS 来看为一个系统运行的一组构件块。

[0058] 例如，构件块 10 和 10a 作为分区 #A 来进行操作，并且构件块 10b 至 10d 作为分区 #B 来进行操作。在这种情况下，由构件块 10 操作的 OS 将构件块 10 和 10a 识别为一个操作系统，并且由构件块 10b 操作的 OS 将构件块 10b 至 10d 识别为一个操作系统。

[0059] 接下来，将使用图 2 描述构件块的配置示例。图 2 是用于说明根据实施例的构件块的功能配置的图。根据图 2 所示的实施例，构件块 10 具有节点 20、服务处理器 24、XB 连接单元 27 和 27a 以及 PCIe (外围组件互联高速) 连接单元 28。

[0060] 节点 20 具有多个 CPU21 至 21c、多个存储器 22 至 22c 以及通信单元 23。

[0061] 服务处理器 24 具有控制单元 25 和通信单元 26。另外，根据图 2 所示的示例，各 CPU 21 至 21c 相互连接，并与通信单元 23 相连。另外，存储器 22 至 22c 分别与 CPU 21 至 21c 相连。服务处理器 24 通过未示出的网络(诸如，LAN (局域网))连接到服务器的管理终端，并且响应于来自管理终端的指令而执行改变节点或构件块 10 中的各种设置的控制。

[0062] 另外，CPU 21 至 CPU 21c 中的每一个均与 XB 连接单元 27 或 XB 连接单元 27a 相连。另外，XB 连接单元 27 和 27a 可以是相同的 XB 连接单元。另外，CPU 21 至 CPU 21c 中

的每一个均与 PCIe 连接单元 28 相连。此外，通信单元 23 与服务处理器 24 的通信单元 26 相连。另外，控制单元 25、通信单元 26、通信单元 23 以及各 CPU 21 至 CPU 21c 通过例如 I2C（集成电路之间）相连。

[0063] CPU 21 至 CPU 21c 是执行应用程序的计算处理装置。另外，CPU 21 至 CPU 21c 分别与存储器 22 至 22c 相连。此外，当执行的应用程序要求请求分配共享存储器时，各 CPU 21 至 CPU 21c 相互通信，并且分配由应用程序使用的共享存储器。另外，各 CPU 21 至 CPU 21c 使用各存储器 22 至 22c 或其它构件块 10a 至 10e 中的存储器的一部分作为共享存储器。

[0064] 图 3 是示出在另一节点附接到分配有共享存储器实体的节点时的存储器映射的示例的图。根据图 3 中的示例，当共享存储器被分配给直接连接到其中存在共享存储器的物理存储器的节点（被称为“主节点（home node）”）时，主节点将共享存储器分割成一定的区域大小。尽管该分割单位被称为“段（segment）”，但是将共享存储器分割成段不是不可缺少的。当另一节点请求分配主节点所具有的共享存储器时，可以通过附接共享存储器来使用主节点的共享存储器。该远程节点使用的存储器区域被称为“共享存储器镜像区域”。该共享存储器镜像区域可通过单个远程节点或多个远程节点附接。

[0065] 回到图 2，CPU 21 至 21c 中的每一个均具有使存储器的物理地址和 CPUID（标识）相关联的节点映射，其中 CPUID 是与存储器相连的 CPU 的标识符。另外，该 CPUID 是由系统 1 唯一确定的并且不会重复。

[0066] CPU 21 至 21c 中的每一个均使用节点映射来与另一 CPU 通信。例如，当与访问目标物理地址相关联的 CPUID 表示不同于 CPU 21 至 21c 的 CPU 时，CPU 21 通过 XB 连接单元 27 或 XB 连接单元 27a 和 XB 2 将存储器访问请求发送至另一节点。另外，当从另一节点接收到对与 CPU 21 连接的存储器的请求时，CPU 21 从与 CPU 21 连接的存储器 22 读取请求目标数据，并且将数据发送至请求源。其它 CPU 21a 至 21c 也执行相同处理。

[0067] 另外，CPU 21 至 21c 中的每一个均具有如下功能：使用 TLB（转换后备缓冲器）来转换地址，并且当 TLB 失误发生时，执行与在传统 CPU 中相同的处理（诸如，陷阱处理）。

[0068] 存储器 22 至 22c 是信息处理系统 1 的所有 CPU 共享的存储器。另外，在信息处理系统 1 中，构件块 10 至 10e 中的每一个的服务处理器将映射到同一物理地址空间的物理地址分给所有构件块 10 至 10e 的存储器。即，具有没有赋值的值的物理地址被分配给信息处理系统 1 的所有存储器当中至少用作共享存储器的存储器。

[0069] 另外，存储器 22 至 22c 使用存储器区域的一部分作为信息处理系统 1 的所有 CPU 共享的共享区域，并且使用其它部分作为访问存储器 22 至 22c 的 CPU 21 至 21c 将内核数据和用户数据存储在其中的本地区域以及在通过共享存储器与另一节点交换时无关的且 I/O 装置使用的 I/O 区域。

[0070] 控制单元 25 控制构件块 10。例如，控制单元 25 管理构件块 10 的电源，或者监视并控制构件块 10 中的异常。另外，控制单元 25 通过未示出的网络与其它构件块 10a 至 10e 的服务处理器连接，并且执行构件块 10a 与 10e 之间协同的控制。另外，控制单元 25 可以与各 CPU 21 至 21c 执行的 OS 通信。

[0071] 另外，控制单元 25 通过通信单元 26 和通信单元 23 访问 CPU 21 至 21c 中的每一个。而且，控制单元 25 更新并控制构件块 10 至 10e 中的每一个的节点映射。

[0072] 另外，服务处理器 24 的通信单元 26 接收从控制单元 25 发送的控制信号，并且将所接收到的控制信号发送至节点 20 的通信单元 23。另外，通信单元 23 接收从通信单元 26 发送的控制信号，并且将所接收到的控制信号发送至 CPU 21 至 21c 中的每一个。另外，XB 连接单元 27 和 27a 将 CPU 21 至 21a 中的每一个与 XB 2 相连，并且中继构件块 10 至 10e 的 CPU 之间的通信。另外，PCIe 连接单元 28 中继通过 CPU 21 至 21c 对 I/O 装置的访问。

[0073] 接下来，将使用图 4 描述 CPU 21 至 21c 中的每一个的功能配置。图 4 是用于说明根据实施例的 CPU 的功能配置的图。另外，CPU 21a 至 21c 具有与 CPU 21 的功能相同的功能，因此，将不进行描述。另外，根据图 4 所示的示例，未示出连接服务处理器 24 和 CPU 21 的通信单元 23 和 26。

[0074] 根据图 4 所示的示例，CPU 21 具有计算处理单元 30、路由器 40、存储器访问单元 41 以及 PCIe 控制单元 42。另外，计算处理单元 30 具有计算单元 31、L1（一级（level 1））高速缓存器 32、L2（二级（level 2））高速缓存器 33、节点映射 34、地址转换单元 35、高速缓存目录管理单元 36 以及分组控制单元 37。另外，路由器 40、存储器访问单元 41 以及 PCIe 控制单元 42 中的每个单元都不必包括在单个 CPU 21 内。

[0075] 另外，分组控制单元 37 具有分组生成单元 37a 和分组接收单元 37b。此外，PCIe 控制单元 42 具有请求生成单元 42a 和 PCIe 总线控制单元 42b。

[0076] 首先，将描述计算处理单元 30 的节点映射 34。节点映射 34 是表示存储器的存储器区域的物理地址的范围和与存储器连接的 CPU 的 CPUID 相关联地被登记的表。将使用附图描述登记在节点映射 34 中的信息的示例。

[0077] 图 5 是用于说明根据实施例的节点映射的数据配置的示例的图。根据图 5 所示的示例，节点映射 34 具有将各项（诸如，“地址”、“有效”、“节点 ID”以及“CPUID”）的登记内容关联的条目。同时，在每个条目的“地址”项中，存储了表示包括连续的多个物理地址的地址区域的信息。

[0078] 例如，信息处理系统 1 将分给所有存储器的物理地址空间分割成大小相等的地址区域，并且为每个地址区域分配诸如 #0、#1 和 #2 的标识符。另外，信息处理系统 1 在节点映射 34 的每个条目的“地址”中登记表示每个地址区域的标识符。例如，图 5 示出在第一条目的“地址”项中登记标识符 #0 的情况。另外，图 5 示出例如在第二条目的“地址”项中登记标识符 #1 的情况。此外，图 5 示出例如在第三条目的“地址”项中登记标识符 #2 的情况。

[0079] 另外，在每个条目的“有效”项中登记有效位，其表示每个 CPU 是否可以访问由物理地址表示的存储器区域。例如，当由物理地址表示的存储器区域是 CPU 之间共享的共享区域时，登记表示每个 CPU 可以访问存储器区域的有效位（例如，“1”）。图 5 示出例如在第一条目的“有效”项中登记有效位“1”的情况。另外，图 5 示出例如在第二条目的“有效”项中登记有效位“1”的情况。此外，图 5 示出例如在第三条目的“有效”项中登记有效位“0”的情况，其中该有效位“0”表示每个 CPU 不可以访问由物理地址表示的存储器区域。

[0080] 另外，在每个条目的“节点 ID”项中，登记了表示具有被分有物理地址的存储器的节点的标识符。图 5 示出例如在第一条目的“节点 ID”项中登记表示节点的标识符“1”的情况。另外，图 5 示出例如在第二条目的“节点 ID”项中登记表示节点的标识符“1”的情况。

[0081] 此外，在每个条目的“CPUID”项中，登记了表示与被分有物理地址的存储器连接的CPU的标识符。即，节点映射34表示访问目标物理地址是与哪个CPU连接的存储器的物理地址。图5示出例如在第一条目的“CPUID”项中登记表示CPU的标识符“4”的情况。另外，图5示出了在第二条目的“CPUID”项中登记表示CPU的标识符“5”的情况。

[0082] 另外，只要可以表示访问目标物理地址是与哪个CPU连接的物理地址，就可以在节点映射34中以除了本示例的格式之外的任意格式登记信息。

[0083] 回到图4，计算单元31是执行计算处理的计算装置的核，并且执行OS（操作系统）和应用程序。另外，当读取或写入数据时，计算单元31将存储作为读取目标或写入目标的数据的存储器区域的虚拟地址（VA）输出至地址转换单元35。

[0084] L1高速缓存器32是临时存储在计算单元31中频繁使用的数据的高速缓存存储器。尽管与L1高速缓存器32类似，L2高速缓存器33临时存储频繁使用的数据，但是L2高速缓存器33是以低速读取和写入数据的高速缓存存储器。同时，目录信息36a存储在高速缓存目录管理单元36中，并且是表示高速缓存有存储器22的每个存储器区域中所存储的数据CPU或所高速缓存的数据的更新状态的信息。在以下描述中，在一些情况下将“目录信息”简称为“目录”。基于该目录的高速缓存存储器管理方法是在ccNUMA（高速缓存一致性非均衡存储器访问）系统中常用的技术。ccNUMA技术和目录技术都是公知技术，因此，将不进行详细描述。另外，尽管在图4的高速缓存目录管理单元36中构建目录36a，但是还可以将目录信息36a记录在存储器22的存储器区域的一部分中。

[0085] 地址转换单元35具有TLB 35a。在TLB 35a中，登记将虚拟地址和物理地址相关联的条目。地址转换单元35使用TLB 35a将从计算单元31输出的虚拟地址转换为物理地址。例如，地址转换单元35从TLB 35a中搜索与从计算单元31获得的虚拟地址相关联的物理地址，并且当获得作为搜索结果的物理地址时，将所获得的物理地址输出至高速缓存目录管理单元36。另外，当TLB失误发生时，地址转换单元35执行陷阱处理。同时，诸如OS的系统软件将TLB失误发生的一组物理地址与虚拟地址的组登记在TLB 35a中。同时，对于组中登记被禁止的物理地址，即使当TLB失误发生时，诸如OS的系统软件也不将一组物理地址与虚拟地址登记在TLB 35a中。

[0086] 同时，当计算单元31执行的应用程序请求对共享存储器的分配时，OS和地址转换单元35执行以下处理。即，当TLB失误发生时，诸如OS的系统软件将条目登记在TLB 35a中。另外，当TLB失误未发生时，已经在TLB 35a中登记条目，并且地址转换单元35将虚拟地址转换为物理地址。

[0087] 另外，当应用程序或OS请求本地区域的分配时，地址转换单元35和OS执行以下处理。即，当TLB失误发生时，诸如OS的系统软件将如下条目登记在TLB 35a中：其将允许应用程序或OS访问CPU 21专用的本地区域的虚拟地址与分配给本地区域的物理地址的范围相关联。

[0088] 高速缓存目录管理单元36管理高速缓存数据和目录。高速缓存目录管理单元36获取从地址转换单元35输出的物理地址。

[0089] 另外，当从地址转换单元35获取物理地址时，高速缓存目录管理单元36执行以下处理。即，高速缓存目录管理单元36使用目录36a来判定存储在所获取的物理地址中的数据是否高速缓存在L1高速缓存器32和L2高速缓存器33中。

[0090] 另外，当判定存储在所获取的物理地址中的数据被高速缓存时，高速缓存目录管理单元 36 将所高速缓存的数据输出至计算单元 31。另外，当存储在所获取的物理地址中的数据没有高速缓存在 L1 高速缓存器 32 和 L2 高速缓存器 33 中时，高速缓存目录管理单元 36 执行以下处理。首先，高速缓存目录管理单元 36 参考节点映射 34，并且识别包括所获取的物理地址的条目的范围。另外，高速缓存目录管理单元 36 判定所识别的条目的 CPUID 是否是 CPU 21 的 CPUID。随后，当所识别的条目的 CPUID 是 CPU 21 的 CPUID 时，高速缓存目录管理单元 36 将物理地址输出至存储器访问单元 41。

[0091] 另外，当所识别的条目的 CPUID 不是 CPU 21 的 CPUID 时，高速缓存目录管理单元 36 执行以下处理。即，高速缓存目录管理单元 36 获取所识别的条目的 CPUID 和物理地址。另外，高速缓存目录管理单元 36 将所获取的 CPUID 和物理地址输出至分组控制单元 37。

[0092] 此外，当获取由从存储器访问单元 41 或分组控制单元 37 所输出的物理地址表示的存储器区域中所存储的数据时，高速缓存目录管理单元 36 将所获取的数据存储在 L1 高速缓存器 32 和 L2 高速缓存器 33 中。另外，高速缓存目录管理单元 36 将高速缓存在 L1 高速缓存器 32 中的数据输出至计算单元 31。

[0093] 此外，当从分组控制单元 37 获取物理地址时，即，当从另一 CPU 或 I/O 装置获取存储器访问的请求目标物理地址时，高速缓存目录管理单元 36 执行以下处理。即，高速缓存目录管理单元 36 参考节点映射 34，并且判定所获取的物理地址是否是分给本地区域的物理地址。

[0094] 当另一分区是请求源并且所获取的物理地址是分给本地区域的物理地址时，高速缓存目录管理单元 36 指示分组控制单元 37 向请求源发送拒绝响应(访问错误)。

[0095] 另外，当所获取的物理地址是分给共享区域的物理地址时，高速缓存目录管理单元 36 获取由所获取的物理地址表示的存储器区域中所存储的数据，将所获取的数据输出至分组控制单元 37，并且指示分组控制单元 37 将所获取的数据发送至请求源。

[0096] 另外，高速缓存目录管理单元 36 使用目录系统来执行保持所高速缓存的数据的一致性的处理。例如，当将对存储在存储器 22 中的数据的请求发送至请求发送源的 CPU 时，高速缓存目录管理单元 36 判定除发送源的 CPU 之外的 CPU 是否高速缓存该数据。

[0097] 另外，当另一 CPU 未高速缓存请求目标数据时，高速缓存目录管理单元 36 从 L1 高速缓存器 32、L2 高速缓存器 33 和存储器 22 获取请求目标数据。然后，高速缓存目录管理单元 36 将所获取的数据输出至分组控制单元 37。

[0098] 同时，当另一 CPU 高速缓存请求目标数据时，高速缓存目录管理单元 36 使用诸如 Illinois 协议的方法来执行用于保持高速缓存一致性的处理。例如，高速缓存目录管理单元 36 判定所高速缓存的数据的状态是 MESI（修改的 / 排他的 / 共享的 / 无效的）中的哪一个。

[0099] 另外，高速缓存目录管理单元 36 根据判定结果向另一 CPU 的高速缓存目录管理单元发送用于保持一致性的请求或命令(指令)或者从另一 CPU 的高速缓存目录管理单元接收用于保持一致性的请求或命令(指令)，并且执行与所高速缓存的数据的状态匹配的处理。同时，“修改的”表示一个 CPU 高速缓存数据并且所高速缓存的数据被更新的状态。另外，当所高速缓存的数据的状态是“修改的”时，需要执行写回。

[0100] 另外，“排他的”表示一个 CPU 高速缓存数据并且所高速缓存的数据不被更新的状

态。此外，“共享的”表示多个 CPU 高速缓存数据并且所高速缓存的数据不被更新的状态。另外，“无效的”表示未登记高速缓存状态。

[0101] 将描述具体示例。高速缓存目录管理单元 36 指示分组生成单元 37a 发送指示高速缓存具有状态 M（修改的）的数据的 CPU 进行写回的命令。另外，高速缓存目录管理单元 36 更新数据状态，并且执行与更新后的状态匹配的处理。另外，以下将描述高速缓存目录管理单元 36 发送和接收的请求或命令的类型。

[0102] 当从高速缓存目录管理单元 36 获取物理地址和 CPUID 时，分组生成单元 37a 生成存储有所获取的物理地址和 CPUID 的分组，即，用作存储器访问请求的分组。另外，分组生成单元 37a 将所生成的分组发送至路由器 40。

[0103] 图 6 是用于说明根据实施例从 CPU 发送的分组的图。根据图 6 所示的示例，分组生成单元 37a 生成包括 CPUID、物理地址以及表示请求内容的数据的请求，并且将所生成的请求输出至路由器 40。在这种情况下，路由器 40 通过 XB 连接单元 27 将分组生成单元 37a 生成的请求输出至 XB 2。然后，XB 2 将请求传送至由包括在请求中的 CPUID 表示的 CPU。

[0104] 另外，当从高速缓存目录管理单元 36 接收到发出用于保持一致性的请求或命令的指令时，分组生成单元 37a 生成所指示的请求或命令。另外，分组生成单元 37a 通过路由器 40、XB 连接单元 27 以及 XB 2 将所生成的请求或命令发送至所指定的 CPU。另外，当从 I/O 装置获取数据时，分组生成单元 37a 将对 I/O 装置的访问请求输出至路由器 40。

[0105] 当接收到除本地节点之外的另一 CPU 或另一 I/O 装置输出的分组时，分组接收单元 37b 获取包括在所接收到的分组中的物理地址。另外，当通过 PCIe 控制单元 42 和路由器 40 接收到从本地节点的 I/O 装置输出的分组时，分组接收单元 37b 获取包括在所接收到的分组中的物理地址。此外，分组接收单元 37b 将所获取的物理地址输出至高速缓存目录管理单元 36。另外，当接收到从另一 CPU 发送的数据时，分组接收单元 37b 将所接收到的数据输出至高速缓存目录管理单元 36。

[0106] 另外，当接收到用于保持一致性的请求或命令时，分组接收单元 37b 将所接收到的请求或命令输出至高速缓存目录管理单元 36。此外，当从路由器 40 接收到对 I/O 装置的访问请求的响应或数据时，分组接收单元 37b 将所接收到的响应或数据输出至高速缓存目录管理单元 36。在这种情况下，例如，高速缓存目录管理单元 36 将所获取的数据输出至存储器访问单元 41。通过这种手段，存储器访问单元 41 将数据存储在存储器 22 中。

[0107] 当接收到从分组生成单元 37a 输出的分组时，路由器 40 将所接收到的请求输出至 XB 连接单元 27。另外，路由器 40 通过 XB 连接单元 27 接收到从另一 CPU 发送的分组和数据，并且将所接收到的分组和数据输出至分组接收单元 37b。此外，路由器 40 将从分组控制单元 37 输出至例如 I/O 装置的分组输出至 PCIe 控制单元 42。另外，当从 PCIe 控制单元 42 接收到例如来自 I/O 装置的请求时，路由器 40 例如将所接收到的请求输出至分组控制单元 37 或 XB 连接单元 27。此外，当通过 XB 连接单元 27 或分组控制单元 37 接收到对 I/O 装置的响应时，路由器 40 将所接收到的响应输出至 PCIe 总线控制单元 42b。

[0108] 存储器访问单元 41 是所谓的 MAC（存储器访问控制器），并且控制对存储器 22 的访问。例如，当从高速缓存目录管理单元 36 接收到物理地址时，存储器访问单元 41 获取存储在由所接收到的物理地址表示的存储器 22 的区域中的数据，并且将所获取的数据输出至高速缓存目录管理单元 36。另外，存储器访问单元 41 使用存储器镜像功能来使得共享存

储器冗余。

[0109] 另外，存储器访问单元 41 具有 ECC 检查单元 41a、CE 地址寄存器 41b 以及 ICE 发生次数计数器 41c。

[0110] ECC 检查单元 41a 每个预定循环或者每当高速缓存目录管理单元 36 对存储器 22 提出读取访问请求时执行以下处理。即，ECC 检查单元 41a 判定是在存储器 22 的所有存储器区域的数据中还是在访问目标存储器区域的数据中发生了 CE。基于该判定，ECC 检查单元 41a 检测 CE。当检测到 CE 时，ECC 检查单元 41a 读取所检测到的 CE 发生的存储器区域中的数据，校正所读取的数据的错误，并将错误校正后的数据写回到所检测到的 CE 发生的存储器区域中。另外，ECC 检查单元 41a 再次读取数据被写回的存储器区域中的数据，并且再次判定在所读取的数据中是否发生 CE。当基于第二次判定而判定发生了 CE 时，ECC 检查单元 41a 判定 PCE 发生。以这样的方式，ECC 检查单元 41 检测到 PCE。

[0111] 另外，ECC 检查单元 41a 将在预定时间内的 CE 发生次数记录在 ICE 发生次数计数器 41c 中，并且当在预定时间内 CE 发生超过预定次数 α 时，判定 ICE 发生。以这样的方式，ECC 检查单元 41a 检测 ICE。另外，存储器访问单元 41 中的处理单元(诸如，微型计算机)可根据程序处理来执行例如对在预定时间内的 CE 发生次数进行计数的计数操作。

[0112] 另外，当检测到 ICE 或 PCE 时，ECC 检查单元 41a 为 CE 地址寄存器 41b 设置所检测到的 ICE 或 PCE 发生的存储器 22 的物理地址。

[0113] 当通过路由器 40 获取对 I/O 装置的访问请求时，请求生成单元 42a 生成发送至作为访问请求目标的 I/O 装置的请求，并且将所生成的请求输出至 PCIe 总线控制单元 42b。另外，当从 I/O 装置获取物理地址和 CPUID 时，请求生成单元 42a 生成存储有所获取的物理地址和 CPUID 的分组，即，用作存储器访问请求的分组。这样的请求的类型包括 I/O 装置读取与 CPU 21 或其它 CPU 连接的存储器的请求。另外，当从 I/O 装置获取物理地址、CPUID 以及写入数据时，请求生成单元 42a 生成存储有所获取的物理地址、CPUID 以及写入数据的分组，即，用作存储器访问请求的分组。这样的请求的类型包括 I/O 装置将数据写入连接到 CPU 21 或其它 CPU 的存储器中的请求。另外，请求生成单元 42a 将所生成的分组发送至路由器 40。

[0114] 当请求生成单元 42a 获取所生成的请求时，PCIe 总线控制单元 42b 通过 PCIe 连接单元 28 将请求发送至 I/O 装置。另外，当通过 PCIe 连接单元 28 从 I/O 装置获取物理地址和 CPUID 时，PCIe 总线控制单元 42b 将所获取的物理地址和 CPUID 发送至请求生成单元 42a。此外，当通过 PCIe 连接单元 28 从 I/O 装置获取物理地址、CPUID 以及写入数据时，PCIe 总线控制单元 42b 将所获取的物理地址、CPUID 以及写入数据发送至请求生成单元 42a。

[0115] 接下来，将使用图 7 描述从 CPU 21 向另一 CPU 发送请求的处理的示例。图 7 是用于说明根据实施例从 CPU 发送请求的处理的示例的图。如由例如图 7 中(A)所示，服务处理器 24 对节点映射 34 设置如下条目：其将访问被分有物理地址的存储器的 CPU 的 CPUID 和存储器的物理地址相关联。

[0116] 另外，计算单元 31 执行计算处理，并且将访问目标虚拟地址输出至地址转换单元 35，如由图 7 中的(B)所示。然后，地址转换单元 35 将虚拟地址转换为物理地址，并且将转换后的物理地址输出至高速缓存目录管理单元 36，如由图 7 中的(C)所示。

[0117] 同时，如由图 7 中的(D)所示，当从地址转换单元 35 获取物理地址时，高速缓存目

录管理单元 36 参考节点映射 34，并且获取与所获取的物理地址相关联的 CPUID。另外，如由图 7 中的(E)所示，当所获取的 CPUID 不是 CPU 21 的 CPUID 时，高速缓存目录管理系统 36 将所获取的 CPUID 和物理地址输出至分组控制单元 37。

[0118] 在这种情况下，分组生成单元 37a 生成存储有从高速缓存目录管理单元 36 获取的物理地址和 CPUID 的分组，并且将所生成的分组输出至路由器 40，如由图 7 中的(F)所示。接下来，如由图 7 中的(G)所示，路由器 40 将从分组生成单元 37a 获取的分组输出至 XB 连接单元 27。随后，如由图 7 中的(H)所示，XB 连接单元 27 将所获取的分组输出至 XB 2。然后，XB 2 将分组发送至由分组中所存储的 CPUID 表示的 CPU。

[0119] 接下来，将使用图 8 描述当 CPU 21 从另一 CPU 接收到分组时所执行的处理的示例。图 8 是用于说明根据实施例当 CPU 接收分组时所执行的处理的示例的图。例如，如由图 8 中的(J)所示，分组接收单元 37b 从另一 CPU 接收到存储有 CPU 21 的 CPUID 和被分给存储器 22 的物理地址的分组、或者响应分组。

[0120] 在这种情况下，分组接收单元 37b 从所接收到的分组获取物理地址，并且如由图 8 中的(K)所示，将所获取的物理地址连同表示所获取的物理地址的请求源是否是本地分区的信息一起输出至高速缓存目录管理单元 36。然后，高速缓存目录管理单元 36 判定由物理地址表示的存储器区域是否是共享区域或本地区域。

[0121] 当请求源是另一分区时，高速缓存目录管理单元 36 检查是否访问了共享区域，并且在本地区域的情况下请求分组控制单元 37 对错误作出响应。在其它情况下，如由图 8 中的(L)所示，高速缓存目录管理单元 36 判定由物理区域表示的存储器区域中的数据是否高速缓存在 L1 高速缓存器 32 和 L2 高速缓存器 33 中。

[0122] 另外，当判定没有高速缓存该数据时，高速缓存目录管理单元 36 将物理地址输出至存储器访问单元 41，如由图 8 中的(M)所示。然后，如由图 8 中的(N)所示，存储器访问单元 41 从存储器 22 获取由物理地址表示的存储器区域中的数据，并且将该数据输出至高速缓存目录管理单元 36。另外，当物理地址被输入至存储器访问单元 41 时，如果在由所输入的物理地址表示的存储器区域中所存储的数据中检测到 IEC 和 PEC，则 ECC 检查单元 41a 为 CE 地址寄存器 41b 设置所输入的物理地址，如由图 8 中的(O)所示。

[0123] 另外，当从 L1 高速缓存器 32、L2 高速缓存器 33 或存储器访问单元 41 获取数据时，高速缓存目录管理单元 36 将所获取的数据输出至分组控制单元 37，并且指示分组控制单元 37 将该数据发送至请求源的 CPU。

[0124] 接下来，将使用图 9 描述从 I/O 装置向除 CPU 21 之外的 CPU 发送读取或写入请求的处理的示例。图 9 是用于说明根据实施例从 I/O 装置发送请求的处理的示例的图。例如，当从 I/O 装置获取物理地址和 CPUID 时，PCIe 连接单元 28 将所获取的物理地址和 CPUID 输出至 PCIe 总线控制单元 42b，如由图 9 中的(P)所示。另外，当从 I/O 装置获取物理地址、CPUID 以及写入数据时，PCIe 连接单元 28 将所获取的物理地址、CPUID 以及写入数据输出至 PCIe 总线控制单元 42b，如由图 9 中的(P)所示。

[0125] 此外，当从 PCIe 连接单元 28 获取物理地址和 CPUID 时，PCIe 总线控制单元 42b 将所获取的物理地址和 CPUID 输出至请求生成单元 42a，如由图 9 中的(Q)所示。另外，当从 PCIe 连接单元 28 获取物理地址、CPUID 以及写入数据时，如由图 9 中(Q)的所示，PCIe 总线控制单元 42 将所获取的物理地址、CPUID 以及写入数据发送至请求生成单元 42a。

[0126] 当从 PCIe 总线控制单元 42b 获取物理地址和 CPUID 时，请求生成单元 42a 生成用作包括所获取的物理地址和 CPUID 的读取请求的分组。另外，当从 PCIe 总线控制单元 42b 获取物理地址、CPUID 以及写入数据时，请求生成单元 42a 生成用作包括所获取的物理地址、CPUID 以及写入数据的写入请求的分组。另外，如由图 9 中的(R)所示，请求生成单元 42a 将生所成的分组输出至路由器 40。

[0127] 接下来，如由图 9 中的(T)所示，路由器 40 将从请求生成单元 42a 获取的请求输出至 XB 连接单元 27。随后，如由图 9 中的(U)所示，XB 连接单元 27 将所获取的请求输出至 XB 2。然后，XB 2 将分组发送至由存储在该请求中的 CPUID 表示的 CPU。

[0128] 接下来，将使用图 10 描述在 I/O 装置从除 CPU 21 之外的 CPU 接收响应的处理的示例。图 10 是用于说明根据实施例在 I/O 装置接收响应的处理的示例的图。例如由图 10 中的(V)所示，XB 连接单元 27 从除 CPU21 之外的 CPU 接收对 I/O 装置的响应。

[0129] 当接收到响应时，如由图 10 中的(W)所示，XB 连接单元 27 将所接收到的响应输出至路由器 40。当接收到响应时，如由图 10 中的(X)所示，路由器 40 将所接收到的响应输出至请求生成单元 42a。另外，如由图 10 中的(Y)所示，请求生成单元 42a 将响应输出至 PCIe 总线控制单元 42b。当接收到响应时，如由图 10 中的(Z)所示，PCIe 总线控制单元 42b 将所接收到的响应输出至 PCIe 连接单元 28。通过这样的手段，响应被从 PCIe 连接单元 28 发送至 I/O 装置。

[0130] 通信单元 23、服务处理器 24、XB 连接单元 27、XB 连接单元 27a 以及 PCIe 连接单元 28 是电子电路。同时，作为电子电路的示例，集成电路(诸如，ASIC（专用集成电路）和 FGPA（现场可编程门阵列））、CPU 或 MPU（微处理单元）是可应用的。另外，替代 CPU 21 至 21c，集成电路(诸如，ASIC 和 FGPA）或 MPU 是可应用的。

[0131] 另外，存储器 22 至 22a 是半导体存储器元件，诸如 RAM(随机存取存储器)、ROM(只读存储器）或闪存。另外，L1 高速缓存器 32 和 L2 高速缓存器 33 是高速半导体存储元件，诸如 SRAM（静态随机存取存储器)。

[0132] 接下来，将描述在 CPU 21 至 21c 的每一个中保持高速缓存一致性的处理。另外，在以下描述中，信息处理系统 1 的每个 CPU 均使用 Illinois 协议来保持一致性。

[0133] 另外，在以下描述中，信息处理系统 1 的每个存储器被识别为具有所有 CPU 可以高速缓存数据的空间的存储器。此外，在以下描述中，通过 CPU 中的 MAC 直接地且物理地连接到存储所请求的数据的存储器的 CPU 被称为“主 CPU”，并且请求数据以将其存储在其高速缓存器中的 CPU 被称为“本地 CPU”。

[0134] 另外，已经将请求发送至主 CPU 并已高速缓存数据的 CPU 被称为“远程 CPU”。另外，在一些情况下，本地 CPU 和主 CPU 是相同的，并且在一些情况下，本地 CPU 和远程 CPU 是相同的。

[0135] 本地 CPU 参考本地 CPU 的节点映射来判定访问目标物理地址被分给访问主 CPU 的存储器。另外，本地 CPU 向主 CPU 发出存储物理地址的请求。另外，由本地 CPU 发出的请求包括多种类型的请求。因此，主 CPU 的高速缓存目录管理单元根据所获得的请求的类型控制高速缓存一致性。

[0136] 由本地 CPU 发出的请求的类型例如包括共享的提取(fetch）访问、排他的提取访问、高速缓存无效请求以及高速缓存替换请求。共享的提取访问是例如“移入共享(MoveIn

to Share)”的执行请求，并且是在主 CPU 从要访问的存储器读取数据时发出的请求。

[0137] 另外，排他的提取访问是例如“排他地移入(MoveIn Exclusively)”的执行请求，并且是在主 CPU 将数据存储在要访问的存储器中时将数据加载到高速缓存器时发出的。此外，高速缓存无效请求是例如“移出(MoveOut)”的执行请求，并且是在请求主 CPU 使高速缓存线无效时发出的。另外，当接收到高速缓存无效请求时，在一些情况下主 CPU 向远程 CPU 发出高速缓存无效请求，并且在一些情况下发出执行高速缓存器的“无效(Invalidation)”的命令。

[0138] 高速缓存替换请求是例如“写回(WriteBack)”的执行请求，并且是在主 CPU 将更新后的高速缓存数据(即，处于“已修改(Modified)”状态的高速缓存数据)写入要访问的存储器时发出的。另外，高速缓存替换请求是例如“回刷(FlushBack)”的执行请求，并且是在放弃没有更新的高速缓存数据(即，处于“共享的”或“排他的”的状态的高速缓存数据)时发出。

[0139] 当从本地 CPU 或远程 CPU 接收到以上请求时，主 CPU 向本地 CPU 或远程 CPU 发出命令以处理请求。同时，主 CPU 根据所获得的请求的类型发出多种类型的命令以控制高速缓存一致性。例如，主 CPU 向本地 CPU 发出“移出并旁路以共享(MoveOut and Bypass to Share)”以加载远程 CPU 高速缓存的数据。

[0140] 另外，例如，主 CPU 使除本地 CPU 之外的所有远程 CPU 的高速缓存器无效，然后，主 CPU 向本地 CPU 发出“排他地移出并旁路(MoveOut and Bypass Exclusively)”以发送数据。此外，主 CPU 发出用于请求远程 CPU 使高速缓存器无效的“移出与无效(MoveOut WITH Invalidatio)”。另外，当主 CPU 发出“移出与无效”时，所有 CPU 高速缓存器进入目标地址的“无效”状态。另外，当事务完成时，本地 CPU 高速缓存数据。

[0141] 另外，主 CPU 发出用于请求远程 CPU 使高速缓存线无效的“移出以刷新(MovOut for Flush)”。另外，当主 CPU 发出“移出以刷新”时，目标数据仅存储在主 CPU 的存储器中。另外，当目标数据状态是“共享的”时，主 CPU 发出用于请求远程 CPU 放弃高速缓存器的“缓冲器无效(Buffer Invalidation)”。

[0142] 主 CPU 根据请求的类型发出以上命令，并且转变每个 CPU 高速缓存的数据的状态。另外，当接收到命令时，本地 CPU 和远程 CPU 执行由命令表示的处理，并且转变由本地 CPU 和远程 CPU 高速缓存的数据的状态。

[0143] 随后，本地 CPU 和远程 CPU 将对命令的完成的响应或具有数据的完成响应发送至主 CPU。另外，主 CPU 和远程 CPU 执行有序处理，然后，将具有数据的请求响应发送至本地 CPU。

[0144] 处理的流程

[0145] 接下来，将使用图 11 描述控制控制共享区域的信息处理系统 1 的处理的流程。图 11 是用于说明控制共享区域的处理的流程的流程图。首先，信息处理系统 1 根据来自应用程序的请求执行在节点之间分配共享存储器的处理(步骤 S101)。接下来，信息处理系统 1 执行附接在节点之间共享的共享存储器的处理(步骤 S102)。

[0146] 随后，信息处理系统 1 的每个 CPU 执行的应用程序使用每个存储器(步骤 S103)。接下来，信息处理系统 1 执行拆卸共享存储器的处理(步骤 S104)。随后，信息处理系统 1 执行释放共享存储器的处理(步骤 S105)，并且结束处理。另外，仅该共享存储器的主节点的应

用程序可以执行步骤 S101 和步骤 S105，或者当实际处理是 nop（无操作）时，除该共享存储器的主节点之外的节点的应用程序也可以执行步骤 S101 和步骤 S105。

[0147] 接下来，将使用图 12 描述图 11 中的步骤 S101 中分配共享存储器的处理的流程。图 12 是用于说明分配共享存储器的处理的流程图。根据图 12 所示的示例，CPU 21 执行的应用程序请求 OS 执行在节点之间分配共享存储器的处理（步骤 S201）。

[0148] 然后，CPU 21 执行的 OS 分配具有从用于共享区域的物理地址区域请求的大小的存储器（步骤 S202）。接下来，将由 OS 分配的共享存储器的管理 ID 传递到应用程序（步骤 S203），并且结束共享存储器分配处理。

[0149] 接下来，将使用图 13 描述图 11 中的步骤 S102 中在节点之间附接共享存储器的处理的流程。图 13 是用于说明共享存储器附接处理的流程图。首先，应用程序将管理 ID 传递到 OS，并且请求在节点之间附接共享存储器的处理（步骤 S301）。在这种情况下，OS 与另一节点执行的 OS 通信，并且获取与管理 ID 相关联的物理地址（步骤 S302）。

[0150] 同时，当 OS 与另一节点执行的 OS 通信时，使用通过 LAN 的通信或通过服务处理器 24 在节点之间的通信。另外，每个节点执行的 OS 将特定的共享存储器设置为用于节点之间的通信的区域，并且存储或读取关于所设置的区域的信息以执行通信。

[0151] 接下来，OS 确定并分配与物理地址相关联的虚拟地址（步骤 S303）。例如，CPU 21 执行的 OS 为地址转换单元 35 设置物理地址和虚拟地址的 TLB 35a。

[0152] 另外，CPU 21 至 21c 中的每一个使用的虚拟地址可以是重叠的范围，或可以按每个 CPU 变化的范围。另外，可以由应用程序对于 OS 指定 CPU21 至 21c 中的每一个使用的虚拟地址。随后，将虚拟地址的值传递到应用程序（步骤 S304），并且结束处理。

[0153] 接下来，将使用图 14 描述图 11 中的步骤 S103 中在应用程序中使用节点之间的共享存储器的处理的流程。图 14 是用于说明在应用程序中使用共享存储器的处理的流程图。例如，CPU 21 执行的应用程序发出虚拟地址，并且访问由虚拟地址表示的存储器区域（步骤 S401）。

[0154] 然后，CPU 21 判定 TLB 失误是否发生（步骤 S402）。另外，当 TLB 失误发生时（在步骤 S402 中为是），CPU 21 执行陷阱处理，并且为 TLB 设置作为一组虚拟地址与物理地址的条目（步骤 S403）。

[0155] 接下来，应用程序再次发出虚拟地址，使用 TLB 将虚拟地址转换为物理地址，然后正常地访问共享存储器（步骤 S404），并结束处理。同时，当 TLB 失误没有发生时（在步骤 S402 中为否），正常地执行对共享存储器的访问（步骤 S405），并且结束处理。

[0156] 接下来，将使用图 15 描述图 11 中的步骤 S104 中拆卸节点之间的共享存储器的处理的流程。图 15 是用于说明拆卸节点之间的共享存储器的处理的流程图。例如，CPU 21 执行的应用程序对 OS 指定节点之间的共享存储器的虚拟地址或管理 ID，并且请求拆卸处理（步骤 S501）。

[0157] 然后，CPU 21 执行的 OS 刷新高速缓存器（步骤 S502）。即，当在释放（解除分配）共享存储器的分配之后 OS 再次分配共享存储器时，如果在没有分配共享存储器的同时共享存储器的主节点重新启动，则涉及高速缓存器的状态和实存储器的状态冲突。因此，OS 刷新高速缓存器，并且防止高速缓存器的状态和实存储器的状态冲突的状态。

[0158] 另外，OS 取消节点间共享存储器（即，应用程序使用的虚拟地址的范围）的分配，并

且删除与所取消的虚拟地址相关的 TLB 35a 的条目(步骤 S503)。另外,随后,即使当在该节点上在所拆卸的存储器地址中发生 TLB 失误时(在步骤 S402 中为是),OS 也不为 TLB 35a 设置与所拆卸的虚拟地址相关联的物理地址。因此,步骤 S404 没有正常地结束,并且发生了访问错误。不同于步骤 S302 中的处理,在完成拆卸之后,OS 执行节点之间的通信,并且该应用程序通知完成对该共享存储器的 PA 的访问(步骤 S504)。如果在主节点上释放该共享存储器并且该应用程序是使用该共享存储器的最后应用程序,则请求主节点执行释放处理(步骤 S505),并且结束处理。

[0159] 接下来,将使用图 16 描述图 11 中的步骤 S105 中释放节点间共享存储器的处理的流程。图 16 是用于说明释放节点间共享存储器的处理的流程图。例如,CPU 21 执行的应用程序请求 OS 执行释放节点间共享存储器的处理(步骤 S601)。然后,当所有用户拆卸所指定的共享存储器时,OS 释放分配(步骤 S602),并且结束处理。如果未完成拆卸,则完成处理而不执行分配释放处理。另外,在步骤 S505 中执行完成实际分配的处理。

[0160] 接下来,将使用图 17 描述从 CPU 21 向另一 CPU 发送存储器访问请求的处理的流程。图 17 是用于说明发出请求的处理的流程的流程图。例如,CPU 21 的计算单元 31 发出虚拟地址(步骤 S701)。

[0161] 然后,地址转换单元 35 将虚拟地址转换为物理地址(步骤 S702)。接下来,高速缓存目录管理单元 36 获取物理地址,并且管理高速缓存目录(步骤 S703)。即,高速缓存目录管理单元 36 转变由所获取的物理地址表示的存储器区域的高速缓存状态。

[0162] 接下来,高速缓存目录管理单元 36 参考节点映射 34,并且判定所获取的物理地址是否是被分给另一节点(另一分区)的存储器的物理地址(步骤 S704)。另外,当判定所获取的物理地址不是被分给另一节点(另一分区)的存储器的物理地址时(在步骤 S704 中为否),高速缓存目录管理单元 36 使用所获取的物理地址执行存储器访问(步骤 S705)。此外,结束处理。

[0163] 同时,当所获取的物理地址是被分给另一节点(另一分区)的存储器的物理地址时(在步骤 S704 中为是),高速缓存目录管理单元 36 从节点映射获取与物理地址相关联的 CPUID (步骤 S706)。另外,分组发送单元生成存储有 CPUID 和物理地址的分组(即,存储器访问请求)并将分组发送至 XB 2 (步骤 S707),并且结束处理。

[0164] 接下来,将使用图 18 描述当 CPU 21 从另一 CPU 接收存储器访问请求时所执行的处理的流程。图 18 是用于说明当接收到请求时所执行的处理的流程的流程图。另外,根据图 18 所示的示例,当 CPU 21 从另一 CPU 接收到“移入共享”或“排他地移入”时所执行的处理的流程。例如,CPU 21 通过 XB 2 从另一 CPU 接收请求(步骤 S801)。

[0165] 在这种情况下,CPU 21 使用节点映射来判定请求目标物理地址是否是本地区域(步骤 S802)。另外,当请求目标物理地址是本地区域时(在步骤 S802 中为是),CPU 21 将拒绝响应返回给请求源 CPU (步骤 S803),并且结束处理。

[0166] 此外,当请求目标物理地址不是本地区域时(步骤 S802 中为否),CPU 21 管理保持一致性的高速缓存目录(步骤 S804)。另外,CPU 21 判定由物理地址表示的存储器区域的状态(步骤 S805)。

[0167] 另外,CPU 21 向另一 CPU 发出与所判定的状态匹配的命令(步骤 S806),并且转变状态(步骤 S807)。随后,CPU 21 对用于将由物理地址表示的存储器区域中的数据发送至请

求源 CPU 作出应答(步骤 S808),并且结束处理。

[0168] 接下来,将使用图 19 描述在每个预定循环或每当高速缓存目录管理单元 36 向存储器 22 提出读取访问请求时由 ECC 检查单元 41a 执行的处理的流程。图 19 是用于说明由 ECC 检查单元执行的处理的流程的流程图。如图 19 所示,ECC 检查单元 41a 检查存储器 22 的所有存储器区域中的数据和访问目标存储器区域中的数据的错误(步骤 S901)。另外,ECC 检查单元 41a 判定 CE 是否发生(步骤 S902)。

[0169] 当 CE 没有发生时(在步骤 S902 中为否),即,例如当数据正常或 PE 发生时,结束处理。另外,在这种情况下,存储器访问单元 41 将在读取访问目标存储器区域中的数据发送至高速缓存目录管理单元 36。同时,当 CE 发生时(在步骤 S902 中为是),ECC 检查单元 41a 读取在 CE 发生的存储器区域中的数据,校正所读取的数据的错误,并将错误校正后的数据写回到 CE 发生的存储器区域中(步骤 S903)。另外,ECC 检查单元 41a 再次读取数据被写回的存储器区域中的数据(步骤 S904),并且针对所读取的数据再次(第二次)检查错误(步骤 S905)。

[0170] 随后,ECC 检查单元 41 判定在再次检查错误的数据中是否发生了 CE(步骤 S906)。当 CE 发生时(在步骤 S906 中为是),ECC 检查单元 41a 向 CE 地址寄存器 41b 设置 CE 发生的存储器 22 的物理地址(步骤 S907),并且结束处理。通过这样的手段,能够检测 ICE。

[0171] 同时,当在第二次错误检查中没有发生 CE 时(在步骤 S906 中为否),ECC 检查单元 41a 使与 CE 发生的存储器 22 的物理地址对应的 ICE 发生次数计数器 41a 的值加 1 (步骤 S908)。同时,按存储器 22 的每个预定存储器区域设置 ICE 发生次数计数器 41c。例如,当可以根据 ECC 按每 64 字节执行一位的错误校正时,则每 64 字节设置 ICE 发生次数计数器 41c。另外,可以按通过分割存储器 22 的存储器区域而获得的每页设置 ICE 发生次数计数器 41c。

[0172] 另外,ECC 检查单元 41a 判定 ICE 发生次数计数器的值是否是阈值 α 或更小(步骤 S909)。当 ICE 发生次数计数器的值大于阈值 α 时(在步骤 S909 中为否),该步骤进行至步骤 S907。同时,当 ICE 发生次数计数器 41c 的值是阈值 α 或更小时(在步骤 S909 中为是),完成处理。

[0173] 接下来,将使用图 20 描述每个预定循环由 ECC 检查单元 41a 执行的处理的流程。图 20 是用于说明由 ECC 检查单元执行的处理的流程的流程图。在比执行图 19 所示的处理的循环长的循环内执行该处理。如图 20 所示,ECC 检查单元 41a 将 ICE 发生次数计数器 41c 的值清零(步骤 S1001),并且结束处理。

[0174] 另外,ECC 检查单元 41a 对发生的 CE 进行计数并计算通过将 CE 次数除以计数所花费的时间而获得的值(每单位时间 CE 的发生次数),并且当所算出的值超过阈值 β 时可以判定 ICE 发生。将描述由 ECC 检查单元 41a 以这样的方式执行的检测 ICE 发生的处理的示例。

[0175] 图 21 至图 27 是用于说明 ECC 检查单元执行的检测 ICE 的发生的处理的示例的图。图 21 至图 27 示出了 ECC 检查单元 41a 使用的表的示例。通过图 21 至图 27 的示例所示的表包括登记每单位时间的 CE 发生次数的“平均值”项、登记 CE 发生次数的计数开始的时间的“开始时间”项以及登记 CE 最后一次发生的时间的“最终发生时间”项。

[0176] 例如,在图 21 所示的表中所登记的内容表示 CE 发生次数的计数开始于 2011 年 1

月 1 日 0 :00，并且 CE 最后一次发生的时间是 2011 年 1 月 1 日 3 :30。另外，在图 21 所示的表中所登记的内容表示通过将从 2011 年 1 月 1 日 0 :00 至 3 :30 的 210 分钟内发生的 CE 的次数除以 210 分钟所获得的值(即，在一分钟的单位时间内发生的 CE 的次数的平均值)是 0.1 (次 / 分)。

[0177] 将根据图 21 中的示例说明下述情况 :ECC 检查单元 41a 在 2011 年 1 月 1 日 3 :50 检测到新的 CE。在这种情况下，根据通过图 21 中的示例所示的表，ECC 检查单元 41a 计算从 2011 年 1 月 1 日 0 :00 至 3 :30 的 210 分钟内发生的 CE 的次数“21”(0.1×210)。另外，ECC 检查单元 41a 计算从 2011 年 1 月 1 日 0 :00 至 3 :50 的 230 分钟内发生的 CE 的次数“22”(21+1)。随后，ECC 检查单元 41a 计算通过将在从 2011 年 1 月 1 日 0 :00 至 3 :50 的 230 分钟内发生的 CE 的次数“22”除以 230 分钟所获得的值，即，在一分钟的单位时间内发生的 CE 的次数的平均值“0.095”。另外，如在图 22 的示例中所示，ECC 检查单元 41a 将表的“平均值”项和“最终发生时间”项分别更新为“0.095”和“2011/1/1 03:50”。随后，ECC 检查单元 41a 判定平均值“0.095”是否超过阈值 β，在平均值超过阈值 β 时检测到 ICE 的发生，并且为 CE 地址寄存器 41b 设置发生了 CE 的存储器 22 的物理地址。

[0178] 另外，如在图 23 中的示例所示，在每项中登记“0”作为表缺省值。另外，在“开始时间”项和“最终发生时间”项中的“0”是指系统中的特定时间，诸如在 Unix (注册商标)系统中的 1970 年 1 月 1 日 0 :00。

[0179] 另外，ECC 检查单元 41a 在 CE 第一次发生时执行以下处理。将根据图 23 中的示例描述以下情况 :ECC 检查单元 41a 在 2011 年 1 月 1 日 0 :00 检测到第一次 CE。在这样的情况下，如在图 24 中的示例所示，ECC 检查单元 41 将表中的“平均值”项、“开始时间”项以及“最终发生时间”项分别更新为“1”、“2011/1/1 00:00”以及“2011/1/1 00:00”。

[0180] 另外，ECC 检查单元 41a 在第二次 CE 发生时执行以下处理。将根据图 24 中的示例描述以下情况 :ECC 检查单元 41a 在 2011 年 1 月 1 日 0 :05 检测到第二次 CE。在这样的情况下，ECC 检查单元 41a 计算从 2011 年 1 月 1 日 0:00 至 0 :05 的 5 分钟内发生的 CE 的次数“2”(1+1)。随后，ECC 检查单元 41a 计算通过将从 2011 年 1 月 1 日 0 :00 至 0 :05 的 5 分钟内发生的 CE 的次数“2”除以 5 分钟所获得的值，即，在一分钟的单位时间内发生的 CE 的次数的平均值“0.4”。另外，如在 25 的示例中所示，ECC 检查单元 41a 将表的“平均值”项和“最终发生时间”项分别更新为“0.5”和“2011/1/1 00:05”。

[0181] 另外，当在登记“最终发生时间”项的时间之后经过了预定时间(诸如，一个小时以上)后检测到 CE，ECC 检查单元 41a 可以检测到该 CE 的发生为第一次 CE 的发生。将根据图 26 中的示例描述以下情况 :在表的“最终发生时间”项中登记表示在 2011 年 1 月 1 日 3 :30 最终检测到 CE 的信息。在这种情况下，当在从 2011 年 1 月 1 日 3 :30 起经过预定时间(一个小时)或更多时间之后在 2011 年 1 月 1 日 5 :30 检测到 CE 时，ECC 检查单元 41a 检测到第一次检测到 CE。另外，如在图 27 中所示，ECC 检查单元 41a 将表中的“平均值”项、“开始时间”项以及“最终发生时间”项分别更新为“1”、“2011/1/1 05:30”以及“2011/1/1 05:30”。通过这样的手段，当 CE 发生频率快速上升时，能够防止注意到 CE 发生频率由于过去的发生频率低而上升。

[0182] 在下文中，将描述由 CPU 21 执行的 OS 执行的处理。图 28A 是示出根据实施例由 CPU 执行的 OS 的功能配置的示例的图。图 28B 是示出在 OS 的处理中所参考的表的数据配

置的示例的图。图 29 至图 33、图 36 和图 39 是用于说明由 OS 执行的处理的流程的流程图。

[0183] 图 28A 示出了主节点 50 和远程节点 70 中的 OS 的功能配置。根据图 28A 中的示例，主节点 50 具有检测单元 51、停止单元 52、停止请求单元 53、发送单元 54、接收单元 55、判定单元 56、重映射单元 57、重映射请求单元 58、重新开始单元 59 以及重新开始请求单元 60。另外，远程节点 70 具有接收单元 71、停止单元 72、完成通知创建单元 73、发送单元 74、重映射单元 75 以及重新开始单元 76。

[0184] 如图 29 所示，检测单元 51 通过判定是否为 CE 地址寄存器 41b 设置存储器 22 的物理地址来判定是否发生了 ICE 或 PCE（步骤 S1101）。当 ICE 或 PCE 没有发生时（在步骤 S1101 中为否），检测单元 51 再次执行步骤 S1101 中的处理。另外，当 ICE 或 PCE 发生时（在步骤 S1101 中为是），检测单元 51 判定在共享存储器中是否发生了 ICE 或 PCE（步骤 S1102）。例如，检测单元 51 判定为 CE 地址寄存器 41b 设置的物理地址是否是共享存储器的存储器区域的地址。通过这样的手段，检测单元 51 可以判定在共享存储器中是否发生了 ICE 或 PCE。另外，当为 CE 地址寄存器 41b 设置虚拟地址时，检测单元 51 参考登记有将虚拟地址转换为物理地址的等式的表，根据虚拟地址计算物理地址，并且判定物理地址是否是共享存储器的存储器区域的地址。同时，这样的表包括将表示地址区域的信息、由地址区域表示的物理地址的范围、将虚拟地址转换为物理地址的等式以及将物理地址转换为虚拟地址的等式关联并注册的条目。例如，检测单元 51 参考图 28B 中所示的表来根据虚拟地址计算物理地址。

[0185] 当在共享存储器中没有发生 ICE 或 PCE 时（在步骤 S1102 中为否），检测单元 51 执行预定处理（步骤 S1103），并且结束处理。同时，预定处理参考例如退化或保留目标页的退化。另外，存储器访问单元 41 可以通过忽略 CE 的发生来执行将数据发送至高速缓存目录管理单元 36 的处理。

[0186] 同时，当在共享存储器中发生了 ICE 或 PCE 时（在步骤 S1102 中为是），停止单元 52 执行访问停止处理（步骤 S1104）。另外，停止请求单元 53 将指示使用共享存储器的另一节点（远程节点 70）OS 停止访问共享存储器的指令（访问停止请求）发送至发送单元 54（步骤 S1105）。通过这样的手段，发送单元 54 将访问停止请求发送至使用共享存储器的另一节点。另外，接收到访问停止请求的远程节点 70 的 OS 执行访问停止处理，并且停止对共享存储器的访问。另外，远程节点 70 的 OS 向主节点 50 通知完成了共享存储器的访问停止处理。另外，发送单元 54 将访问停止请求发送至与本地节点连接的所有其它节点，或者参考在图 28C 的示例中所示的表，指定使用共享存储器的节点并将访问停止请求发送至所指定的节点。在图 28C 的示例中所示的表包括使共享存储器的物理地址、共享存储器的虚拟地址、共享存储器的存储器长度、使用共享存储器的节点的标识符和表示下一条目的指针关联的条目。

[0187] 随后，判定单元 56 判定本地节点（主节点 50）和其它节点是否停止访问共享存储器（步骤 S1106）。例如当检查到停止单元 52 完成访问停止处理时，判定单元 56 判定本地节点停止访问共享存储器。另外，当接收单元 55 接收到完成了对共享存储器的访问停止处理的通知时，判定单元 56 判定已发送通知的另一节点停止访问共享存储器。

[0188] 当判定本地节点（主节点 50）和另一节点停止访问共享存储器时（在步骤 S1106 中为是），重映射单元 57 执行页面重映射处理（步骤 S1107）。

[0189] 同时，根据具体示例将描述页面重映射处理的处理内容。例如，根据页面重映射处理，包括发生了 ICE 或 PCE 的存储器区域的页面首先被分割成多个页面，以使发生了 ICE 或 PCE 的存储器区域包括在具有最小大小的分割页面中。当例如一个页面具有 256MB 时，包括发生了 ICE 或 PCE 的存储器区域的一个页面（256MB）被如下分割成多个页面，以使包括发生了 ICE 或 PCE 的存储器区域的页面包括在最小大小为 8KB 的页面中。例如，一个页面（256MB）被分割成 128MB （1 页）、8KM （8 页）、64KB （63 页）以及 4MB （31 页）的总共 103 个页面。通过这样的手段，能够将复制量从 256KB 抑制为 8KB。执行该页面分割处理以减少复制处理时间，并且不一定是不可缺少的处理。即，不必要执行该页面分割处理。替选地，仅当可以使用（未被应用程序使用）的节点间共享存储器具有一定容量或更少容量、或者不足时，可以执行页面分割处理。这是因为即使不能获得 256MB 的页面，也可以获得 8KB 的页面。另外，根据页面重映射处理，获得复制目的地页面。另外，在启动 OS 或启动应用程序仅可以获得由用户指定的或系统固定的复制目的地页面的大小，以使得可以获得复制目的地页面而不会失败。随后，根据页面重映射处理，包括发生了 ICE 或 PCE 的存储器区域的页面（8KB）的页面被复制到复制目的地页面。另外，当没有执行页面分割处理时，根据上述示例复制诸如整个 256MB 的整个原始页面。随后，根据页面重映射处理，为 OS 的管理区域设置新的页面配置。

[0190] 随后，重映射请求单元 58 向发送单元 54 发送将页面重映射到使用共享存储器的另一节点的 OS 的指令（页面重映射请求）（步骤 S1108）。通过这样的手段，发送单元 54 将页面重映射请求发送至使用共享存储器的另一节点。另外，已接收到页面重映射请求的远程节点 70 的 OS 执行页面重映射处理，并且重映射页面。另外，远程节点 70 的 OS 向主节点 50 通知完成了页面重映射处理。另外，发送单元 54 将页面重映射请求发送至与本地节点连接的所有其它节点，或者参考图 28C 的示例中所示的表，指定使用共享存储器的节点并将页面重映射请求发送至所指定的节点。

[0191] 随后，判定单元 56 判定本地节点和其它节点是否完成了对页面重映射（步骤 S1109）。例如，当检查到重映射单元 57 完成了页面重映射处理时，判定单元 56 判定本地节点完成了对页面重映射。另外，当接收单元 55 接收到完成了页面重映射处理的通知时，判定单元 56 判定发送了通知的另一节点完成了对页面重映射。

[0192] 当判定本地节点和另一节点完成了对页面重映射时（在步骤 S1109 中为是），重新开始单元 59 执行访问重新开始处理（步骤 S1110）。另外，重新开始请求单元 60 将指示使用共享存储器的另一节点的 OS 重新开始对共享存储器的访问的指令（访问重新开始请求）发送至发送单元 54 （步骤 S1111），并且结束处理。通过这样的手段，发送单元 54 将访问重新开始请求发送至使用共享存储器的另一节点。另外，接收到访问重新开始请求的远程节点 70 的 OS 执行访问重新开始处理，并且重新开始对共享存储器的访问。另外，远程节点 70 的 OS 向主节点 50 通知完成了访问重新开始处理。另外，发送单元 54 将访问重新开始请求发送至与本地节点连接的所有其它节点，或者参考在图 28C 的示例中所示的表，指定使用共享存储器的节点并将访问重新开始请求发送至所指定的节点。

[0193] 接下来，将使用图 30 描述由接收到访问停止请求的远程节点 70 的 OS 执行的处理。图 30 是用于说明由接收到访问停止请求的远程节点的 OS 执行的处理的流程图。如图 30 所示，停止单元 72 执行访问停止处理（步骤 S1201）。另外，完成通知创建单元 73 向发送

单元 74 通知完成了访问停止处理(步骤 S1202),并且结束处理。同时,接收到完成了访问停止处理的通知的发送单元 74 将完成了访问停止处理的通知发送至主节点 50。

[0194] 接下来,将使用图 31 描述由接收到页面重映射请求的远程节点 70 的 OS 执行的处理。图 31 是用于说明由接收到页面重映射请求的远程节点的 OS 执行的处理的流程图。如图 31 所示,重映射单元 75 执行页面重映射处理(步骤 S1301)。根据该页面重映射处理,为 OS 的管理区域设置新的页面配置(VA 与 RA 之间的新对应关系)。另外,完成通知创建单元 73 向发送单元 74 通知完成了页面重映射处理(步骤 S1302),并且结束处理。同时,接收到完成了页面重映射处理的通知的发送单元 74 将完成了页面重映射处理的通知发送至主节点 50。

[0195] 接下来,将使用图 32 描述由接收到访问重新开始请求的远程节点 70 的 OS 执行的处理。图 32 是用于说明由接收到访问重新开始请求的远程节点的 OS 执行的处理的流程图。如图 32 所示,重新开始单元 76 执行访问重新开始处理(步骤 S1401),并且结束处理。另外,完成通知创建单元 73 也可以在步骤 S1401 的处理之后向发送单元 74 通知完成了访问停止处理。同时,接收到完成了访问停止处理的通知的发送单元 74 将完成了访问停止处理的通知发送至主节点 50。

[0196] 接下来,将使用图 33 描述图 29 中的步骤 S1104 和图 30 中的步骤 S1201 的访问停止处理的流程。图 33 是用于说明访问停止处理的流程图。如图 33 所示,停止单元 52 (停止单元 72) 获取用于所指定的共享存储器的存储器管理表(页面管理表) 的 I/O 处理锁定(步骤 S1501)。通过这样的手段,挂起对 I/O 装置的访问。

[0197] 同时,将描述存储器管理表的数据配置的示例。图 34 是示出存储器管理表的数据配置的示例的图。图 34 的示例中的存储器管理表包括登记了表示访问停止标记的开 / 关状态的值的“访问停止标记”项和登记了表示 I/O 处理锁定的状态的值的“I/O 处理锁定”项。另外,图 34 的示例中的存储器管理表包括登记了指向另一页面管理表的指针的“指向另一页面管理表的指针”项和登记了指向地址转换表的指针的“指向地址转换表的指针”项。另外,图 34 的示例中的存储器管理表包括登记了各种类型的其它管理信息项的“其它管理信息”项。

[0198] 另外,将描述由在存储器管理表中所登记的“指向地址转换表的指针”表示的地址转换表。图 35 是示出地址转换表的数据配置的示例的图。图 35 的示例中的地址转换表包括登记了存储器 22 的物理地址的“PA”项,登记了与物理地址相关联的虚拟地址的“VA”项、以及登记了由存储器 22 的物理地址表示的存储器区域的大小的“区域长度”项。另外,图 35 的示例中的地址转换表包括登记了指向页面管理表的指针的“指向页面管理表的指针”项和登记了指向另一地址转换表的指针的“指向另一地址转换表的指针”项。另外,图 35 的示例中的地址转换表包括登记了各种类型的其它管理信息项的“其它管理信息”项。

[0199] 另外,停止单元 52 (停止单元 72) 为所指定的共享存储器设置存储器管理表的访问停止标记(步骤 S1502)。通过这样的手段,停止对共享存储器的访问。随后,停止单元 52 (停止单元 72)参考地址转换表,并且当在 TLB 35a 中登记将共享存储器的虚拟地址与物理地址相关联的条目时执行以下处理。即,停止单元 52 (停止单元 72) 从 TLB 35a 删除条目(步骤 S1503),并且结束处理。

[0200] 同时,当从 TLB 35a 删除将共享存储器的虚拟地址与物理地址相关联的条目时,

如果应用程序访问共享存储器，则 TLB 失误发生。根据本示例，当这样的 TLB 失误发生时，作为中断处理，执行 TLB 失误处理，并且阻止应用程序对共享存储器的访问。

[0201] 图 36 是用于说明 TLB 失误处理的流程的流程图。如图 36 所示，OS 根据中断发生的进程计数器指定进程（步骤 S1601）。随后，OS 根据发生地址搜索图 35 中所示的地址转换表。当发现对应的地址转换表时，检查由该地址转换表中的页面管理表的指针指示的页面管理表。判定由访问目的地地址表示的存储器区域是停止了访问的存储器区域，是发生换出（swap out）的存储器区域（其中记录有已执行换出的信息），还是异常发生的存储器区域（不具有地址转换表的存储器区域）（步骤 S1602）。

[0202] 当访问停止标记为开时（即，当存储器区域是停止了访问的存储器区域时）（步骤 S1602：停止访问），OS 从正执行进程列表中移除目标进程信息，并且将正执行进程列表移动至访问重新开始等待列表（步骤 S1603）。同时，将描述访问重新开始等待列表的数据配置的示例。图 37 是示出访问重新开始等待列表的数据配置的示例的图。图 37 的示例中所示的访问重新开始等待列表包括登记了用于识别进程的标识符的“进程标识符”项、以及登记了表示撤销了登记信息（诸如，重新开始进程计数器）的区域的指针的“进程恢复信息”。另外，图 37 的示例中所示的访问重新开始等待列表包括登记了重新开始等待共享存储器的虚拟地址的“重新开始等待共享存储器地址”项、以及登记了指向下一列表的指针的“指向下一列表的指针”项。

[0203] 随后，OS 请求 OS 中的调度模块启动另一进程（步骤 S1606），并且结束处理。

[0204] 另外，在换出发生的存储器区域的情况下（步骤 S1602：换出），OS 在对访问目的地页面的处理中启动交换（步骤 S1604）。另外，OS 从正执行进程列表中移除目标进程信息并将目标进程信息移动到调度等待列表（步骤 S1605），并且进行至步骤 S1606。同时，将描述调度等待列表的数据配置的示例。图 38 是示出访问重新开始等待列表的数据配置的示例的图。图 38 的示例中所示的访问重新开始等待列表包括登记了用于识别进程的标识符的“进程标识符”项、以及登记了表示撤销了登记信息（诸如，重新开始进程计数器）的区域的指针的“进程恢复信息”项。另外，图 38 的示例中所示的访问重新开始等待列表包括登记了指向下一列表的指针的“指向下一列表的指针”项。

[0205] 此外，在异常发生的存储器区域的情况下（步骤 S1602：异常地址），OS 执行引起在访问进程中的访问错误的“异常地址访问处理”（步骤 S1607），并且结束处理。

[0206] 接下来，将使用图 39 描述在图 29 的步骤 S1110 和图 32 的步骤 S1401 中的访问重新开始处理的流程。图 39 是用于说明访问重新开始处理的流程图。如图 39 所示，重新开始单元 59（重新开始单元 76）使用于所指定的共享存储器的存储器管理表的访问停止标记清零（步骤 S1701）。通过这样的手段，CPU 重新开始对共享存储器的访问。

[0207] 另外，重新开始单元 59（重新开始单元 76）释放用于所指定的共享存储器的存储器管理表的 I/O 处理锁定（步骤 S1702）。通过这样的手段，I/O 重新开始访问。随后，重新开始单元 59（重新开始单元 76）检查访问重新开始等待进程列表并将进程移动到调度等待列表（步骤 S1703），并且结束处理。

[0208] 另外，例如由停止单元 52、停止单元 72、重新开始单元 59 以及重新开始单元 76 执行的处理可通过进程来执行。图 40 是示出根据实施例的 OS 的功能配置和由 CPU 执行的进程的示例的图。图 40 的示例与图 28A 的示例的区别在于与图 28A 的示例相比，进程包括停

止单元 52、停止单元 72、重新开始单元 59 以及重新开始单元 76。同时，OS 创建预先将共享存储器的物理地址与各种类型的事件处理器相关联的信息。例如，当应用程序请求 OS 将与该事件相关的“共享存储器的地址”、“需要接收的事件类型”（访问停止请求和访问重新开始请求）和“事件处理器程序的地址“相关联时创建该信息。应用程序通常仅知道“共享存储器的虚拟地址（VA）”，并且不知道物理地址（PA）。然后，OS 侧将 VA 转换为 PA，并且记录 PA。当与该 PA 相关的事件（例如，访问停止请求）发生时，启动与该事件和 PA 相关联的事件处理器程序。（登记事件处理器程序的开始地址，并且从该开始地址起开始程序操作）。

[0209] 另外，当检测单元 51 检测到 ICE 或 PCE 时，停止单元 52 和停止单元 72 参考创建的信息，并且对与为 CE 地址寄存器 41b 设置的共享存储器的物理地址相关联的第一事件处理器进行读取。更具体地，当在应用程序进行操作的同时事件发生时，将在该时间点进行操作的寄存器信息撤销到堆栈，并且从第一事件处理器程序的开始地址起开始事件处理器程序的操作。然后，第一事件处理器停止从应用程序对共享存储器的每次访问（读取 / 写入以及 I/O 访问）。停止访问的方法包括（1）在程序中准备并创建“用于共享存储器的访问停止标记”，以使得仅当标记不为开时应用程序继续访问。该方法还包括在第一事件处理器中开启该标记，以及（2）在第一事件处理器中停止应用程序的操作并且停止应用程序处理的每个进程。

[0210] 另外，当判定单元 56 判定本地节点和另一节点重映射页面时，重新开始单元 59 参考所创建的信息，并且对与为 CE 地址寄存器 41b 设置的共享存储器的物理地址相关联的第二事件处理器进行读取。然后，应用程序（进程）重新开始对所指定的共享存储器的所有停止的访问（读取 / 写入以及 I/O 访问）。另外，当接收到访问重新开始请求时，重新开始单元 76 参考创建的信息，并且对与为 CE 地址寄存器 41b 设置的共享存储器的物理地址相关联的第二事件处理器进行读取。更具体地，当在应用程序进行操作的同时事件发生时，将在该时间点进行操作的寄存器信息撤销到堆栈，并且从第一事件处理器程序的开始地址起开始事件处理器程序的操作。然后，第二事件处理器重新开始从应用程序（进程）对所指定的共享存储器的所有停止的访问。重新开始访问的方法包括（1）在程序中准备并创建“共享存储器的访问停止标记”，以使得仅当标记不为开时应用程序继续访问。该方法还包括在第二事件处理器中关闭标记，以及（2）由于应用程序停止在第一事件处理器中进行操作，因此对 PC 重新写入并且从第一事件处理器返回应用程序处理。另外，从重新开始单元 59 直接调用的本地节点中的事件处理器和从重新开始单元 76 调用的另一节点中的事件处理器可以是包括单个指令序列的程序或包括不同指令序列的程序。（这是程序员的偏好，并且可以执行两者）。

[0211] 实施例的效果

[0212] 如上所述，信息处理系统 1 包括每一个均具有存储器的多个构件块和连接在多个构件块之间的 XB 2。多个构件块的至少一个构件块 10 执行对如下数据的处理：该数据包括在构件块 10 或其它构件块的存储器中，并且存储在构件块 10 或其它构件块访问的共享存储器区域中。即，构件块 10 检测在预定时间内发生了超过预定次数的 ICE，或者检测在共享存储器区域中的单个位置处发生的 PCE。当检测到错误时，构件块 10 执行控制以阻止构件块 10 和其它构件块访问共享存储器。构件块 10 在不同于共享存储器区域的存储器中的恢复数据。构件块 10 向另一构件块通知不同的存储器区域。构件块 10 执行控制以重新开

始从构件块 10 或其它构件块对共享存储器的访问。因此，信息处理系统 1 可以抑制信息处理系统 1 发生故障的可能性。

[0213] 另外，信息处理系统 1 基于所接收到的物理地址判定访问目标是共享区域还是本地区域，并且可以为存储在本地区域中的内核数据或用户数据保持高的安全级别。此外，信息处理系统 1 使得所有存储器可高速缓存，并且可以容易地掩盖在存储器访问时的潜伏时间（latency）。

[0214] 另外，根据与对存储器 22 的访问相同的方法，CPU 21 访问另一 CPU 访问的存储器的共享区域。即，即使当在存储器 22 或其它存储器上存在访问目标存储器区域时，CPU 21 的计算单元 31 仅需要输出虚拟地址。

[0215] 因此，CPU 21 可以容易地访问共享区域而不执行例如诸如 I/O 的排他控制的处理和编程，并且因此，改善了存储器访问性能。另外，CPU 21 可以适当地使用共享存储器而不修改要执行的程序或 OS，结果，执行与传统的方法类似的预取处理，使得能够改善存储器访问性能。

[0216] 此外，当来自另一 CPU 的存储器访问目标是对本地区域的访问时 CPU 21 返回拒绝响应。因此，信息处理系统 1 阻止对除共享区域之外的区域的访问，结果，可以阻止错误。

[0217] 另外，高速缓存目录管理单元 36 使用节点映射 34 来将物理地址转换为与节点映射 34 相关联地存储的 CPUID。因此，CPU 21 可以识别访问被分有访问目标物理地址的存储器的 CPU。

[0218] 另外，CPU 21 使用用于管理高速缓存存储器 22 中所存储的数据的 CPU 的目录来控制高速缓存一致性。结果，即使当信息处理系统 1 的 CPU 的数量增加时，信息处理系统 1 可以有效地保持高速缓存一致性而不增加 XB 2 的流量（traffic）。

[0219] 更具体地，在信息处理系统 1 中，CPU 之间的通信将被限制在远程 CPU 与主 CPU 之间或在远程 CPU、主 CPU 与高速缓存更新后的数据的本地 CPU 之间。结果，信息处理系统 1 可以有效地保持高速缓存一致性。

[0220] 尽管描述了本发明的示例，但是除了上述示例之外，还可以实现各种不同的示例。在下文中，将描述其它实施例。

[0221] （1）关于构件块

[0222] 以上信息处理系统 1 包括具有 4 个 CPU 的构件块 10 至 10e。然而，示例并不限于此，并且构件块 10 至 10e 可以具有任意数量的 CPU 和每个 CPU 访问的存储器。另外，CPU 和存储器不必一对一地相关联，并且直接访问存储器的 CPU 可以是整体的一部分。

[0223] （2）关于 CPU 发送的分组

[0224] 上述 CPU 21 发送包括 CPUID 和 PA（物理地址）的分组作为存储器访问请求。然而，示例并不限于此。即，如果 CPU 21 可以唯一地识别访问访问目标存储器的 CPU，则 CPU 21 可输出存储任意信息的分组。

[0225] 另外，例如，CPU 21 可将 CPUID 转换为 VC（虚拟连接）ID，并且存储 VCID。此外，CPU 21 可将诸如表示数据长度的长度的信息存储在分组中。

[0226] （3）关于 CPU 发出的命令（指令）

[0227] 如上所述，CPU 21 至 21c 中的每一个发出请求或命令，并且保持高速缓存一致性。然而上述请求或命令一贯地是示例，并且例如，CPU 21 至 21c 可发出 CAS（比较并交换）指

令。

[0228] 因此，当 CPU 21 至 21c 发出 CAS 指令时，即使在多个 CPU 之间频繁发生排他控制的争用，也能对每个 CPU 的高速缓存器执行处理。结果，CPU 21 至 21c 可以防止由于存储器访问的发生引起的延迟，并且防止 CPU 之间的事务拥堵。

[0229] (4)关于通过管理程序的控制

[0230] 描述了 OS 访问作为信息处理系统 1 中硬件的地址转换单元 35 的示例。然而，示例不限于此，并且例如操作虚拟机的管理程序(HVP)可访问地址转换单元 35。

[0231] 即，在管理程序进行操作的节点中，OS 请求管理程序执行操作而不直接地操作 CPU 21 至 21c 的硬件资源(诸如，高速缓存器和 MMU)。因此，当通过管理程序控制 CPU 21 至 21c 中的每一个时，CPU 21 至 21c 中的每一个将虚拟地址转换为实际地址(RA)，然后，将实际地址转换为物理地址。

[0232] 另外，在管理程序进行操作的节点中，根据中断处理中断 HPV 而不直接中断 OS。在这种情况下，管理程序通过读取 OS 的中断处理处理器来执行中断。另外，通过上述管理程序执行的处理是为了操作虚拟机器执行的公知处理。

[0233] (5)使用分区的处理

[0234] 在上述信息处理系统 1 中，CPU 21 至 21c 中的每一个均使用一个节点映射来发送存储器访问。然而，示例不限于此。例如，构件块 10 至 10e 中的每一个均可作为多个节点组来进行操作，并且按每个节点组配置操作单个固件(管理程序)的一个虚拟分区。

[0235] 在这种情况下，CPU 21 至 21c 中的每一个均包括表示访问目的地 CPU 的节点映射和表示在单个虚拟分区中的 CPU 的节点映射。因此，CPU21 至 21c 中的每一个均包括表示包括在单个虚拟分区中的 CPU 的节点映射，并且可以识别可能不能传递到虚拟分区之外的特殊分组的传递范围，诸如错误发生通知、向下请求或重置请求分组。

[0236] (6)通过服务处理器的控制

[0237] 根据以上信息处理系统 1 描述了服务处理器 24 访问作为硬件的节点映射 34 的示例。然而，示例不限于此，并且除服务处理器 24 之外的单元可被配置成访问节点映射 34。例如，由 CPU 21 至 21c 上的一个或所有 CPU 或 HPV 操作的基本固件 BIOS(基本输入 / 输出系统)可被配置成访问节点映射 34。

[0238] 根据实施例，能够抑制信息处理设备发生故障的可能性。

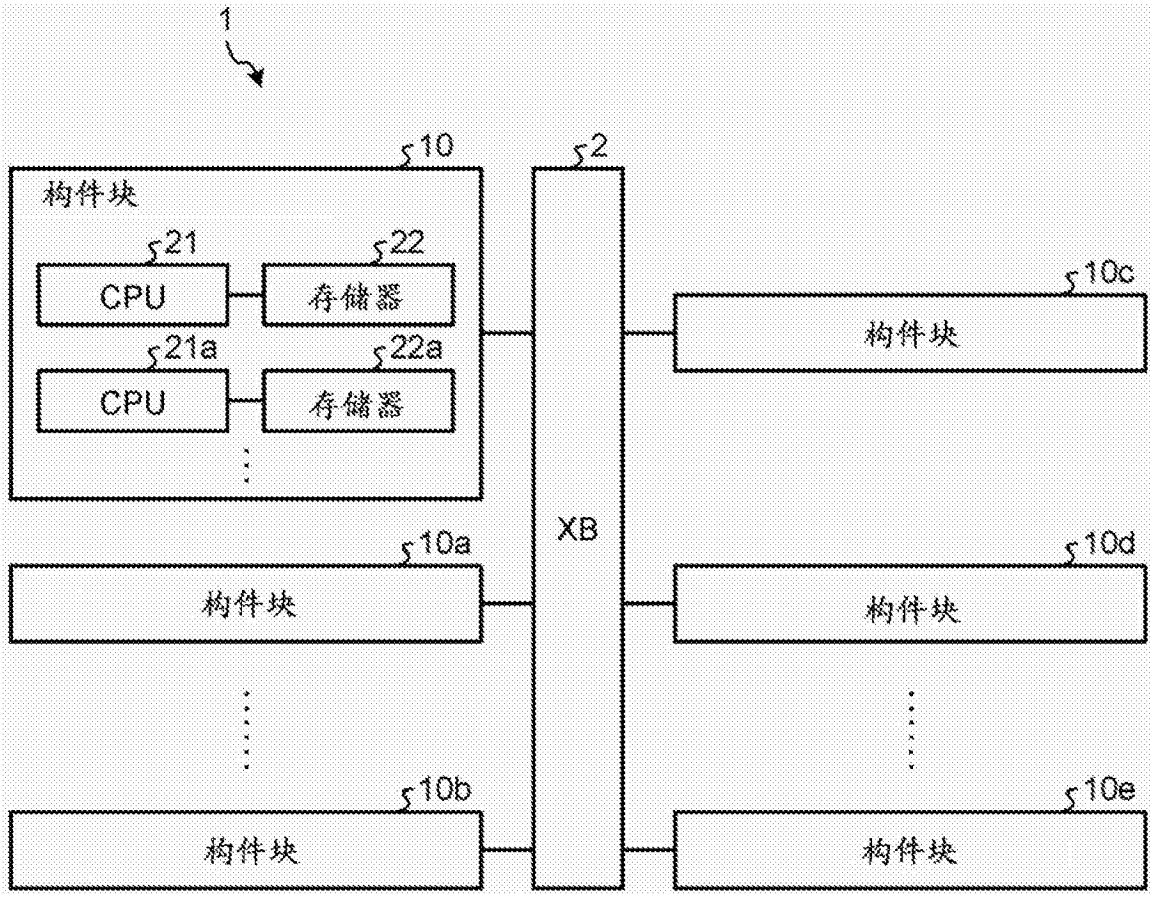

1
10
构件块
21
CPU
22
存储器
21a
CPU
22a
存储器
2
XB
10a
构件块
10b
构件块
10c
构件块
10d
构件块
10e
构件块

图 1

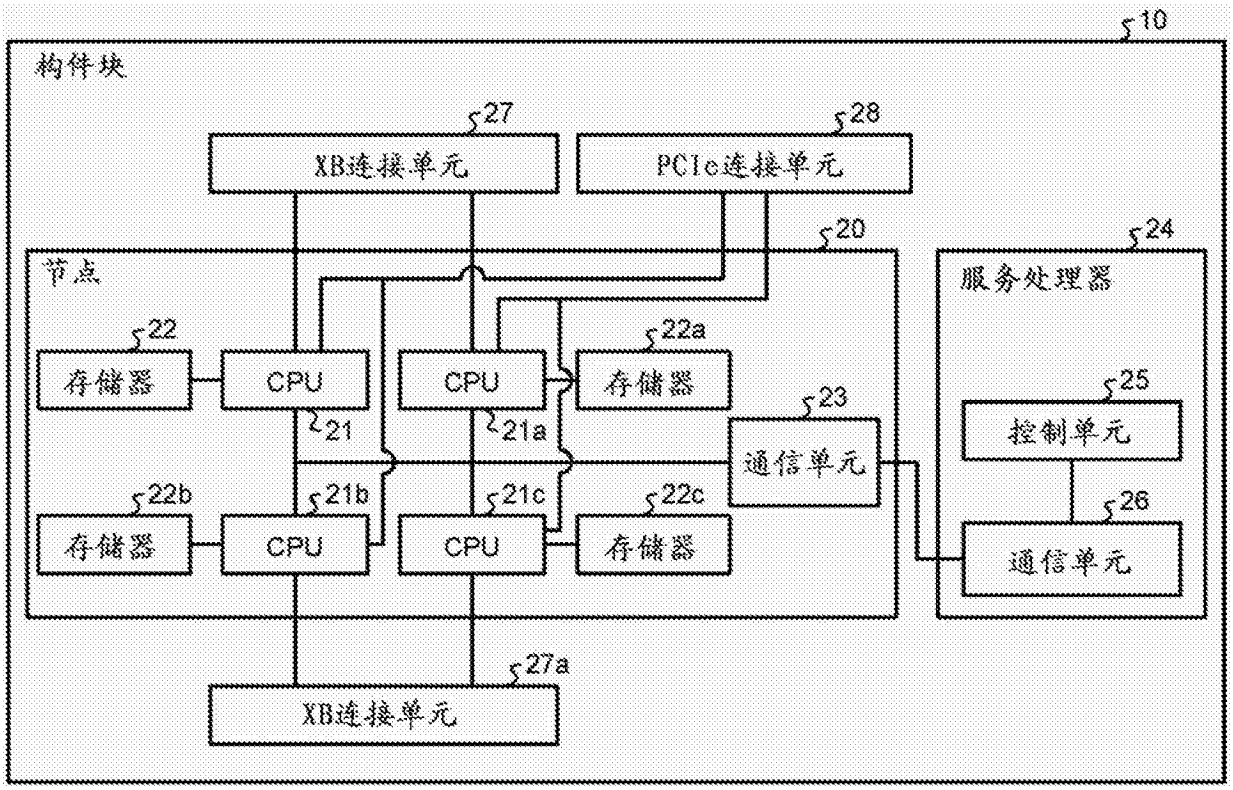

10
构件块
27
XB连接单元
28
PCIe连接单元
20
节点
24
服务处理器
22
存储器
CPU
CPU
22a
存储器
21
21a
23
通信单元
25
控制单元
22b
21b
21c
22c
26
存储器
CPU
CPU
存储器
通信单元
27a
XB连接单元

图 2

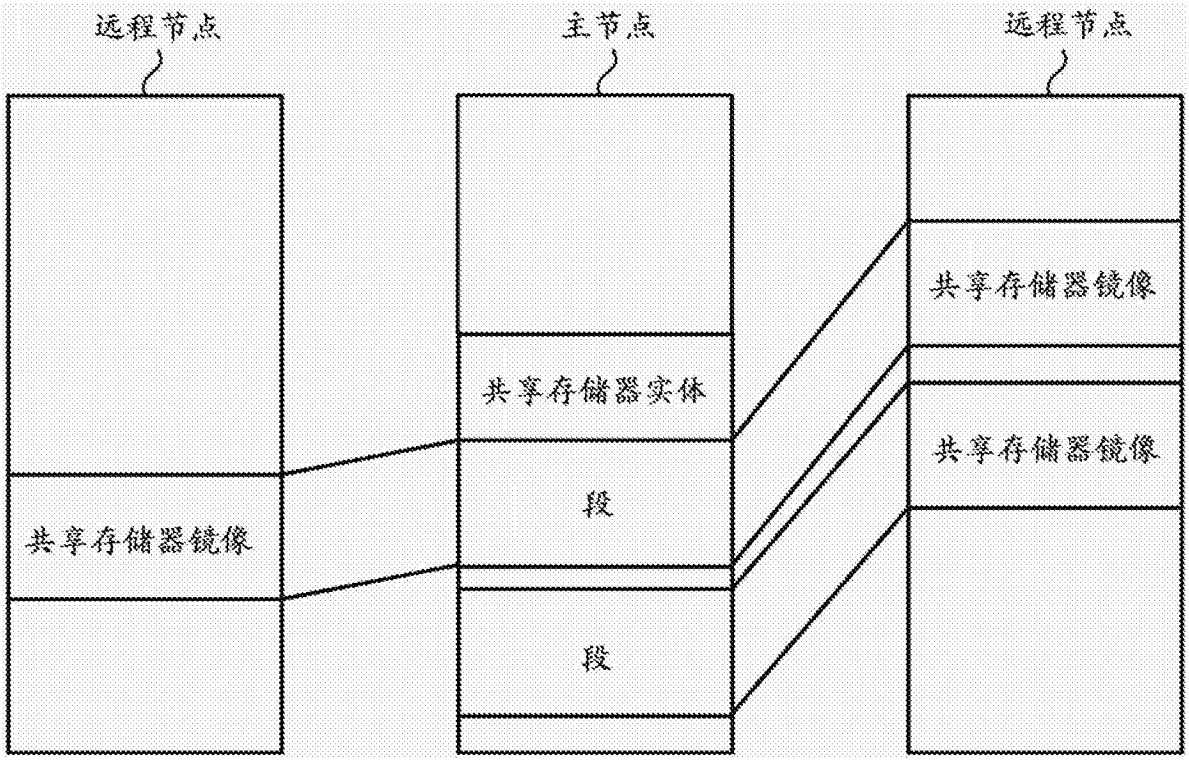

远程节点
主节点
远程节点
共享存储器镜像
共享存储器实体
段
段
共享存储器镜像
共享存储器镜像

图 3

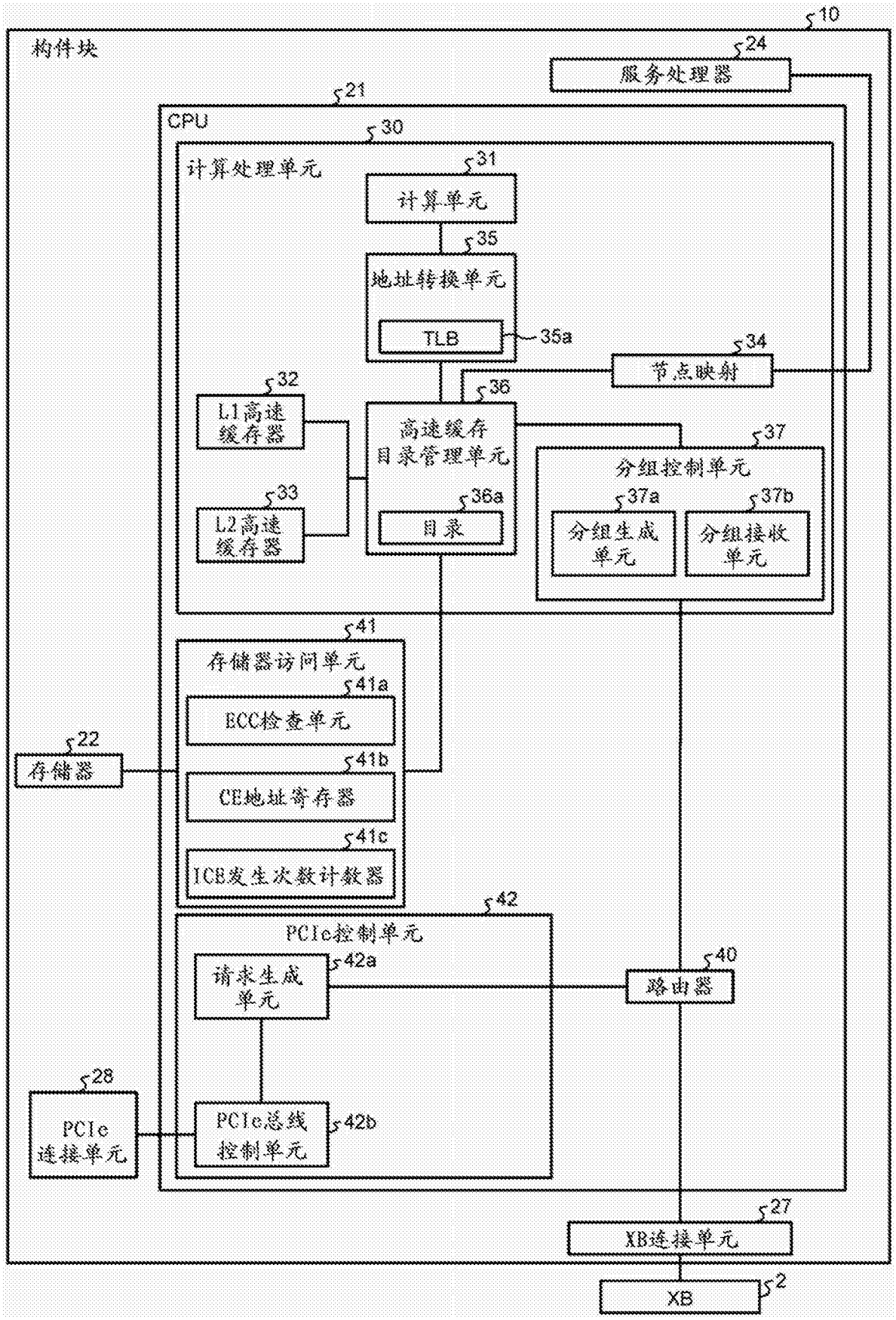

10
构件块
24
服务处理器
21
CPU
30
计算处理单元
31
计算单元
35
地址转换单元
TLB
35a
34
节点映射
32
L1高速缓存器
36
高速缓存目录管理单元
37
分组控制单元
33
L2高速缓存器
36a
目录
37a
分组生成单元
37b
分组接收单元
41
存储器访问单元
41a
ECC检查单元
22
存储器
41b
CE地址寄存器
41c
ICE发生次数计数器
42
PCIe控制单元
42a
请求生成单元
40
路由器
28
PCIe连接单元
PCIe总线控制单元
42b
27
XB连接单元
XB
2

图 4

地址	有效	节点 ID	CPU ID
#0	1	1	4
#1	1	1	5
#2	0		
⋮			

图 5

CPU ID	PA	数据

图 6

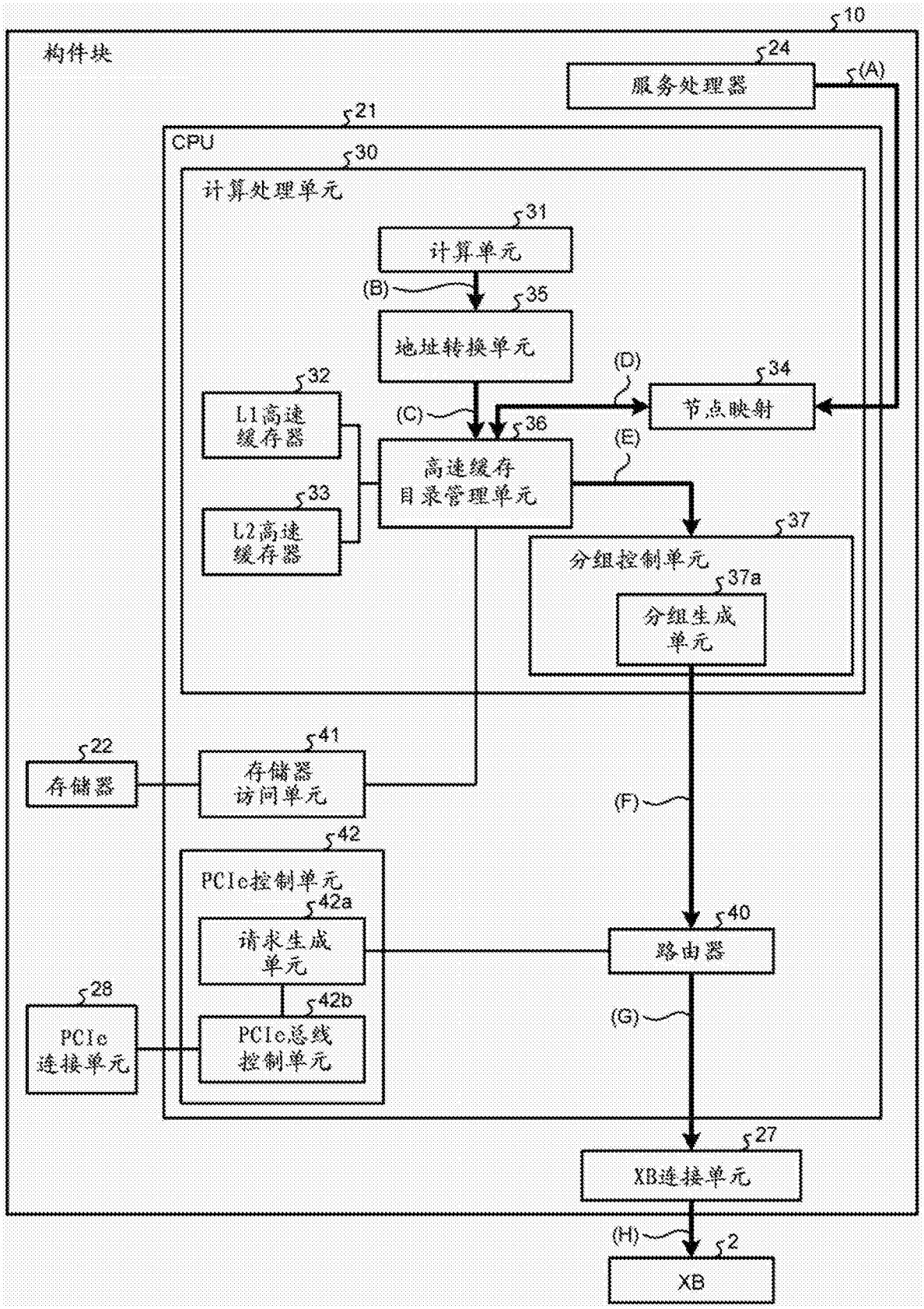

构件块
10
服务处理器
24
(A)
CPU
21
计算处理单元
30
计算单元
31
(B)
地址转换单元
35
L1高速缓存器
32
L2高速缓存器
33
(C)
高速缓存目录管理单元
36
(D)
节点映射
34
(E)
分组控制单元
37
分组生成单元
37a
存储器
22
存储器访问单元
41
(F)
PCIe控制单元
42
请求生成单元
42a
PCIe总线控制单元
42b
PCIe连接单元
28
路由器
40
(G)
XB连接单元
27
(H)
XB
2

图 7

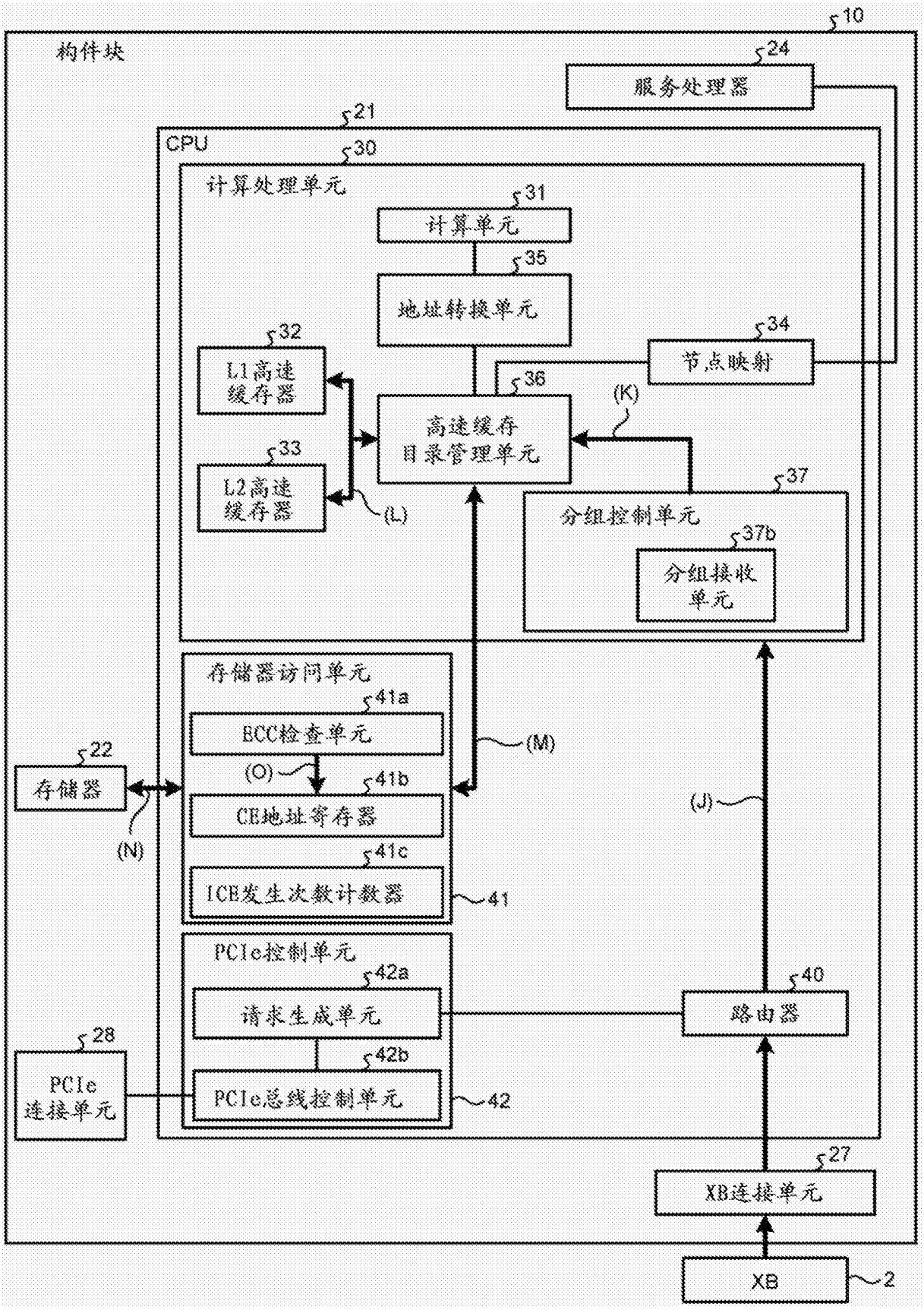

10
构件块
24
服务处理器
21
CPU
30
计算处理单元
31
计算单元
35
地址转换单元
32
L1高速缓存器
34
节点映射
36
(K)
高速缓存目录管理单元
33
L2高速缓存器
(L)
37
分组控制单元
37b
分组接收单元
存储器访问单元
41a
ECC检查单元
(M)
22
存储器
(O)
41b
CE地址寄存器
(N)
(J)
41c
ICE发生次数计数器
41
PCIe控制单元
42a
请求生成单元
40
路由器
28
PCIe连接单元
42b
PCIe总线控制单元
42
27
XB连接单元
XB
2

图 8

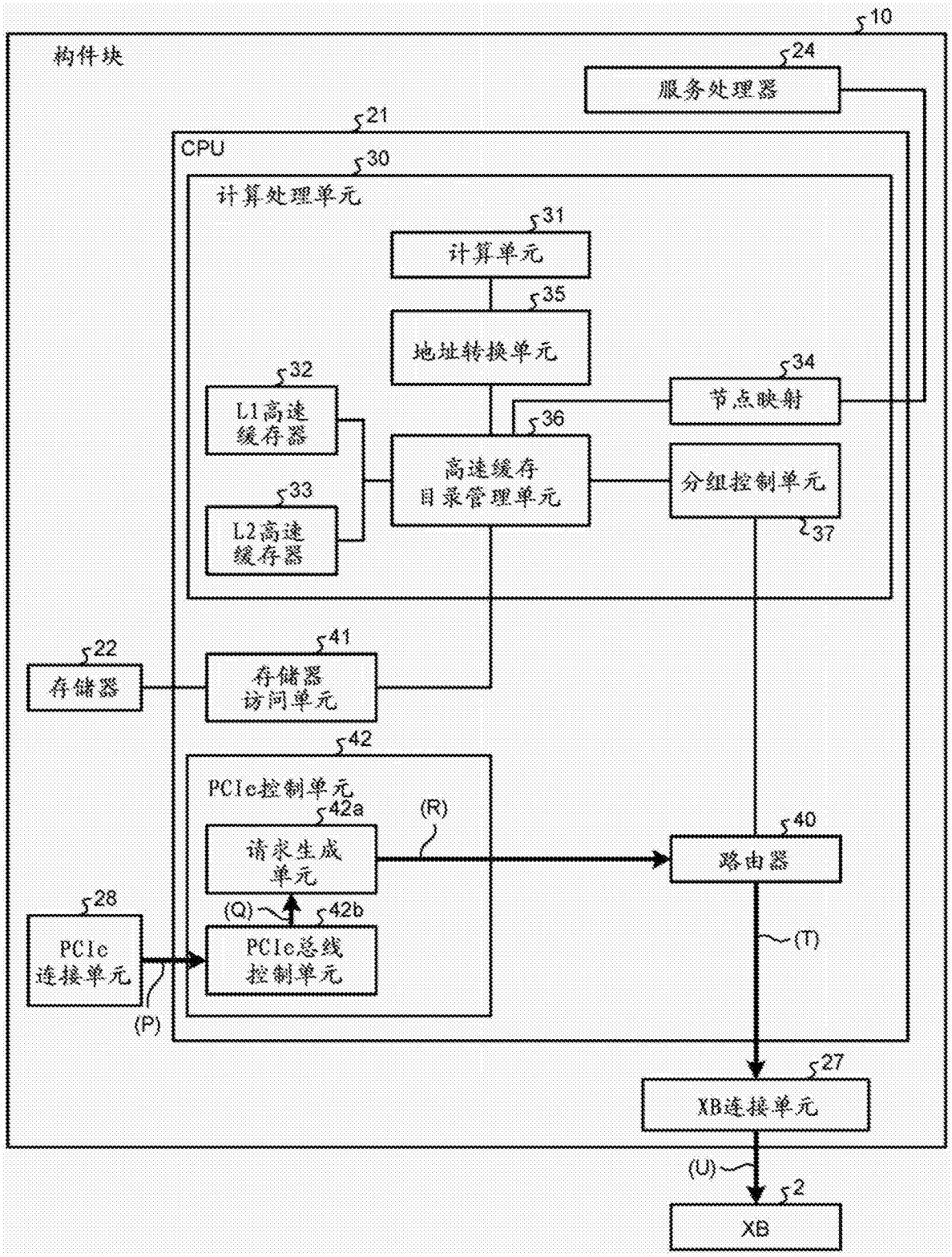

10
构件块
24
服务处理器
21
CPU
30
计算处理单元
31
计算单元
35
地址转换单元
32
L1高速缓存器
34
节点映射
36
高速缓存目录管理单元
分组控制单元
37
33
L2高速缓存器
22
存储器
41
存储器访问单元
42
PCIe控制单元
42a
请求生成单元
(R)
40
路由器
28
PCIe连接单元
(Q)
42b
PCIe总线控制单元
(T)
(P)
27
XB连接单元
(U)
2
XB

图 9

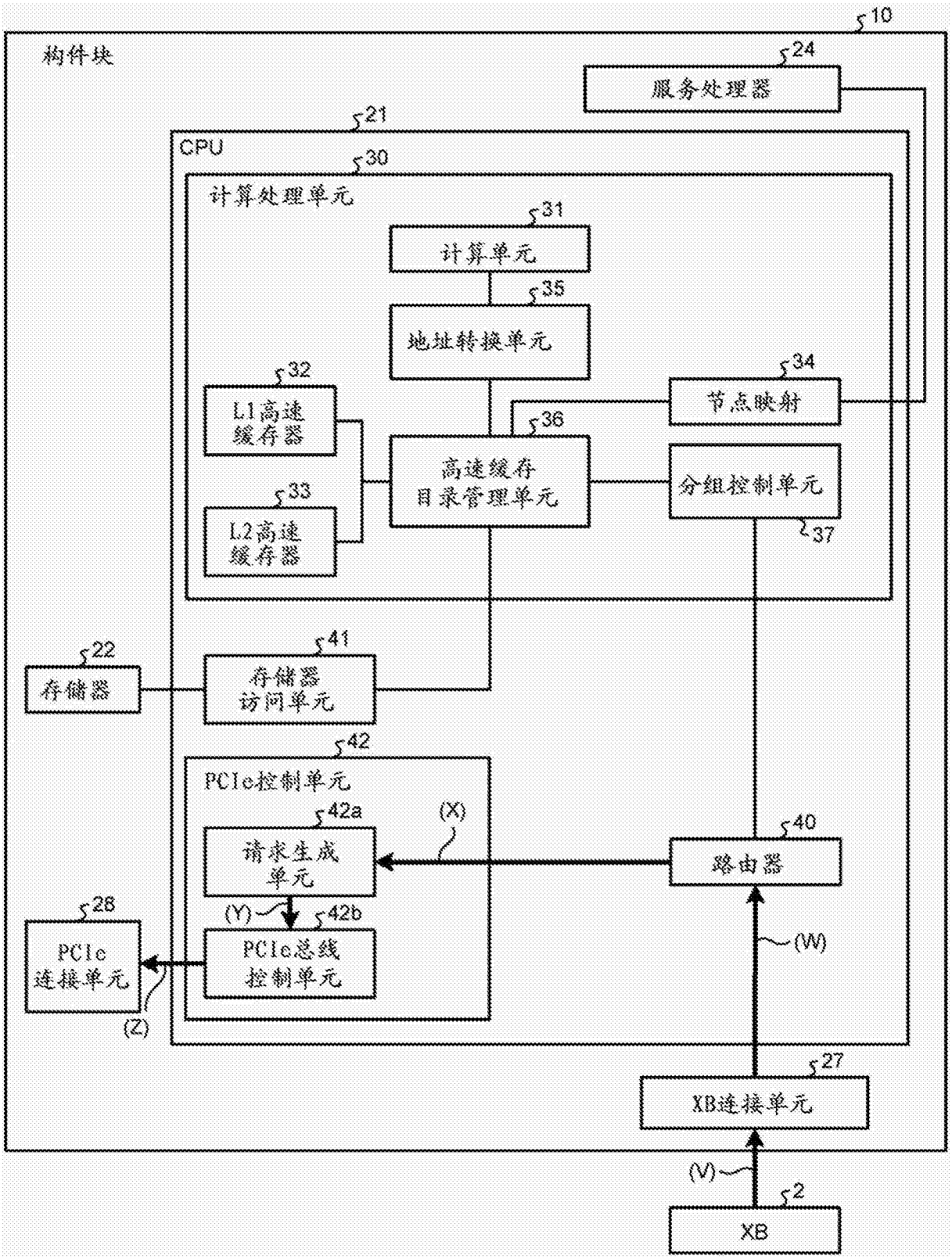

10
构件块
24
服务处理器
21
CPU
30
计算处理单元
31
计算单元
35
地址转换单元
32
L1高速缓存器
33
L2高速缓存器
34
节点映射
36
高速缓存目录管理单元
分组控制单元
37
22
存储器
41
存储器访问单元
42
PCIe控制单元
42a
请求生成单元
(X)
40
路由器
28
PCIe连接单元
(Y)
42b
PCIe总线控制单元
(W)
(Z)
27
XB连接单元
(V)
2
XB

图 10

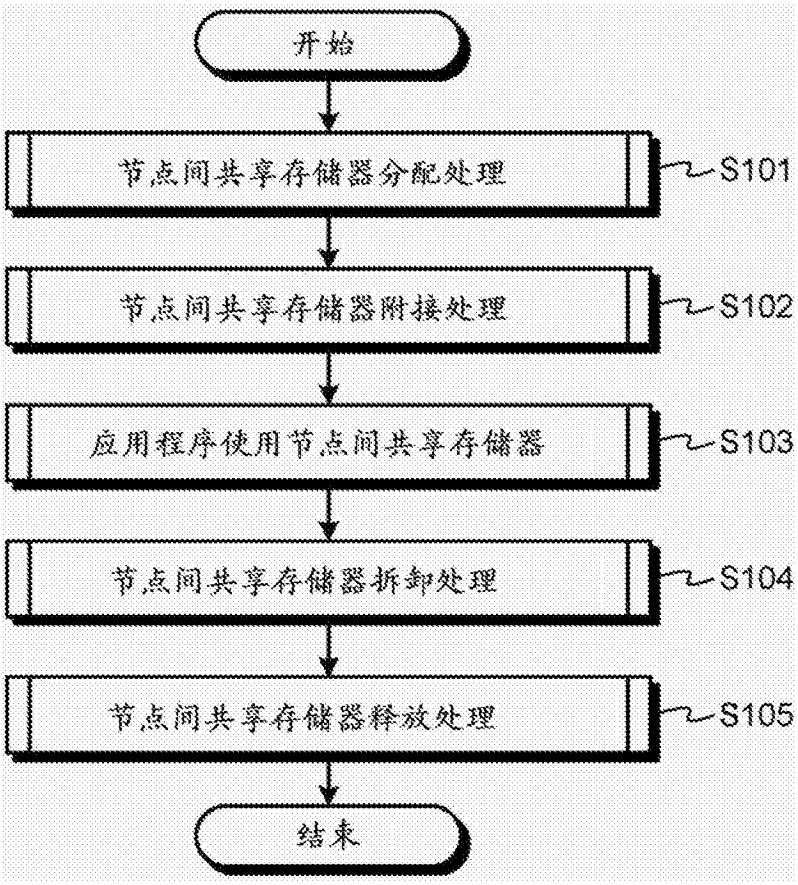

开始
节点间共享存储器分配处理
S101
节点间共享存储器附接处理
S102
应用程序使用节点间共享存储器
S103
节点间共享存储器拆卸处理
S104
节点间共享存储器释放处理
S105
结束

图 11

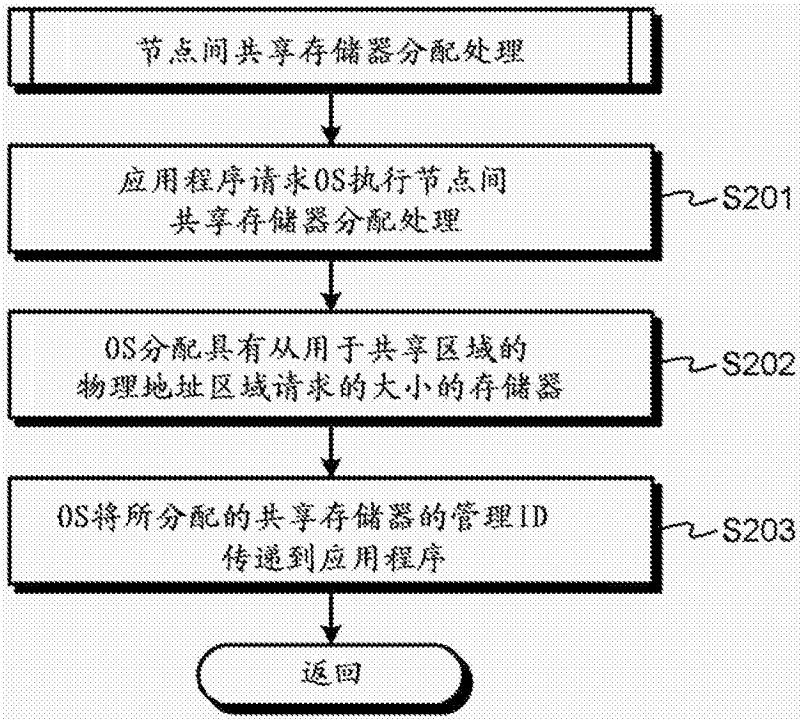

节点间共享存储器分配处理
应用程序请求OS执行节点间
共享存储器分配处理
S201
OS分配具有从用于共享区域的
物理地址区域请求的大小的存储器
S202
OS将所分配的共享存储器的管理ID
传递到应用程序
S203
返回

图 12

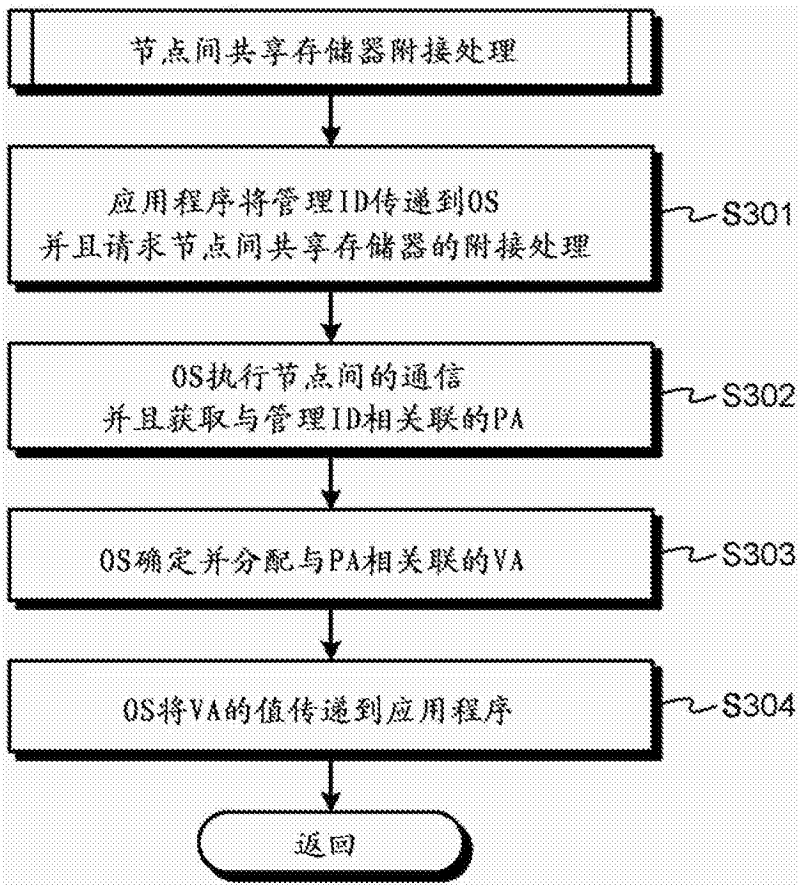

节点间共享存储器附接处理
应用程序将管理ID传递到OS
并且请求节点间共享存储器的附接处理
S301
OS执行节点间的通信
并且获取与管理ID相关联的PA
S302
OS确定并分配与PA相关联的VA
S303
OS将VA的值传递到应用程序
S304
返回

图 13

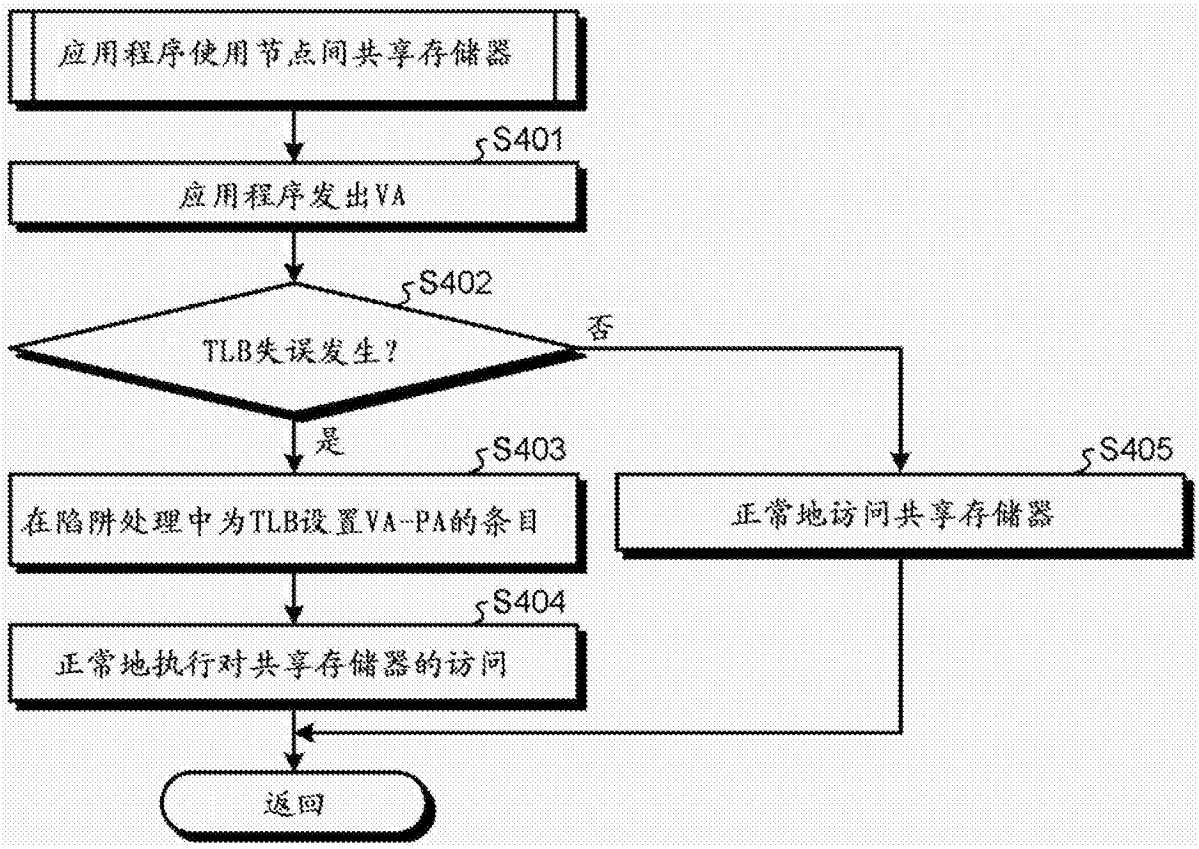

应用程序使用节点间共享存储器
S401
应用程序发出VA
S402
TLB失误发生？
否
是
S403
在陷阱处理中为TLB设置VA-PA的条目
S404
正常地执行对共享存储器的访问
S405
正常地访问共享存储器
返回

图 14

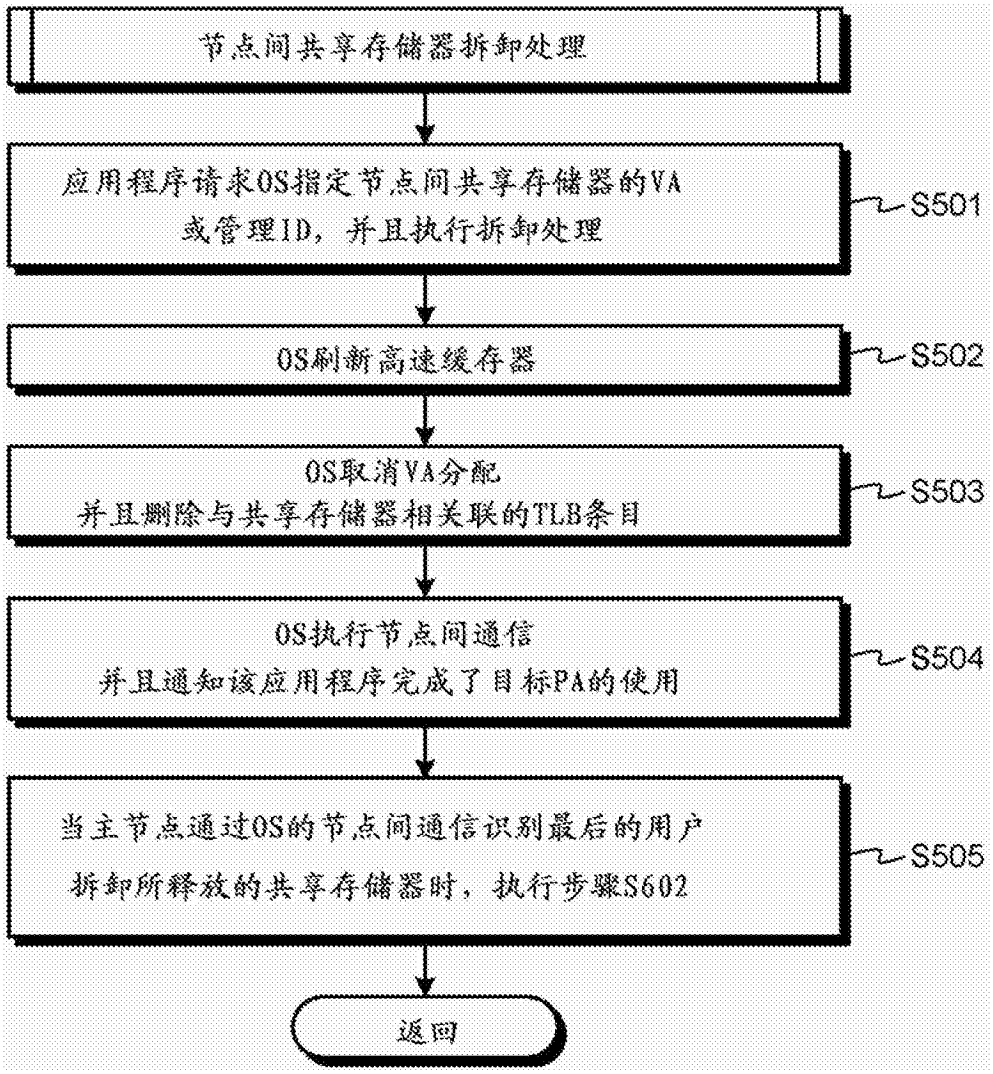

节点间共享存储器拆卸处理
应用程序请求OS指定节点间共享存储器的VA
或管理ID，并且执行拆卸处理
S501
OS刷新高速缓存器
S502
OS取消VA分配
并且删除与共享存储器相关联的TLB条目
S503
OS执行节点间通信
并且通知该应用程序完成了目标PA的使用
S504
当主节点通过OS的节点间通信识别最后的用户
拆卸所释放的共享存储器时，执行步骤S602
S505
返回

图 15

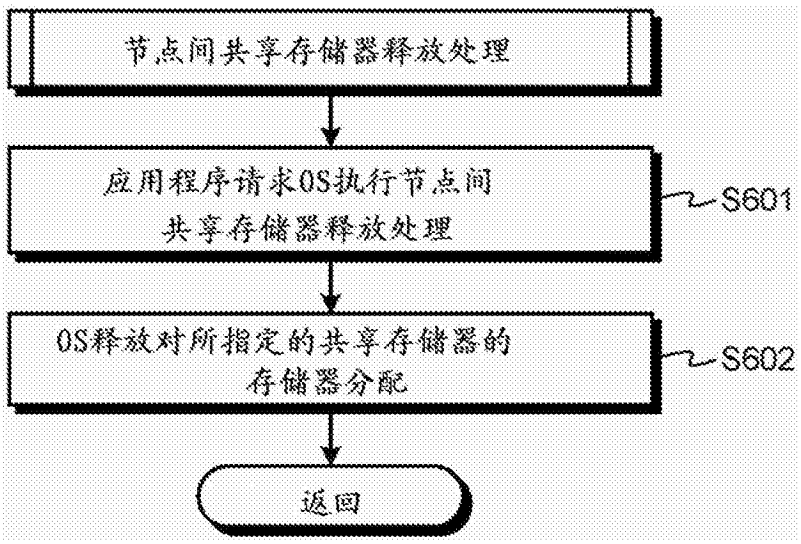

节点间共享存储器释放处理
应用程序请求OS执行节点间
共享存储器释放处理
S601
OS释放对所指定的共享存储器的
存储器分配
S602
返回

图 16

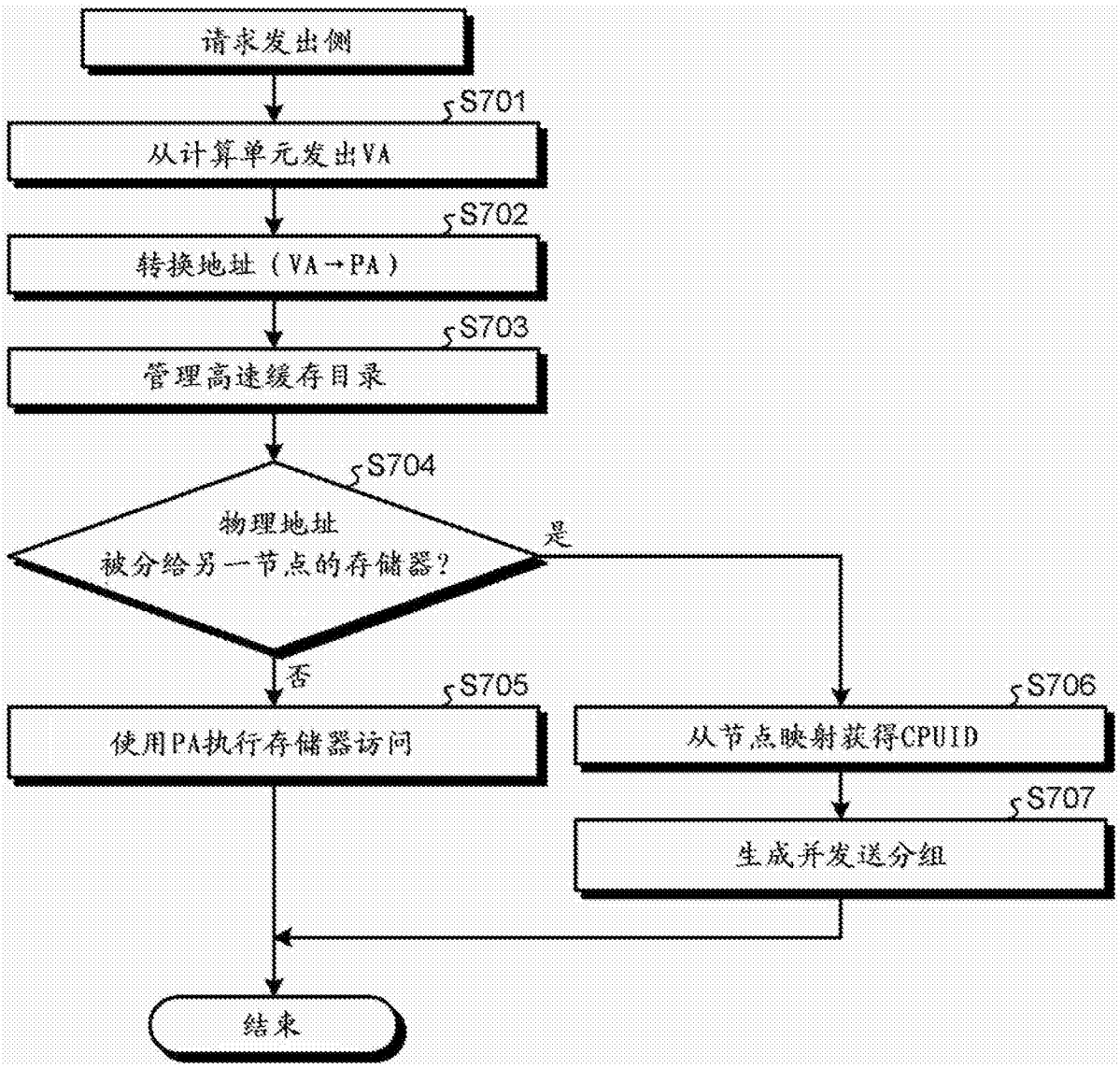

请求发出侧
S701
从计算单元发出VA
S702
转换地址（VA→PA）
S703
管理高速缓存目录
S704
物理地址
被分给另一节点的存储器？
是
否
S705
使用PA执行存储器访问
S706
从节点映射获得CPUID
S707
生成并发送分组
结束

图 17

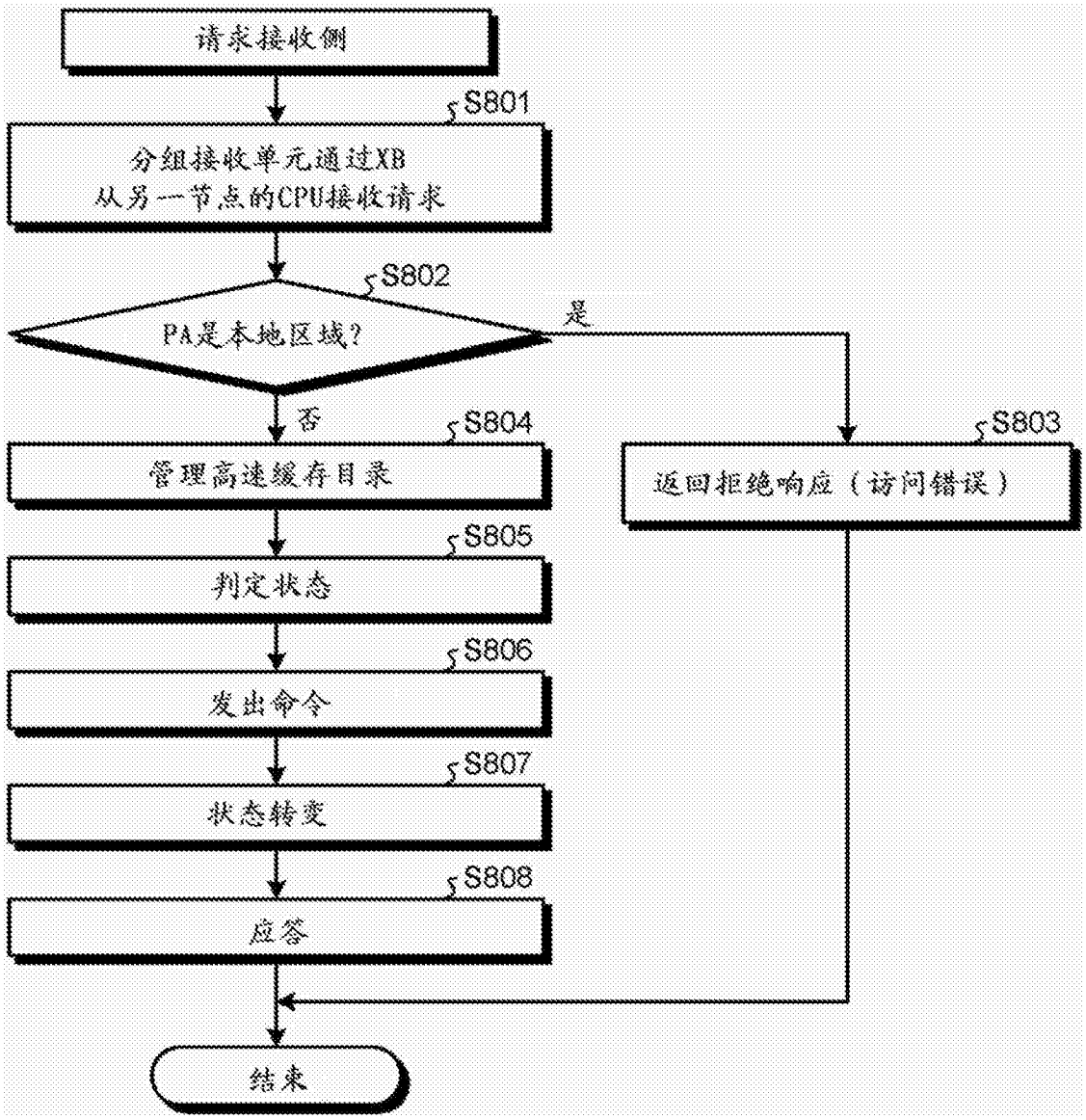

请求接收侧
S801
分组接收单元通过XB
从另一节点的CPU接收请求
S802
PA是本地区域?
是
否
S803
返回拒绝响应（访问错误）
S804
管理高速缓存目录
S805
判定状态
S806
发出命令
S807
状态转变
S808
应答
结束

图 18

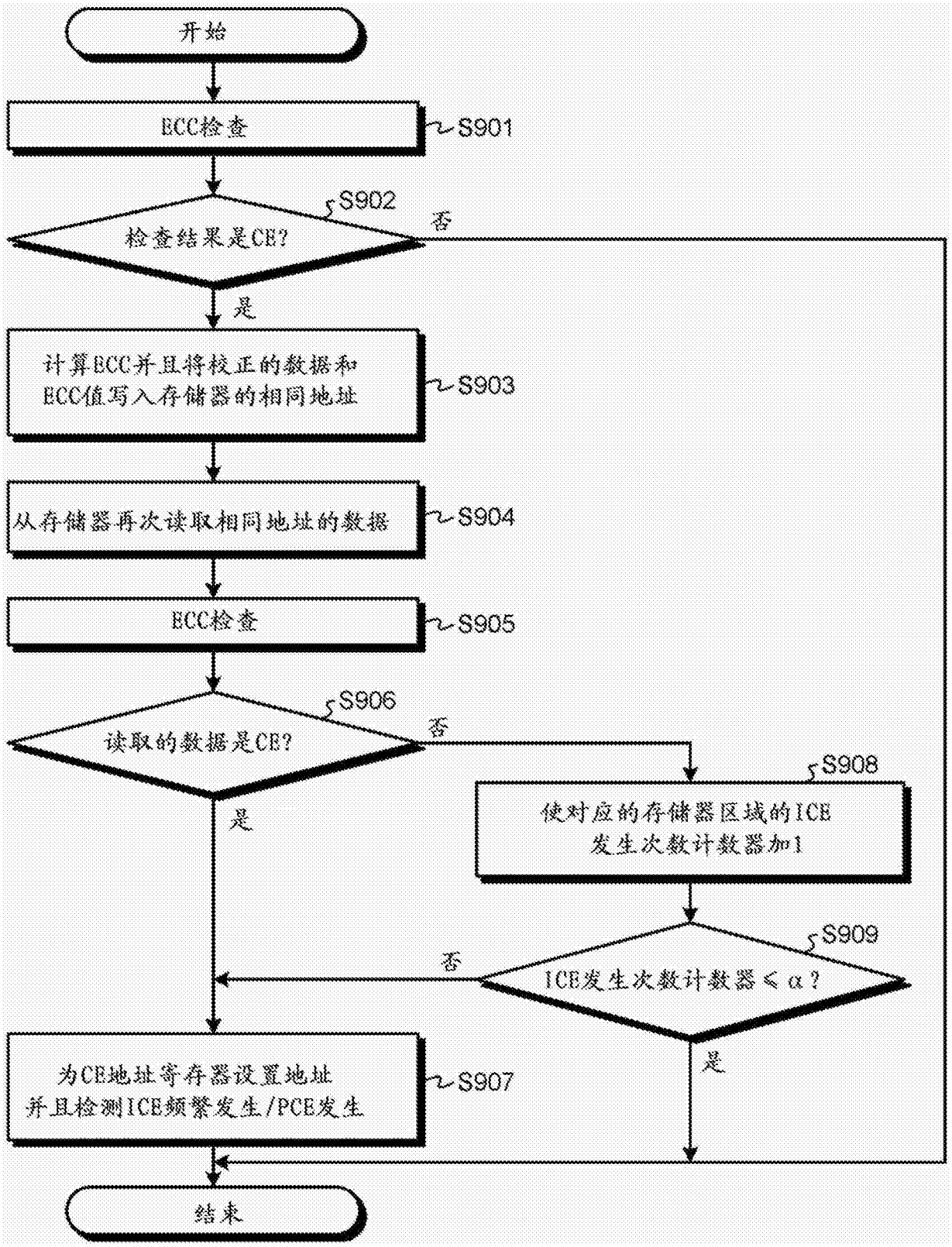

开始
ECC检查
S901
检查结果是CE?
S902
否
是
计算ECC并且将校正的数据和ECC值写入存储器的相同地址
S903
从存储器再次读取相同地址的数据
S904
ECC检查
S905
读取的数据是CE?
S906
否
是
使对应的存储器区域的ICE发生次数计数器加1
S908
ICE发生次数计数器≤α?
S909
否
是
为CE地址寄存器设置地址并且检测ICE频繁发生/PCE发生
S907
结束

图 19

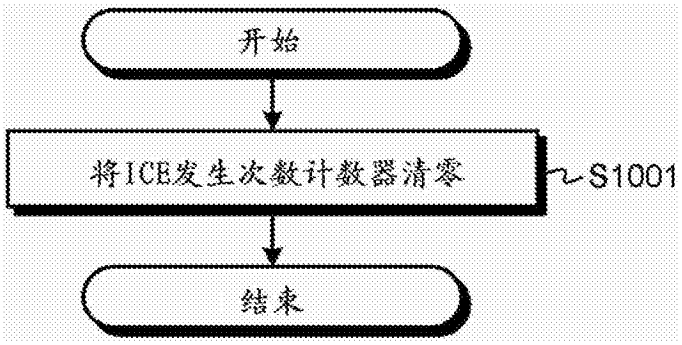


图 20

平均值（次/分）	开始时间	最终发生时间
0.1	2011/1/1 00:00	2011/1/1 03:30

图 21

平均值（次/分）	开始时间	最终发生时间
0.095	2011/1/1 00:00	2011/1/1 03:50

图 22

平均值（次/分）	开始时间	最终发生时间
0	0	0

图 23

平均值（次/分）	开始时间	最终发生时间
1	2011/1/1 00:00	2011/1/1 00:00

图 24

平均值（次/分）	开始时间	最终发生时间
0.4	2011/1/1 00:00	2011/1/1 00:05

图 25

平均值（次/分）	开始时间	最终发生时间
0.1	2011/1/1 00:00	2011/1/1 03:30

图 26

平均值（次/分）	开始时间	最终发生时间
1	2011/1/1 05:30	2011/1/1 05:30

图 27

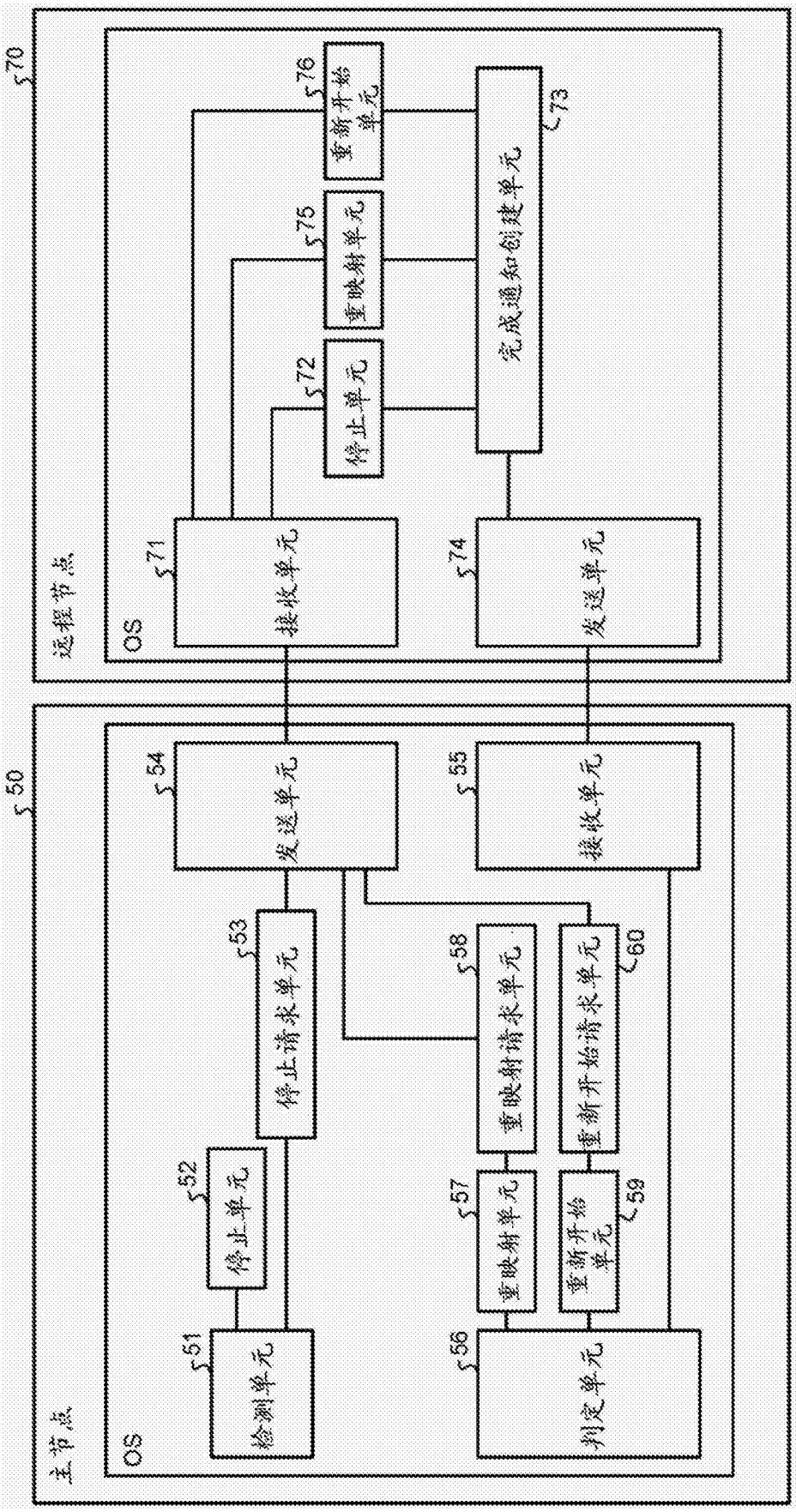

70
远程节点
OS
71
接收单元
74
发送单元
72
停止单元
75
重映射单元
76
重新开始单元
73
完成通知创建单元
50
主节点
OS
54
发送单元
55
接收单元
53
停止请求单元
52
停止单元
51
检测单元
58
重映射请求单元
60
重新开始请求单元
57
重映射单元
59
重新开始单元
56
判定单元

图 28A

地址	PA范围	DA→PA 转换等式	PA→DA 转换等式
#0	0x0000-0x3fff	PA=DA	DA=PA
#1	0x4000-0x7fff	PA=DA+0x4000	DA=PA-0x4000

图 28B

节点间共享 存储器PA	节点间共享 存储器VA	节点间共享 存储器长度	使用节点 信息	指向下一条 目的指针

图 28C

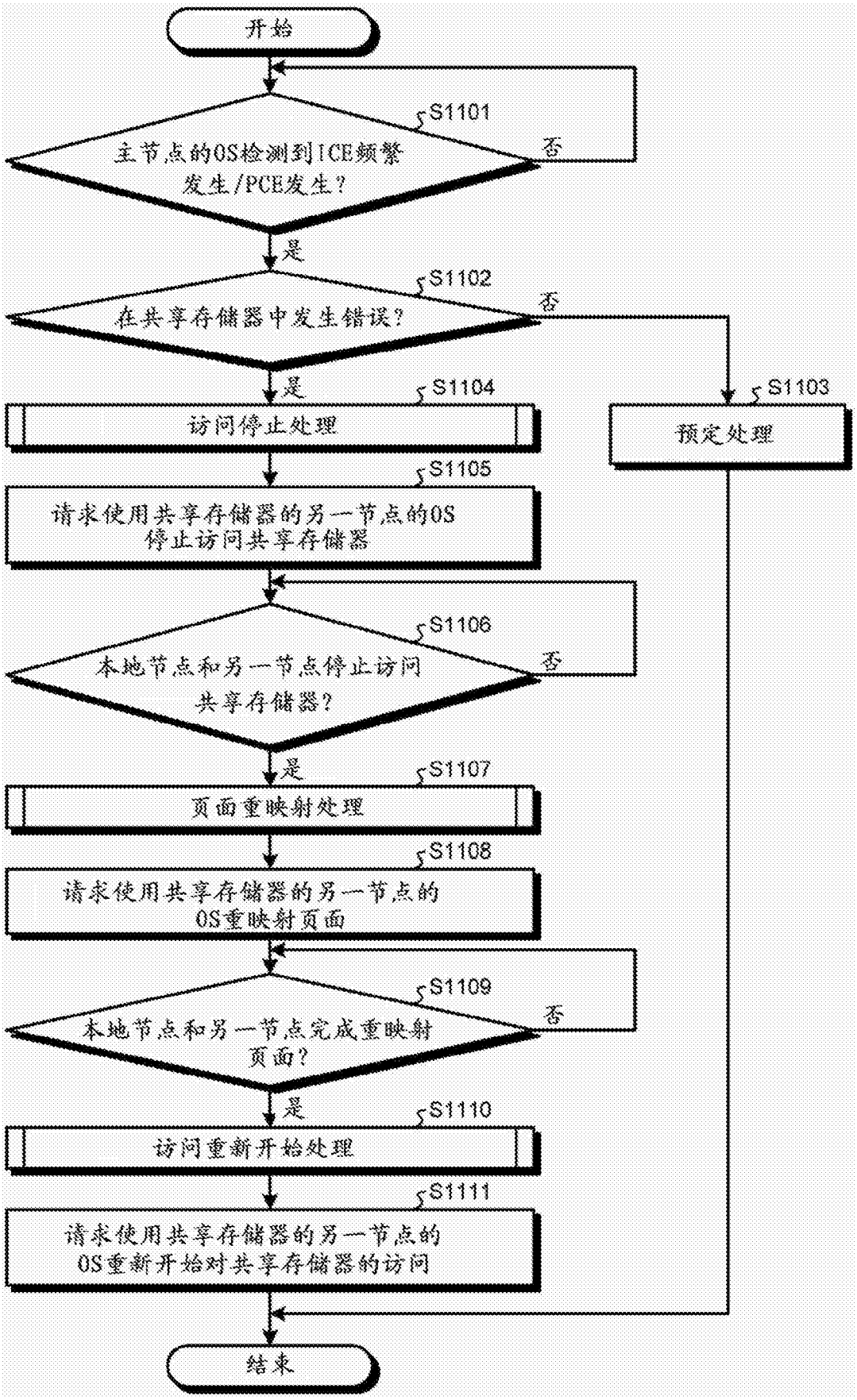

开始
S1101
主节点的OS检测到ICE频繁发生/PCE发生？
否
是
S1102
在共享存储器中发生错误？
否
是
S1104
访问停止处理
S1103
预定处理
S1105
请求使用共享存储器的另一节点的OS停止访问共享存储器
S1106
本地节点和另一节点停止访问共享存储器？
否
是
S1107
页面重映射处理
S1108
请求使用共享存储器的另一节点的OS重映射页面
S1109
本地节点和另一节点完成重映射页面？
否
是
S1110
访问重新开始处理
S1111
请求使用共享存储器的另一节点的OS重新开始对共享存储器的访问
结束

图 29

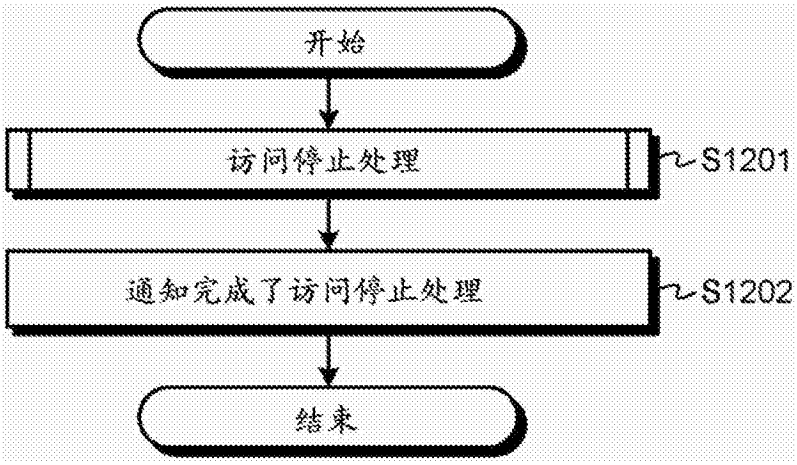

开始
访问停止处理
S1201
通知完成了访问停止处理
S1202
结束

图 30

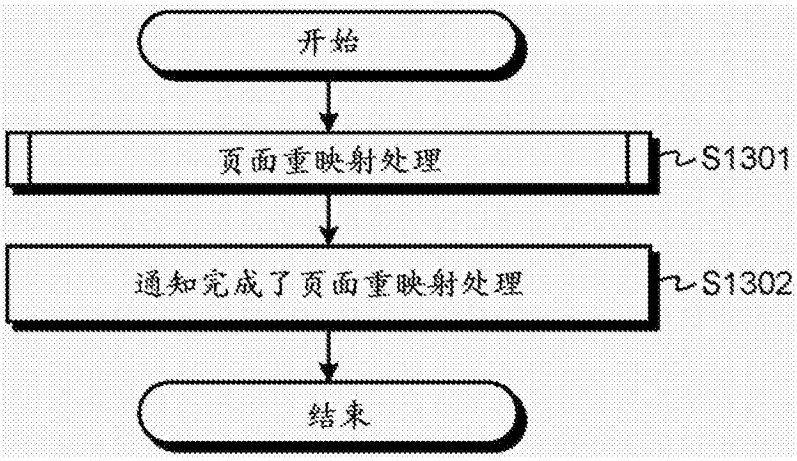

开始
页面重映射处理
S1301
通知完成了页面重映射处理
S1302
结束

图 31

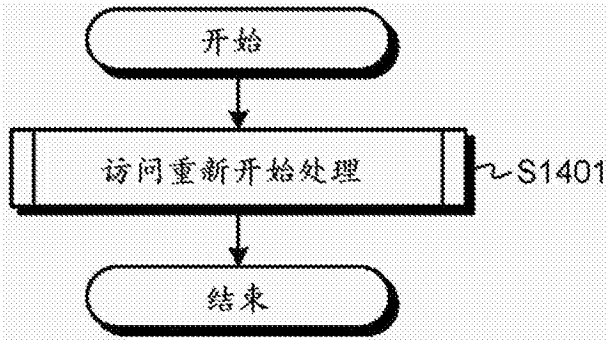

开始
访问重新开始处理
S1401
结束

图 32

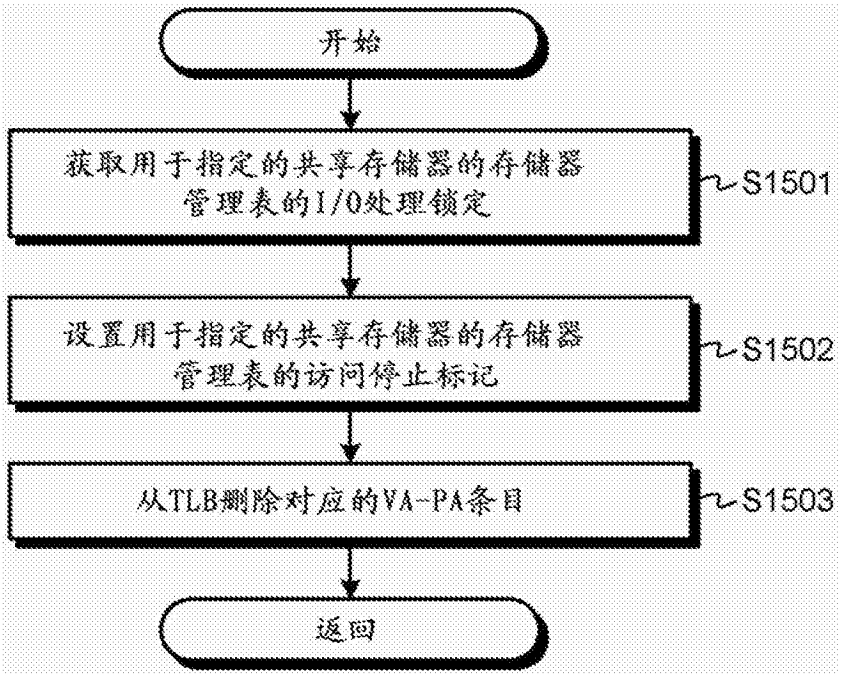


图 33

访问停止标记	I/O处理锁定	指向另一页面管理表的指针	指向地址转换表的指针	其他管理信息

图 34

VA	PA	区域长度	指向页面管理表的指针	指向另一地址转换表的指针	其他管理信息

图 35

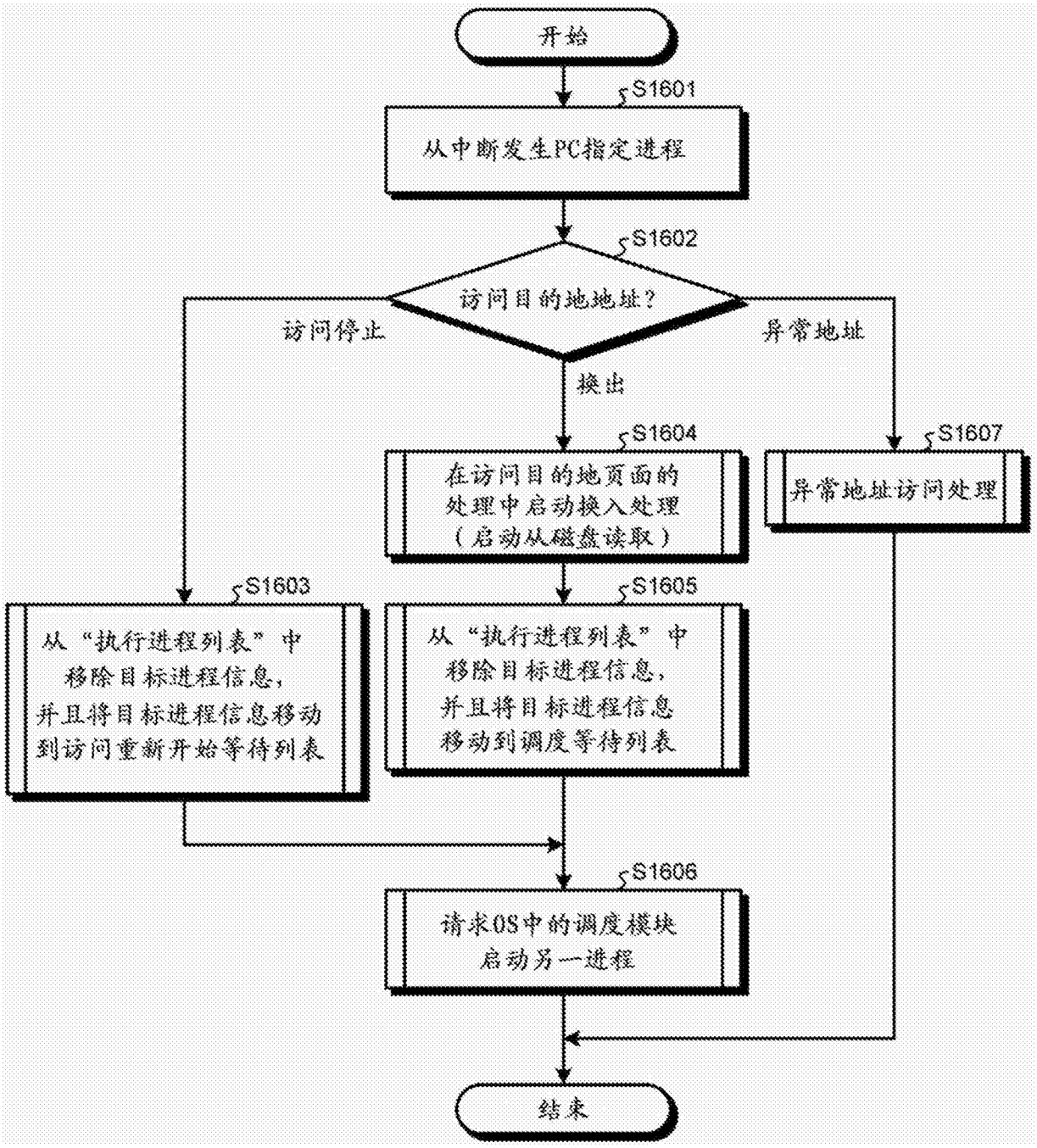


图 36

进程标识符	进程恢复信息	重新开始等待共享存储器地址	指向下一列表的指针

图 37

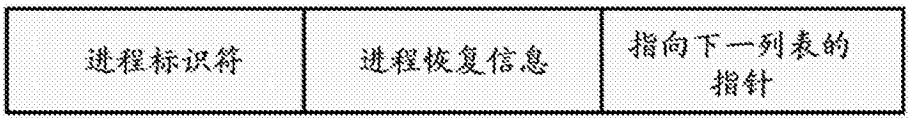

进程标识符
进程恢复信息
指向下一列表的指针

图 38

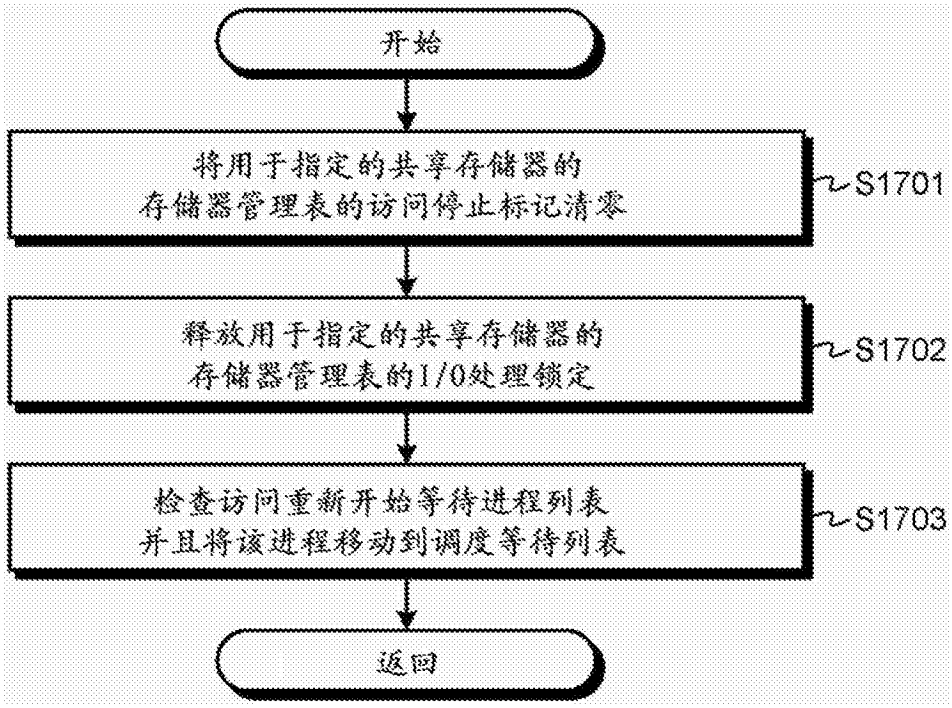

开始
将用于指定的共享存储器的存储器管理表的访问停止标记清零
S1701
释放用于指定的共享存储器的存储器管理表的I/O处理锁定
S1702
检查访问重新开始等待进程列表并且将该进程移动到调度等待列表
S1703
返回

图 39

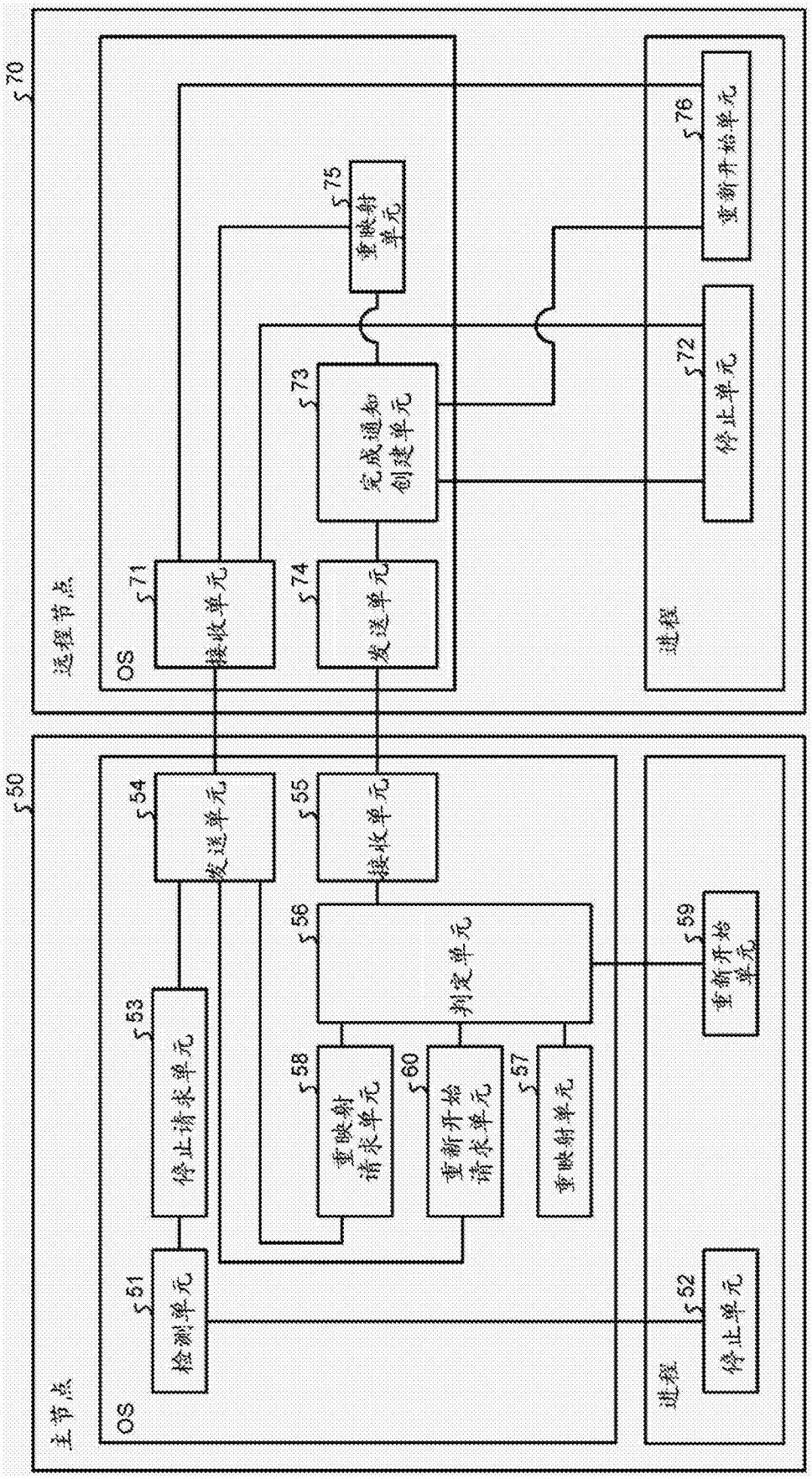

70
远程节点
OS
71
接收单元
74
发送单元
73
完成通知创建单元
75
重映射单元
76
重新开始单元
72
停止单元
进程
50
主节点
OS
54
发送单元
55
接收单元
56
判定单元
53
停止请求单元
58
重映射请求单元
60
重新开始请求单元
57
重映射单元
59
重新开始单元
51
检测单元
52
停止单元
进程

图 40