



(12) 发明专利申请

(10) 申请公布号 CN 103297490 A

(43) 申请公布日 2013. 09. 11

(21) 申请号 201210575905. 2

(22) 申请日 2012. 12. 26

(30) 优先权数据

2012-015693 2012. 01. 27 JP

(71) 申请人 富士通株式会社

地址 日本神奈川县

(72) 发明人 越智亮 小池康夫 前田敏之

古田智德 伊藤史昭 宫路忠宏

藤田和久

(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 康建峰 贾萌

(51) Int. Cl.

H04L 29/08 (2006. 01)

G06F 12/08 (2006. 01)

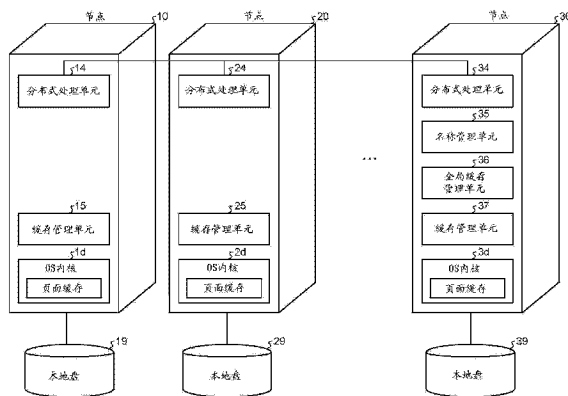
权利要求书2页 说明书10页 附图9页

(54) 发明名称

信息处理装置、分布式处理系统和分布式处理方法

(57) 摘要

提供一种信息处理装置、分布式处理系统和分布式处理方法。该信息处理装置包括：接收单元，其从信息处理装置以分布式方式执行处理的分布式处理系统中的多个信息处理装置之一接收数据的访问请求；查询发出单元，其在由接收单元接收到数据的访问请求时关于数据是否存储在由信息处理装置中的每一个上的操作系统管理的页面缓存中向信息处理装置中的每一个发出查询；以及响应单元，其对访问请求做出响应，该响应作为访问目的地指定已经响应由查询发出单元发出的查询的信息处理装置。



1. 一种信息处理装置,包括:

接收单元,其从信息处理装置以分布式方式执行处理的分布式处理系统中的多个信息处理装置之一接收到数据的访问请求;

查询发出单元,其在所述接收单元接收到数据的访问请求时,关于所述数据是否存储在由所述信息处理装置中的每一个上的操作系统管理的页面缓存中,向所述信息处理装置中的每一个发出查询;以及

响应单元,其对所述访问请求做出响应,所述响应将已响应由所述查询发出单元发出的查询的信息处理装置指定为访问目的地。

2. 如权利要求 1 所述的信息处理装置,其中,所述响应单元通过指示所述数据存储在与已响应所述查询的信息处理装置上的操作系统管理的页面缓存中,做出指定所述已响应所述查询的信息处理装置的响应。

3. 如权利要求 2 所述的信息处理装置,其中,当多个信息处理装置已通过指示所述数据存储在与由所述操作系统管理的页面缓存中进行了响应时,所述响应单元做出指定响应的信息处理装置中准备好首先开始处理的信息处理装置的响应。

4. 如权利要求 2 所述的信息处理装置,还包括:存储单元,其通过将用于标识由所述信息处理装置中的每一个存储在盘中的数据的信息与所述信息处理装置相关联而存储所述信息,其中,

当不存在已通过指示所述数据存储在与由所述操作系统管理的页面缓存中进行了响应的信息处理装置时,所述响应单元从所述存储单元指定存储有所述数据的信息处理装置,并利用指定所指定的信息处理装置的信息进行响应。

5. 一种信息处理装置,包括:

接收单元,其从自信息处理装置以分布式方式执行处理的分布式处理系统中的多个信息处理装置之一接收到数据的访问请求的信息处理装置,接收关于数据是否存储在由操作系统管理的页面缓存中的查询;

确定单元,其在所述接收单元接收到所述查询时,参考由自身信息处理装置中的操作系统管理的页面缓存,并确定所述数据是否存储在由所述操作系统管理的页面缓存中;以及

响应单元,其通过发送由所述确定单元确定的结果响应接收的查询。

6. 如权利要求 5 所述的信息处理装置,其中,

所述确定单元根据所述数据的大小计算页面的总数,并且当页面的总数与由自身信息处理装置中的操作系统管理的页面缓存中存储的页面数量的比率等于或大于预定值时,确定所述数据存储在与所述操作系统管理的页面缓存中。

7. 一种分布式处理系统,包括以分布式方式执行处理的多个信息处理装置,其中,

来自所述信息处理装置之中的第一信息处理装置包括:

请求接收单元,其从另一信息处理装置接收数据的访问请求,

查询发出单元,其在所述请求接收单元接收到所述访问请求时,关于所述数据是否存储在由所述信息处理装置中的每一个上的操作系统管理的页面缓存中,向所述信息处理装置中的每一个发出查询,以及

请求响应单元,其对访问请求做出响应,所述响应将已响应由所述查询发出单元发出

的查询的信息处理装置指定为访问目的地,以及

所述信息处理装置中的每一个包括:

查询接收单元,其从所述第一信息处理装置接收关于所述数据是否存储在由所述操作系统管理的页面缓存中的查询,

确定单元,其在所述查询接收单元接收到所述查询时,参考由自身信息处理装置中的操作系统管理的页面缓存,并确定所述数据是否存储在由所述操作系统管理的所述页面缓存中,以及

结果响应单元,其通过发送由所述确定单元确定的结果响应接收的查询。

8. 一种分布式处理方法,在多个信息处理装置以分布式方式执行处理的分布式处理系统中执行,所述分布式处理方法包括:

由来自所述信息处理装置之中的第一信息处理装置从所述信息处理装置之一接收数据的访问请求;

由所述第一信息处理装置向所述信息处理装置中的每一个发出关于访问请求目的地上的数据是否存储在由所述信息处理装置中的每一个上的操作系统管理的页面缓存中的查询;

由所述信息处理装置中的每一个接收关于所述数据是否存储在由操作系统管理的页面缓存中的查询;

由所述信息处理装置中的每一个参考由信息处理装置中的每一个中的操作系统管理的页面缓存,并确定所述数据是否存储在由所述操作系统管理的页面缓存中;

由所述信息处理装置中的每一个通过发送在所述确定中确定的结果响应接收的查询;以及

由所述第一信息处理装置对访问请求做出响应,所述响应指定已响应查询的信息处理装置。

信息处理装置、分布式处理系统和分布式处理方法

技术领域

[0001] 本文中讨论的实施例涉及一种信息处理装置、分布式处理系统以及分布式处理方法。

背景技术

[0002] 已知使用多个节点的传统分布式处理框架(诸如,Hadoop)作为用于高速处理大量数据的分布式处理技术。Hadoop 划分文件并把划分的文件存储在多个节点中,并且并行地允许管理划分的文件的每个节点执行处理指令,从而执行分布式处理。

[0003] 使用 Hadoop 分布式文件系统(HDFS)作为 Hadoop 数据结构的基础。HDFS 是提供在多个节点之中统一的名称空间的双层结构文件系统;然而,在实践中双层结构文件系统使用每个节点中的本地文件系统管理数据。具体地,HDFS 是由管理名称空间的用户层应用程序和管理物理文件输入和输出的操作系统构建的双层结构文件系统。

[0004] 图 9 是例示传统分布式处理系统总体配置实例的示意图。图 9 中示出的分布式处理系统包括三个节点,即,节点 A、节点 B 以及节点 C。在节点中的每个节点中执行使用分布式处理框架的应用程序。节点 A 连接到本地盘 A,节点 B 连接到本地盘 B,节点 C 连接到本地盘 C。节点 C 是执行管理哪个文件存储在哪个本地盘中的名称管理的主节点。

[0005] 上述分布式处理系统创建文件的副本并在其中存储原始文件和副本文件。例如,在分布式处理系统中,如果文件 C 存储在节点 C 中,则作为文件 C 副本的副本文件 C 也存储在节点 B 或节点 A 中。以此方式,分布式处理系统实施文件冗余。

[0006] 以下,将通过作为实例使用作为分布式处理一部分的引用(reference)过程具体描述处理实例。在这一点上,假定节点 A 执行的应用程序 A 请求节点 C 引用文件 A。在这种情形中,节点 C 使用名称管理指定文件 A 存储在本地盘 B 和本地盘 C 这二者中。然后,节点 C 向应用程序 A 响应要使用较贴近作为请求源的节点 A 的本地盘 B 作为文件 A 的存储目的地。接收响应的应用程序 A 请求连接到本地盘 B 的节点 B 读取文件 A,并然后引用从本地盘 B 读取的文件 A。

[0007] 此外,在分布式处理系统中,在由每个节点为应用程序预留的存储器区域中管理文件。当应用程序做出对要引用文件的请求时,从存储器读取主题文件,这使得可以减少处理时间。

[0008] 专利文献 1:日本专利公开 2005-234919 号公报

[0009] 专利文献 2:日本专利公开 11-15718 号公报

[0010] 专利文献 3:日本专利公开 07-182220 号公报

[0011] 然而,在上述分布式处理系统的情况下,问题在于:因为未高效使用由 OS 内核管理的页面缓存和因而出现本地盘的输入/输出,所以未改进处理性能。

[0012] 例如,如果在图 9 中例示的分布式处理系统中出现分布式处理,则节点 C 确定要处理的文件是来自在其中存储要处理文件的本地盘之中最贴近请求源节点的本地盘中存储的文件。相应地,会存在如下这种情形:代替把主题文件存储在由 OS 内核管理的页面缓存

中的节点,请求未把主题文件存储在由 OS 内核管理的页面缓存中的节点处理主题文件。

[0013] 在这种情形中,请求处理文件的节点从处理速度比由 OS 内核管理的页面缓存的处理速度低的本地盘读取文件。换言之,在整个分布式处理系统方面,即使在由 OS 内核管理的页面缓存中存储的文件是要读取的文件时,有时也会通过执行低速盘的输入 / 输出处理读取文件。

[0014] 此外,执行了低速盘输入 / 输出处理的节点把从本地盘读取的文件存储在由 OS 内核管理的页面缓存中。在这一点上,在丢弃来自由 OS 内核管理的页面缓存的其它文件之后,节点缓存读取的文件,而致使其它文件缓存命中率的减小。

[0015] 如上所述,在传统分布式处理系统的情况下,未高效使用页面缓存,因而难以提高处理性能。此外,用于通过在为应用程序预留的存储器区域中存储数据提高吞吐量的方法因为需要在存储器中存储大量文件所以不切实际。

[0016] 相应地,在本发明实施例的一个方面中目的是提供可以提高处理性能的信息处理装置、分布式处理系统、缓存管理程序以及分布式处理方法。发明内容

[0017] 根据实施例的方面,信息处理装置包括:接收单元,从信息处理装置以分布式方式执行处理的分布式处理系统中的多个信息处理装置之一接收数据的访问请求;查询发出单元,在接收单元接收到数据的访问请求时关于数据是否存储在由信息处理装置的每一个上的操作系统管理的页面缓存中向信息处理装置中的每一个发出查询;以及响应单元,用于对访问请求做出响应,该响应作为访问目的地已经指定响应由查询发出单元发出的查询的信息处理装置。

附图说明

[0018] 图 1 是例示根据第一实施例的分布式处理系统总体配置实例的示意图;

[0019] 图 2 是例示根据第一实施例的分布式处理系统中主节点配置的功能框图;

[0020] 图 3 是例示根据第一实施例的分布式处理系统中节点配置的功能框图;

[0021] 图 4 是例示根据第一实施例的由请求源节点执行的处理流程的流程图;

[0022] 图 5 是例示根据第一实施例的由主节点执行的处理流程的流程图;

[0023] 图 6 是例示根据第一实施例的由每个节点执行的处理流程的流程图;

[0024] 图 7 是例示使用共享盘时分布式处理系统总体配置实例的示意图;

[0025] 图 8 是例示执行缓存管理程序的计算机硬件配置实例的示意图;以及

[0026] 图 9 是例示传统分布式处理系统总体配置实例的示意图。

具体实施方式

[0027] 将参照附图解释本发明的优选实施例。本发明不限于实施例。

[0028] [a] 第一实施例

[0029] 总体配置

[0030] 图 1 是例示根据第一实施例的分布式处理系统总体配置实例的示意图。如图 1 中所示,通过多个节点构建分布式处理系统。图 1 例示节点 10、节点 20 以及节点 30。这些节点经由例如网络彼此相连。此外,可以任意设置节点的数量。在实施例中,将通过使用文件作为实例给出描述;然而,实施例不限于此。例如,也可以使用各种类型的数据。

[0031] 在分布式处理系统中,在每个节点中执行使用分布式处理框架(诸如, Hadoop)的分布式处理应用程序,并使用 HDFS 等作为数据结构的基础。此外,在分布式处理系统中,除了文件之外创建并存储副本文件。此外,在分布式处理系统中,把大文件划分成具有预定大小的文件,并向节点分发并在节点中存储与原始文件的一部分相对应的获得的划分文件。

[0032] 构建这种分布式处理系统的节点 10、节点 20 以及节点 30 是通过与其它节点协作执行分布式处理的节点。如图 1 中所示,节点 10 连接到本地盘 19,包括分布式处理单元 14 和缓存管理单元 15,并在操作系统(OS)内核 1d 中执行页面缓存。本地盘 19 在其中存储例如上述文件、文件的一部分以及副本文件。分布式处理单元 14 执行使用分布式处理框架的分布式处理应用程序,并通过与其它节点协作执行分布式处理。

[0033] 节点 20 的配置与节点 10 一样。具体地,节点 20 连接到本地盘 29,包括分布式处理单元 24 和缓存管理单元 25,并且在 OS 内核 2d 中执行页面缓存。本地盘 29 在其中存储例如上述文件、文件的一部分以及副本文件。分布式处理单元 24 执行使用分布式处理框架的分布式处理应用程序,并通过与其它节点协作执行分布式处理。

[0034] 节点 30 是除了节点 10 或节点 20 的配置或功能之外,还执行分布式处理系统的名称管理的主节点。节点 30 连接到本地盘 39 ;包括分布式处理单元 34、名称管理单元 35、全局缓存管理单元 36 以及缓存管理单元 37 ;并在 OS 内核 3d 中执行页面缓存。与本地盘 29 类似地,本地盘 39 在其中存储例如上述文件、文件的一部分以及副本文件。此外,每个本地盘中存储的信息并非总是相同。分布式处理单元 34 执行使用分布式处理框架的分布式处理应用程序,并通过与其它节点协作执行分布式处理。名称管理单元 35 管理哪个数据存储存储在哪个节点中的本地盘中。全局缓存管理单元 36 是查询关于要处理文件是否被缓存的处理单元。分别在节点 10、节点 20 以及节点 30 中的 OS 内核 1d、OS 内核 2d 以及 OS 内核 3d 构建 OS 的一部分,且是作为 OS 的核心运作的软件。OS 内核 1d、OS 内核 2d 以及 OS 内核 3d 各自管理系统资源并管理软件组件和硬件的交换。

[0035] 如果在分布式处理系统中出现文件的处理请求,则在存储有要处理文件的节点中执行请求的处理。然后,执行了处理的节点通过发送执行结果响应节点,即,处理请求源。此外,如果出现对于通过划分和存储在分布式处理系统中获得的文件的处理请求,则在其中存储划分的文件的每个节点中执行请求的处理,然后响应于节点(即,请求源)发送处理的结果。

[0036] 以下,将描述节点 10 做出对文件的引用请求的情形。如果由于执行应用程序而访问文件,则节点 10 中的分布式处理单元 14 向对应于主节点的节点 30 传输对文件的访问请求。如果节点 30 中的分布式处理单元 34 接收到访问请求,则节点 30 中的全局缓存管理单元 36 执行每个节点关于要访问的文件是否存储在由相应 OS 内核管理的页面缓存中的查询。在这一点上,全局缓存管理单元 36 在包括它自身节点的节点上执行查询。

[0037] 随后,在接收了查询的相应节点中的缓存管理单元的每一个中,执行缓存管理,其中,通过引用(refer to)由相应 OS 内核管理的页面缓存确定要访问的文件是否存储在由 OS 内核管理的页面缓存中。然后,每个节点通过把是否出现了缓存命中告知节点 30 来响应节点 30。

[0038] 然后,节点 30 中的全局缓存管理单元 36 通过发送指定通过发送指示文件数据被存储在由 OS 内核管理的页面缓存中的信息响应查询的节点的信息,经由分布式处理单元

34 响应作为请求源的节点 10。指定节点的信息的实例包括互联网协议(IP) 地址。

[0039] 此后, 节点 10 中的分布式处理单元 14 向由节点 30 通知的节点传输对文件的引用请求。接收了引用请求的节点从由 OS 内核管理的页面缓存读取文件并响应节点 10。随后, 由节点 10 执行的应用程序可以引用文件。

[0040] 如上所述, 分布式处理系统中的主节点做出每个节点关于请求的文件是否存储在由相应 OS 内核管理的页面缓存中的查询。然后, 主节点通过告知缓存数据的节点响应请求源; 因此, 高效使用了由 OS 内核管理的页面缓存, 并因而可以提高处理性能。

[0041] 主节点的配置

[0042] 以下, 将对构建分布式处理系统的主节点给出描述。在图 1 中, 因为执行名称管理的节点 30 是主节点, 所以在下面的描述中把节点称作主节点 30。

[0043] 图 2 是例示根据第一实施例的分布式处理系统中主节点配置的功能框图。如图 2 中所示, 主节点 30 包括连接到本地盘 39 的通信控制单元 31、页面缓存 32 以及控制单元 33。图 2 中示出的处理单元只是实例; 因此, 实施例不限于此。

[0044] 通信控制单元 31 是控制与其它节点通信的处理单元, 其例如是网络接口卡。例如, 通信控制单元 31 从其它节点接收各种请求, 诸如访问请求或引用请求。此外, 通信控制单元 31 向其它节点传输对各种请求的响应。

[0045] 页面缓存 32 是在其中存储由程序使用的数据或由控制单元 33 执行的各种应用程序的存储设备。此外, 页面缓存 32 存储例如由控制单元 33 执行的应用程序从本地盘 39 读取的数据。具体地, 页面缓存 32 缓存文件。可以对用于控制缓存的方法使用各种已知方法, 诸如最近最少使用(LRU) 方法; 因此, 此处将略去其描述。

[0046] 控制单元 33 是管理由主节点 30 执行的整个处理并执行应用程序等的处理单元。控制单元 33 是电子电路, 诸如中央处理单元(CPU)。控制单元 33 包括分布式处理单元 34、名称管理单元 35、全局缓存管理单元 36、缓存管理单元 37 以及处理执行单元 38。

[0047] 分布式处理单元 34 是使用分布式处理框架执行分布式处理应用程序并通过与其它节点协作执行分布式处理的处理单元。分布式处理单元 34 与执行分布式处理的其它节点和主节点 30 中包括的各种处理单元协作。例如, 分布式处理单元 34 经由名称管理单元 35 向各种处理单元输出从其它节点接收的各种请求。此外, 分布式处理单元 34 经由名称管理单元 35 把从各种处理单元输出的处理结果输出到作为请求源的其它节点。

[0048] 名称管理单元 35 是管理每个节点中的本地盘中存储的数据的处理单元。例如, 名称管理单元 35 通过使用文件管理表执行名称管理, 在文件管理表中, 用于标识节点的标识符与用于标识每个节点中的本地盘中存储的数据的信息相关联。此外, 名称管理单元 35 通过使用地址管理表执行节点管理, 在地址管理表中, 用于标识每个节点的标识与由每个节点使用的 IP 地址相关联。文件管理表和地址管理表被存储在诸如存储器或硬盘(未示出)的存储单元中。

[0049] 全局缓存管理单元 36 是包括请求接收单元 36a、查询发出单元 36b 以及请求响应单元 36c, 并通过使用这些单元管理每个节点中缓存状态的处理单元。请求接收单元 36a 是从构建分布式处理系统的节点接收对文件的访问请求的处理单元。请求接收单元 36a 接收由它自身的节点或其它节点经由分布式处理单元 34 和名称管理单元 35 传输的访问请求, 并把访问请求输出到查询发出单元 36b。

[0050] 当请求接收单元 36a 接收对文件的访问请求时,查询发出单元 36b 是关于由 OS 内核管理的页面缓存是否在其中存储请求的文件向每个节点发出查询的处理单元。在图 1 中示出的实例中,如果节点 10 传输对文件 A 的访问请求,则查询发出单元 36b 向节点 10、节点 20 以及节点 30 发出查询。

[0051] 请求响应单元 36c 是这样的处理单元:其通过发送用于指定通过发送指示文件存储在由 OS 内核管理的页面缓存中的信息响应来自查询发出单元 36b 的查询的节点的信息来响应访问请求源节点。在上述实例中,当请求响应单元 36c 接收指示缓存文件 A 的来自节点 B 的响应时,请求响应单元 36c 通过发送节点 B 的 IP 地址(即,由查询发出单元 36b 在查询时使用的 IP 地址信息)经由分布式处理单元 34 和名称管理单元 35 响应作为请求源的节点 A。

[0052] 此外,如果多个节点已经响应指示文件存储在由 OS 内核管理的页面缓存中,则请求响应单元 36c 通过作为访问目的地指示最贴近请求源节点的节点的 IP 地址进行响应。例如,请求响应单元 36c 创建分布式处理系统的拓扑,并通过指示至请求源的跳数最低的节点的 IP 地址进行响应。此外,请求响应单元 36c 也可以预先确定节点的优先级次序,并指定响应请求源的节点。此外,除了预先确定的信息之外,也可以通过参考动态加载信息灵活确定优先级的次序。

[0053] 此处将通过返回参照图 2 给出描述。缓存管理单元 37 是这样的处理单元:其包括查询接收单元 37a、确定单元 37b 以及查询响应单元 37c,并通过使用这些单元确定文件数据缓存状态。查询接收单元 37a 是从分布式处理系统中的主节点 30 接收关于要访问的文件是否存储在由 OS 内核管理的页面缓存中的查询的处理单元。在主节点 30 的情形中,查询接收单元 37a 从同样节点中包括的全局缓存管理单元 36 接收查询,并把接收的查询输出到确定单元 37b。

[0054] 确定单元 37b 是通过在查询接收单元 37a 接收到查询时参考其自身节点中的页面缓存 32 确定查询的文件是否存储在页面缓存 32 中的处理单元。具体地,确定单元 37b 确定是否缓存查询的文件,并把确定结果输出到查询响应单元 37c。例如,如果整个文件是查询的目标,则确定单元 37b 参考页面缓存 32,并确定整个文件是否存储在页面缓存 32 中。

[0055] 此外,如果文件的一部分存储在页面缓存 32 中,则确定单元 37b 按照存储的页面数量确定文件是否被缓存。具体地,确定单元 37b 根据文件大小计算页面的总数量。如果页面缓存 32 中存储的页面的数量与页面总数量的比率等于或大于预定值,则确定单元 37b 确定文件存储在页面缓存中。例如,通过使用查询文件的 i 节点(索引节点:文件和目录的独特标识符)作为密钥,确定单元 37b 从 OS 内核获取页面缓存(Pc)的数量。此外,确定单元 37b 通过使用“文件的整体大小(字节)÷4096”计算文件页面的总数量(Pa)。然后,如果计算的值大于通过“(Pc×100)/Pa”获得的预定值“R(%)”,则确定单元 37b 确定文件存储在页面缓存中。

[0056] 查询响应单元 37c 通过发送由确定单元 37b 确定的结果响应查询源节点。在主节点 30 的情形中,查询响应单元 37c 作为由确定单元 37b 确定的结果向全局缓存管理单元 36 输出指示文件被缓存或指示文件未被缓存的信息。

[0057] 处理执行单元 38 是执行经由分布式处理单元 34 接收的分布式处理中出现或其自身的节点中出现的分布式处理的处理单元。例如,如果接收对文件的引用请求,则处理执行

单元 38 通过使用引用请求中包含的文件名称或 i 节点作为密钥查找页面缓存 32。如果在页面缓存 32 中存在主题文件,则处理执行单元 38 从页面缓存 32 读取主题文件,并把该文件传输给请求源节点。相比而言,如果在页面缓存 32 中不存在主题文件,则处理执行单元 38 从本地盘 39 读取主题文件并把该文件传输给请求源节点。

[0058] 节点的配置

[0059] 以下,将对构建分布式处理系统的节点之中除了主节点以外的节点给出描述。下述节点是图 1 中示出的节点 10 和节点 20。节点 10 和节点 20 的配置一样;因此,下面将只描述节点 10 作为实例。

[0060] 图 3 是例示根据第一实施例的分布式处理系统中节点配置的功能框图。如图 3 中所示,节点 10 包括连接到本地盘 19 的通信控制单元 11、页面缓存 12 以及控制单元 13。图 3 中示出的处理单元只是实例;因此,实施例不限于此。

[0061] 通信控制单元 11 是控制与其它节点通信的通信接口,例如是网络接口卡。例如,通信控制单元 11 从其它节点接收各种请求,诸如引用请求。此外,通信控制单元 11 向其它节点传输对各种请求的响应。

[0062] 页面缓存 12 是在其中存储由程序使用的数据或由控制单元 13 执行的各种应用程序的存储设备。此外,页面缓存 12 在其中存储例如由控制单元 13 执行的应用程序从本地盘 19 读取的数据。

[0063] 控制单元 13 是管理由节点 10 执行的整个处理并执行应用程序等的处理单元。例如,控制单元 13 是电子电路,诸如 CPU。控制单元 13 包括分布式处理单元 14、缓存管理单元 15 以及处理执行单元 16。与图 2 中示出的分布式处理单元 34 类似地,分布式处理单元 14 是使用分布式处理框架执行分布式处理应用程序并通过与其它节点协作执行分布式处理的处理单元。分布式处理单元 14 从主节点或从其它节点接收访问请求或引用请求,并把接收的请求输出到各种处理单元。此外,分布式处理单元 14 向主题节点传输从处理单元输出的响应或处理结果。

[0064] 缓存管理单元 15 执行与图 2 中示出的缓存管理单元 37 执行的一样的处理;因此,此处将略去其详细描述。请注意:查询接收单元 15a、确定单元 15b 以及查询响应单元 15c 分别对应于查询接收单元 37a、确定单元 37b 以及查询响应单元 37c。再者,查询源并非其自身的节点而是主节点,这与图 2 中示出的配置不同。相应地,缓存管理单元 15 响应的目的地是主节点。此外,处理执行单元 16 执行与图 2 中示出的处理执行单元 38 执行的一样的处理;因此,此处将略去其详细描述。

[0065] 处理的流程

[0066] 以下,将描述由分布式处理系统执行的各种处理的流程。在下面的描述中在实例中假定引用请求源节点是节点 10,且主节点是节点 30。主节点执行与节点 10 执行的一样的处理。此外,节点 30 也执行图 4 或 6 中例示的处理的流程。

[0067] 请求源节点

[0068] 图 4 是例示根据第一实施例的由请求源节点执行的处理流程的流程图。如图 4 中所示,当在节点 10 中执行应用程序并访问文件(步骤 S101)时,分布式处理单元 14 向主节点传输对文件的访问请求(步骤 S102)。

[0069] 然后,当分布式处理单元 14 从主节点接收响应(在步骤 S103,是)时,处理执行单

元 16 向做出了响应的节点的 IP 地址传输对文件的引用请求(步骤 S104)。

[0070] 然后,当处理执行单元 16 从传输了引用请求的节点接收文件(在步骤 S105,是)时,处理执行单元 16 引用文件并继续执行在步骤 S101 执行的应用程序。

[0071] 主节点

[0072] 图 5 是例示根据第一实施例的由主节点执行的处理流程的流程图。如图 5 中所示,当节点 30 中的请求接收单元 36a 经由分布式处理单元 34 接收到访问请求(在步骤 S201,是)时,查询发出单元 36b 关于文件是否被缓存向每个节点发出查询(S202)。

[0073] 此后,当请求响应单元 36c 从执行了查询的节点接收响应(在步骤 S203,是)时,请求响应单元 36c 确定是否存在缓存文件的节点(步骤 S204)。

[0074] 如果请求响应单元 36c 确定存在缓存文件的节点(在步骤 S204,是),则请求响应单元 36c 通过使用对在步骤 S202 执行的查询使用的 IP 地址指定由 OS 内核管理的页面缓存中存储的节点的 IP 地址(步骤 S205)。然后,请求响应单元 36c 通过使用指定的 IP 地址作为访问目的地响应请求源节点(步骤 S206)。在这一点上,如果请求响应单元 36c 确定多个节点缓存文件,则请求响应单元 36c 通过发送处理首先开始的节点的 IP 地址响应请求源。

[0075] 相比而言,如果请求响应单元 36c 确定不存在缓存文件的节点(在步骤 S204,否),则请求响应单元 36c 通过使用名称管理单元 35 等指定把文件存储在其本地盘中的节点(步骤 S207)。

[0076] 如果存在多个指定节点(在步骤 S208,是),则请求响应单元 36c 通过使用例如名称管理单元 35 指定首先开始处理的节点的 IP 地址(步骤 S209),并响应请求源节点(步骤 S210)。

[0077] 相比而言,如果不存在多个指定节点(在步骤 S208,否),则请求响应单元 36c 指定存储有通过使用名称管理单元 35 等指定的文件的节点的 IP 地址(步骤 S211),并响应请求源节点(步骤 S212)。

[0078] 每个节点

[0079] 图 6 是例示根据第一实施例的由每个节点执行的处理流程的流程图。以下,将对作为实例的节点 20 给出描述。向节点 20 中的处理单元中的每一个分配附图标记的方式与节点 10 的一样,例如,节点 20 中的分布式处理单元 24。

[0080] 如图 6 中所示,当节点 20 中的查询接收单元 25a 从主节点 30 接收查询(在步骤 S301,是)时,确定单元 25b 搜索页面缓存 22 (步骤 S302),并确定是否出现缓存命中(步骤 S303)。

[0081] 如果出现了缓存命中(在步骤 S303,是),则查询响应单元 25c 通过指示存在缓存命中响应主节点 30 (步骤 S304)。相比而言,如果未出现缓存命中(在步骤 S303,否),则查询响应单元 25c 通过指示不存在缓存命中响应主节点 30 (步骤 S305)。

[0082] 此后,当分布式处理单元 24 接收到对文件的引用请求(在步骤 S306,是)时,处理执行单元 26 参考页面缓存 22,并确定在页面缓存 22 中是否存在文件(步骤 S307)。如果未接收到对文件的引用请求,则处理结束。

[0083] 如果出现了缓存命中(在步骤 S308,是),则处理执行单元 26 从页面缓存 22 读取文件(步骤 S309),并通过发送文件响应引用请求源节点(步骤 S310)。

[0084] 相比而言,如果未出现缓存命中(在步骤 S308,否),则处理执行单元 26 从本地盘

29 读取文件(步骤 S311) 并响应指示文件的引用请求源节点(步骤 S312)。

[0085] 如果在页面缓存 22 中存在自由空间(在步骤 S313, 是), 则处理执行单元 26 执行对其进行了响应的文件中的缓存(步骤 S314)。具体地, 处理执行单元 26 把从本地盘 29 读取的文件存储在页面缓存 22 中的自由空间中。

[0086] 相比而言, 在页面缓存中不存在自由空间(在步骤 S313, 否), 在处理执行单元 26 执行文件外的缓存(步骤 S315) 之后, 处理执行单元 26 执行其进行了响应的文件中的缓存(步骤 S316)。具体地, 处理执行单元 26 从页面缓存 22 丢弃满足预定条件的文件, 然后把从本地盘 29 读取的文件存储在存储了丢弃文件的区域中。可以使用各种已知方法(诸如最近最少使用(LRU) 方法) 设置条件。

[0087] 如上所述, 图 1 中示出的分布式处理系统即使在使用传统应用程序时也可以在不使用每个应用程序的专用缓存管理功能的情况下高效使用操作系统中的页面缓存。具体地, 在图 1 中示出的分布式处理系统的情况下, 如果节点中的任何一个节点缓存主题文件, 则总是可以在分布式处理中使用页面缓存。相应地, 可以减少用于从本地盘读取文件的 I/O 处理并增进高速存储器访问, 从而提高处理性能。

[0088] [b] 第二实施例

[0089] 在以上解释中, 对根据本发明的实施例给出了描述; 然而, 实施例不限于此, 可以在除了上述实施例以外各种实施例的情况下实施本发明。因此, 下面将描述另一实施例。

[0090] 共享盘

[0091] 在第一实施例中, 对每个节点包括本地盘的情形给出了描述; 然而, 实施例不限于此。例如, 每个节点也可以连接到共享盘。图 7 是例示使用共享盘时分布式处理系统总体配置实例的示意图。如图 7 中所示, 每个节点的配置与第一实施例中描述的一样。第二实施例与第一实施例的不同之处在于连接到每个节点的盘是共享盘 50 而非本地盘。

[0092] 即使在图 7 中示出的情形中, 每个节点也具有与第一实施例中执行的一样的功能和执行同样处理。具体地, 每个节点指定把目标数据缓存在由 OS 内核管理的页面缓存中的节点, 且指定的节点从由 OS 内核管理的页面缓存读取主题文件。当使用共享盘时, 可以在同样成本的情况下从服务器中的任一个访问数据。相应地, 可以在连续执行数据的写入和读取时做出缓存命中。结果是, 当与第一实施例相比较时, 可以提高缓存命中的概率, 因而可以提高处理性能。

[0093] 节点的配置

[0094] 在第一实施例中, 对通过单个服务器构建每个节点的情形给出了描述; 然而, 实施例不限于此。例如, 也可以通过多个机架装配服务器构建或通过包括多个系统板的单个服务器构建每个节点。例如, 也可以通过这样的三个机架装配服务器构建单个节点: 包括分布式处理单元 34、名称管理单元 35 以及处理执行单元 38 的机架装配服务器; 包括全局缓存管理单元 36 的机架装配服务器; 以及包括缓存管理单元 37 的机架装配服务器。此外, 单个节点也可以是包括多个系统板的服务器, 每个系统板也可以执行单个处理单元。此外, 可以在每个节点内部布置或在节点外部布置本地盘。

[0095] 处理内容

[0096] 在第一实施例中, 对引用处理给出了描述作为实例; 然而, 实施例不限于此。例如, 也可以通过与上述的类似的方式执行写入处理等。此外, 在存储有主题文件的节点因为使

用分布式处理框架所以有时会执行写入处理、更新处理等。此外,如果划分和在多个节点中存储要处理的文件,则在存储有文件一部分的每个节点并行地执行处理。然后,主节点批处理每个处理并响应请求源。可以使用通常方法以确定主节点。例如,确定最贴近请求源的节点是主节点。

[0097] 系统

[0098] 在实施例中描述的处理中,也可以手动执行提到的作为自动执行处理的全部或一部分,也可以使用已知方法自动执行作为手动执行提到的处理的全部或一部分。此外,如非另行说明则可以任意改变处理的流程、控制过程、具体名称以及包含以上说明书和图中指示的各种数据或参数的信息。

[0099] 图中示出的每个单元的组件只用于在概念上示出其功能和并非总是在物理上被配置成如图中所示。换言之,单独或整合设备的具体形状不限于图。具体地,可以通过依据各种负载或使用条件在功能上或物理上分割或整合单元中的任何单元配置设备中的所有设备或一部分。例如,全局缓存管理单元 36 可以与缓存管理单元 37 整合。此外,可以通过 CPU 和通过 CPU 分析和执行的程序实施或通过布线逻辑作为硬件实施每个设备执行的处理功能中的所有处理功能或一部分。

[0100] 硬件配置

[0101] 可以通过事先准备的程序实施和通过诸如个人计算机或工作站的计算机系统执行上述实施例中执行的各种处理。相应地,以下,将描述执行功能与以上实施例中描述的一样的程序的计算机系统作为实例。

[0102] 图 8 是例示执行缓存管理程序的计算机硬件配置实例的示意图。如图 8 中所示,计算机 100 包括 CPU102、输入设备 103、输出装置 104、通信接口 105、介质读取器 106、硬盘驱动器(HDD)107、随机访问存储器(RAM)108 以及存储器 109。图 8 中示例的单元经由总线 101 彼此相连。

[0103] 输入设备 103 是鼠标或键盘;输出装置 104 是例如显示器;通信接口 105 是接口,诸如网络接口卡(NIC)。HDD107 与缓存管理程序 107a 一起在其中存储参照图 2 描述的表。作为记录介质的实例提到 HDD107;然而,本发明不限于此。例如,也可以在另一计算机可读记录介质(诸如只读存储器(ROM)、RAM、CD-ROM 或者固态驱动器(SSD))中存储各种程序,并也可以通过计算机读取各种程序。此外,也可以通过在远程站点处布置存储介质和通过访问存储介质的计算机获得和使用程序。此外,此时,获得的程序也可以存储在计算机中的记录介质中。

[0104] CPU102 读取缓存管理程序 107a 和在 RAM108 中加载它,因而缓存管理程序 107a 作为执行以上参照图 2 和 3 描述的每个功能的缓存管理处理 108a 运作。具体地,如果计算机 100 是主节点,则缓存管理处理 108a 执行图 2 中示例的分布式处理单元 34、名称管理单元 35、全局缓存管理单元 36、缓存管理单元 37 以及处理执行单元 38。此外,如果计算机 100 是公共节点,则缓存管理处理 108a 执行图 3 中示例的分布式处理单元 14、缓存管理单元 15 以及处理执行单元 16。以此方式,通过读取和执行程序,计算机 100 作为执行缓存管理方法的信息处理装置工作。

[0105] 此外,计算机 100 通过使用介质读取器 106 从记录介质读取缓存管理程序和执行读取的缓存管理程序,从而实施实施例中描述的同样功能。实施例中提到的程序不限于计

算机 100 执行的程序。例如,也可以在另一计算机或服务器在与计算机 100 协作的情况下协作执行程序的情形中使用本发明。

[0106] 根据实施例,可以提高处理性能。

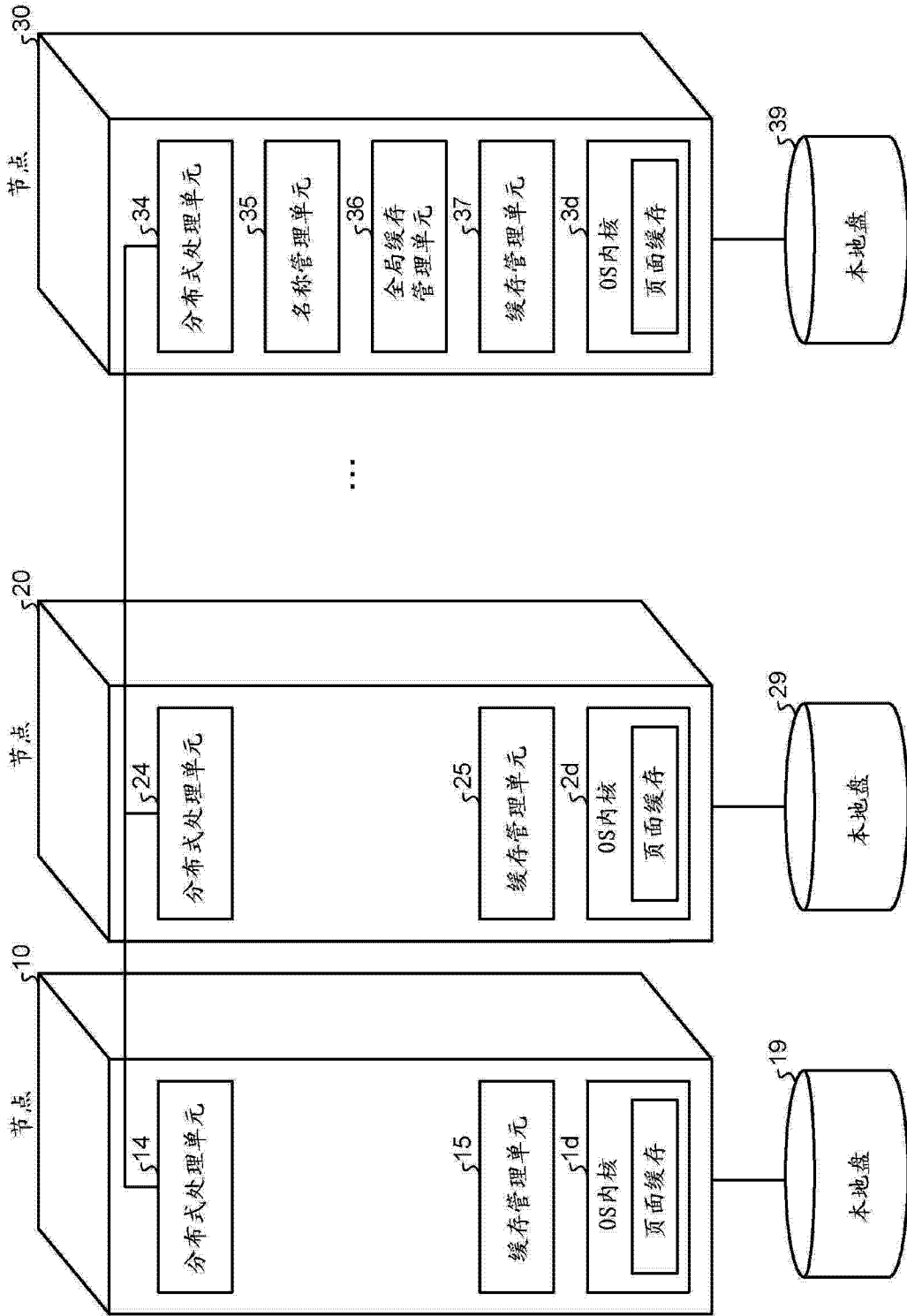


图 1

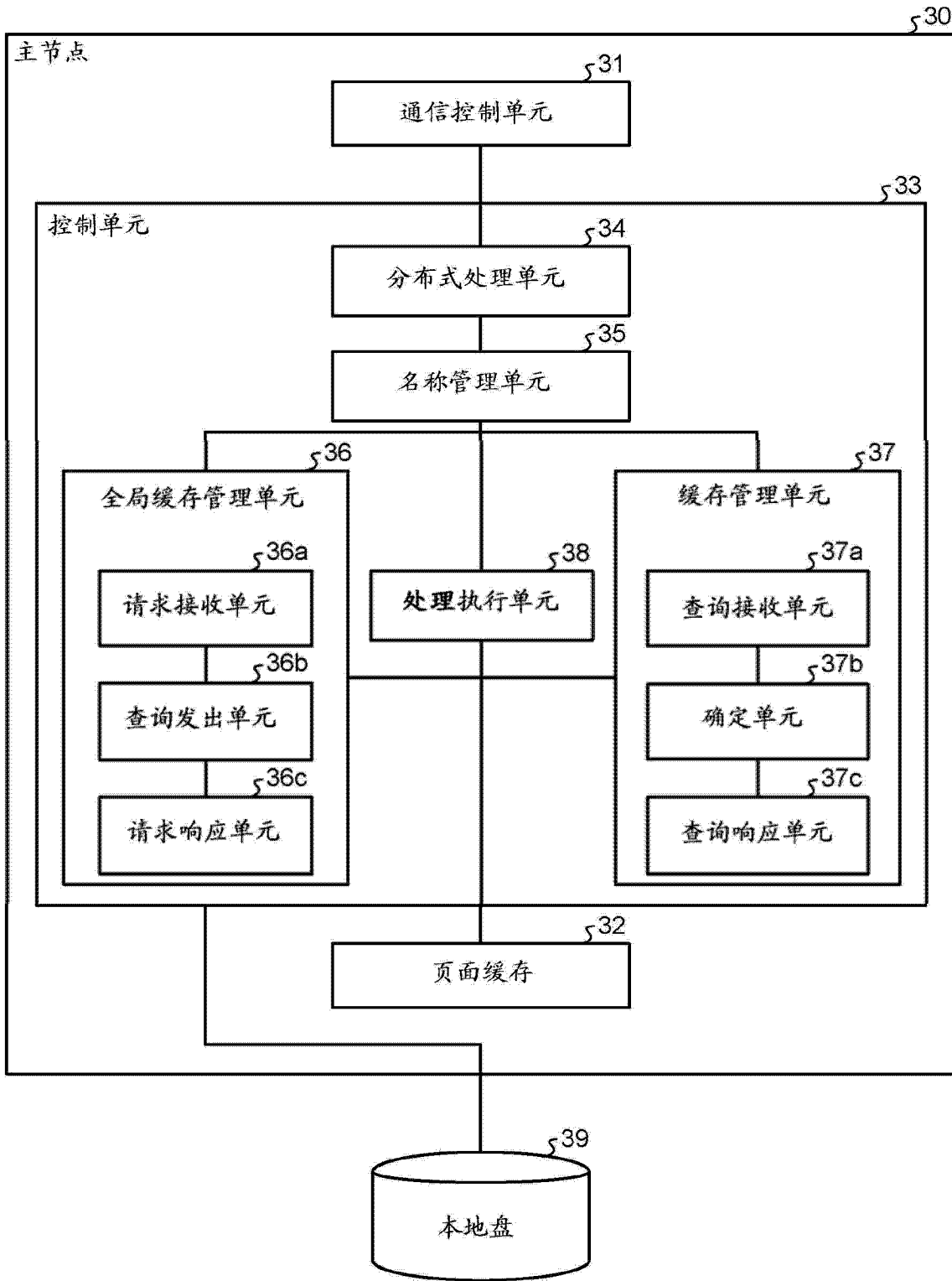


图 2

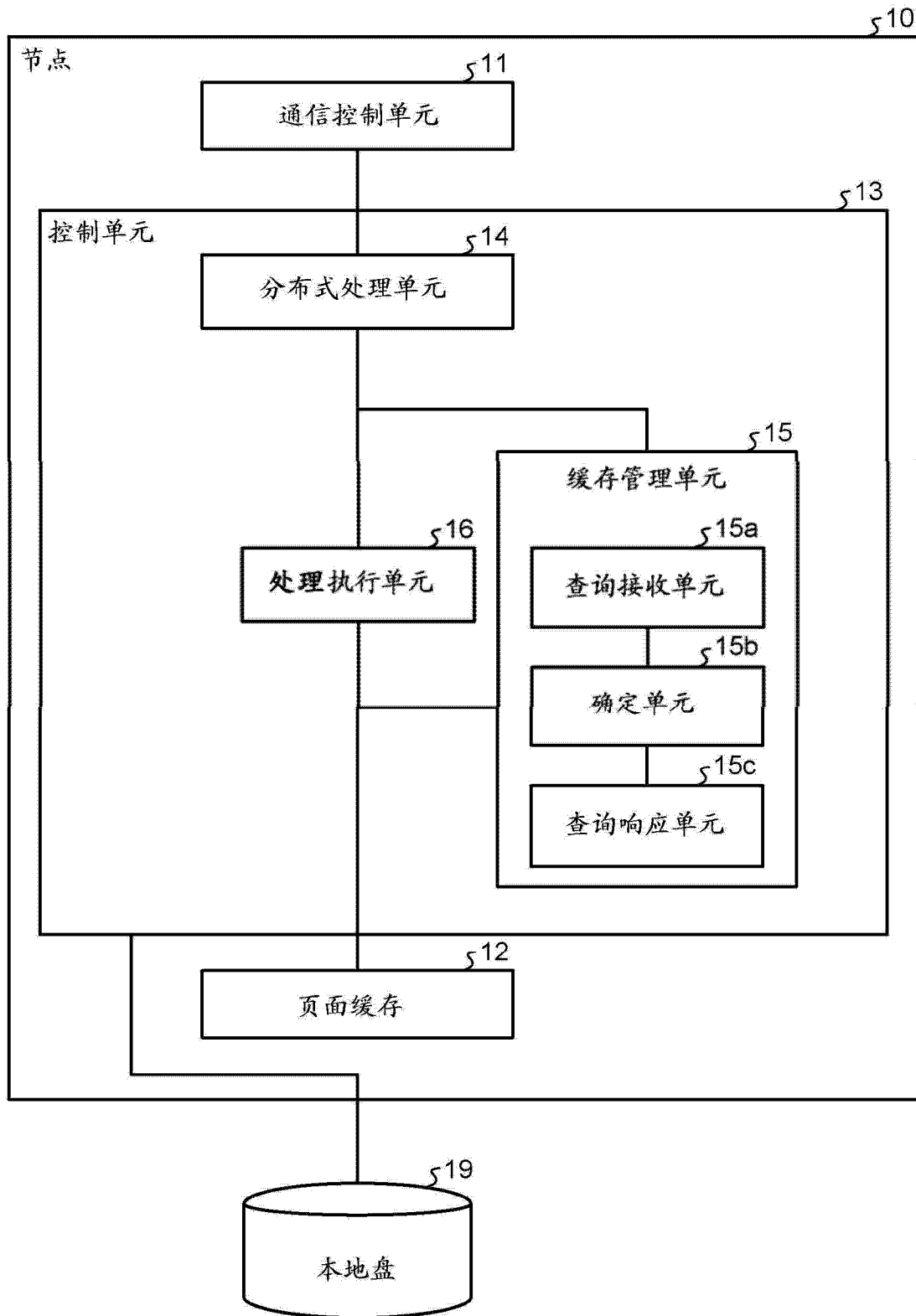


图 3

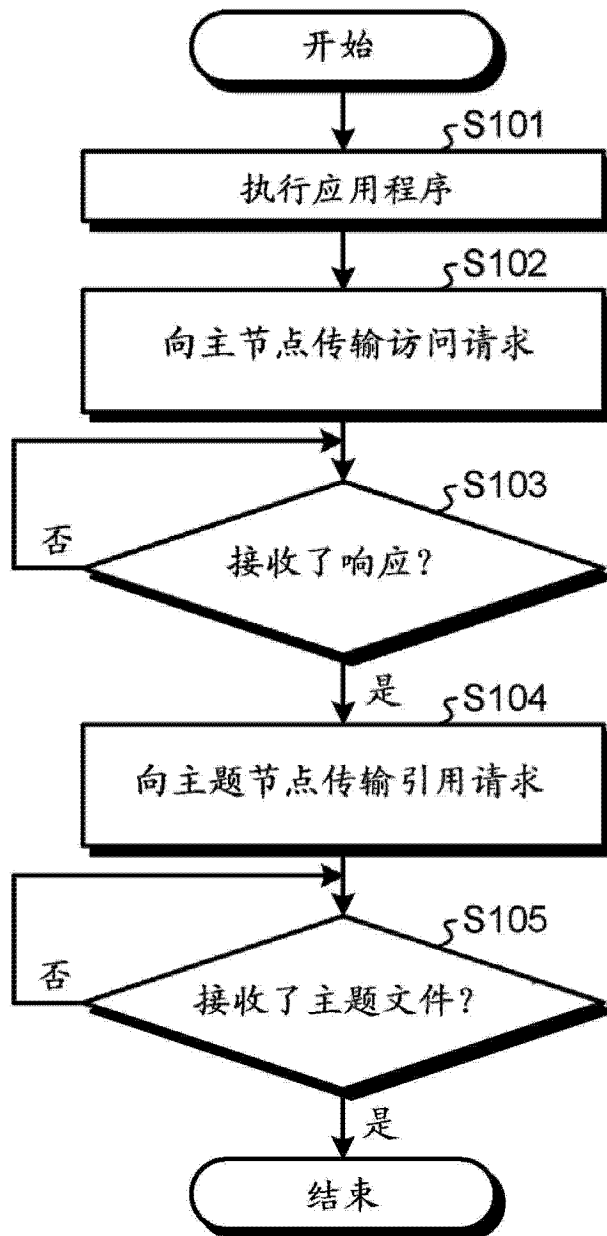


图 4

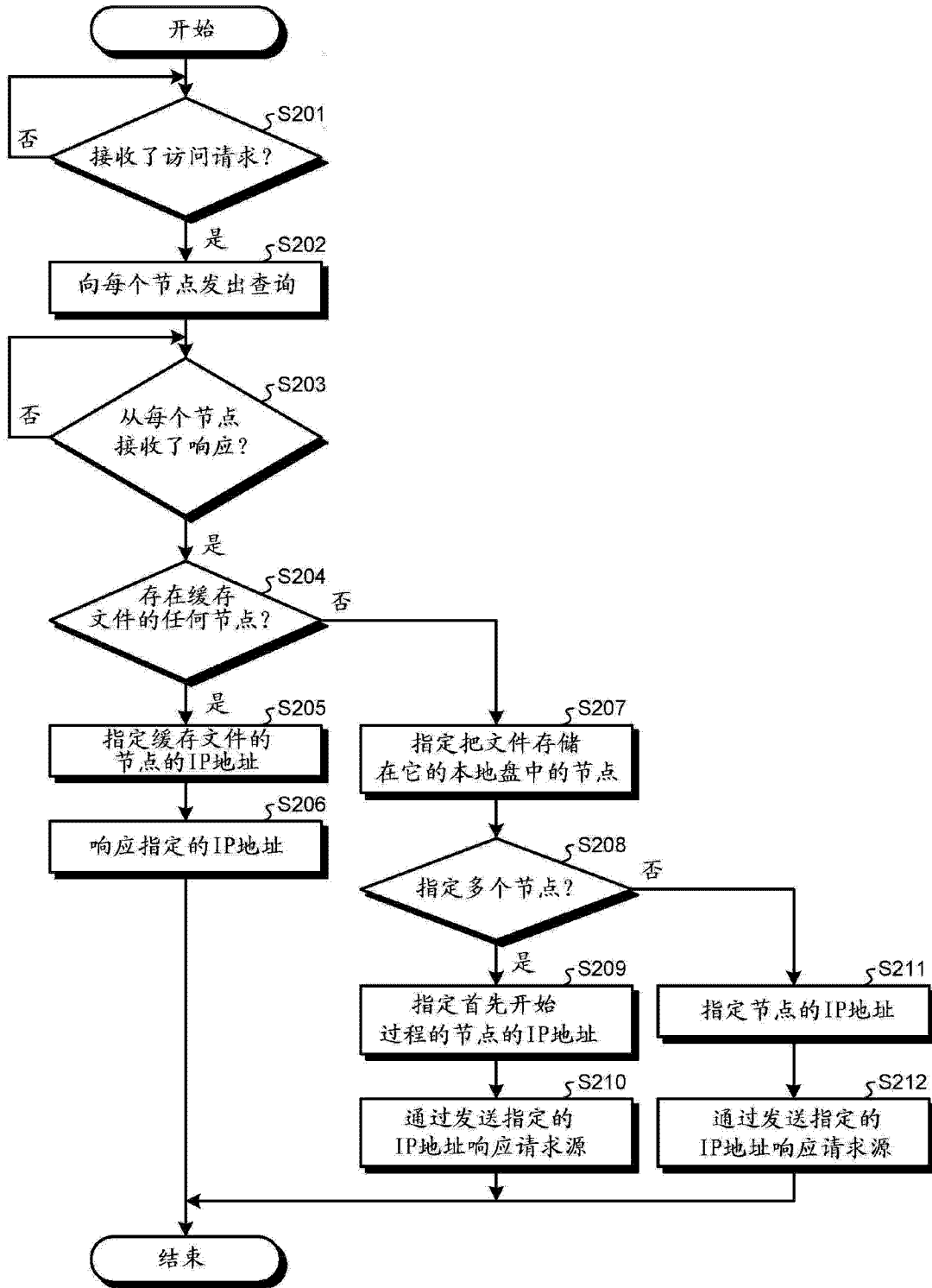


图 5

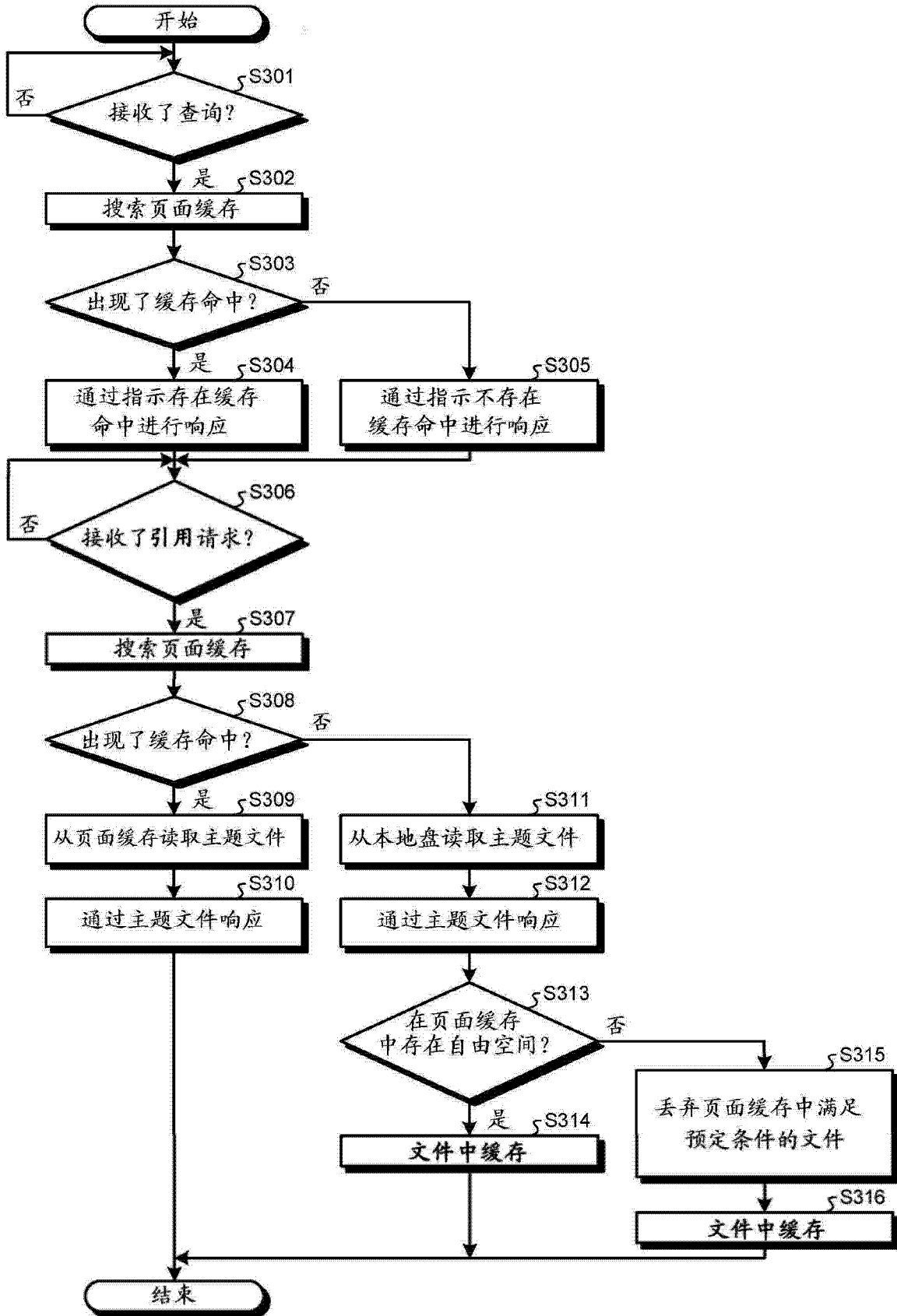


图 6

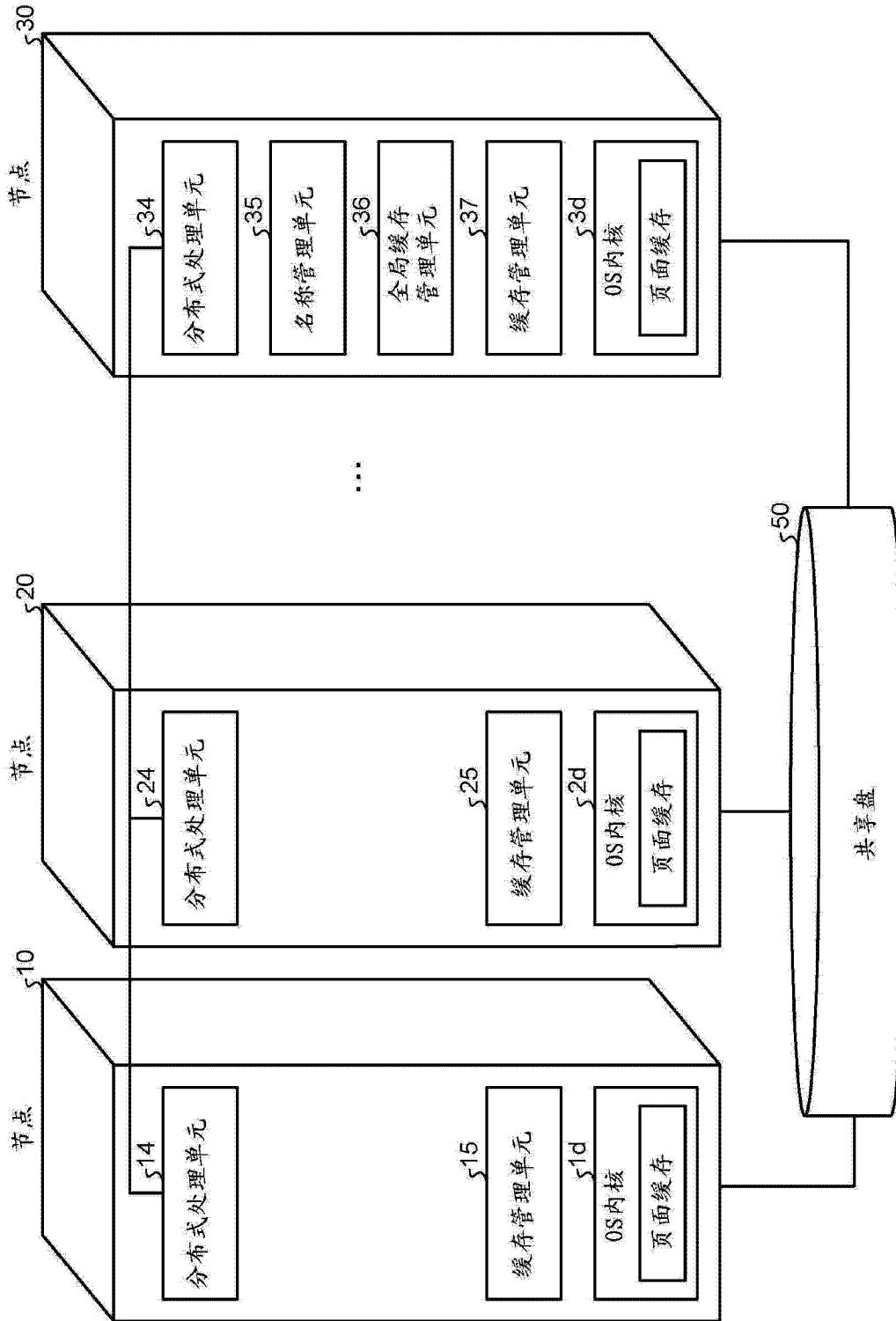


图 7

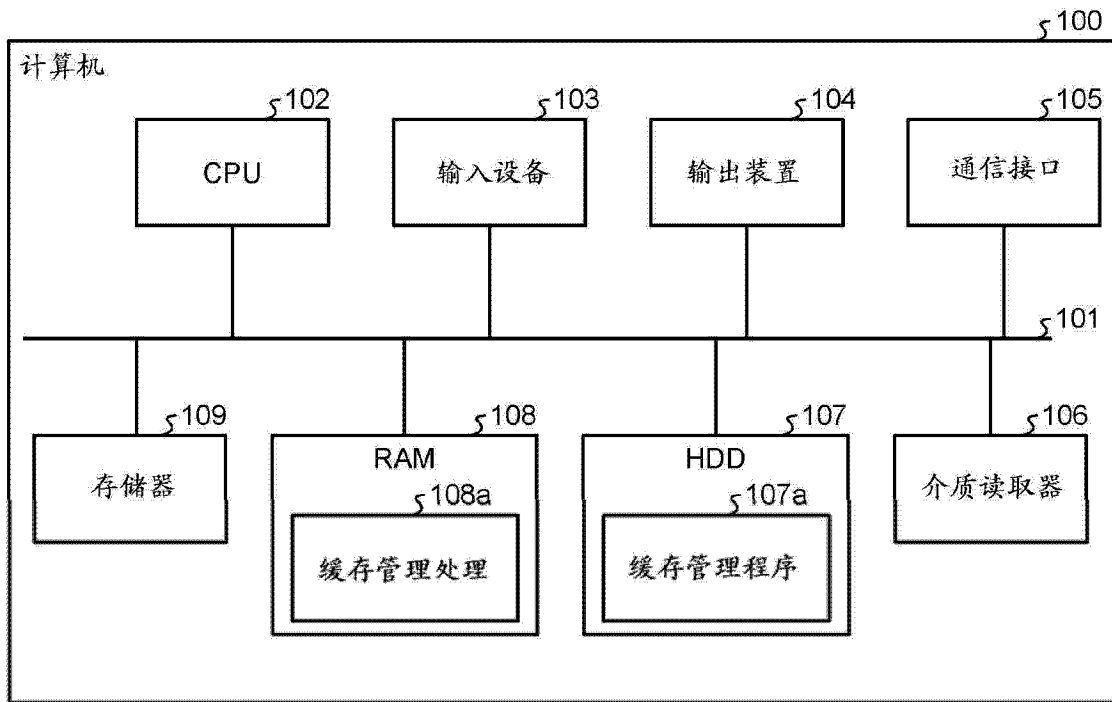


图 8

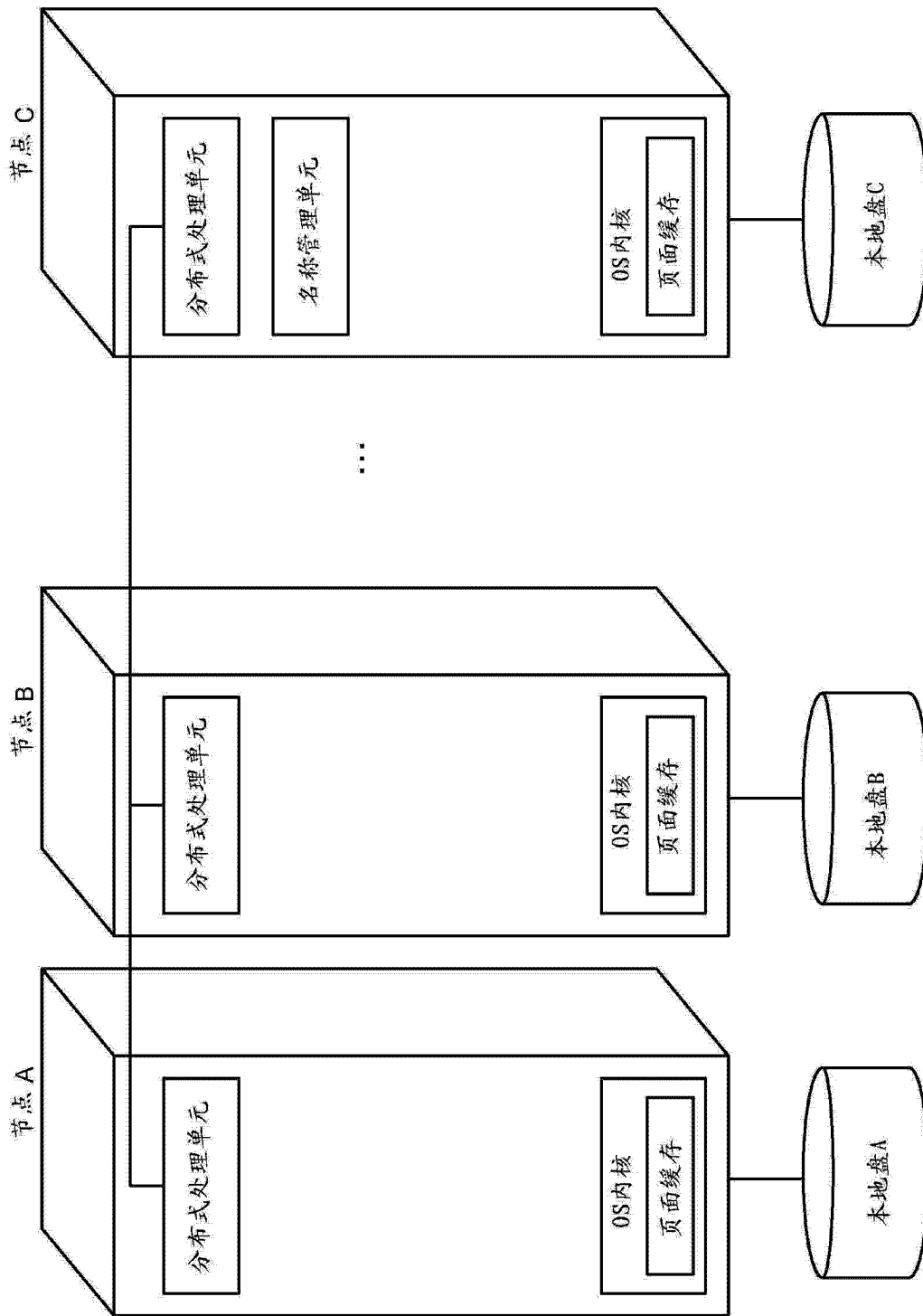


图 9