



(12) 发明专利申请

(10) 申请公布号 CN 103984638 A

(43) 申请公布日 2014. 08. 13

(21) 申请号 201410047792. 8

(22) 申请日 2014. 02. 11

(30) 优先权数据

13/765, 253 2013. 02. 12 US

(71) 申请人 LSI 股份有限公司

地址 美国加利福尼亚州

(72) 发明人 E·T·科恩

(74) 专利代理机构 上海专利商标事务所有限公

司 31100

代理人 李玲

(51) Int. Cl.

G06F 12/02 (2006. 01)

G06F 3/06 (2006. 01)

H04L 29/06 (2006. 01)

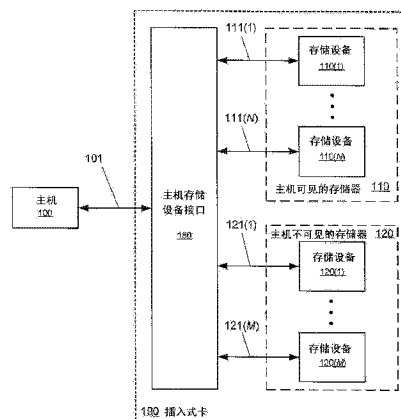
权利要求书2页 说明书8页 附图5页

(54) 发明名称

链接的可缩放的存储设备

(57) 摘要

揭示了一种链接的可缩放的存储设备。所描述的实施例访问在链接的可缩放的存储系统中的数据。一个或多个存储设备的主要代理从耦合到主要代理的主机那儿接收包括逻辑地址的主机请求。主要代理基于逻辑地址确定至少一个存储设备中的对应的物理地址，并且基于物理地址生成对存储设备中的每一个确定的物理地址的子请求。主要代理通过可与主机无关地操作的存储设备接口网络向存储设备发送子请求。存储设备接口网络是将存储设备耦合到主要代理的对等网络。主要代理响应于子请求，接收子状态，并确定总体状态。主要代理将总体状态提供到主机以便主机耦合到存储设备而无需交换机。



1. 一种访问在链接的可缩放的存储系统中的数据的方法,所述方法包括:

由一个或多个存储设备的主要代理通过主机接口网络接收来自耦合到所述主要代理的主机设备的主机请求,所述请求访问所述一个或多个存储设备的逻辑地址;

由所述主要代理基于所述逻辑地址确定在所述一个或多个存储设备中的至少一个中的对应的物理地址;

由所述主要代理基于所述物理地址生成对应于所述主机请求和所述一个或多个存储设备中的至少一个存储设备中的每一个所确定的对应的物理地址的子请求;

由所述主要代理通过可与所述主机设备无关地操作的存储设备接口网络向所述至少一个存储设备发送所述子请求,所述存储设备接口网络是将所述存储设备耦合到所述主要代理的对等网络;以及

由所述主要代理响应于所述子请求从所述至少一个存储设备接收相应的子状态,基于每一个相应的子状态确定总体状态,并将所述总体状态提供到所述主机设备,

其中,所述主机设备耦合到所述一个或多个存储设备而不使用网络交换机。

2. 如权利要求 1 所述的方法,其特征在于,所述存储设备接口网络并不直接可被所述主机接口网络访问,所述方法还包括:

由所述存储设备中的每一个通过带有与所述存储设备接口网络分开的所述主机的相应的单独的数据通信路径来发送数据通信,

由此,所述主机设备和所述存储设备之间的控制流量只是在所述主机设备和所述主要代理之间,而数据通信带宽随着所述存储设备的数量而缩放。

3. 如权利要求 1 所述的方法,其特征在于,所述主机接口网络和所述存储设备接口网络包括其中包括下列各项中的至少一项的传输介质:底板、一根或多根铜电缆、一根或多根光纤、一根或多根同轴电缆、一根或多根双绞线铜线。

4. 如权利要求 3 所述的方法,其特征在于,还包括:

有选择地向所述一个或多个存储设备的子集提供较高带宽存储设备接口网络连接,所述一个或多个存储设备的所述子集包括位于所述主机设备附近的所述存储设备中的一个或多个。

5. 如权利要求 4 所述的方法,其特征在于,所述主机接口网络包括外围组件高速互连(PCI-E) Gen4 网络,而所述存储设备互连网络包括下列各项中的一项或多项:PCI-E Gen3 网络、以太网、串行外接小型计算机系统接口(SAS)网络以及串行高级技术附件(SATA)网络。

6. 如权利要求 1 所述的方法,还包括:

在独立磁盘的冗余阵列(RAID)系统中使用所述一个或多个存储设备,其中,所述一个或多个存储设备包括下列各项中的至少一项:固态硬盘(SSD)、硬盘驱动器(HDD)、磁阻随机存取存储器(MRAM)、磁带库以及混合型磁性和固态存储系统。

7. 如权利要求 1 所述的方法,还包括:

将带宽提供到与所述一个或多个存储设备的综合可提供的带宽相关的所述主机接口网络,

其中:

所述存储设备接口网络包括一个或多个物理链路,每一个链路都具有独立带宽,以及

所述一个或多个物理链路中的每一个都包括(i)用于传输控制数据的相对较低的带宽边带耦合以及(ii)用于传输用户数据的相对较高的带宽主要频带耦合。

8. 如权利要求7所述的方法,其特征在于,所述提供包括给所述存储设备中的每一个提供所述主机接口网络的单独的物理链路。

9. 一种链接的可缩放的存储系统,包括:

多个存储设备,所述存储设备中的至少一个,所述多个存储设备中的一个或多个的主要代理;

通过主机接口网络耦合到所述至少一个主要代理的主机设备,其中,所述至少一个主要代理被配置成:

接收来自所述主机设备的主机请求,所述请求访问所述多个存储设备中的一个或多个存储设备的逻辑地址;

基于所述逻辑地址确定所述多个存储设备中的所述一个或多个存储设备中的至少一个存储设备中的对应的物理地址;

基于所述物理地址生成对应于所述主机请求和所述多个存储设备中的一个或多个存储设备中的至少一个存储设备中每一个所确定的对应的物理地址的子请求;

通过可与所述主机设备无关地操作的存储设备接口网络,向所述至少一个存储设备发送所述子请求,所述存储设备接口网络是将所述存储设备耦合到所述主要代理的对等网络;以及

响应于所述子请求从所述至少一个存储设备接收相应的子状态,基于每一个相应的子状态确定总体状态,并将所述总体状态提供到所述主机设备,

其中,所述主机设备耦合到所述一个或多个存储设备而不使用网络交换机,其中,所述存储设备接口网络并不直接可被所述主机接口网络访问。

10. 如权利要求9所述的系统,其特征在于:

所述主机设备和所述存储设备之间的控制流量只是在所述主机设备和所述至少一个主要代理之间,而数据通信带宽随着所述存储设备的数量而缩放;以及

所述存储设备接口网络被配置成下列各项中的至少一项:

有选择地将较高带宽连接提供到所述一个或多个存储设备的子集;以及

将带宽提供到与所述一个或多个存储设备的综合可提供的带宽相关的所述主机接口网络。

链接的可缩放的存储设备

[0001] 相关申请的交叉引用

[0002] 本申请是 2012 年 12 月 7 提交的美国专利申请 13/702,976 的部份继续申请,并要求申请日期的权益,而美国专利申请 13/702,976 要求 2011 年 6 月 16 日提交的美国临时申请 61/497,525、2011 年 6 月 17 日提交的国际专利申请 PCT/US2011/040996、以及 2010 年 6 月 18 日提交的美国临时申请 61/356,443 的申请日期的权益,所有这些申请的内容都包括在本申请中作为参考。

背景技术

[0003] 存储区域网络(SAN)是向耦合到 SAN 的一个或多个主机设备提供对诸如磁盘阵列和磁带库之类的合并的、块级别的存储器的访问的系统。SAN 将多个存储设备表示为到主机设备的单一逻辑接口,在概念上,将由存储设备中的每一个实现的存储器集合为单一逻辑存储空间。典型的 SAN 可以是可缩放的,就是说,在不同的 SAN 系统中,可以根据需要改变存储空间的数量(例如,存储设备的数量)。如所指出的,SAN 提供块级别的访问,意味着,文件系统通常是由主机设备管理的。典型的 SAN 可以使用块级别的协议,诸如光纤通道(FC)、以太网上的高级技术附接(ATA)(AoE)、因特网小型计算机系统接口(iSCSI)或 HyperSCSI。SAN 直接在存储设备和主机设备之间传输数据。

[0004] 网络附属存储(NAS)是提供对耦合到 NAS 的一个或多个主机设备的文件级别的访问的系统。与 SAN 不同,NAS 系统为其附接的存储设备提供文件系统,基本上充当访问一个或多个本地块级别的存储设备的文件服务器。典型的 NAS 可以使用文件级别的协议,诸如网络文件系统(NFS)或服务器消息块/通用因特网文件系统(SMB/CIFS)。SAN-NAS 混合型系统是从同一个存储系统给主机提供文件级别的访问(类似于 NAS 设备)和块级别的访问(类似于 SAN 系统)的系统。

[0005] 在 SAN、NAS 和 SAN-NAS 混合型系统中,希望使用多个存储设备,以便整个系统存储器的大小可以通过将多个存储设备分组在一起增大。这样的对存储设备的组通常需要带有交换机的通信层次结构,以便存储设备对主机可用,无论是分别地还是集合地。

发明内容

[0006] 提供本发明内容是为了以精简的形式介绍将在以下详细描述中进一步描述的一些概念。本发明内容并不旨在标识所要求保护的主题的关键特征或必要特征,也不旨在用于限制所要求保护的主题的范围。

[0007] 所描述的实施例访问在链接的、可缩放的存储系统中的数据。一个或多个存储设备的主要代理从耦合到该主要代理的主机那儿接收包括逻辑地址的主机请求。该主要代理基于该逻辑地址确定至少一个存储设备中的对应的物理地址,并且基于该物理地址生成对该存储设备中的每一个确定的物理地址的子请求。该主要代理通过可与主机无关地操作的存储设备接口网络向该存储设备发送该子请求。该存储设备接口网络是将该存储设备耦合到该主要代理的对等网络。该主要代理响应于该子请求而接收子状态,并确定总体状态。该

主要代理将该总体状态提供到该主机,以便该主机耦合到该存储设备而无需交换机。

附图说明

[0008] 通过下面的详细描述,所附的权利要求,以及附图,所描述的实施例的其他方面、特点、以及优点将变得更加完全显而易见,在附图中,类似的参考编号标识类似或相同的元素。

[0009] 图 1 示出了根据示例性实施例的可缩放的存储系统的框图;

[0010] 图 2 示出了根据示例性实施例的可缩放的存储系统的框图;

[0011] 图 3 示出了根据示例性实施例的可缩放的存储系统的框图;

[0012] 图 4 示出了根据示例性实施例的可缩放的存储系统的框图;以及

[0013] 图 5 示出了根据示例性实施例的可缩放的存储系统的框图。

具体实施方式

[0014] 所描述的实施例访问在链接的可缩放的存储系统中的数据。一个或多个存储设备的主要代理从耦合到该主要代理的主机那儿接收包括逻辑地址的主机请求。该主要代理基于该逻辑地址确定至少一个存储设备中的对应的物理地址,并且基于该物理地址生成对该存储设备中的每一个确定的物理地址的子请求。该主要代理通过可与主机无关地操作的存储设备接口网络向该存储设备发送该子请求。该存储设备接口网络是将该存储设备耦合到该主要代理的对等网络。该主要代理响应于该子请求而接收子状态,并确定总体状态。该主要代理将该总体状态提供到该主机,以便该主机耦合到该存储设备而无需交换机。

[0015] 表 1 定义了为有助于理解所描述的实施例在整个说明书中使用的缩写词列表:

[0016]

表 1			
AoE	以太网上的高级技术附接 (ATA)	CD	压缩光盘
CIFS	普通因特网文件系统	DVD	数字多功能盘
FC	光纤通道	HDD	硬盘驱动器
HIF	主机接口	IC	集成电路
I/O	输入/输出	iSCSI	因特网SCSI
MRAM	磁阻随机存取存储器	NAS	网络附属存储
NFS	网络文件系统	PCI-E	外围组件高速互连
PHY	物理层	RAID	独立磁盘冗余阵列
RF	射频	SAN	存储区域网络
SAS	串行外接SCSI	SATA	串行高级技术附件
SCSI	小型计算机系统接口	SMB	服务器消息块
SoC	芯片上的系统	SRIO	串行快速输入/输出

[0017]

SSD	固态磁盘	USB	通用串行总线
-----	------	-----	--------

[0018] 在某些 SAN、NAS、或 SAN-NAS 混合型系统中，存储设备可以让设备的主要代理通过主机接口 (HIF) 协议接受从主机设备接收到的存储请求。主要代理处理主机请求，并通过对等协议，向每一个存储设备的辅助代理，生成一个或多个子请求。辅助代理接受并处理子请求，并将每一个子请求的子状态信息报告给主要代理和 / 或主机。主要代理可任选地将子状态累积到主机请求的总体状态中。代理之间的对等通信可任选地用于在主机访问和 / 或故障恢复过程中传递冗余信息。各种故障恢复技术可以通过冗余信息，重新分配存储器、重新分配代理并恢复数据。

[0019] 图 1 示出了例如如 2012 年 12 月 7 日提交的相关的美国专利申请 13/702, 976 所描述的示例性可缩放的存储系统的框图，该专利申请以引用的方式并入本文中。如图 1 所示，可缩放的存储系统包括至少一台通过耦合 101 耦合到可插入的存储模块 190 的主机设备 (100)。耦合 101 可以实现为传输介质，诸如底板、铜电缆、光纤、一根或多根同轴电缆、一根或多根双绞线铜线，和 / 或一个或多个射频 (RF) 通道。例如，耦合 101 可以实现为 FC、AoE、iSCSI，或 HyperSCSI 链路 (例如，在 SAN 系统中) 或实现为 NFS 或 SMB/CIFS 链路 (例如，在 NAS 系统中)。

[0020] 可插入的存储模块 190 包括至少一个主机 / 存储设备接口 (示为 180)。虽然在图 1 中被示为与可插入的存储模块 190 集成，但是，在某些实施例中，主机 / 存储设备接口 180 可以与每一个主机设备 100 集成。在某些实施例中，可插入的存储模块 190 可以实现为插入式卡。如图 1 所示，可插入的存储模块 190 包括主机可见的存储器 110，该存储器 110 包

括一个或多个存储设备 110 (1)–110 (N)。主机可见的存储器 110 实现存储器,其部分或全部被配置成被主机设备 100 通过主机到存储设备接口 180 进行访问。可插入的存储模块 190 还包括主机不可见的存储器 120,该存储器 120 包括一个或多个存储设备 120 (1)–120 (M)。主机不可见的存储器 120 实现未被直接报告的存储器,如此,对主机设备 100 “不可见”。然而,对主机不可见的存储器被报告,并可以通过主机可见的存储器 110 的元件,例如,通过对等协议,间接地被主机设备 100 访问。例如,存储元件的主要代理报告主要代理和与主要代理进行通信的任何辅助代理的组合的存储容量,尽管辅助代理对主机设备 100 不可见。在某些实施例中,存储设备 110 和 120 是物理存储设备,诸如固态硬盘(SSD)、硬盘驱动器(HDD)、磁带库、混合型磁性和固态存储系统,或其某种组合。

[0021] 耦合 101、111 和 121 的组合一起在主机设备 100 和主机可见的存储器 110(以及通过主机可见的存储器 110,主机不可见的存储器 120)之间启用请求、状态、和数据传输。例如,耦合中的一个或多个通过主机接口协议来实现传输,例如,通过主机设备 100 中的一个作为主控设备来操作,主机可见的存储器 110 的存储元件中的一个作为从属设备来操作。进一步地,耦合中的一个或多个通过对等协议来实现传输,例如,通过主机可见的存储器 110 的元件中的一个作为主要代理来操作,主机不可见的存储器 120 的元件中的一个或主机可见的存储器 110 的元件中的另一个作为辅助代理来操作。耦合 111 和 121 可以实现为自定义设计的通信链路,或可以实现为符合诸如,例如,小型计算机系统接口(SCSI)链路、串行外接 SCSI (SAS)链路、串行高级技术附件(SATA)链路、通用串行总线(USB)、光纤通道(FC)链路、以太网链路(例如,10GE 链路)、IEEE802. 11 链路、IEEE802. 15 链路、IEEE802. 16 链路、外围组件高速互连(PCI-E)链路、串行快速 I/O (SRIO)链路、InfiniBand 链路之类的标准通信协议的链路,或其他类似的接口链路。

[0022] 在某些实施例中,主机 / 存储设备接口 180 通常可以实现为一个或多个 PCI-E 或 InfiniBand 交换机,以便主机设备 100、耦合 101 和主机 / 存储设备接口 180 实现统一的交换机。在更进一步的实施例中,统一交换机可作为相对于主机可见的存储器 110 透明的交换机来操作,同时还可作为相对于主机不可见的存储器 120 非透明的交换机来操作。如图 1 所示,PCI-E 交换机(例如,主机 / 存储设备接口 180)是不同于存储设备 110 和 120 中的每一个的单独的元件。

[0023] 如此,以引用的方式并入本文中的 2012 年 12 月 7 日提交的相关的美国专利申请 13/702, 976 描述了包括一个或多个 PCI-E 或 InfiniBand 交换机(例如,主机 / 存储设备接口 180)的可缩放的存储系统。如果 PCI-E 交换机是非透明的交换机,交换机下面的拓扑的细节以及单个存储设备的配置的细节对主机设备不可见(例如,在附属的设备的主机初始化发现中)。如此,通过使用非透明的交换机,所描述的实施例可以选择存储设备中的一个来充当主控设备(例如,主要代理)来处理与所有存储设备的所有主机通信,并选择其余存储设备来充当对主机设备不可见的从属设备(例如,作为辅助代理),尽管所有存储设备 110 和 120 可以是重复的设备。进一步地,存储设备的聚合的组可以对主机设备而言作为单一存储设备来出现。

[0024] 其他所描述的实施例可以通过使用“近邻 – 到 – 近邻”通信,提供可缩放的功能,而不使用单独的 PCI-E 交换机,以便通信使用每一个存储设备之间的点对点链路,无需较高的级别(例如,PCI-E 层次结构)。通过诸如路由或交换之类的技术,所有存储设备都能够

在彼此之间进行通信,尽管所有连接在存储设备之间都是点对点的。

[0025] 图 2 示出了示例性存储设备 110 的框图。主机设备 100 通过耦合 101 耦合到存储设备 110。耦合 100 与 PHY 接口 202 进行通信。如图 2 所示,PHY 接口 202 包括一个或多个上游物理层链路或端口(PHY)(示为 101)和一个或多个下游 PHY(示为 218(1) - 218(N))。如图 2 所示,存储设备 110 包括大容量存储设备 216,包括下列各项中的一项或多项:固态存储器 210(例如,SSD)、磁存储器 212(例如,HDD 或磁带库)和光学存储器 214(例如,CD 或 DVD)。存储设备 110 包括与每一个单个存储设备 210、212 和 214 进行通信的存储器接口 206。逻辑/物理转换模块 204 在从主机设备 100 接收到的操作的逻辑地址和大容量存储器 216 上的物理地址之间转换。存储设备 110 还包括子状态模块 222 和子请求模块 220,这两者都与 PHY 接口 202 进行通信。

[0026] 在所描述的实施例中,上游 PHY(例如,101)通过 PCI-E 层次结构,与主机设备(例如,100)进行通信,而下游 PHY(例如,218)与其他存储设备(例如,多个 110)进行通信。示例性实施例可以使用固定数量的可配置的 PHY,例如,总计 8 个可配置的 PHY,其中,给定 PHY 可被配置成上游链路或下游链路。具有可配置的 PHY 可以在提供给主机 101 的带宽(例如,上游连接)和可缩放的存储系统的容量(例如,下游连接)之间取得折衷。其他实施例可以使用固定数量的上游 PHY 和固定数量的下游 PHY,例如,2 个上游 PHY 和 6 个下游 PHY。

[0027] 在各种实施例中,存储设备(例如,110)的 PHY101 和 218 中的某些或全部可以以相同速度(例如,相同最大速度)操作或可以每一个都以不同的速度操作。例如,一些实施例可以使 PHY101 和 218 中的每一个独立地支持下列各项中的任何一项或多项:PCI-E Gen1、Gen2、Gen3 或 Gen4、10GE、InfiniBand、SAS、SATA 或用于与一个或多个存储设备进行通信的非标准的协议。PHY101 和 218 中的每一个都耦合到集成在每一个存储设备 110 内的一个或多个相应的 PHY 接口。当,例如,PHY 接口 202 是 PCI-E 接口时,PCI-E 接口是可配置的,以作为下列各项中的一项或多项:根综合体;转发点;以及端点。转发点与根综合体的相似之处在于,转发点可以在一个或多个 PCI-E 接口之间发送和接收流量。根综合体另外还是单独的 PCI-E 层次结构的根。由于耦合到一个或多个存储设备(例如,110)的主机设备(例如,100)本身是根综合体,因此,如果耦合到主机的存储设备中的一个或多个也是根综合体,那么,创建了多根 PCI-E 层次结构。

[0028] 多个存储设备 110 可以以任意数量的不同的方式连接。图 3-5 示出了根据示例性实施例的可缩放的存储系统中的多个存储设备的示例性点到点连接的框图。如图所示,在各种实施例中,PHY 和 PHY 控制器可以通过下列各项来耦合:如图 3 所示的菊花链(或可选地,环);到主机设备的固定的、1 到 1 互连(图 4 所示出的);全交叉开关拓扑;部分交叉开关拓扑;多路复用器网络;其组合;或用于耦合多个硬件设备的任何其他技术。在某些实施例中,存储设备之间的连接网络是交换网络,而在其他实施例中,存储设备之间的连接网络是路由网络。进一步地,在某些实施例中,存储设备 110 中的至少某些具有不同的配置的 PHY,或一种或多种不同类型的 PHY(例如,PCI-E、10GE、InfiniBand、SAS、SATA 等等)。

[0029] 如图 3 和 4 所示,图 3 的存储设备 110(A)-110(N),以及图 4 的存储设备 110(1)-110(N)具有被配置为转发点的内部 PHY 接口。图 5 示出了分层的耦合,其中,全部存储设备 110 都具有被配置为转发点的 PHY 接口,其 PHY 接口被配置为端点的存储设备 110。Z1 到 110。ZN 除外。如此,在所描述的实施例中,存储设备 110 中的一个或多个(图 3 的存储

设备 110 (A)、图 4 的 110 (1)–110 (N)、图 5 的 110. A) 耦合到主机设备 100, 存储设备中的全部都直接耦合到主机设备 100 (例如, 如图 4 所示), 或通过其他存储设备, 间接地耦合到主机设备 100, 而不使用, 例如, PCI-E 交换机。

[0030] 存储设备 110 中的至少一个充当主要代理, 而存储设备 110 中的至少一个或多个充当辅助代理。在各种实施例中, 一个或多个主要代理比辅助代理具有直接、更直接、较短的、和 / 或较低延迟的连接。例如, 如图 3 所示, 存储设备 110 (A) 可以充当存储设备 110 (B)–110 (N) 的主要代理, 因为, 例如, 存储设备 110 (A) 具有到主机设备 100 的直接连接, 而存储设备 110 (B)–110 (N) 以菊花链彼此耦合。如图 4 所示, 存储设备 110 (1)–110 (N) 中的全部都能够充当它们自己的主要代理, 因为每一个存储设备 110 (1)–110 (N) 都具有到主机设备 100 的直接连接。具有到主机的直接连接的每一个存储设备有利地启用往返于主机的带宽, 以随着存储设备的数量线性地缩放。进一步地, 让存储设备的子集, 诸如存储设备中的仅仅一个, 充当主要代理, 让其他的充当辅助代理, 实现可缩放的容量, 无需主机控制多个单独的存储设备。如图 5 所示, 存储设备 110. A 可以充当存储设备 110. B1–110. Bn 的主要代理, 因为, 例如, 存储设备 110. A 具有到主机设备 100 的直接连接, 而存储设备 110. B1 可以充当通过耦合 218 (C1) 耦合的存储设备 (未示出) 的主要代理, 依次类推。

[0031] 在所描述的实施例中, 主要代理和辅助代理之间所有通信都作为对主机设备 100 不可见的 (如此, 对主机设备 100 的 PCI-E 层次结构不可见) 近邻到近邻流量来执行。例如, 如图 3 所示, 对耦合 218 (1)–218 (N) 执行近邻到近邻流量中的全部, 而不对将存储设备 110 耦合到主机设备 100 的连接 101 执行近邻到近邻流量。类似地, 如图 4 所示, 对耦合 218 (1)–218 (N) 执行近邻到近邻流量中的全部, 而不对将存储设备 110 (1)–110 (N) 耦合到主机设备 100 的耦合 101 (1)–101 (N) 执行近邻到近邻流量。类似地, 如图 5 所示, 对耦合 218 (B1)–218 (Zn) 执行近邻到近邻流量中的全部, 而不对将存储设备 110. A 耦合到主机设备 100 的耦合 101 执行近邻到近邻流量。

[0032] 在所描述的实施例中, 近邻到近邻流量是控制流量, 诸如由主要代理接收到的命令从主机设备 100 转发到存储设备 110 中的特定的一个以及从存储设备 110 中的特定一个回到主要代理的响应 (例如, 完成), 从主机设备 100 那儿接收到的命令中所导出的信息, 诸如同步或心跳之类的维护流量; RAID 或其他数据冗余性控制或数据流量 (RAID 的增量), 及其他流量。例如, 当写入命令更新存储设备 110 中的特定一个上的 RAID 条带的一部分时, 特定存储设备将 RAID 增量发送到其他存储设备中的一个或多个 (例如, 存储设备中的一个存储条带的 RAID 奇偶校验), 作为近邻到近邻流量。

[0033] 耦合 101 以及 218, 如图 3–5 所示, 可任选地或有选择地是不同的带宽和 / 或不同的协议。例如, 到主机设备 100 的上游连接 (例如, 耦合 101) 通常可以是 PCI-E Gen4, 而各种存储设备 110 之中的下游连接 (例如, 耦合 218) 通常可以是 PCI-E Gen3 或不同的协议, 诸如 10GE、InfiniBand、SAS 等等。耦合中的任何一个都可以具有彼此不同的带宽或不同数量的物理链路。在某些实施例中, 耦合 101 和 218 中的任何一个的控制流量都可以通过相对较低的带宽边带耦合来传输, 而数据流量可以通过相对较高带宽的主要频带耦合来传输。如此, 在某些实施例中, 耦合 101 和 218 中的任何一个都可以实现为自定义设计的通信链路, 或可以实现为符合诸如, 例如, SCSI、SAS、SATA、USB、FC、以太网以太网 (例如, 10GE)、IEEE802. 11、IEEE802. 15、IEEE802. 16、PCI-E、SRI0、InfiniBand 之类的标准通信协议的链

路,或其他类似的接口链路。

[0034] 在某些实施例中,诸如如图 4 所示出的,主机设备 100 上游的带宽基本上等于各种存储设备 110 的综合可提供的带宽。在某些实施例中,诸如如图 5 所示出的,可通信地更靠近主机设备 100(例如,存储设备 110.A)的存储设备 110 的存储设备 110 比可通信地远离主机设备 100 的存储设备(例如,存储设备 110.Z1)被配置成用于较高带宽。在某些实施例中,存储设备 110 中的每一个都可以具有不同的容量、功能,或实现为不同类型的存储介质,诸如固态硬盘(SSD)、硬盘驱动器(HDD)、磁阻随机存取存储器(MRAM)、磁带库、混合型磁性和固态存储系统,或其某种组合。

[0035] 在某些实施例中,存储设备 110 之间的连接网络使用 PCI-E 协议(或其他标准协议),但是,以非标准的方式,诸如通过具有循环的互连(例如,如由图 3 和 4 中的可选耦合 218(N)所指示的)。在更进一步的实施例中,存储设备 110 之间的连接网络被允许使用非标准的带宽,信令、命令或协议扩展,以有利地提高性能。一般而言,存储设备 110 之间的连接网络被允许以在带宽、延迟,以及电源中的一项或多项有效率的方式提供设备间的通信。

[0036] 如此,如此处所描述的,所描述的实施例访问链接的可缩放的存储系统中的数据。一个或多个存储设备的主要代理从耦合到主要代理的主机接收包括逻辑地址的主机请求。主要代理基于逻辑地址确定至少一个存储设备中的对应的物理地址,并且基于物理地址生成对存储设备中的每一个确定的物理地址的子请求。主要代理通过可与主机无关地操作的存储设备接口网络向存储设备发送子请求。存储设备接口网络是将存储设备耦合到主要代理的对等网络。主要代理响应于子请求,接收子状态,并确定总体状态。主要代理将总体状态提供到主机,以便主机耦合到存储设备,无需交换机。

[0037] 此处所提及的“一个实施例”是指参考实施例所描述的特定功能、结构或特征可以包括在至少一个实施例中。在本说明书中的不同位置出现短语“在一个实施例中”不一定都是指同一个实施例,也不是指与其他实施例互相排斥的单独的或备选实施例。同理也适用于术语“实现”。

[0038] 如本申请中所使用的,在此使用词语“示例性”意指用作示例、实例或说明。本文中被描述为“示例性”的任何方面或设计不一定被解释为优选于或更优于其他方面或设计。相反,使用词语“示例性”旨在以具体的方式呈现各个概念。

[0039] 尽管参考软件程序中的处理框,包括作为数字信号处理器、微控制器,或通用计算机的可能的实现描述了示例性实施例,但是,所描述的实施例不仅限于此。对那些精通本技术的人显而易见的是,软件的各种功能也可以电路的进程来实现。这样的电路可以用于,例如,单一集成电路、多芯片模块、单一卡、或多卡电路封装。

[0040] 所描述的实施例还可以以方法的形式以及用于实施那些方法的设备来实现。所描述的实施例还可以以在非瞬时的有形的介质(如磁记录介质、光学记录介质、固态存储器、软盘、CD-ROM、硬盘驱动器或任何其他非瞬时的机器可读的存储介质)中实现的程序代码的形式来实现,其中,当程序代码被加载到诸如计算机之类的机器并由它们执行时,机器变为用于实施所描述的实施例的设备。所描述的实施例也可以以程序代码的形式实现,例如,无论存储在非瞬时的机器可读的存储介质中,加载到机器中和/或由机器执行,或通过某些传输介质或载体传输,如通过电线或电缆、通过光纤,或通过电磁辐射,其中,当程序代码被加载到中并由诸如计算机之类的机器执行时,机器变为用于实施所描述的实施例的设备。

当在通用处理器上实现时，程序代码段与处理器结合，以提供与特定逻辑电路类似地工作的唯一设备。所描述的实施例还可以以比特流或通过介质以电的方式或以光学方式传输的，在磁记录介质等等中存储的磁场变体、使用所描述的实施例的方法和 / 或设备生成的信号值的其他序列的形式来实现。

[0041] 应该理解，此处阐述的示例性方法的步骤不一定按所描述的顺序执行，这样的方法的步骤的顺序应该被理解为只是示例性的。同样，在这样的方法中也可以包括额外的步骤，在根据所描述的实施例的方法中，也可以省略或组合某些步骤。

[0042] 如此处参考元件和标准所使用的，术语“兼容的”是指，元件以完全或部分地由标准所指定的方式与其他元件进行通信，将被其他元件识别为充分能够以由标准所指定的方式与其他元件进行通信。兼容的元件不需要以由标准所指定的方式在内部操作。除非明确地换声明，每一个数字值和范围都应该被解释为与值或范围前面的单词“大约”或“差不多”近似。

[0043] 此外，对于本说明书，术语“耦合”、“连接”是指当前技术已知的或以后开发的允许在两个或更多元件之间传输能量的任何方式，预期可以插入一个或多个额外的元件，虽然不是必需的。相反，术语“直接耦合”、“直接连接”等等，意味着不存在这样的额外元件。信号和对应的节点或端口可以通过相同名称来引用，并为这里的目的而可互换。

[0044] 应该进一步地理解，在不偏离如下面的权利要求书所表达的本发明的范围的情况下，本领域的技术人员可以对这里所描述的以便说明所描述的实施例的本质的那些部分的细节、材料、布局进行许多更改。

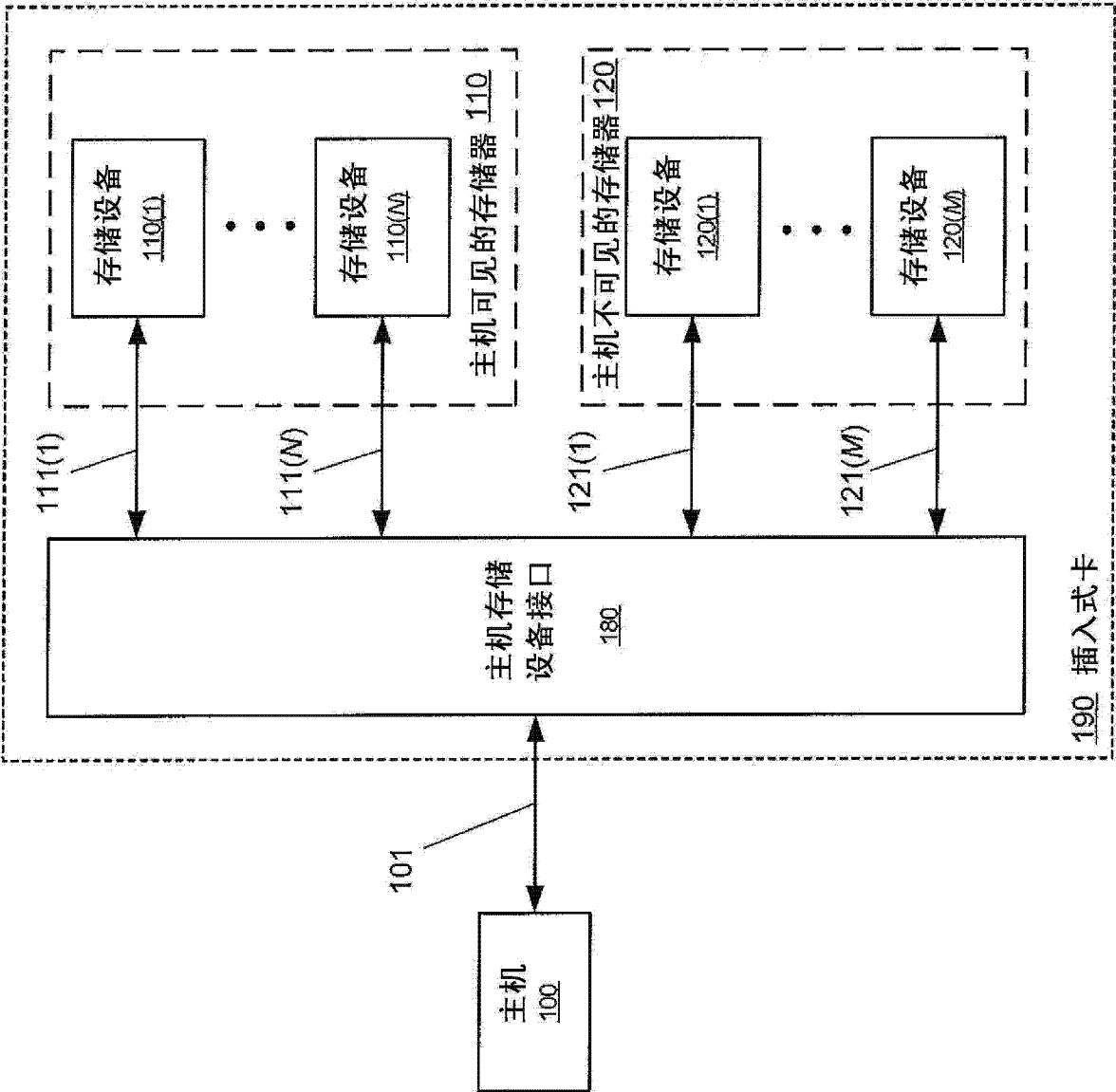


图 1

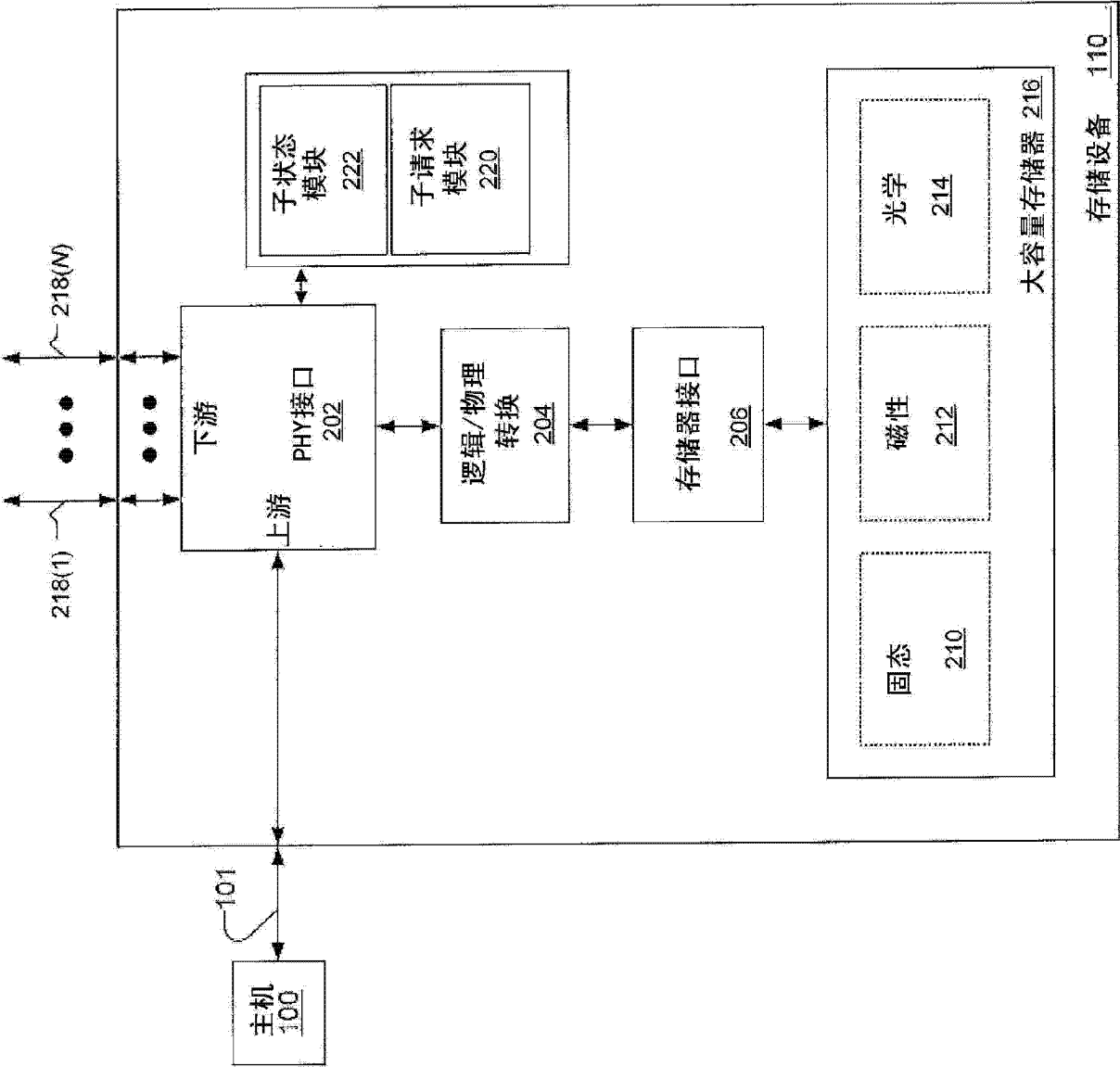


图 2

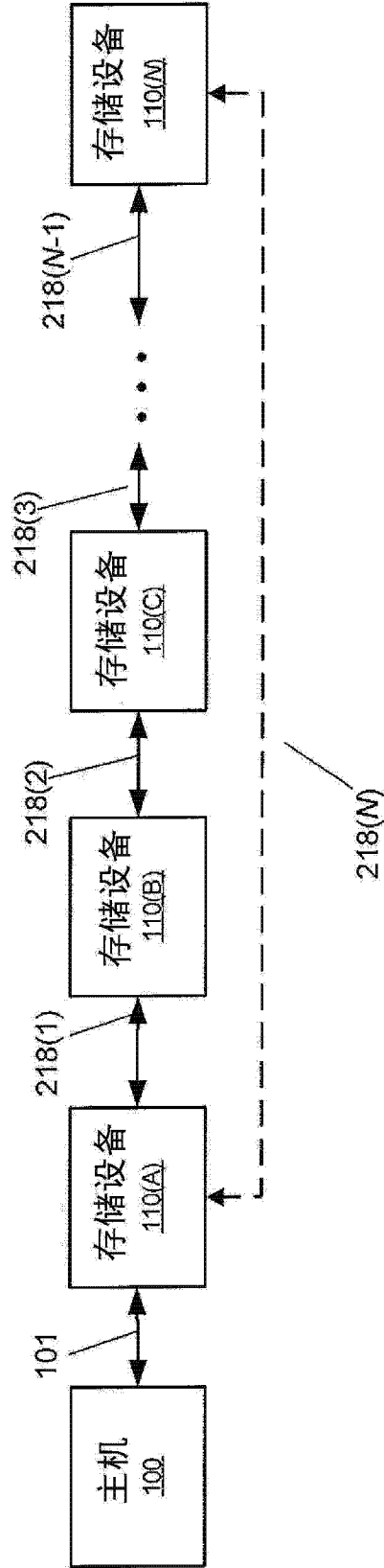


图 3

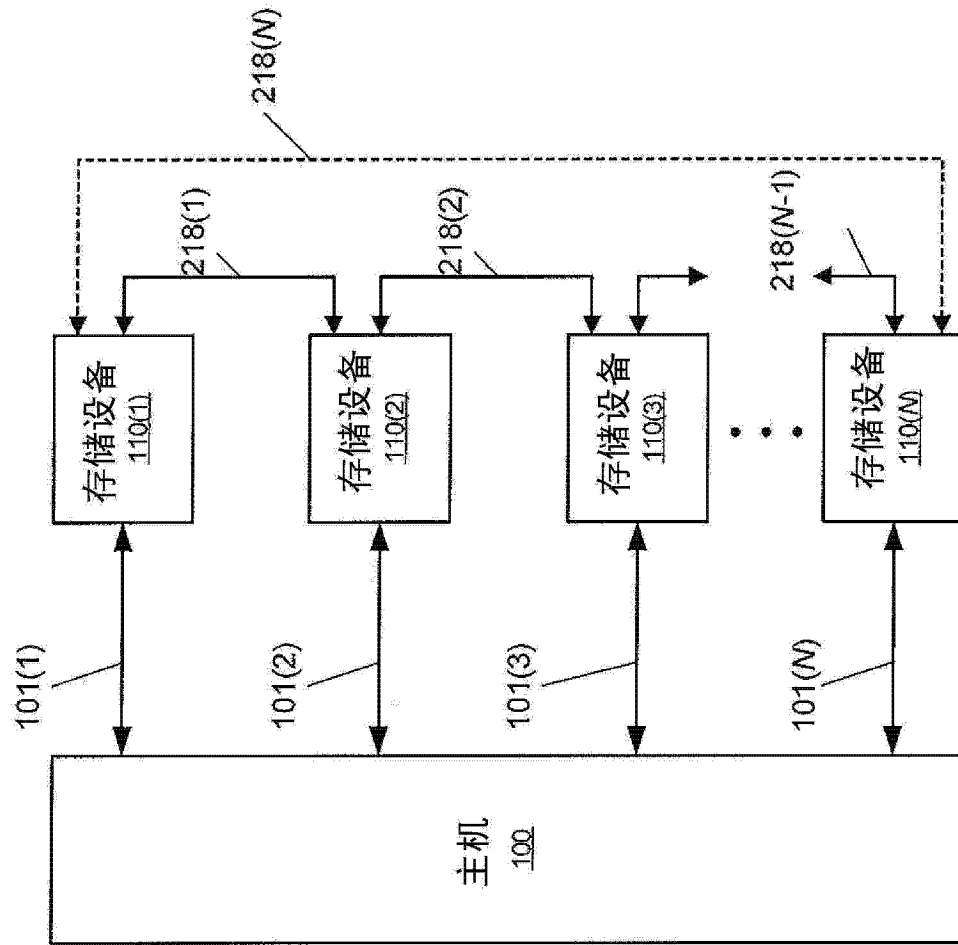


图 4

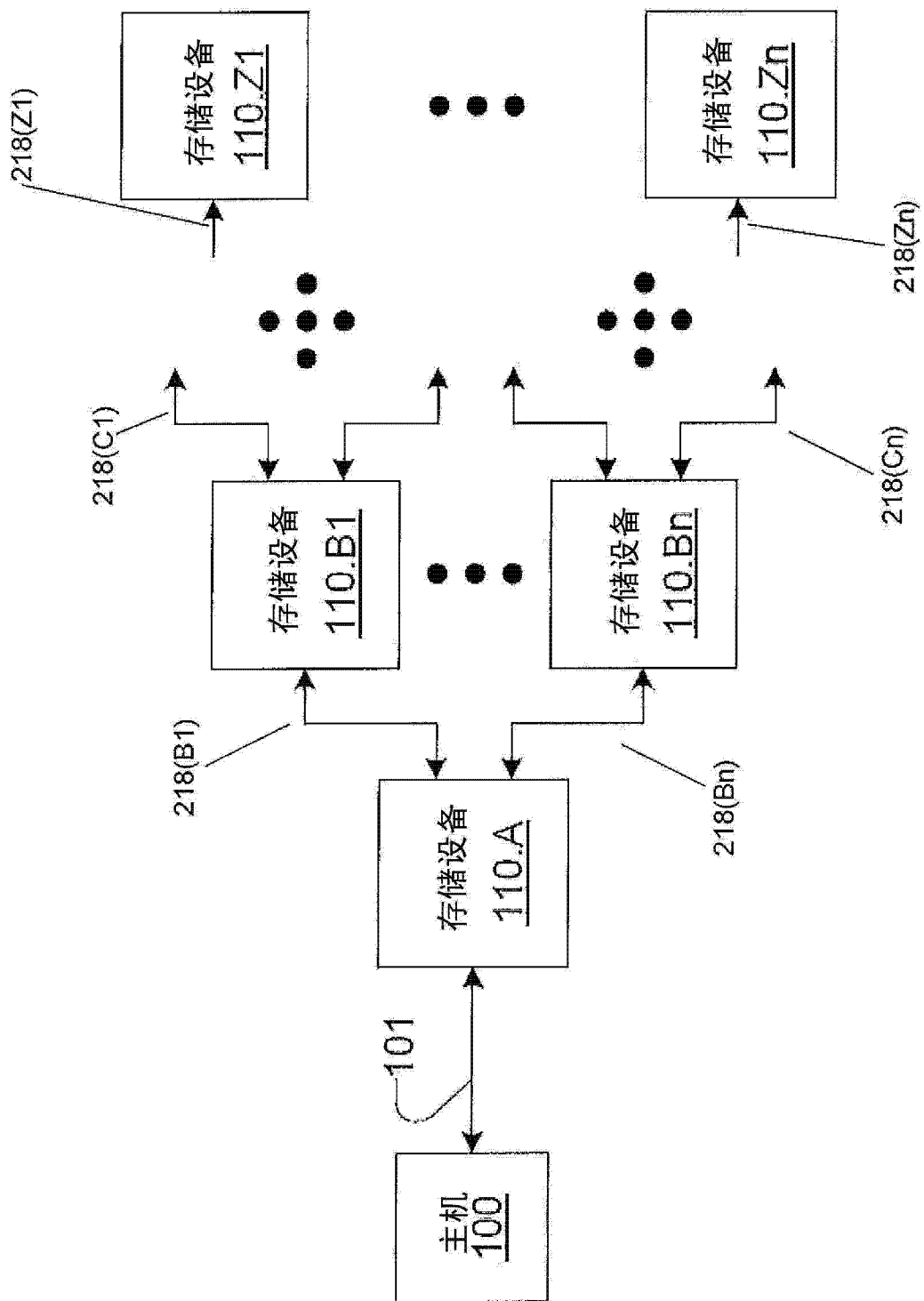


图 5