



(12) 发明专利申请

(10) 申请公布号 CN 103095796 A

(43) 申请公布日 2013. 05. 08

(21) 申请号 201210427832. 2

(22) 申请日 2012. 10. 31

(30) 优先权数据

13/289, 617 2011. 11. 04 US

(71) 申请人 LSI 公司

地址 美国加利福尼亚州米尔皮塔斯市

(72) 发明人 鲁伊兹·D·瓦其维特契科

詹森·A·昂瑞恩 里德·A·考夫曼

(74) 专利代理机构 北京纽乐康知识产权代理事

务所 11210

代理人 田磊

(51) Int. Cl.

H04L 29/08 (2006. 01)

G06F 3/06 (2006. 01)

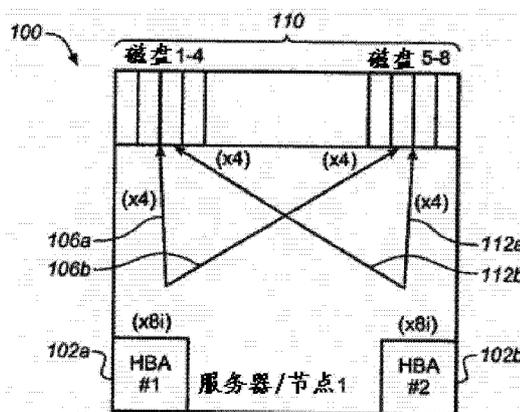
权利要求书3页 说明书5页 附图4页

(54) 发明名称

通过 SAS 扩展器共享的服务器直连存储

(57) 摘要

本发明提供了一种数据存储系统,包括第一服务器和第二服务器,所述第一服务器包括一被配置用以储存数据的第一存储磁盘组;以及一第一主机总线适配器,所述第一主机总线适配器包括一第一处理器;所述第二服务器包括一被配置用以储存数据的第二存储磁盘组,以及一第二主机总线适配器,所述第二主机总线适配器包括一第二处理器。其中,所述第一服务器的第一主机总线适配器通过一个串行连接的小型计算机系统接口(SAS)连接与所述第二服务器的第二主机总线适配器耦合,并且所述第一存储磁盘组和第二存储磁盘组的每个磁盘均通过第一服务器和第二服务器中的每一个访问。本发明能够减少延迟并增加系统/驱动器的可用性。



1. 一种数据存储系统,包括:

—第一服务器,包括:

—第一存储磁盘组,其被配置用以储存数据;以及

—第一主机总线适配器,其包括一第一处理器,该第一处理器被配置用以提供一第一虚拟扩展器和一第一逻辑组件;以及

—第二服务器,包括:

—第二存储磁盘组,其被配置用以储存数据;以及

—第二主机总线适配器,其包括一第二处理器,该第二处理器被配置用以提供一第二虚拟扩展器和一第二逻辑组件;

其中,所述第一服务器的所述第一主机总线适配器通过一串行连接的小型计算机系统接口(SAS)连接与所述第二服务器的所述第二主机总线适配器耦合,并且其中所述第一存储磁盘组和所述第二存储磁盘组中的每个磁盘均可通过所述第一服务器和所述第二服务器中的每一个访问。

2. 根据权利要求1所述的系统,其特征在于:所述SAS连接为一SAS电缆。

3. 根据权利要求1所述的系统,其特征在于:所述第一服务器的所述第一主机总线适配器的所述第一虚拟扩展器通过所述SAS连接与所述第二服务器的所述第二主机总线适配器的所述第二虚拟扩展器耦合。

4. 根据权利要求1所述的系统,其特征在于:所述所述第一服务器和所述第二服务器中的每一个还包括一个二级主机总线适配器。

5. 根据权利要求4所述的系统,其特征在于:所述第一服务器的所述二级主机总线适配器与所述第二服务器的所述二级主机总线适配器耦合。

6. 根据权利要求4所述的系统,进一步包括一总线,该总线与所述第一服务器的所述第一主机总线适配器和所述二级主机总线适配器中的每一个相连接。

7. 根据权利要求1所述的系统,进一步包括一第三服务器,所述第三服务器包括:

—第三存储磁盘组,其被配置用以储存数据;以及

—第三主机总线适配器,其包括一第三处理器,该第三处理器被配置用以提供一第三虚拟扩展器和一第三逻辑组件;

其中,所述第一服务器与所述第三服务器相连接,以及所述第二服务器与所述第三服务器相连接。

8. 根据权利要求7所述的系统,包括设在所述第一服务器与所述第二服务器、所述第一服务器与所述第三服务器或者所述第二服务器与所述第三服务器之间的一故障转移连接。

9. 根据权利要求1所述的系统,其特征在于:所述第一存储磁盘组和所述第二存储磁盘组均被配置为独立磁盘冗余阵列(RAID)结构。

10. 一种数据存储系统,包括:

—第一服务器,包括:

—第一存储磁盘组,其被配置用以储存数据;以及

—第一主机总线适配器,其包括一第一多核处理器,该第一多核处理器中的一个核被配置用以提供一第一虚拟扩展器;以及

一第二服务器,包括:

一第二存储磁盘组,其被配置用以储存数据;以及

一第二主机总线适配器,其包括一第二多核处理器,该第二多核处理器中的一个核被配置用以提供一第二虚拟扩展器;

其中,所述第一服务器的所述第一主机总线适配器通过一个串行连接的小型计算机系统接口(SAS)连接与所述第二服务器的所述第二主机总线适配器耦合,并且其中所述第一存储磁盘组和所述第二存储磁盘组的每个磁盘均可通过所述第一服务器和所述第二服务器中的每一个访问。

11. 根据权利要求10所述的系统,其特征在于:所述SAS连接为一SAS电缆。

12. 根据权利要求10所述的系统,其特征在于:所述第一服务器的所述第一主机总线适配器的所述第一虚拟扩展器通过所述SAS连接与所述第二服务器的所述第二主机总线适配器的所述第二虚拟扩展器耦合。

13. 根据权利要求10所述的系统,其特征在于:所述第一服务器和所述第二服务器中的每一个均包括一个二级主机总线适配器。

14. 根据权利要求13所述的系统,其特征在于:所述第一服务器的所述二级主机总线适配器与所述第二服务器的所述二级主机总线适配器耦合。

15. 根据权利要求13所述的系统,进一步包括一总线,该总线与所述第一服务器的所述第一主机总线适配器和所述二级主机总线适配器中的每一个相连接。

16. 根据权利要求10所述的系统,进一步包括一第三服务器,所述第三服务器包括:

一第三存储磁盘组,其被配置用以储存数据;以及

一第三主机总线适配器,其包括一第三多核处理器,该第三多核处理器中的一个核被配置用以提供一第三虚拟扩展器;

其中,所述第一服务器与所述第三服务器相连接,所述第二服务器与所述第三服务器相连接。

17. 根据权利要求16所述的系统,包括设在所述第一服务器与所述第二服务器、所述第一服务器与所述第三服务器或者所述第二服务器与所述第三服务器之间的一故障转移连接。

18. 根据权利要求10所述的系统,其特征在于:所述第一存储磁盘组和所述第二存储磁盘组均被配置为独立磁盘冗余阵列(RAID)结构。

19. 一种数据存储系统,包括

至少四个服务器,该至少四个服务器中的每一个均包括:

一存储磁盘组,其被配置用以储存数据;

一第一主机总线适配器,其包括一第一处理器,该第一处理器被配置用以提供一第一虚拟扩展器;以及

一第二主机总线适配器,其包括一第二处理器,该第二处理器被配置用以提供一第二虚拟扩展器,

其中,所述至少四个服务器中的每个均包括一第一连接配置,该第一连接配置连接该至少四个服务器中的一个服务器的所述第一虚拟扩展器与该至少四个服务器中的两个其它服务器的不同的第一虚拟扩展器,所述至少四个服务器中的每个服务器均包括一第二连

接配置,该第二连接配置连接该至少四个服务器中的一个服务器的所述第二虚拟扩展器与该至少四个服务器中的两个服务器的不同的第二虚拟扩展器,以及其中所述至少四个服务器中的至少一个服务器的所述第一连接配置不同于所述至少四个服务器中的所述至少一个服务器的所述第二连接配置,取决于与所述第一连接配置和所述第二连接配置相关联的那些服务器。

通过 SAS 扩展器共享的服务器直连存储

技术领域

[0001] 本发明涉及数据存储系统领域,特别是通过共享虚拟 SAS (串行连接 SCSI (小型计算机系统接口)) 扩展器的服务器直连存储。

背景技术

[0002] 云计算的兴起,提供了一个能够减少大量工作量的点播网络访问配置的计算资源共享池(如网络、服务器、存储、应用程序和服务),这个计算资源共享池的快速配置和轻度的管理需求使得其能够得到进一步的广泛应用。云计算综合利用这个计算资源共享池中的冗余,这些冗余可以体现为很多方式,其中四个即为下文中提及的,且每一个均包括一些存在问题的功能:

(1) 每个节点可以被连接到一个共同的 SAN (存储区域网络) 结构,此结构能够提供一个低延迟块的接口以用来存储;(2) 每个节点都可以连接到以太网,并且可以利用文件访问共享存储;(3) 外部的 JBODs (“磁盘簇”);以及(4) 直接连接的磁盘(内部)。

[0003] 配置(1)和(2)可能需要额外的外部组件,如光纤或以太网交换机,以连接节点与公用存储,用于形成一个集群(cluster)。这样的外部组件是完全没有必要的,因为其有可能导致单点故障。其结果是,当需要使用冗余成分的配置以提高可用性时,将产生纳入到系统内的额外成本。

[0004] 配置(3)从控制成本的角度来讲十分有效,但该配置将集群节点的数量限定在 JBOD 上的连接器的数量上,这导致了过分的限制并且限制了其可扩展性。此外,配置(1)-(3)一般要求存储系统设置在外壳中,这带来了额外的电力、空间和维护成本。

[0005] 配置(4)比较经济,然而由于没有共享存储的连接磁盘存在,导致没有规定高可用性集群。因此,这些配置往往都存在着各种各样的成本上的问题,并带来了额外的复杂性,不存在理想的解决方案,以满足高可用性集群的存储需求(如冗余和公共访问)。

发明内容

[0006] 本发明的一个实施方案中提供一种数据存储系统,包括第一服务器和第二服务器,所述第一服务器包括:一被配置用以储存数据的第一存储磁盘组;以及一第一主机总线适配器,所述第一主机总线适配器包括一第一处理器,该第一处理器被配置用以提供一第一虚拟扩展器和一第一逻辑组件;以及所述第二服务器包括一被配置用以储存数据的第二存储磁盘组,以及一第二主机总线适配器,所述第二主机总线适配器包括一第二处理器,该第二处理器被配置用以提供一第二虚拟扩展器和一第二逻辑组件,其中,所述第一服务器的第一主机总线适配器通过一 SAS 连接与所述第二服务器的第二主机总线适配器耦合,并且所述第一存储磁盘组和第二存储磁盘组的每个磁盘通过第一服务器和第二服务器中的每一个访问。

[0007] 在本发明的另一个实施方案中提供一种数据存储系统,包括第一服务器和第二服务器,所述第一服务器包括一被配置用以储存数据的第一存储磁盘组;以及一第一主机总

线适配器,所述第一主机总线适配器包括一第一多核处理器,该第一多核处理器中的一个核(core)被配置用以提供一第一虚拟扩展器;所述第二服务器包括一被配置用以储存数据的第二存储磁盘组,以及一第二主机总线适配器,所述第二主机总线适配器包括一第二多核处理器,该第二多核处理器中的一个核(core)被配置用以提供一第二虚拟扩展器,其中,所述第一服务器的第一主机总线适配器通过一 SAS 连接与所述第二服务器的第二主机总线适配器耦合,并且所述第一存储磁盘组和第二存储磁盘组的每个磁盘均通过第一服务器和第二服务器中的每一个访问。

[0008] 在本发明的另一种实施方案中还提供了一种数据存储系统,包括至少四个服务器,该至少四个服务器中的每一个均包括一被配置用以储存数据的存储磁盘组;一第一主机总线适配器,包括一第一处理器,该第一处理器被配置用以提供一第一虚拟扩展器;以及一个二级主机总线适配器,包括一第二处理器,该第二处理器被配置用以提供一第二虚拟扩展器,其中,所述至少四个服务器中的每个均包括一第一连接配置,该连接配置连接该至少四个服务器中的一个服务器中的第一虚拟扩展器与该至少四个服务器中的两个其它服务器中的不同的第一虚拟扩展器;所述至少四个服务器中的每个均包括一第二连接配置,该连接配置连接该至少四个服务器中的一个服务器中的第二虚拟扩展器与该至少四个服务器中的两个其它服务器中的不同的第二虚拟扩展器;根据与所述第一连接配置和第二连接配置相关联的哪一个服务器,所述至少四个服务器中的至少一个服务器中的第一连接配置不同于所述至少四个服务器中的至少一个服务器中的第二连接配置。

[0009] 应当理解,以上一般说明和以下具体说明都仅仅是示范性和解释性的且不是对本发明要求权利的限制。被并入且构成本说明书一部分的附图,描述了本发明的实施方案,并和一般说明一起用于解释原理。

[0010]

附图说明

[0011] 通过参考附图,本领域技术人员可更好地理解本发明的众多目标和优点,其中:

- 图 1 是一服务器内部配置示意图;
- 图 2 是一主机总线适配器的结构示意图;
- 图 3A 是一种级联的 DAS (直连存储) 集群的结构示意图;
- 图 3B 是另一种级联的 DAS 集群的结构示意图;
- 图 4 是图 3A 的一种级联的 DAS 集群的部分结构示意图;以及
- 图 5 是一种级联的 DAS 集群的实施方式结构示意图。

具体实施方式

[0012] 现在将对在附图中描述的公开主题进行具体说明。本发明的范围不仅仅限于权利要求;包含了众多替换,修改和等同体。为了说明清楚,尚未对与这些实施方案相关的技术领域已知技术材料进行具体说明以避免不必要地使本说明不清楚。

[0013] 本发明公开提供服务器执行以属于节点集群(如服务器),这些服务器共享存储,不使用外部组件如开关或外部存储。一般来说,利用 SAS 技术与直接连接的磁盘,是通过每个节点与各个节点之间的连接,从而通过级联 SAS 拓扑模拟 SAN 环境。现代计算服务器可

以包括通过 SAS 嵌入的多个磁盘,这使得一台服务器的内部存储可以被其它相连接的服务器共享。而当内部存储可共享时,外部存储则无需大量的数据访问。因此 SAS HBA(主机总线适配器)就有足够的能力使所有其它节点和相应的附加磁盘做到双向通信。

[0014] 图 1 是本发明的一个实施例的服务器 100 内部配置示意图,服务器 100 可以并入多个节点的集群中。服务器 100 可以包括一个或多个 HBA(如 SAS HBA),图 1 中描述了两个 HBA,102a 和 102b。图 2 是服务器 100 中的 HBA 102a 的结构示意图。如图所示,HBA 102a 包括一对四个外部连接器 104a 和 104b,和一对四个内部连接器 106a 和 106b,用于总共 16 个 phys。HBA102 还包括一个处理器,如用于管理操作 HBA 102a 的双核 CPU108。如图 1 所示,该对四个内部连接器 106a 和 106b 连接 HBA 102a 与服务器 110 上可用作存储的多个磁盘 110。同样的,HBA 102b 包括连接器 112a 和 112b,连接器 112a 和 112b 连接 HBA 102b 与服务器 110 上的多个磁盘 110。

[0015] HBA 102a 和 HBA 102b 上的外部连接器(如 104a 和 104b)用作将服务器 100 与作为集群的部分的其它服务器相连接。每台服务器至少包括一个 HBA 用来与集群中的其它服务器相连接,其中每个服务器中不止一个 HBA 允许冗余。例如,每个服务器/节点可以包括与其它两个节点相连接的 SAS 连接(通过每个服务器/节点的 HBA)以用于处理冗余。如图 3A 所示的一种级联的 DAS 的结构,该配置包括五个服务器/节点 100、200、300、400 和 500,其中服务器/节点 100 为第一节点,服务器/节点 500 为最后节点。服务器/节点 100 通过连接器 104a 和 104b 连接到服务器/节点 200。服务器/节点 200 通过连接器 204a 和 204b 连接到服务器/节点 300。服务器/节点 300 通过连接器 304a 和 304b 连接到服务器/节点 400。服务器/节点 400 通过连接器 404a 和 404b 连接到服务器/节点 500。第一节点和最后节点之间也可以相互连接,但该连接也可被禁用以防止循环(如一个无效的 SAS 拓扑)。如图 3A 所示,服务器/节点 100 通过连接器 504a 和 504b 连接到服务器/节点 500,此连接即处于禁用状态,直到集群中的一个节点不可用时连接打开。在一个节点或连接不再运行(如节点发生故障)的情况下,被禁用的第一节点和最后节点之间的连接(如连接器 504a 和 504b)可通过固件立即启用,以确保所有可用的节点可以被不间断的访问。系统的每个服务器/节点均可以包括可访问所有节点的本地 SAS(或 SATA(串行高级技术附件))存储,如多个磁盘 110。每一个节点可以包括到两个其它节点的冗余连接,即所有的终端设备均具有双路径可以用于冗余,然而,在本发明公开的所有实施例中冗余连接可以不是必须的。

[0016] 如图 3B 所示的另一种级联的 DAS 集群结构,所述结构包括两个不同的布线图案。例如,连接器 104a、204a、304a、404a 和 504a 之间的连接结构与图 3A 所描述的结构相同。而图 3B 中连接器 104b、204b、304b、404b 和 504b 之间的连接结构的与图 3A 中连接器 104b、204b、304b、404b 和 504b 之间的连接结构不同。由于连接结构中包括图 3B 中的布线图案的不同结构,此种布线图案可以降低延迟并增加系统/驱动器的可用性,效果胜于服务器/节点的每个 HBA 均被连接到相同的服务器/节点。当每个服务器/集群的节点均可操作时,图 3B 中的连接器 104b 和 504a 可以是禁用的故障转移连接,但当集群中的一个节点或连接不可操作(如节点故障)时被激活。当集群中的节点或连接不可操作时,固件可以立即激活连接器 104b 和 / 或 504a,以提供所述集群中所有可用的节点可以被不间断的访问。

[0017] 参考图 4,为图 3A 所示的一种级联的 DAS 的局部结构示意图。图 4 中,所示的每个服务器/节点的每个 HBA 可以包括两个主要组件:(1)PCI(外设组件互连标准)逻辑和

HBA 逻辑以提供 HBA 的运行和系统 100 上多个 HBA 之间的通信 ;以及(2)虚拟扩展器以处理驱动器与 HBA 逻辑组件之间以及 HBA 逻辑组件与外部 Phys 之间流量路由。例如, HBA 102a 服务器 / 节点 100 包括 PCI / HBA 逻辑组件 114a 和虚拟扩展器 116a, 而 HBA 102b 的服务器 / 节点 100 包括 PCI / HBA 逻辑组件 114b 和虚拟扩展器 116b。连接器 106a 和 106b 可以将多个磁盘 110 耦合到 HBA 102a 的虚拟扩展器 116a, 连接器 112a 和 112b 可以将多个驱动器 110 耦合到 HBA 102b 的虚拟扩展器 116b。类似的配置可能存在的其它服务器 / 节点的集群中得以实施, 例如, 服务器 / 节点的 HBA 202a 包括 PCI / HBA 的逻辑组件 214a 和虚拟扩展器 216a, 而服务器 / 节点 200 的 HBA 202B 包括的 PCI / HBA 逻辑组件 214b 和虚拟扩展器 216b, 并通过多个驱动器 210 与虚拟扩展器 216a 和 216b 之间的连接相连接。

[0018] 每个服务器 / 节点可以包括总线用以提供服务器 / 节点的组件之间的通信。例如, 服务器 / 节点 100 可以包括 PCI 总线 118, 其可以耦合 HBA 102a 和 102b 中的每一个, 而服务器 / 节点 200 可以包括 PCI 总线 218, 其耦合 HBA 202a 和 202b 中的每一个。而且, 每个服务器 / 节点均可以如图 3A 和 3B 所示连接到两个其它服务器 / 节点上。每个服务器 / 节点之间的连接可以为 SAS 连接器, 如 SAS 电缆 406, 其提供每个服务器 / 节点之间的外部耦合。如图 4 中所示, 服务器 / 节点 100 包括两个环状的外部 SAS 电缆 406, 其与集群中的最后一台装置(如端节点)相连接。一个或多个 SAS 电缆可能会被禁用, 以防止其在无效的 SAS 拓扑中充当故障转移电缆。

[0019] 图 5 是本发明的一个实施例的一种级联的 DAS。总的来说, 图 5 中所示的连接方式与图 4 中的服务器 / 节点的系统不同。如图所示, 服务器 / 节点 100 的 HBA 102a 通过连接器 502a 与服务器 / 节点 400 的 HBA 402a 耦合并通过连接器 504a 与服务器 / 节点 200 的 HBA 202a 耦合, 而 HBA 102b 通过连接器 502b 与服务器 / 节点 400 的 HBA 402b 耦合并通过连接器 504b 与服务器 / 节点 300 的 HBA 302b 耦合 ; 服务器 / 节点 200 的 HBA 202a 通过连接器 504a 与服务器 / 节点 100 的 HBA 102a 耦合并通过连接器 506a 与服务器 / 节点 300 的 HBA 302a 耦合, 而 HBA 202B 通过连接器 506b 与服务器 / 节点 400 的 HBA 402b 耦合并通过连接器 508b 与服务器 / 节点 300 的 HBA 302b 耦合 ; 以及服务器 / 节点 300 的 HBA 302a 通过连接器 506a 与服务器 / 节点 200 的 HBA 202a 耦合并通过连接器 508a 与服务器 / 节点 400 的 HBA 402a 耦合, 而 HBA 302b 通过连接器 508b 与服务器 / 节点 100 的 HBA 102b 相耦合并通过连接器 504b 与服务器 / 节点 100 的 HBA 102b 耦合。此种耦合方案相比于将可每个 HBA 连接到相同的服务器 / 节点的方式, 能够减少延迟并增加系统 / 驱动器的可用性。

[0020] 当每个服务器 / 集群的节点均可操作时, 连接器 502a 和 508b 可以是禁用的故障转移连接, 但当集群中的节点或连接不可操作时(如节点故障)被激活。当集群中的节点或连接不可操作时, 固件可以立即激活连接器 502a 和 508b, 以确保集群中所有可用的节点可以被不间断的访问。

[0021] 为了加快数据访问 / 处理的速度, 其中的 IO (输入 / 输出) 可以通过利用 HBA 上的多核处理器的有效路由算法处理, 如采用图 2 中的双核处理器 CPU 108。这样的用法可以减少 HBA 的虚拟扩展器(例如, 虚拟扩展器 116a)的延迟。例如, 当 HBA 包括双核处理器, 第二核(core)可以专用于虚拟扩展器。

[0022] 图 5 中所示的一种级联的 DAS 的实际结构图的直连式存储(DAS)可以被配置为 RAID (独立磁盘冗余阵列)。比如, 集群的服务器 / 节点中的多个驱动器 110, 210, 310, 410

可以被置于 RAID 配置中(如图 5 所示的那些),以提供增加可用性的集群,诸如通过减轻一个或多个驱动器故障、系统故障、BHA 故障或电缆故障。

[0023] 应当相信,通过前面的说明,将理解本发明和许多其伴随的优点,应当清楚,在其组件的形式,构造和设置中可做出各种改变而不背离本发明的范围和精神或不牺牲其所有实质优点。此处之前所述的形式仅仅是其的解释性的实施方案,打算的是,后续权利要求要包含和包含这些改变。

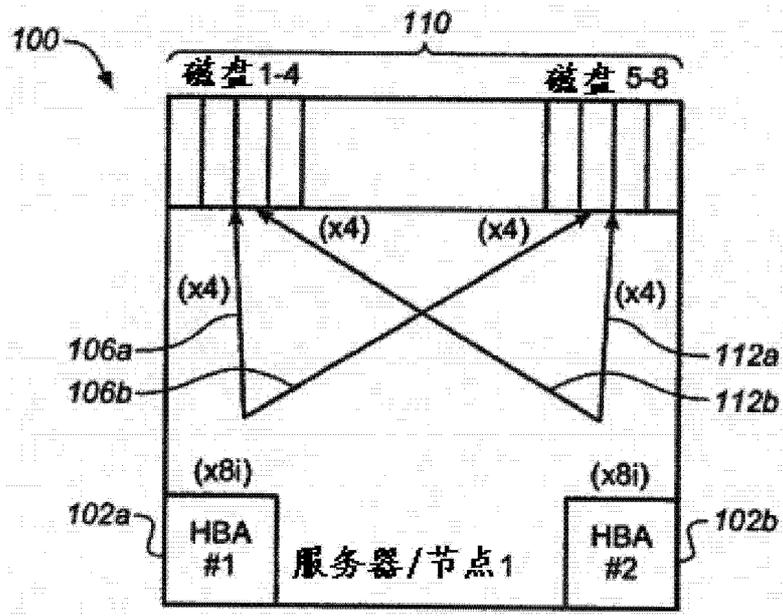


图 1

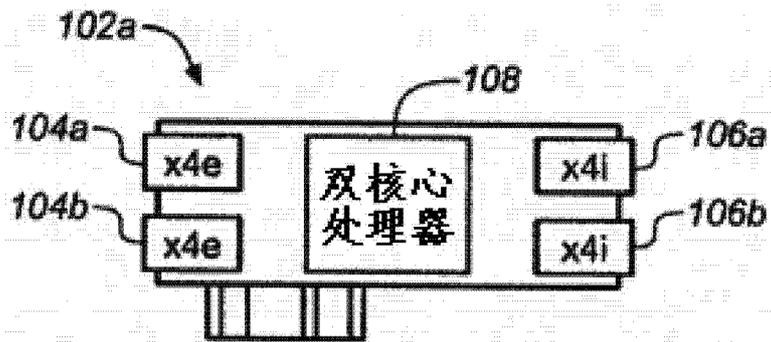


图 2

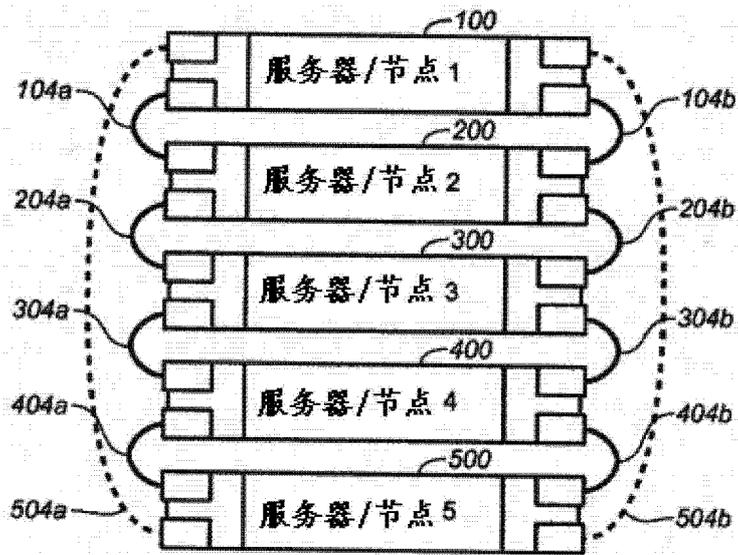


图 3A

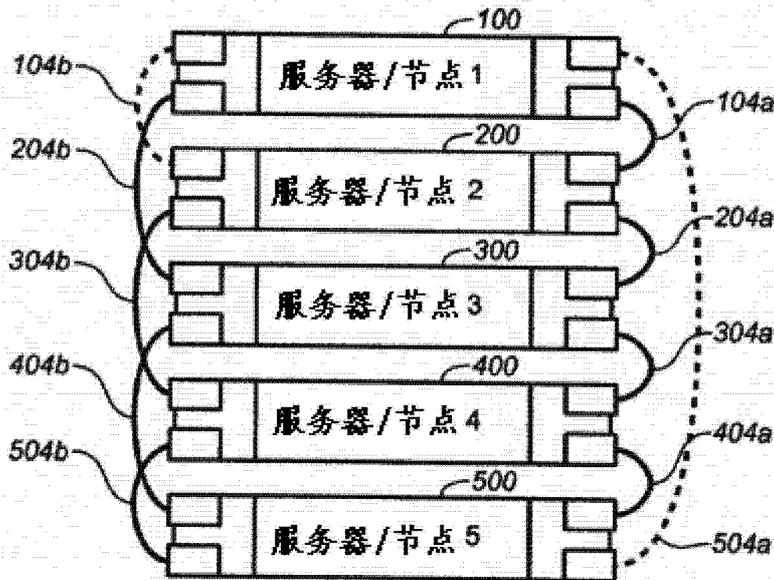


图 3B

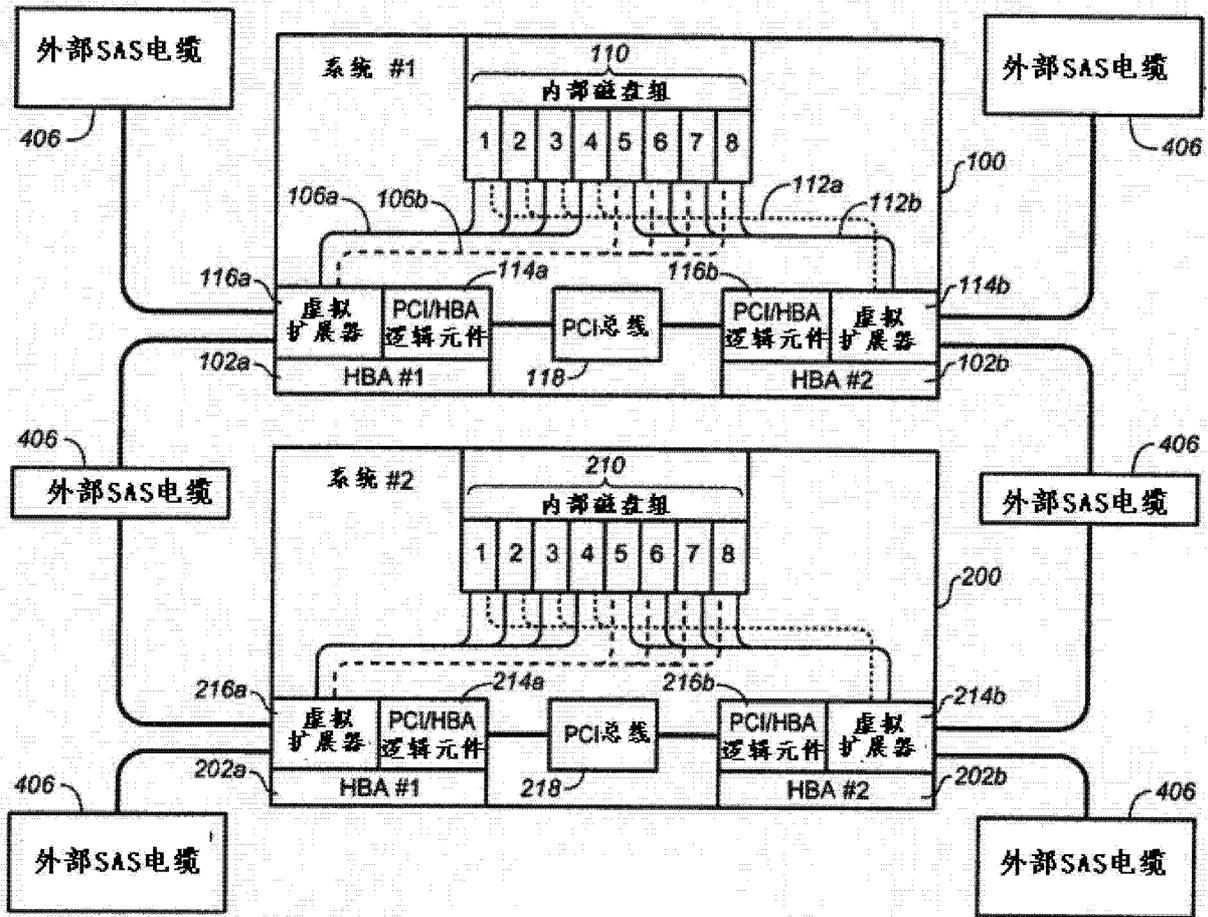


图 4

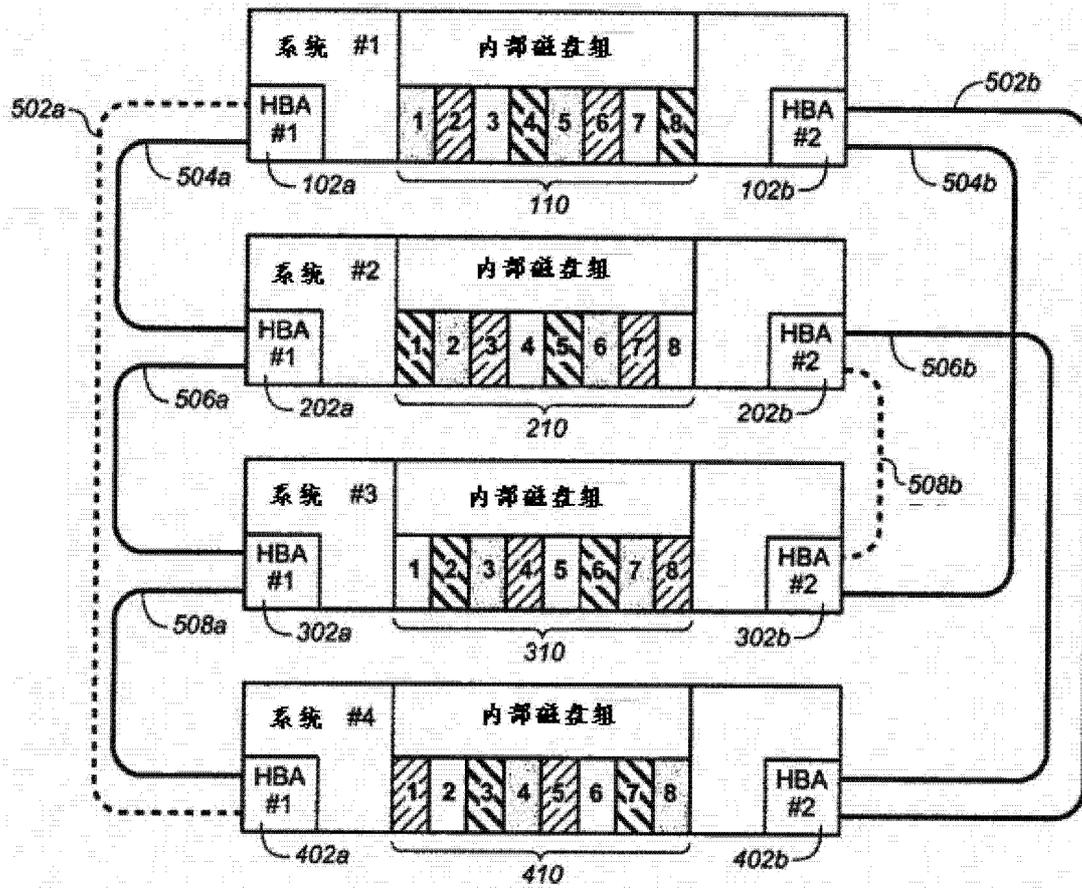


图 5