(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2021 年 1 月 21 日 (21.01.2021)

WIPO | PCT

(10) 国际公布号
WO 2021/008197 A1

(51) 国际专利分类号:
G06F 9/50 (2006.01)

(21) 国际申请号: PCT/CN2020/088784

(22) 国际申请日: 2020 年 5 月 6 日 (06.05.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910646566.4 2019年7月17日 (17.07.2019) CN
201910955603.X 2019年10月9日 (09.10.2019) CN

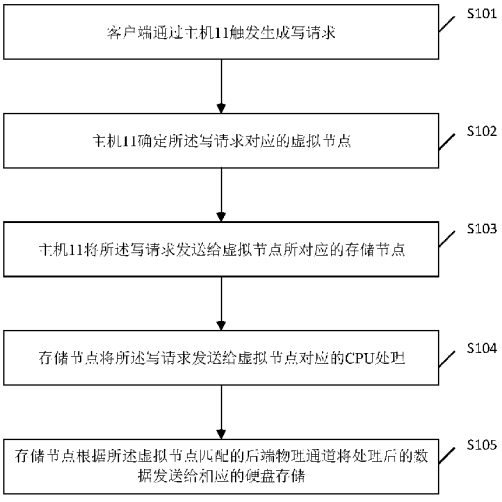
(71) 申请人: 华 为 技 术 有 限 公 司 (**HUAWEI TECHNOLOGIES CO., LTD.**) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(72) 发明人: 董 浩 (**DONG, Hao**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 肖庆航 (**XIAO, Qinghang**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 周 臣 (**ZHOU, Chen**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(81) 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

(54) **Title:** RESOURCE ALLOCATION METHOD, STORAGE DEVICE, AND STORAGE SYSTEM

(54) 发明名称: 资源分配方法、存储设备和存储系统



S101 Client generates write request by means of trigger of host 11
S102 Host 11 determines virtual node corresponding to write request
S103 Host 11 sends write request to storage node corresponding to virtual node
S104 Storage node sends write request to CPU corresponding to virtual node for processing
S105 According to back-end physical channel matched by virtual node, storage node sends processed data to corresponding hard disk for storage

图 7

(57) **Abstract:** Provided are a resource allocation method and a storage device. The storage device comprises a hard disk frame and a plurality of controllers; each controller comprises a plurality of processors; each processor comprises a plurality of processor cores; each of the plurality of controllers is coupled to the hard disk frame; the hard disk frame comprises a plurality of hard disks. The plurality of processors are used for providing computing resources. The plurality of hard drives are used for providing a storage space; the logical address corresponding to the storage space is divided into a plurality of address segment sets; each address segment set contains one or a plurality of address segments; each address segment set is allocated one part of the computing resources among said computing resources; said part of the computing resources is used for executing a data access request to access the address segments included in the address segment set, the computing resources used for processing different address segment sets coming from different processors, or, the computing resources used for processing different address segment sets coming from different processor cores. The invention can reduce the bandwidth resources between processors.

WO 2021/008197 A1

[见续页]

ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要：一种资源分配方法和存储设备。存储设备包括硬盘框和多个控制器，每个控制器包括多个处理器，每个处理器包括多个处理器核，所述多个控制器分别和所述硬盘框耦合，所述硬盘框包括多个硬盘。所述多个处理器，用于提供计算资源。所述多个硬盘，用于提供存储空间，所述存储空间所对应的逻辑地址被划分为若干个地址段集合，每个地址段集合包含一个或多个地址段，每个地址段集合分配有所述计算资源中的一部分计算资源，所述一部分计算资源用于执行访问所述地址段集合所包含的地址段的数据访问请求，其中，用于处理不同地址段集合的计算资源来自不同的处理器，或者，用于处理不同地址段集合的计算资源来自不同的处理器核。可以减少处理器之间的带宽资源。

资源分配方法、存储设备和存储系统

技术领域

本申请涉及存储领域，并且更具体地，涉及一种资源分配方法、存储设备和存储系统。

背景技术

当前的存储节点通常具有多个控制器，而每个控制器又包括多个处理器。当存储节点执行一个业务请求时，往往需要多个处理器并行处理由这个业务请求拆分而来的多个子请求，由于这些子请求之间具有关联性，因此会涉及到子请求在多个处理器之间转发和数据交互，占用了处理器之间的带宽资源。

发明内容

本申请主要是解决如何节省处理器之间的带宽资源的问题。

第一方面提供了一种存储设备，包括硬盘框和多个控制器，每个控制器包括多个处理器，每个处理器包括多个处理器核，所述多个控制器分别和所述硬盘框耦合，所述硬盘框包括多个硬盘。所述多个处理器，用于提供计算资源。所述多个硬盘，用于提供存储空间，所述存储空间所对应的逻辑地址被划分为若干个地址段集合，每个地址段集合包含一个或多个地址段，每个地址段集合分配有所述计算资源中的一部分计算资源，所述一部分计算资源用于执行访问所述地址段集合所包含的地址段的数据访问请求，其中，用于处理不同地址段集合的计算资源来自不同的处理器，或者，用于处理不同地址段集合的计算资源来自不同的处理器核。

关于第一方面中的地址段，由于所述存储空间被虚拟化为逻辑单元 LUN，所述 LUN 由所述存储设备提供给用户使用，所述 LUN 所对应的地址被划分而成的若干个逻辑块地址 LBA 区间，所以每个地址段为一个 LBA 区间。

第一方面实际上提供了两种具体实现，一种是用于处理不同地址段集合的计算资源来自不同的处理器，由于用于访问某个地址段的数据访问请求都由分配的处理器执行，不同地址段集合分配不同的 CPU，可以避免在多个处理器之间转发数据访问请求，节省处理器之间的带宽资源。另一种是用于处理不同地址段集合的计算资源来自不同的处理器核。由于用于访问某个地址段的数据访问请求都由分配的处理器核执行，相对于前一种实现方式，从更细的粒度隔离了不同地址段集合的数据访问请求，数据访问请求在自己分配的处理器核上串行执行，减少了互斥操作，并且可以实现免锁设计。

在一种可选的实现方式中，每个处理器具有内存，所述计算资源还包括所述多个处理器包含的多个内存，其中，一个内存分配给一个地址段集合，且不同地址段集合分配的内存不同。

在一种可选的实现方式中，每个地址段集合分配的内存是该地址段集合分配的处理器的本地内存。在实际应用场景中，处理器具有自己的内存，也称为处理器的本地内存，本地内存往往和处理器集成在一个组件中，或者与处理器直接或间接耦合。那么在分配内存时可以优先将该地址段集合所分配的处理器的本地内存分配给该地址段集合，就近使用处理器的本地内存可以避免跨 CPU 传输数据。

在一种可选的实现方式中，所述存储设备还包括网络资源，所述网络资源由所述多

个控制器与所述硬盘框之间的多个连接提供，每个地址段集合分配有所述网络资源中的一部分网络资源，其中，不同地址段集合分配的连接不同。

在一种可选的实现方式中，所述存储空间包括多个块组，所述多个块组提供存储资源，每个地址段集合分配有所述存储资源中的一部分存储资源，其中，不同地址段集合分配的块组不同。

在一种可选的实现方式中，所述存储设备还包括前端接口卡，所述多个控制器分别与所述前端接口卡耦合，所述前端接口卡中保存映射表，所述映射表用于指示每个地址段集合与所分配的计算资源之间的映射关系。

可选的，所述映射表还用于指示每个地址段集合与所分配的网络资源之间的映射关系。

可选的，所述映射表还用于指示每个地址段集合与所分配的硬盘资源之间的映射关系。

在一种可选的实现方式中，所述存储设备中创建有若干个虚拟节点，一个虚拟节点通过预设算法映射至一个地址段集合，所述映射表记录有每个虚拟节点与该虚拟节点所映射的地址段集合所分配的计算资源之间的映射关系。

可选的，所述映射表还记录有每个虚拟节点与该虚拟节点所映射的地址段集合所分配的网络资源之间的映射关系。

可选的，所述映射表还记录有每个虚拟节点与该虚拟节点所映射的地址段集合所分配的硬盘资源之间的映射关系。

在一种可选的实现方式中，所述存储设备还包括新加入的控制器，所述新加入的控制器包括处理器，所述新加入的控制器与所述硬盘框耦合，所述新加入的控制器，用于将所述新加入的控制器所包含的处理器作为计算资源，分配给一个地址段集合中的第一地址段，并解除所述第一地址段与所述地址段集合所分配的计算资源之间的分配关系。当存储系统增加新的资源时，这样的实现可以减少地址段与原有资源之间的映射关系的改变。

本申请第二方面提供了一种资源分配方法，该方法应用于第一方面任意一种实现所提供的存储设备中，用于实现所述存储设备的功能。

本申请第三方面提供了一种计算机程序产品，当其在存储设备上运行时，使得所述存储设备执行第二方面所提供的资源分配方法。

本申请第四方面提供了一种计算机可读存储介质，所述计算机可读存储介质存储有指令，当其在存储设备上运行时，使得所述存储设备执行第二方面所提供的资源分配方法。

本申请第五方面提供了一种存储节点，所述存储节点包括多个控制器。其中，所述存储节点中创建有多个虚拟节点，每个虚拟节点预先分配有计算资源，所述计算资源来自计算资源池；所述存储节点接收的若干个数据访问请求被路由到所述多个虚拟节点，其中，每个数据访问请求对应一个虚拟节点，并且每个数据访问请求利用对应的虚拟节点所分配的计算资源执行。所述多个控制器，用于提供计算资源给所述计算资源池，所述计算资源由所述多个控制器所包含的多个处理器提供，其中，一个处理器分配给一个虚拟节点，且每个虚拟节点分配的处理器不同。或者，一个处理器分配给多个虚拟节点，每个虚拟节点分配的处理器核不同。

在第五方面的一种可选的实现方式中，每个处理器具有内存，所述计算资源池还包括所述多个处理器包含的多个内存，其中，一个内存分配给一个虚拟节点，每个虚拟节点分配的内存不同。

在第五方面的一种可选的实现方式中，每个虚拟节点分配的内存是该虚拟节点分配的处理器的本地内存。

在第五方面的一种可选的实现方式中，所述多个控制器分别与硬盘框耦合，所述硬盘框包括多个硬盘，每个虚拟节点还预先分配有网络资源，所述网络资源来自网络资源池，所述网络资源池由所述多个控制器与所述硬盘框之间的连接提供，每个虚拟节点对应一条或多条连接，每个虚拟节点分配的连接不同。

在第五方面的一种可选的实现方式中，每个虚拟节点还预先分配有硬盘资源，所述硬盘资源来自硬盘资源池，所述硬盘资源池由所述多个硬盘提供，每个虚拟节点对应一个或多个块组，每个虚拟节点分配的块组不同。

在第五方面的一种可选的实现方式中，所述存储节点还包括前端接口卡，所述前端接口卡与所述多个控制器耦合，所述前端接口卡中保存有每个虚拟节点与分配的计算资源的对应关系，和/或每个虚拟节点与分配的网络资源的对应关系，和/或每个虚拟节点与分配的硬盘资源之间的对应关系。

在第五方面的一种可选的实现方式中，所述前端接口卡还用于：接收主机发送的第一数据访问请求，所述数据访问请求包括虚拟地址；根据所述虚拟地址确定所述第一数据访问请求对应的第一虚拟节点；根据所述保存的对应关系确定所述第一虚拟节点分配的第一处理器；将所述第一数据访问请求转发给所述第一处理器；所述第一处理器，用于对所述第一数据访问请求进行处理。

在第五方面的一种可选的实现方式中，所述对所述第一数据访问请求进行处理时所涉及的数据或元数据临时保存在所述第一虚拟节点对应的第一内存中。

在第五方面的一种可选的实现方式中，所述第一处理器还用于，通过所述第一虚拟节点分配的第一连接，将所述第一数据访问请求要求写入的数据发送给所述硬盘框使得所述数据被写入分配给所述第一虚拟节点的第一块组中，或者通过所述第一连接将所述第一数据访问请求要求读取的数据写入所述第一内存中。

本申请第六方面提供另一种资源管理方法，所述方法应用于存储节点中，所述存储节点包括多个控制器，其中，每个控制器包括多个处理器。该方法包括创建多个虚拟节点；为每个虚拟节点分配计算资源，所述存储节点接收的若干个数据访问请求被路由到所述多个虚拟节点，其中，每个数据访问请求对应一个虚拟节点，并且每个数据访问请求利用对应的虚拟节点所分配的计算资源执行；其中，所述多个处理器中的每个处理器分配给所述多个虚拟节点中的一个虚拟节点，每个节点分配的处理器不同。

在第六方面一种可选的实现方式中，所述计算资源来自计算资源池，其中，所述计算资源池中的计算资源由所述多个控制器提供，所述计算资源池包括所述多个处理器。

在第六方面一种可选的实现方式中，所述创建多个虚拟节点包括：根据所述存储节点包含的处理器的数量创建所述多个虚拟节点，所述多个虚拟节点的数量小于或等于所述多个存储节点包含的处理器的数量。

本申请第七方面提供了一种计算机程序产品，当其在存储节点的控制器上运行时，使得所述控制器执行第六方面所提供的资源分配方法。

本申请第八方面提供了一种计算机可读存储介质，所述计算机可读存储介质存储有指令，当其在存储节点的控制器上运行时，使得所述控制器执行第七方面所提供的资源分配方法。

本申请第九方面提供了一种存储系统，包括主机和多个存储节点，所述多个存储节点提供的存储空间被划分为若干个地址段集合，每个地址段集合包括一个或多个地址段。所述主机用于生成数据访问请求，所述数据访问请求包括虚拟地址；根据所述虚拟地址确定所述目标数据对应的地址段集合，所述地址段集合指示了处理所述数据访问请求的目标存储节点，所述目标存储节点是所述多个存储节点的其中一个；将所述数据访问请求发送给所述目标存储节点；所述目标存储节点利用预先为所述目标数据对应的地址段集合分配的资源对所述数据访问请求进行处理。其中，所述目标存储节点对所述数据访问请求的处理与第一方面类似。

本申请第十方面提供了一种数据处理方法，该方法可以应用在第一方面提供的存储设备中，也可以应用在第五方面提供的存储节点中，也可以应用在存储系统中，该方法包括：接收数据访问请求，该请求中携带虚拟地址。根据虚拟地址可以唯一确定一个地址段集合，或者唯一确定一个虚拟节点，所述地址段集合（或者虚拟节点）分配有计算资源，利用分配的计算资源处理所述数据访问请求。

在第十方面的一种可选的实现方式中，所述地址段集合（或者虚拟节点）还分配有网络资源和硬盘资源，所述方法还包括：通过分配的网络资源从所述硬盘资源中读取数据，或者将待写入的数据存储在所述硬盘资源中。

按照本申请提供的存储系统，访问不同地址段集合的数据访问请求分配不同的处理器，可以在一定程度上避免在处理器之间转发请求或数据，节省了处理器之间的带宽资源。或者，访问不同地址段集合的数据访问请求分配不同的处理器核，可以进一步地避免在处理器核之间转发请求或数据。另外，访问不同地址段集合的数据访问请求所分配的资源相互独立，可以减少互斥操作，实现免锁设计。

本申请第十一方面提供了一种存储节点扩容方法，所述存储节点包括第一控制器，所述第一控制器包括多个处理器，当所述存储节点中增加第二控制器时，创建第二虚拟节点集合，所述第二控制器包括多个处理器，所述第二虚拟节点集合包括多个第二虚拟节点，其中，在创建所述第二虚拟节点集合之前，所述存储节点中有第一虚拟节点集合，所述第一虚拟节点集合包括多个第一虚拟节点，所述存储节点接收的若干个数据访问请求被路由到所述多个第一虚拟节点。将所述第二控制器包括的所述多个处理器添加至计算资源池，所述计算资源池包括所述第一控制器提供的多个处理器。为所述多个第二虚拟节点分配计算资源，所述分配的计算资源来自所述计算资源池，每个第二虚拟节点分配一个处理器，每个第二虚拟节点分配的处理器不同；将路由到所述第一虚拟节点集合的多个数据访问请求中的一部分数据访问请求重新分配给所述第二虚拟节点集合。

本申请第十二方面提供了一种存储节点，用于实现第十一方面提供的方法。

按照本申请第十一方面提供的存储节点扩容方法和第十二方面提供的存储节点，由于创建了若干个虚拟节点，每个虚拟节点分配有资源，那么当存储节点中增加计算资源增加时，可以相应地增加虚拟节点的数量，将增加的计算资源分配给新增的虚拟节点，并且路由到原有的虚拟节点的一部分数据访问请求可以重新分配给新增的虚拟节点，那么整个存储节点的处理能力也增加了。因此，从而实现了在硬件资源增加时，存储节点

的处理能力也随着硬件资源的增加线性增长。

附图说明

图 1 是可应用本发明实施例的技术方案的场景的示意图。

图 2 是本发明实施例提供的存储节点的结构示意图。

图 3 是本发明实施例提供的虚拟节点与处理器资源、内存资源的关系示意图。

图 4 是本发明实施例提供的虚拟节点与网络资源的关系示意图。

图 5 是本发明实施例提供的虚拟节点与硬盘资源的关系示意图。

图 6 是本发明实施例提供的虚拟节点所包含的资源的示意图。

图 7 是本发明实施例提供的处理写请求的流程示意图。

图 8 是本发明实施例提供的 CPU 核分组调度的示意图。

图 9 是本发明实施例提供的一种扩容示意图。

图 10 是本发明实施例提供的一种扩容方法的流程示意图。

图 11 是本发明实施例提供的另一种扩容示意图。

图 12 是本发明实施例提供的另一种扩容方法的流程示意图。

具体实施方式

本申请主要解决如何节省处理器之间的带宽资源的问题。下面将结合附图，对本申请中的技术方案进行描述。

图 1 是可应用本申请实施例的技术方案的场景的示意图。如图 1 所示，主机 11 和存储系统通信，存储系统包括多个存储节点（或简称“节点”）100。每个存储节点 100 是一个存储引擎（简称：引擎），其中，每个节点 100 包括多个控制器 103。每个控制器 103 包括多个处理器，每个处理器包括多个处理器核。另外，每个节点 100 具有前端接口卡 101 和后端接口卡 102。前端接口卡 101 用于节点 100 与主机 11 通信，后端接口卡 102 用于节点 100 和多个硬盘框通信，其中每个硬盘框包括多个硬盘 107，硬盘用于存储数据，可以是磁盘或者其他类型的存储介质，例如固态硬盘或者叠瓦式磁记录硬盘等。前端接口卡 101 通过内部网络通道和每个节点所包含的多个控制器 103 直连，后端接口卡 102 也通过内部网络通道和每个节点所包含的多个控制器 103 直连（图 1 中只示出了部分连接情况），确保每个节点 100 内的控制器 103 可以通过前端接口卡 101 或后端接口卡 102 进行业务的收发。另外，每个节点 100 都可以通过后端接口卡 102 连接硬盘框，以实现在节点间共享数据。在某些应用场景中，可以将一个或多个存储节点 100 与硬盘框合称为存储设备。

控制器 103 是一种计算设备，如服务器、台式计算机等。在硬件上，如图 2 所示，控制器 103 至少包括处理器 104、内存 105 和总线 106。其中，处理器 104 是一个中央处理器（英文：central processing unit，CPU），用于处理来自节点 100 外部的 IO 请求，或者控制器 103 内部生成的请求。在一个控制器 103 中，可能包含多个处理器 104，其中每个处理器包括多个处理器核（图中未示出）。内存 105 用于临时存储从主机 11 接收的数据或硬盘读取的数据。控制器 103 接收主机 11 发送的多个写请求时，可以将所述多个写请求中的数据暂时保存在内存 105 中。当内存 105 的容量达到一定阈值时，将内存 105

中存储的数据发送给硬盘存储。内存 105 包括易失性存储器或非易失性存储器或其组合。易失性存储器例如随机访问存储器(英文：random-access memory，RAM)。非易失性存储器例如闪存芯片、软盘、硬盘、固态硬盘（solid state disk，SSD）、光盘等各种可以存储程序代码的机器可读介质。内存 105 可具有保电功能，保电功能是指系统发生掉电又重新上电时，内存 105 中存储的数据也不会丢失。总线 106 用于控制器 103 内部各组件之间的通信。

控制器 103 为主机 11 提供的存储空间来源于多个硬盘 107，然而硬盘所提供的存储空间的实际地址并不直接给暴露给控制器 103 或主机 11。在实际应用中，物理存储空间被虚拟化成若干个逻辑单元（logical unit，LU）提供给主机 11 使用，每个逻辑单元具有唯一的逻辑单元号（logical unit number，LUN）。由于主机 11 能直接感知到逻辑单元号，本领域技术人员通常直接用 LUN 代指逻辑单元。每个 LUN 具有 LUN ID，用于标识所述 LUN。数据位于一个 LUN 内的具体位置可以由起始地址和该数据的长度（length）确定。对于起始地址，本领域技术人员通常称作逻辑块地址（logical block address，LBA）。可以理解的是，LUN ID、LBA 和 length 这三个因素标识了一个确定的地址段。主机 11 在生成数据访问请求，通常在该请求中携带 LUN ID、LBA 和 length，为了方便描述，在本实施例中，将 LUN ID、LBA 和 length 称为虚拟地址。由上面的描述可知，根据所述虚拟地址可以确定该请求所要访问的 LUN 以及该 LUN 的具体位置。控制器 103 中保存有所述虚拟地址与数据存储在硬盘中的地址之间的对应关系，因此控制器 103 在接收所述数据访问请求之后，可以根据所述对应关系，确定对应的物理地址，并指示硬盘读取或写入数据。

为了保证数据均匀存储在各个存储节点 100 中，在选择存储节点时通常采用分布式哈希表(Distributed Hash Table，DHT)方式进行路由，按照分布式哈希表方式，将哈希环均匀地划分为若干部分，每个部分称为一个分区，一个分区对应一个上面描述的地址段。主机 11 向存储系统发送的数据访问请求，都会被定位到一个地址段上，例如从所述地址段上读取数据，或者往所述地址段上写入数据。应理解，在处理这些数据访问请求时都需要利用存储系统中的 CPU 资源、内存资源以及其他资源（其中，业界往往将 CPU 资源和内存资源合并为计算资源）。CPU 资源和内存资源是由控制器 103 提供的，存储节点通常具有多个控制器 103，而每个控制器 103 又包括多个处理器。当存储节点执行一个业务请求时，往往需要多个处理器并行处理由这个业务请求拆分而来的多个子请求，由于这些子请求之间具有关联性，因此会涉及到子请求在多个处理器之间转发和数据交互，占用了处理器之间的带宽资源。

为了解决这个问题，本申请实施例为一个地址段集合分配一个 CPU，或者分配一个 CPU 内的一个或多个 CPU 核，所述地址段集合包含一个或多个地址段，所述地址段之间可以是连续的，也可以是不连续的。只要是用于访问这些地址段的数据访问请求都由所述分配的 CPU 执行，或者由所述分配的一个或多个 CPU 核执行。不同地址段集合分配不同的 CPU，或者分配同一个 CPU 中的不同 CPU 核。

进一步地，本申请实施例中，为每个地址段集合分配一个内存，用于访问该地址段集合所包含的地址段的数据访问请求所涉及的数据（既包括业务数据也包括元数据）都临时保存在所述分配的内存中。具体的，一个内存分配给一个地址段集合，且不同地址段集合分配的内存不同。这里的内存包括但不限于图 2 中的内存 105。在实际应用场景中，

处理器具有自己的内存，也称为处理器的本地内存，本地内存往往和处理器集成在一个组件中，或者与处理器直接或间接耦合。那么在分配内存时可以优先将该地址段集合所分配的处理器的本地内存分配给该地址段集合。需要说明的是，上述内存分配方式只是本申请实施例提供的一种实现方式，也可以采用其他实现方式，例如不预先将内存分配给某个地址段集合，任意一个数据访问请求在需要使用内存时再从所述存储系统所包含多个内存中选取。

用于处理数据访问请求的资源除了 CPU 资源、内存资源之外，还可以包括网络资源和硬盘资源。可选的，网络资源和硬盘资源均可以预先分配给不同的地址段集合。

当存储系统增加新的资源时，可以将新的资源和原有的资源整合之后，重新分配给各个地址段集合，一种实现方式是重新划分地址段集合，保持地址段的数量不变，增加地址段集合的数量，减少每个地址段集合包含的地址段的数量，然后重新将存储系统所拥有的资源分配给调整后的各个地址段集合，另一种实现方式是，保持每个地址段中部分地址段与原有资源的分配关系，将新增的资源分配该地址段集合中的另一部分地址段。这样的实现可以减少地址段与原有资源之间的映射关系的改变。

为了更好地实现针对不同地址段集合的数据访问请求之间的资源隔离，本申请中在存储系统中创建若干个虚拟节点（virtual node），虚拟节点是资源分配的最小单元。存储系统中的资源可以划分为若干个等份，其中每个等份对应一个虚拟节点。具体的，每个虚拟单元均对应有一部分 CPU 资源、一部分内存资源、一部分网络资源和一部分硬盘资源。例如，存储系统有 4 个节点 100，每个节点内有 4 个控制器 103，每个控制器内有 4 个 CPU，每个 CPU 有 48 个 CPU 核，那么一个节点 100 共有 768 个 CPU 核。一个存储系统如果包含 4 个节点，则总核数高达 3072 个。如果每个 CPU 对应 256GB 内存，那么一个控制器就有 1TB 内存，一个节点有 4TB 内存，存储系统总共有 16TB 内存。如果将存储系统包含的所有的硬件资源划分为 256 等份，那么就有 256 个虚拟节点，每个虚拟节点对应的 CPU 资源为 12 个 CPU 核，且每个虚拟节点对应的内存资源是 0.0625TB。如前面的描述，一个分区对应一个地址段，引入虚拟节点之后一个分区集合对应一个虚拟节点，其中，一个分区集合包含多个分区。那么，相应地，一个地址段集合对应一个虚拟节点，其中，一个地址段集合包括多个地址段。换而言之，将一个地址段作为输入，经过预定的算法计算，可以唯一确定到一个分区，并且进一步地唯一确定到一个虚拟节点。假设存储系统中有 1024 个分区，并且创建了 32 个虚拟节点，那么，每个虚拟节点对应 32 个分区集合，其中，每个分区集合包含 32 个分区。通常情况下，存储系统所包含的分区的数量是不变的。即使存储系统增加或减少了虚拟节点，也只是将 1024 个分区在增加或减少后的虚拟节点中重新分配而已。

应理解，创建虚拟节点并非实现资源隔离的唯一方式，如果没有虚拟节点，也可以按照前面的描述直接将资源分配给各个地址段集合。

关于存储系统中虚拟节点的创建，本实施例提供了至少两种创建方式。

一种是存储系统在初始化过程中自动创建，具体的过程如下：

可以根据系统所包括的（1）存储节点的数量（2）控制器的数量（3）CPU 的数量中的任意一种及其组合创建虚拟节点，创建的虚拟节点的数量小于或等于系统所包含的 CPU 的数量。然后，为各个虚拟节点分配资源并创建各个虚拟节点与分配的资源之间的映射关系（该部分内容可参考以下图 3 到图 6 的描述），将创建的映射关系保存在主机

11 和前端接口卡 101 中。

另一种方式是，存储系统在初始化时，存储系统的管理软件为管理员提供界面，管理员通过所述界面选择创建的虚拟节点的数量，存储系统再根据指示创建若干个虚拟节点，为各个虚拟节点分配资源并创建各个虚拟节点与分配的资源之间的映射关系（该部分内容可参考以下图 3 到图 6 的描述），将创建的映射关系保存在主机 11 和前端接口卡 101 中。同样的，管理员在选择虚拟节点的数量时可以根据（1）存储节点的数量（2）控制器的数量（3）CPU 的数量中的任意一种及其组合，或者其他因素来创建。

在以上任意一种创建方式中，都可以存储系统的运行过程中调整虚拟节点的数量，例如在存储系统增加控制器时可以增加虚拟节点的数量，或者在存储系统减少控制器时可以减少虚拟节点的数量，或者在存储系统增加硬盘框时可以增加虚拟节点的数量，或者在存储系统减少硬盘框时可以减少虚拟节点的数量等等。即使没有资源数量的变化，存储系统也可以根据管理员的指定调整虚拟节点的数量。

请参考图 3，图 3 是本申请实施例提供的虚拟节点与 CPU 资源、内存资源的分配关系示意图。如图 3 所示，存储系统中所有的 CPU 和所有的内存组成一个计算资源池，将所述计算资源池中包含的 CPU 资源和内存资源划分为若干个计算资源组，每个计算资源组分配给一个虚拟节点。不同的虚拟节点占用不同的计算资源组。可以是每个计算资源组使用一个 CPU，也可以多个计算资源组共用一个 CPU。例如，计算资源组 0 使用 CPU_0，计算资源组 1 使用 CPU_1，计算资源组 m 和计算资源组 m+1 共用 CPU_m，计算资源组 n 使用 CPU_n，m 和 n 都是大于 1 的整数，且 n 大于 m，可以理解的是计算资源组 n 和计算资源组 m 之间还具有一个或多个计算资源组。对于多个计算资源组共用一个 CPU 的情况，由于一个 CPU 包含若干个 CPU 核（例如 48 个 CPU 核），可以将所述一个 CPU 所包含的若干个 CPU 核划分为多个核组，每个核组（包括一个或多个 CPU 核）分配给一个虚拟节点。另外，每个计算资源组还包括内存资源，每个计算资源组所包含的内存资源可以是该计算资源组所包含的 CPU 本地的内存。这样配置以后，某个虚拟节点对应的数据访问请求运行在分配好的 CPU 上，那么该虚拟节点可以就近使用该 CPU 本地的内存资源。所谓 CPU 本地的内存是指与该 CPU 位于同一个节点内的内存。具体的，举例来说，计算资源组 0 使用的内存资源是 Mem_0，Mem_0 是 CPU_0 本地的内存；计算资源组 1 使用的内存资源是 Mem_1，Mem_1 是 CPU_1 本地的内存；计算资源组 m 和计算资源组 m+1 共用 Mem_m，Mem_m 是 CPU_m 本地的内存；计算资源组 n 使用的内存资源是 Mem_n，Mem_n 是 CPU_n 本地的内存。在本实施例中，每个虚拟节点分配一个计算资源组，因此对应不同虚拟节点的业务请求可以使用不同的 CPU 资源和内存资源，避免了资源竞争。另外，如果按照传统的方式，没有把资源分配给每个虚拟节点，那么执行一个数据访问请求的时候，往往需要多个 CPU 并行处理由这个请求拆分而来的多个子请求，由于这些子请求之间具有关联性，CPU 在处理子请求时会涉及到调度或转发子请求。然而，按照本实施例提供的方式，每个虚拟节点对应一个 CPU，或者多个虚拟节点共用一个 CPU，那么分配到某个虚拟节点的业务请求都会由一个指定的 CPU 执行，减少了 CPU 之间的调度和转发。另外，如果不同的数据访问请求共用一个内存，那么，不可避免地会进行一些互斥操作以实现数据一致性。但按照本实施例的方式，访问不同地址段集合的数据访问请求分配不同的内存，会在一定程度上减少互斥操作。

需要说明的是，上述计算资源池只是本实施例提供的一种实现方式，本实施例还可

以提供其他实现方式，例如存储系统中部分或所有 CPU 组成 CPU 资源，每个虚拟节点分配所述 CPU 资源中的一部分 CPU 资源，又例如存储系统中部分或所有内存组成内存资源，每个虚拟节点分配所述内存资源中的一部分内存资源。

本申请实施例中的网络资源主要包括控制器 103 和硬盘框之间的链路资源。各个后端接口卡 102 上均可以创建多条逻辑链路，每条逻辑链路上可以建立多条连接，这些连接组成网络资源池。请参考图 4，图 4 是本申请提供的虚拟节点与网络资源的分配关系示意图。如图 4 所示，将所述网络资源池所包含的连接划分为若干链接组，每个链接组使用一条或多条连接，例如，链接组 0 使用连接_0，连接_0 是在逻辑链路_0 上建立的；链接组 1 使用连接_1，连接_1 是在逻辑链路_1 上建立的；链接组 P 使用连接_m 至连接_n 之间的所有连接，m 和 n 都是大于 1 的整数，且 n 大于 m，连接_m 与连接_n 之间有一个或多个连接，这些连接是在逻辑链路_n 上建立的。每个链接组分配给一个虚拟节点。由于不同的虚拟节点使用不同的连接，由此避免了系统内节点间交换的网络资源竞争。

需要说明的是，上述网络资源池只是本实施例提供的一种实现方式，本实施例还可以提供其他实现方式，例如控制器 103 与硬盘框之间的部分或所有连接组成网络资源，每个虚拟节点分配所述网络资源中的一部分网络资源。

本实施例中的硬盘资源主要指存储系统所包含的所有硬盘的容量。请参考图 5，图 5 是本申请实施例提供的虚拟节点与硬盘资源的分配关系示意图。如图 5 所示，按照设定粒度将硬盘划分为若干个块（chunk），这些 chunk 组成了存储池，按照独立硬盘冗余阵列(Redundant Array of Independent Disks，RAID)规则，从不同盘上选择一定数量的 chunk，从而组成块组（chunk group，CKG）。例如，块组 0 包含块_0、块_m 和块_n，块_0 来自于硬盘_0，块_m 来自于硬盘_1，块_n 来自于硬盘_n。每个虚拟节点对应一个或多个块组。由于不同虚拟节点使用不同的 CKG，从而实现了后端硬盘资源的隔离。

需要说明的是，上述存储池只是本实施例提供的一种实现方式，本实施例还可以提供其他实现方式，例如存储系统所包含的部分或全部硬盘组成硬盘资源，每个虚拟节点分配所述硬盘资源中的一部分硬盘资源。

综上，每个虚拟节点包含了处理业务所需的 CPU 资源、内存资源、网络资源以及硬盘资源，如图 6 所示，虚拟节点 0 分配的计算资源是计算资源组 0，分配的硬盘资源是块组 1 和块组 2，分配的网络资源是链接组 m 和链接组 n。各个虚拟节点之间分配的资源是相互独立的，随着 CPU 的数量、CPU 核的数量线性增加，只要相应地增加虚拟节点的数量，让单个虚拟节点的性能保持一致，就能实现性能随着物理资源的增加而线性扩展。业界将本实施例介绍的这种技术称为 CoreFarm。

下面介绍存储数据的过程。请参考图 7，图 7 是本实施例提供的处理写请求的流程示意图。如图 7 所示，包括以下步骤。

S101、客户端通过主机 11 触发生成写请求，所述写请求中携带待写入的数据以及所述数据的虚拟地址。虚拟地址是指 LUN ID、LBA 和 length。

S102、主机 11 确定所述写请求对应的虚拟节点。

具体的，主机 11 对所述虚拟地址进行哈希计算，从而获得哈希值。所述哈希值与某个分区对应，然后将所述分区的标识按一定规则映射到多个虚拟节点中的某个虚拟节点（称为目标虚拟节点），规则包括但不限于顺序、随机等算法。为了方便描述，本实施例以目标虚拟节点为图 5 中的虚拟节点 0 为例，根据图 3 至图 5 的描述，虚拟节点 0 所

分配的硬件资源包括计算资源组 0、块组 1、块组 2、链接组 m 和链接组 n。

S103、主机 11 将所述写请求发送给虚拟节点所对应的存储节点。具体的，主机 11 中保存有关于各个虚拟节点资源分配情况的映射表，该映射表中记录了每个虚拟节点与它所分配的各项资源之间的对应关系（如表 1 所示）。

虚拟节点	计算资源组	块组	链接组
虚拟节点 0	计算资源组 0	块组 1 和块组 2	链接组 m 和链接组 n
虚拟节点 1	计算资源组 1	块组 0	链接组 1
……	……	……	……
虚拟节点 p	计算资源组 p	块组 p	链接组 p

表 1

主机 11 根据表 1 中记录的虚拟节点 0 对应的计算资源，确定出该计算资源所在的存储节点。由图 5 的描述可知，虚拟节点 0 对应的 CPU 位于计算资源组 0 中，进一步地，由图 2 可知，计算资源组 0 包括 CPU_0 和 Mem_0，因此，主机 11 将所述写请求发送给 CPU_0 所在的存储节点（例如，存储节点 0）。如果主机 11 与存储节点 0 之间不具有链路，可以通过主机 11 与其他存储节点之间的链路将所述写请求发送给所述其他存储节点，再由所述其他存储节点转发给存储节点 0。如果主机 11 与存储节点 0 之间具有一条链路，则直接通过该链路将所述写请求发送给存储节点 0。如果主机 11 与存储节点 0 之间具有多条链路，则可以通过轮询或者其他方式从多条链路中选择一条链路，通过该选择出的链路将所述写请求发送给存储节点 0。

S104、存储节点接收所述写请求之后，将所述写请求发送给虚拟节点对应的 CPU 处理。具体的，存储节点的前端接口卡 101 中保存有所述关于各个虚拟节点资源分配情况的映射表（如表 1 所示），前端接口卡 101 可以根据写请求中携带的虚拟地址，确定对应的目标虚拟节点，从而进一步确定目标虚拟节点对应的 CPU。继续以目标虚拟节点为虚拟节点 0 为例，所述虚拟节点 0 对应的 CPU 是 CPU_0，因此前端接口卡 101 将所述写请求发送给 CPU_0。所述 CPU 可以对所述写请求中的数据进行相应的处理。处理前和处理后的数据需要暂时存储在内存中，由图 2 可知，计算资源组 0 包含的内存资源是 Mem_0，因此，数据可以利用 Mem_0 所指示的内存空间中进行存储。

S105、存储节点根据所述虚拟节点匹配的后端物理通道将处理后的数据发送给相应的硬盘存储。具体的，当内存 Mem_0 中存储的数据达到一定水位线时，需要将内存 Mem_0 中存储的数据写入硬盘进行持久化存储。存储节点可以根据所述映射表查询所述目标虚拟节点对应的块组，将所述待写入的数据写入目标虚拟节点对应的块组中。例如，由图 5 和表 1 可知，虚拟节点 0 对应的块组为块组 1 和块组 2，说明虚拟节点 0 可以使用块组 1 的硬盘资源也可以使用块组 2 的硬盘资源。由图 5 可知，块组 0 包含块_0、块_m 和块_n，其中，块_0 位于硬盘_0 中，块_m 位于硬盘_1 中，块_n 位于硬盘_n 中。存储节点 0 可以将所述待写入的数据划分为两个数据分片，并计算获得所述两个数据分片的校验分片，然后通过后端接口卡 102 将这两个数据分片和一个校验分片分别发送给硬盘_0、硬盘_1 和硬盘_n。另外，存储节点 0 在发送所述数据分片和校验分片需利用网络资源池中的网络资源。例如，虚拟节点 0 对应的网络资源包括链接组 m 和链接组 n，因此，后端接口卡 102 可以通过链接组 m 和/或链接组 n 包含的多条连接，将所述数据分片和校验分片并

行发送给相应的硬盘。

按照图 7 提供的写请求的处理方法，主机 11 首先确定该写请求对应的虚拟节点，然后利用预先为该虚拟节点分配的硬件资源处理该请求，由于每个虚拟节点分配的资源是相互独立的，所以当主机 11 并行处理多个数据处理请求时，多个数据处理请求之间也不会相互干扰。

图 7 是以处理写请求为例予以说明的，除了处理数据读写之外， 存储系统还处理其他业务请求，例如数据交换、协议解析和数据刷盘等。在多 CPU、多 CPU 核的存储系统中，这些业务请求往往是由多个 CPU 或者一个 CPU 内的多个 CPU 核串行执行的，而影响线性度的关键就是存储系统中的串行执行以及跨 CPU、跨 CPU 核处理带来的开销。在本实施例中，提供了一种 CPU 分组调度方法用于解决该问题。请参考图 8，图 8 是虚拟节点内 CPU 核的分组调度示意图。

首先，虚拟节点与一个 CPU 对应，对应的含义是：一个虚拟节点使用一个 CPU，或者，多个虚拟节点共用一个 CPU。从而保证了针对同一个虚拟节点的业务请求由同一个 CPU 处理，因此业务的调度在虚拟节点间保持独立。

其次，将虚拟节点对应的一个 CPU 所包含的多个 CPU 核按照业务逻辑分成若干业务处理组，每个业务处理组包括一个或多个 CPU 核。如图 7 所示，第一业务处理组专用于 I/O 读写，第二业务处理组专用于数据交换，第三业务处理组用于协议解析，第四业务处理组用于数据刷盘。其中，第三业务处理组和第四业务处理组所包含的 CPU 核可以共用。具体的，以一个 CPU 包含 48 个 CPU 核为例，假设第一业务处理组分配 12 个 CPU 核，第二业务处理组分配 12 个 CPU 核，第三业务处理组和第四业务处理组共同分配 24 个 CPU 核。

按照这样的方式隔离不同的业务请求，在单个业务处理组内，业务请求在自己分配的 CPU 核上串行执行，可以在一定程度上避免该业务请求与其他业务请求争夺资源，由此减少互斥操作并实现免锁设计，在 CPU 所包含的 CPU 核的数量增加时，该 CPU 的处理能力也能得到线性扩展。另外，将业务请求分组处理之后，业务代码相对于分组之前有所减少，那么业务代码所占用的内存空间也会减小，在总的内存空间不变的情况下，内存可以腾挪出更多的空间用于存储业务数据，由此提高数据的内存命中率。

与处理写请求类似，当客户端通过主机 11 触发读请求时，主机 11 可以通过所述读请求中携带的待读取数据的虚拟地址确定该请求所对应的虚拟节点，并进一步确定所述虚拟节点对应的存储节点（与 S103 类似）。主机 11 将所述读请求发送给虚拟节点对应的存储节点。存储节点接收所述读请求之后，将所述读请求发送给虚拟节点对应的 CPU 处理（与 S104 类似）。如果待读取的数据没有在对应的内存中命中，所述虚拟节点对应的 CPU 可以进一步确定所述虚拟节点对应的网络资源和硬盘资源，然后利用所述对应的网络资源，向对应的硬盘发送请求以读取所述待读取的数据。

另外，在实际应用中，通过提升单核能力来提升存储系统能力的代价越来越大，业界目前采用多个节点，每个节点多 CPU 多核方式来提升存储系统的处理能力。例如在单核能力相差不多的情况下，存储系统中核的数量从 48 核扩展到 768 个核，其硬件能力是增长的。但如何让存储系统处理业务的能力能随 CPU 核数和内存等资源增加而线性扩展也是所有存储设备厂家均需要解决的问题。本实施例提供的扩容方法可以使得存储系统处理业务的能力随着硬件资源增加而线性扩展。下面介绍节点扩容的过程，该过程将结

合图 3 至图 6 来举例描述。

请参考图 9 和图 10，图 9 是本申请实施例提供的一种存储系统扩容示意图，图 10 是本申请实施例提供的一种扩容方法的流程示意图。本实施例以在一个节点内扩展控制器的数量为例予以说明，假设该节点在扩容前包含两个控制器，分别是控制器 A 和控制器 B，增加两个控制器（控制器 C 和控制器 D）后，该节点包含四个控制器。无论是扩容前还是扩容后，其前端接口卡 101 和后端接口卡 102 都是各控制器共享的，具体的，如图 10 所示该扩容方法包含如下步骤。

S201，控制器 C 和控制器 D 加入系统后，控制器 C 和控制器 D 分别初始化虚拟节点实例。可以理解的是，控制器的数量增加了，整个节点可提供的 CPU 资源和内存资源也相应地增加，因此，只要增加虚拟节点的数量，将新增的 CPU 资源和内存资源分配给新增的虚拟节点就能提高整个节点的处理能力。

以控制器 C 为例，控制器 C 根据自己所包含的 CPU 的数量创建多个虚拟节点，由于本实施例中一个虚拟节点分配有一个 CPU，因此，虚拟节点的数量可以小于或等于控制器 C 所包含的 CPU 的数量。例如，控制器 C 包含 8 个 CPU，那么控制器 C 最多可以创建 8 个虚拟节点。确定虚拟节点的数量之后，再进一步确定新增的虚拟节点与 CPU 之间的映射关系，以及新增的虚拟节点与内存之间的映射关系。例如，在控制器 C 内部，虚拟节点 x 对应 CPU_x（x 代表一个正整数），虚拟节点 x 需要的内存资源可以就近使用 CPU_ x 本地的内存（例如 Mem_ x），由此 CPU_ x 和 Mem_ x 组成一个计算资源组分配给虚拟节点 x。虚拟节点 x+1 对应 CPU_x+1，虚拟节点 x+1 需要的内存资源可以就近使用 CPU_ x+1 本地的内存（例如 Mem_ x+1），由此 CPU_ x+1 和 Mem_ x+1 组成另一个计算资源组分配给虚拟节点 x+1。控制器 D 创建虚拟节点的方式与控制器 C 类似。

另外，控制器 C 和控制器 D 加入系统之后，控制器 C 和控制器 D 会分别和后端接口卡 102 建立物理链路，在这些物理链路上创建多条逻辑链路，每条逻辑链路上又可以建立多条连接，这些连接被添加到图 4 所示的网络资源池中，以扩充所述网络资源池的网络资源。这些新增的网络资源可以被划分为若干链接组，每个链接组包含一条或多条连接，然后将每个链接组分配给一个新增的虚拟节点，例如虚拟节点 x 对应链接组 x（x 代表一个正整数），虚拟节点 x+1 对应链接组 x+1。

S202，将归属于控制器 A 和控制器 B 的虚拟节点的部分分区迁移至控制器 C 和控制器 D 的虚拟节点上。由前面的描述可知，来自主机 11 的业务请求是根据虚拟地址所对应的分区路由到虚拟节点的，在存储系统所包含的分区的总数不变的情况下，为了让新创建的虚拟节点承担业务请求，就需要将归属于原有的虚拟节点的一部分分区迁移至新创建的虚拟节点上。例如，扩容前一个虚拟节点对应一个分区集合，一个分区集合包含 32 个分区，扩容后，一个虚拟节点对应 24 个分区。一种实现方式是重新建立存储系统中所有分区与各个虚拟节点（既包含原有的虚拟节点也包含新增的虚拟节点）的映射关系，另一种实现方式是将原有的分区集合的部分分区迁移至新增的虚拟节点，保留剩下的分区与原有的虚拟节点之间的对应关系。结合上面的例子，那么就需要将原来的分区集合中的 8 个分区迁移至新增的虚拟节点。需要说明的是，迁移的分区的数量取决于新增的虚拟节点的数量在整个节点所包含的虚拟节点的数量中所占的比例，本实施例不对迁移算法进行限定，只要保证分区在各个虚拟节点中均匀分布即可。

S203，更新映射表，既包含主机 11 中保存的映射表也包含前端接口卡 101 中的映

射表。按照 S201 的描述，新增的虚拟节点分配有 CPU 资源、内存资源和网络资源。这些新增的分配关系需要记录在映射表中供处理业务请求时使用。由于控制器 C 和控制器 D 中没有硬盘，所以新增的虚拟节点所需的硬盘资源仍然来自于图 5 所示的存储池，具体的，可以将归属于原有的虚拟节点的块组迁移至新增的虚拟节点，本实施例并不对迁移算法进行限定，只要满足每个虚拟节点所分配的块组大致相等即可。

更新后的映射表如表 2 所示。

虚拟节点	计算资源组	块组	链接组
虚拟节点 0	计算资源组 0	块组 1 和块组 2	链接组 m 和链接组 n
虚拟节点 1	计算资源组 1	块组 0	链接组 1
……	……	……	……
虚拟节点 p	计算资源组 p	块组 p	链接组 p
……	……	……	……
虚拟节点 x	计算资源组 x	块组 p+1	链接组 x
虚拟节点 x+1	计算资源组 x+1	块组 p+2	链接组 x+1
……	……	……	……

表 1

S204，主机 11 按照新的分区路由关系发送业务请求。业务请求的处理方式请参考图 6 所示的处理写请求的流程示意图，这里不再赘述。

请参考图 11 和图 12，图 11 是本申请实施例提供的另一种存储系统扩容示意图，图 12 是本申请实施例提供的另一种扩容方法的流程示意图。在图 11 和图 12 所示的示例中，不仅对节点内的控制器进行了扩充还对硬盘的数量或者硬盘框的数量进行了扩充。仍然以在一个节点内扩展控制器的数量为例予以说明，假设该节点在扩容前包含两个控制器，分别是控制器 A 和控制器 B，增加两个控制器（控制器 C 和控制器 D）后，该节点包含四个控制器。无论是扩容前还是扩容后，其前端接口卡 101 和后端接口卡 102 都是各控制器共享的。具体的，如图 12 所示该扩容方法包含如下步骤。

S301，控制器 C 和控制器 D 加入系统后，控制器 C 和控制器 D 分别初始化虚拟节点实例。该步骤可参见图 9 所示的 S201。

S302，将归属于控制器 A 和控制器 B 的虚拟节点的部分分区迁移至控制器 C 和控制器 D 的虚拟节点上。该步骤可参见图 9 所示的 S202。

S303，根据选举算法从新增的控制器 C 和控制器 D 选举出一个主控制器，例如控制器 C。

S304，控制器 C 将新加入的硬盘的空间划分为若干个 chunk，并将这些 chunk 加入存储池。当控制器 C 或控制器 D 接收写请求时，所述写请求对应新增的虚拟节点，来自不同硬盘的 chunk 组成一个块组以容纳所述写请求中携带的数据。由此可见，存储池中新增的多个块组可以分配给新增的虚拟节点，其中，每个虚拟节点使用一个或多块组。

S305，更新映射表，既包含主机中保存的映射表也包含前端接口卡 101 中的映射表。该步骤可参见图 9 所示的 S203。与 S203 不同之处在于，在 S203 的示例中，没有增加硬盘资源，所以新增的虚拟节点所对应的块组是由系统中原有的虚拟节点的块组迁移而来的，而在 S305 中的示例中，由于硬盘资源也增加了，所以新增的虚拟节点所对应的块组

来自新增的硬盘资源。

按照图 11 和图 12 所示的扩容方式，新增的硬盘资源分配给新增的虚拟节点，那么对应所述新增的虚拟节点的写请求中携带的数据可写入所述新增的硬盘中。然而仍然有大量的旧数据仍然保存在原有的硬盘中。为了让系统中存储的数据均匀地分布在各个硬盘上，一种方式是，将部分旧数据迁移至新增的硬盘中存储。另一种方式是，不主动迁移旧数据，而是在对旧数据进行垃圾回收时将所述旧数据中的有效数据（没有被修改的数据）迁移至新增的硬盘中。随着系统运行，垃圾数据越来越多，经过若干次垃圾回收操作之后，慢慢也能达到数据均匀分布的目的。这种的方式的好处在于，由于没有主动迁移数据，可以节省节点之间，或者控制器之间的带宽开销。

按照图 9-图 12 所示的两种扩容方法，当系统中的控制器的数量增加时，虚拟节点的数量也相应地增加，并且将新增的资源分配给新增的虚拟节点，在一定程度上减少了对原有的虚拟节点所分配的资源的抢夺，因此，随着硬件资源的数量增加，整个系统的处理能力也相应地增加。

在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时，可以全部或部分地以计算机程序产品的形式实现。所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时，全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一个计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、存储节点或数据中心通过有线（例如同轴电缆、光纤、数字用户线（DSL））或无线（例如红外、无线、微波等）方式向另一个网站站点、计算机、存储节点或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的存储节点、数据中心等数据存储设备。所述可用介质可以是磁性介质，(例如，软盘、硬盘、磁带)、光介质(例如，DVD)、或者半导体介质（例如固态硬盘(Solid State Disk, SSD)）等。

应理解，在本申请实施例中，术语“第一”等仅仅是为了指代对象，并不表示相应对象的次序。

本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统、装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统、装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，存储节点，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（Read-Only Memory，ROM）、随机存取存储器（Random Access Memory，RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

权利要求书

1、一种存储设备，其特征在于，包括硬盘框和多个控制器，每个控制器包括多个处理器，每个处理器包括多个处理器核，所述多个控制器分别和所述硬盘框耦合，所述硬盘框包括多个硬盘；

所述多个处理器，用于提供计算资源；

所述多个硬盘，用于提供存储空间，所述存储空间所对应的逻辑地址被划分为若干个地址段集合，每个地址段集合包含一个或多个地址段，每个地址段集合分配有所述计算资源中的一部分计算资源，所述一部分计算资源用于执行访问所述地址段集合所包含的地址段的数据访问请求，其中，为不同地址段集合分配的计算资源来自不同的处理器，或者，为不同地址段集合分配的计算资源来自不同的处理器核。

2、根据权 1 所述的存储设备，其特征在于，每个处理器具有内存，所述计算资源还包括所述多个处理器包含的内存，其中，一个内存分配给一个地址段集合，且不同地址段集合分配的内存不同。

3、根据权 2 所述的存储设备，其特征在于，每个地址段集合分配的内存是该地址段集合分配的处理器的本地内存。

4、根据权利要求 1 所述的存储设备，其特征在于，所述存储设备还包括网络资源，所述网络资源由所述多个控制器与所述硬盘框之间的多个连接提供，每个地址段集合分配有所述网络资源中的一部分网络资源，其中，不同地址段集合分配的连接不同。

5、根据权利要求 1 所述的存储设备，其特征在于，所述存储空间包括多个块组，所述多个块组提供存储资源，每个地址段集合分配有所述存储资源中的一部分存储资源，其中，不同地址段集合分配的块组不同。

6、根据权利要求 1-5 任一所述的存储设备，其特征在于，所述存储空间被虚拟化为逻辑单元 LUN，所述 LUN 由所述存储设备提供给用户使用，所述 LUN 所对应的地址被划分而成若干个逻辑块地址 LBA 区间，每个地址段为一个 LBA 区间。

7、根据权利要求 1-6 任一所述的存储设备，其特征在于，所述存储设备还包括前端接口卡，所述多个控制器分别与所述前端接口卡耦合，所述前端接口卡中保存映射表，所述映射表用于指示每个地址段集合与所分配的计算资源之间的映射关系，和/或每个地址段集合与所分配的网络资源之间的映射关系，和/或每个地址段集合与所分配的硬盘资源之间的映射关系。

8、根据权利要求 7 所述的存储设备，其特征在于，

所述存储设备中创建有若干个虚拟节点，一个虚拟节点通过预设算法映射至一个地址段集合，所述映射表记录有每个虚拟节点与该虚拟节点所映射的地址段集合所分配的计算资源之间的映射关系，和/或每个虚拟节点与该虚拟节点所映射的地址段集合所分配

的网络资源之间的映射关系，和/或每个虚拟节点与该虚拟节点所映射的地址段集合所分配的硬盘资源之间的映射关系。

9、根据权利要求 1 所述的存储设备，其特征在于，还包括新加入的控制器，所述新加入的控制器包括处理器，所述新加入的控制器与所述硬盘框耦合，

其中，所述新加入的控制器所包含的处理器作为计算资源，被分配给第一地址段集合中的第一地址段，所述第一地址段与所述第一地址段集合所分配的计算资源之间的分配关系被解除。

10、一种资源分配方法，其特征在于，所述方法应用于存储设备中，所述存储设备包括硬盘框和多个控制器，每个控制器包括多个处理器，每个处理器包括多个处理器核，所述多个控制器分别和所述硬盘框耦合，所述硬盘框包括多个硬盘；所述方法包括:

提供计算资源和存储空间，所述存储空间所对应的逻辑地址被划分为若干个地址段集合，每个地址段集合包含一个或多个地址段;

为每个地址段集合分配所述计算资源中的一部分计算资源，所述一部分计算资源用于执行访问所述地址段集合所包含的地址段的数据访问请求，其中，为不同地址段集合分配的计算资源来自不同的处理器，或者，为不同地址段集合分配的计算资源来自不同的处理器核。

11、根据权 10 所述的方法，其特征在于，每个处理器具有内存，所述计算资源还包括所述多个处理器包含的内存，所述方法还包括:

为每个地址段集合分配一个内存，且不同地址段集合分配的内存不同。

12、根据权 10 所述的方法，其特征在于，每个地址段集合分配的内存是该地址段集合分配的处理器的本地内存。

13、根据权 10 所述的方法，其特征在于，所述存储设备还包括网络资源，所述网络资源由所述多个控制器与所述硬盘框之间的多个连接提供，所述方法还包括:

为每个地址段集合分配所述网络资源中的一部分网络资源，其中，不同地址段集合分配的连接不同。

14、根据权 10 所述的方法，其特征在于，所述存储空间包括多个块组，所述多个块组提供存储资源，所述方法还包括:

为每个地址段集合分配所述存储资源中的一部分存储资源，其中，不同地址段集合分配的块组不同。

15、根据权 10-14 任一所述的方法，其特征在于，所述存储空间被虚拟化为逻辑单元 LUN，所述 LUN 由所述存储设备提供给用户使用，所述 LUN 所对应的地址被划分而成若干个逻辑块地址 LBA 区间，每个地址段为一个 LBA 区间。

16、根据权 10-15 任一所述的方法，其特征在于，所述存储设备还包括前端接口卡，所述多个控制器分别与所述前端接口卡耦合，所述前端接口卡中保存映射表，所述映射表用于指示每个地址段集合与所分配的计算资源之间的映射关系，和/或每个地址段集合与所分配的网络资源之间的映射关系，和/或每个地址段集合与所分配的硬盘资源之间的映射关系。

17、根据权 16 所述的方法，其特征在于，所述方法还包括:

接收数据访问请求，所述数据访问请求包括虚拟地址;

根据所述虚拟地址确定所述数据访问请求对应的地址段集合;

在所述映射表中查询为所述数据访问请求对应的地址段集合所分配的计算资源;

利用所述分配的计算资源处理所述数据访问请求。

18、根据权 16 所述的方法，其特征在于，还包括:

创建若干个虚拟节点，一个虚拟节点通过预设算法映射至一个地址段集合，所述映射表记录有每个虚拟节点与该虚拟节点所映射的地址段集合所分配的计算资源之间的映射关系，和/或每个虚拟节点与该虚拟节点所映射的地址段集合所分配的网络资源之间的映射关系，和/或每个虚拟节点与该虚拟节点所映射的地址段集合所分配的硬盘资源之间的映射关系。

19、根据权 10 所述的方法，其特征在于，还包括:

配置新加入的控制器，所述新加入的控制器包括处理器，所述新加入的控制器与所述硬盘框耦合，

将所述新加入的控制器所包含的处理器作为计算资源，分配给一个地址段集合中的第一地址段，并解除所述第一地址段与所述地址段集合所分配的计算资源之间的分配关系。

20、一种存储系统，其特征在于，包括主机和如权利要求 1-9 任一所述的存储设备。

21、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有指令，当其在存储设备运行时，使得所述存储设备执行如权利要求 10-19 任一所述的资源分配方法。

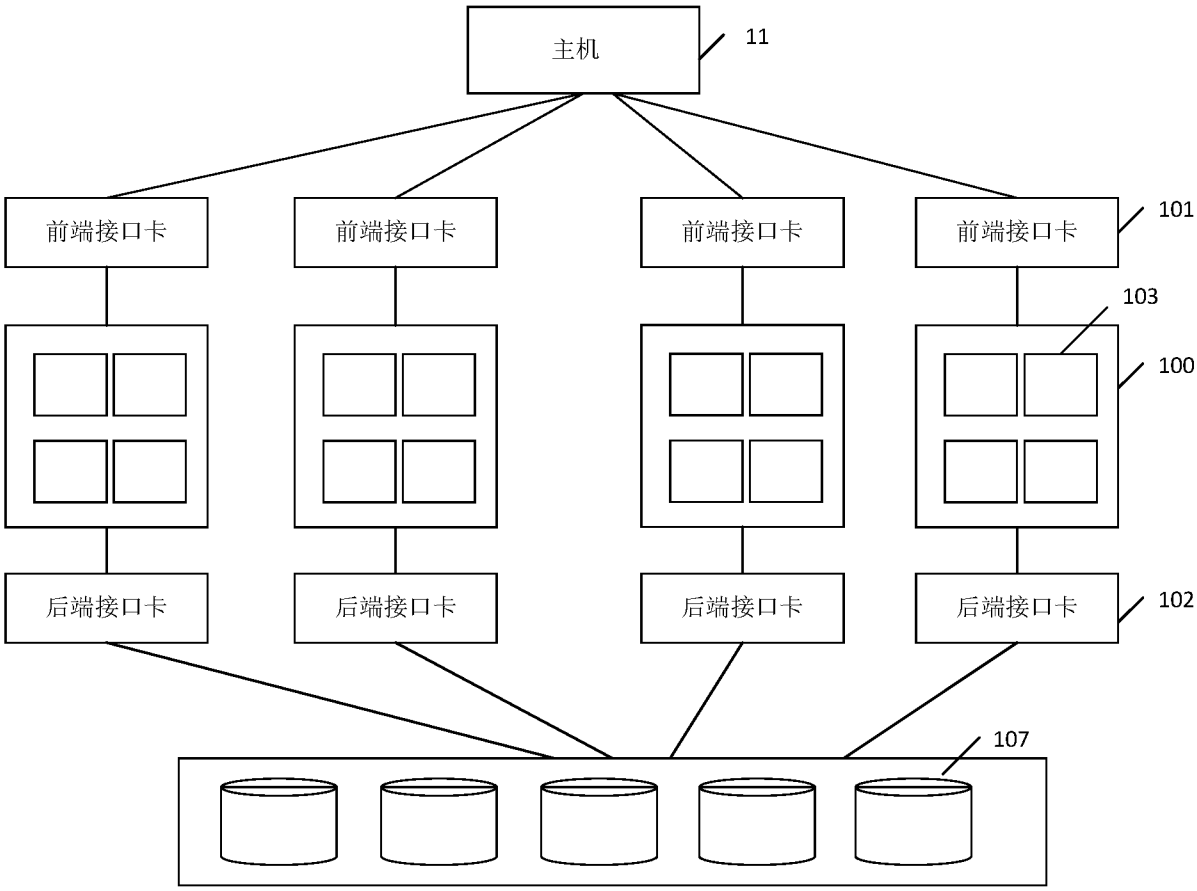

主机
11
前端接口卡
前端接口卡
前端接口卡
前端接口卡
101
103
100
后端接口卡
后端接口卡
后端接口卡
后端接口卡
102
107

图 1

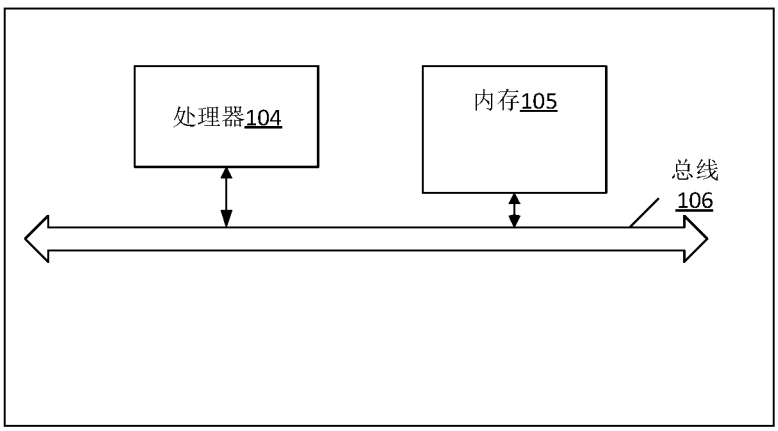

处理器104
内存105
总线
106

图 2

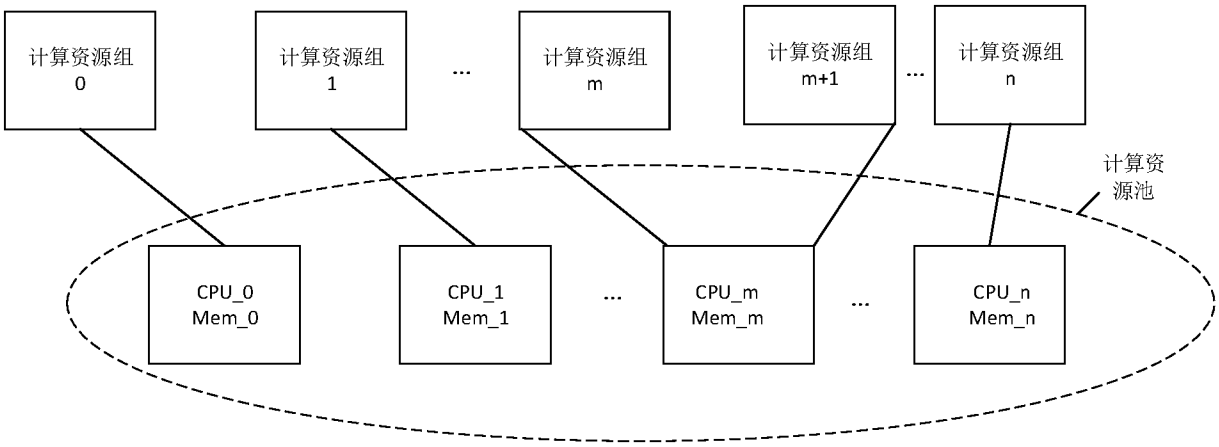

计算资源组
0
计算资源组
1
...
计算资源组
m
计算资源组
m+1
...
计算资源组
n
计算资
源池
CPU_0
Mem_0
CPU_1
Mem_1
...
CPU_m
Mem_m
...
CPU_n
Mem_n

图 3

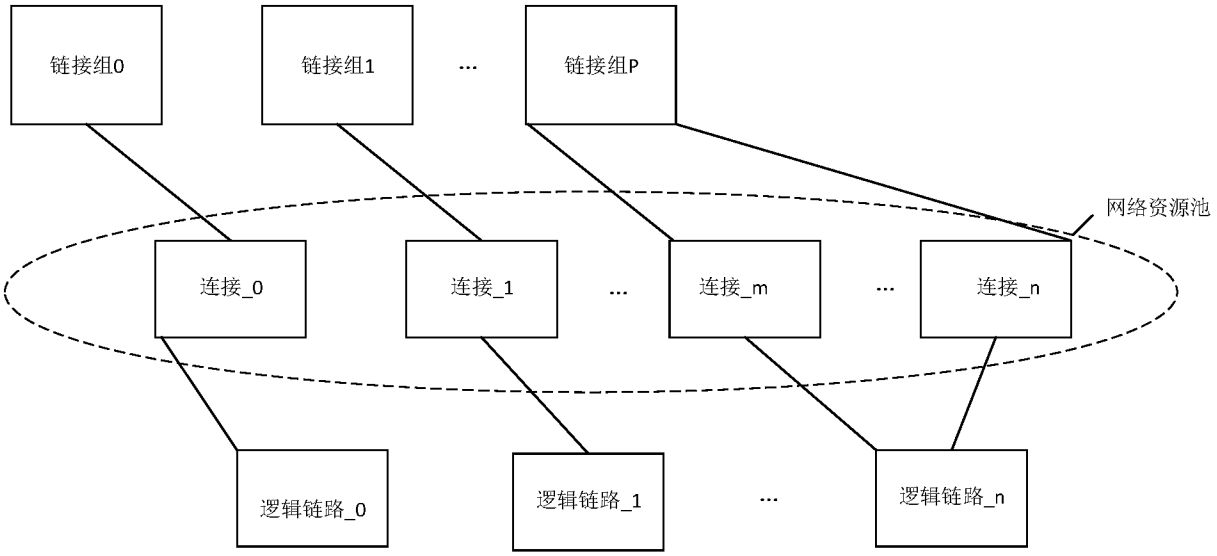

链接组0
链接组1
...
链接组P
网络资源池
连接_0
连接_1
...
连接_m
...
连接_n
逻辑链路_0
逻辑链路_1
...
逻辑链路_n

图 4

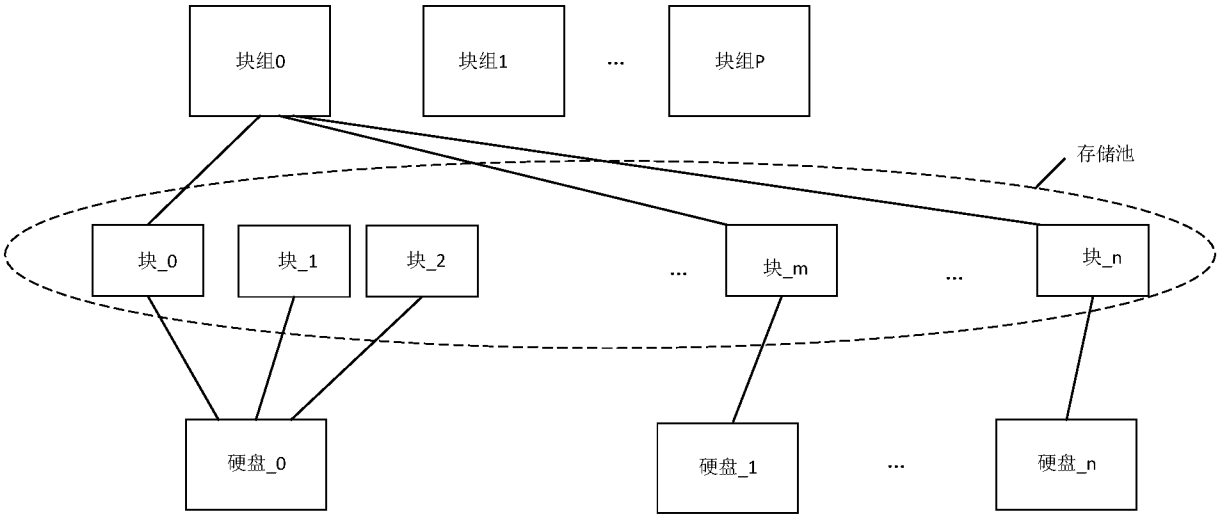

块组0
块组1
...
块组P
存储池
块_0
块_1
块_2
...
块_m
...
块_n
硬盘_0
硬盘_1
...
硬盘_n

图 5

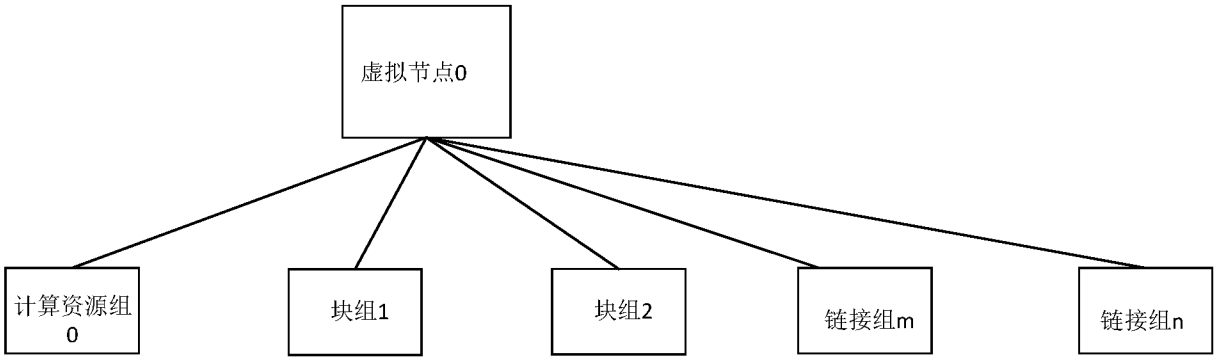

虚拟节点0
计算资源组0
块组1
块组2
链接组m
链接组n

图 6

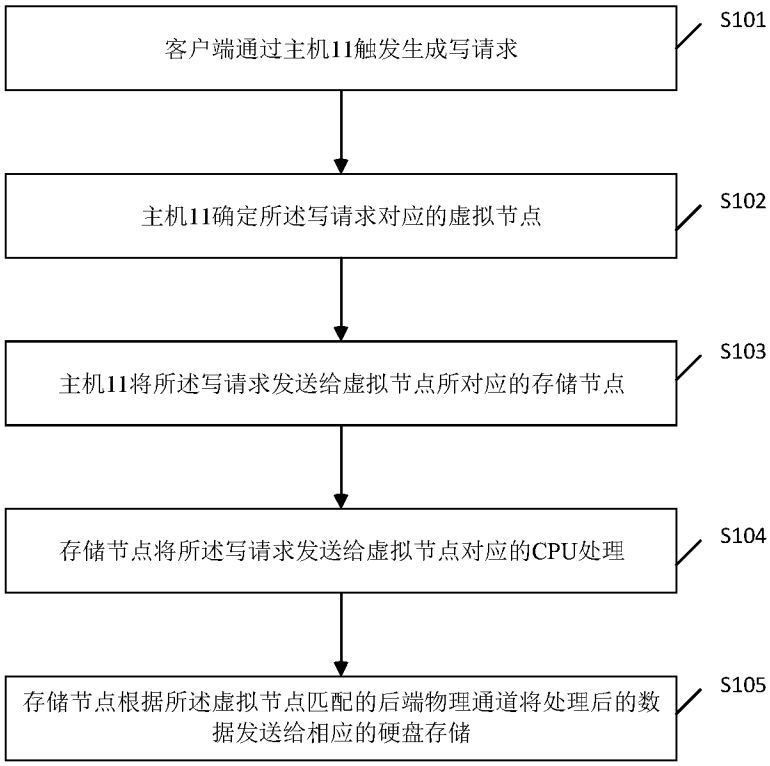

客户端通过主机11触发生成写请求
S101
主机11确定所述写请求对应的虚拟节点
S102
主机11将所述写请求发送给虚拟节点所对应的存储节点
S103
存储节点将所述写请求发送给虚拟节点对应的CPU处理
S104
存储节点根据所述虚拟节点匹配的后端物理通道将处理后的数据发送给相应的硬盘存储
S105

图 7

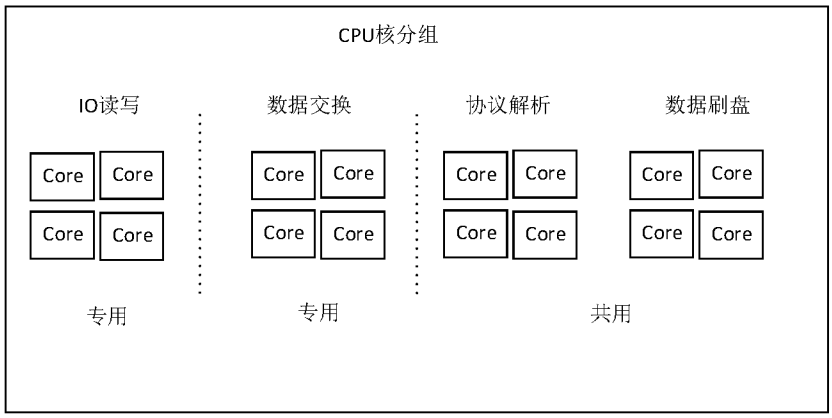

CPU核分组
IO读写
数据交换
协议解析
数据刷盘
Core
专用
专用
共用

图 8

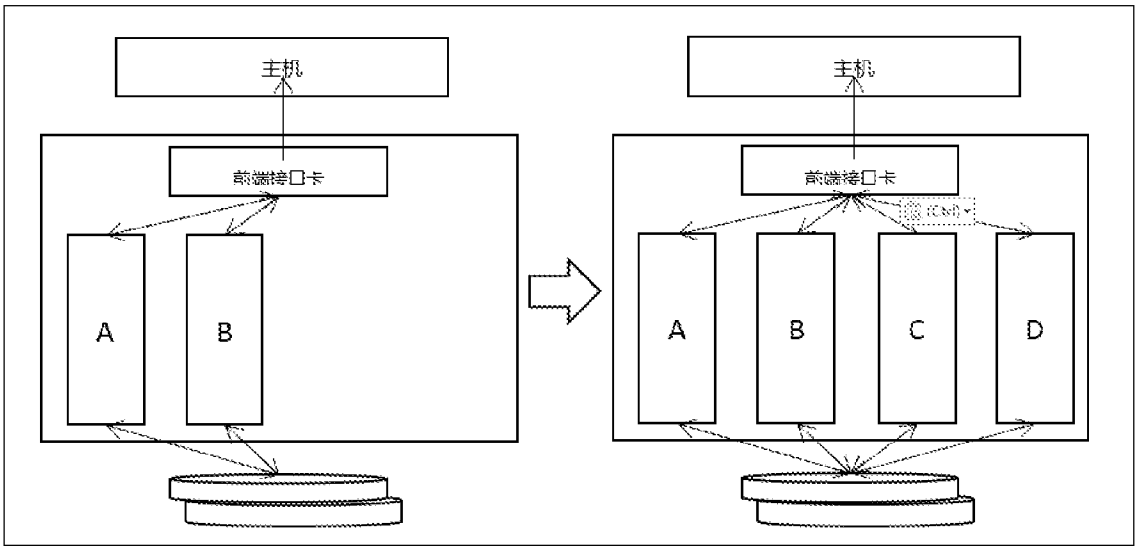

主机
前端接口卡
A
B
主机
前端接口卡
A
B
C
D

图 9

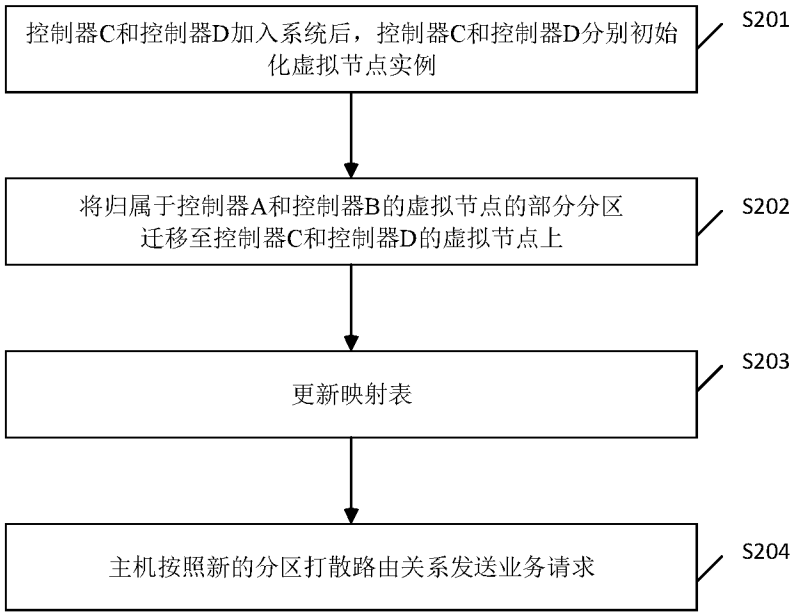

控制器C和控制器D加入系统后，控制器C和控制器D分别初始化虚拟节点实例
S201
将归属于控制器A和控制器B的虚拟节点的部分分区迁移至控制器C和控制器D的虚拟节点上
S202
更新映射表
S203
主机按照新的分区打散路由关系发送业务请求
S204

图 10

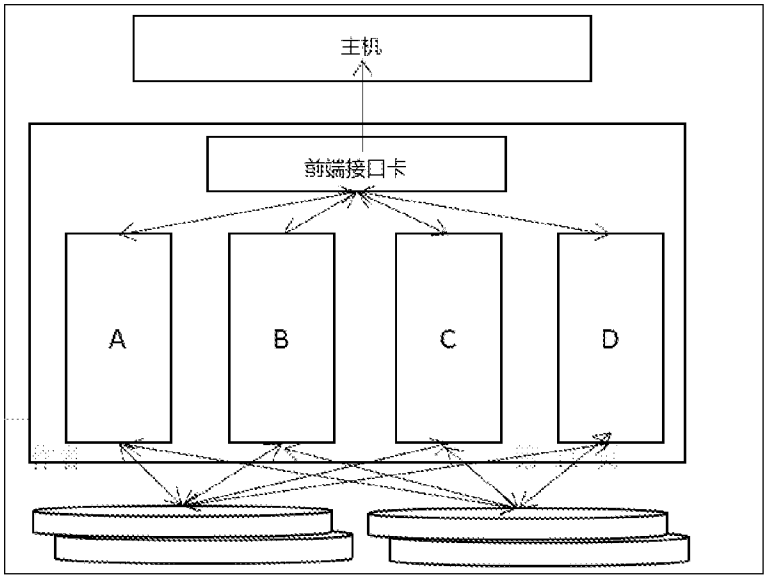

主机
前端接口卡
A
B
C
D

图 11

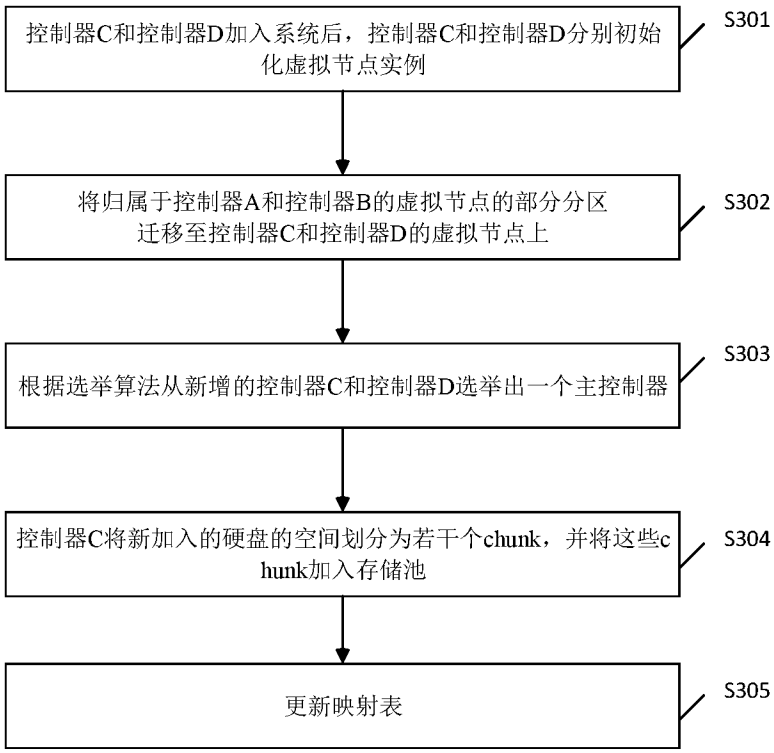

控制器C和控制器D加入系统后，控制器C和控制器D分别初始化虚拟节点实例
S301
将归属于控制器A和控制器B的虚拟节点的部分分区迁移至控制器C和控制器D的虚拟节点上
S302
根据选举算法从新增的控制器C和控制器D选举出一个主控制器
S303
控制器C将新加入的硬盘的空间划分为若干个chunk，并将这些chunk加入存储池
S304
更新映射表
S305

图 12

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2020/088784

A. CLASSIFICATION OF SUBJECT MATTER

G06F 9/50(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC: 存储, 储存, 保存, 缓存, 逻辑单元, 内存, 处理器, 核, core, CPU, 多, 双, 地址, 单独, 独立, 平行, 独占, 交互, 虚拟, storage, catch, ram, rom, memory, processor, more, double, multiple, address, parallel, interact, independ, LBA, LUN, DHT, virtual

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 105677580 A (HANGZHOU HUAWEI DIGITAL TECHNOLOGY CO., LTD.) 15 June 2016 (2016-06-15) description, paragraphs [0029]-[0057], and figures 1-2	1-21
A	CN 105278940 A (BEIHANG UNIVERSITY) 27 January 2016 (2016-01-27) entire document	1-21
A	CN 109407975 A (HUAWEI TECHNOLOGIES CO., LTD.) 01 March 2019 (2019-03-01) entire document	1-21
A	US 2009144531 A1 (HARIKUMAR, Ajay et al.) 04 June 2009 (2009-06-04) entire document	1-21
A	赵勇 (ZHAO, Yong). "基于众核网络处理器的高性能安全存储系统设计与实现 (Research and Implementation of High Performance Secure Storage System Based on Multi-core Network Processor)" *中国优秀硕士学位论文全文数据库 (China Master's Theses Full-text Database)*, No. 2,, 15 February 2019 (2019-02-15), ISSN: 1674--024, entire document	1-21

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 July 2020	**05 August 2020**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2020/088784

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 105677580 A	15 June 2016	None	
CN 105278940 A	27 January 2016	CN 109344112 A	15 February 2019
		CN 109101286 A	28 December 2018
CN 109407975 A	01 March 2019	None	
US 2009144531 A1	04 June 2009	None	

国际检索报告

国际申请号
PCT/CN2020/088784

A. 主题的分类

G06F 9/50(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT, CNKI, WPI, EPODOC: 存储,储存,保存,缓存,逻辑单元,内存,处理器,核，core，CPU，多,双，地址，单独，独立，平行，独占，交互，虚拟，storage, catch, ram, rom, memory, processor, more, double, multiple, address, parallel, interact, independ, LBA, LUN, DHT, virtual

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
X	CN 105677580 A (杭州华为数字技术有限公司) 2016年 6月 15日 (2016 - 06 - 15) 说明书第[0029]-[0057]段，图1-2	1-21
A	CN 105278940 A (北京航空航天大学) 2016年 1月 27日 (2016 - 01 - 27) 全文	1-21
A	CN 109407975 A (华为技术有限公司) 2019年 3月 1日 (2019 - 03 - 01) 全文	1-21
A	US 2009144531 A1 (HARIKUMAR, Ajay 等) 2009年 6月 4日 (2009 - 06 - 04) 全文	1-21
A	赵勇. "基于众核网络处理器的高性能安全存储系统设计与实现" 中国优秀硕士学位论文全文数据库, 第2期，2019年 2月 15日 (2019 - 02 - 15)， ISSN：1674--024, 全文	1-21

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2020年 7月 22日	2020年 8月 5日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 陈晓伟 电话号码 86-(10)-53961673

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2020/088784

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	105677580	A	2016年 6月 15日		无		
CN	105278940	A	2016年 1月 27日	CN	109344112	A	2019年 2月 15日
				CN	109101286	A	2018年 12月 28日
CN	109407975	A	2019年 3月 1日		无		
US	2009144531	A1	2009年 6月 4日		无		