

存储系统、方法以及硬件卸载卡

5 本申请要求于 2022 年 08 月 24 日提交中国专利局、申请号为 202211021514.6、申请名称为“存储系统、方法以及硬件卸载卡”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

10 本说明书实施例涉及计算机技术领域，特别涉及一种存储系统。
背景技术

随着大数据分析、人工智能等技术的飞速发展，客户需要更加高性能、高可用、可扩展且灵活的存储能力。但传统的本地存储技术中，主机上的虚拟机及后台 IO 任务处理占用了比较多的 CPU 资源，使得无法实现高性能 IO 和时延。

15 目前，一些方案使用纯软件方案，实现上较为灵活，通过实现软件接口的标准化，可以将不同设备类型的软硬件交互进行统一。但因为方案的标准化，而忽略了不同设备类型数据传输的特点，在一些大数据量传输的场景，效率比较低下。

发明内容

20 有鉴于此，本说明书实施例提供了一种存储系统。本说明书一个或者多个实施例同时涉及一种存储方法，一种硬件卸载卡，一种计算机可读存储介质以及一种计算机程序，以解决现有技术中存在的技术缺陷。

根据本说明书实施例的第一方面，提供了一种存储系统，包括：硬件卸载卡以及存储设备，所述硬件卸载卡与所述存储设备对等地连接在主机上；所述硬件卸载卡，被配置为从所述主机接收存储任务，执行所述存储任务，将所述存储任务对应的数据访问请求发送给所述存储设备；所述存储设备，被配置为基于所述主机与所述存储设备之间的传输通道，
25 传输所述数据访问请求对应的存储数据。

可选地，所述硬件卸载卡包括可编程片上系统以及专用硬件；所述可编程片上系统，被配置为对所述存储任务中的软件子任务进行识别，调用所述可编程片上系统上运行的软件处理逻辑对所述软件子任务进行处理；所述专用硬件，被配置为执行所述存储任务中的硬件子任务。

30 可选地，所述可编程片上系统，还被配置为识别所述存储设备的介质类型，根据所述介质类型配置对应的交互规则，使所述数据访问请求根据所述交互规则生成。

可选地，所述专用硬件，还被配置为基于虚拟设备仿真技术建立虚拟设备，所述虚拟设备，用于将所述存储设备的物理存储资源进行抽象，面向主机提供虚拟化的存储资源。

可选地，所述虚拟设备，被配置为根据与所述主机的虚拟机协商的内存地址，从所述内存地址获取所述存储任务。

5 可选地，所述专用硬件包括存储协议处理模块，所述存储协议处理模块，被配置为解析发送所述存储任务的虚拟机的通信协议格式，将所述存储任务的通信协议格式转换为通用通信协议格式，使进入所述可编程片上系统的任务为通用通信协议格式。

10 可选地，单台存储设备对应抽象化为多台所述虚拟设备，其中，不同虚拟设备对应主机中的不同虚拟机，多台所述虚拟机共享所述单台存储设备的存储资源；所述可编程片上系统包括多租户共享任务处理模块，所述多租户共享任务处理模块，被配置为将所述多台虚拟机的存储任务分别对应分配所述单台存储设备的不同存储区域，以及，对所述存储任务进行权限校验和访问地址隔离。

可选地，所述主机中包括虚拟机，所述虚拟机内设置有内存；所述虚拟机的内存与所述存储设备之间的传输通道为 DMA 传输通道；所述存储设备，被配置为通过 DMA 直接访问所述虚拟机的内存来传输所述数据访问请求对应的存储数据。

15 可选地，所述硬件卸载卡，被配置为将所述数据访问请求保存在所述硬件卸载卡的内存；所述存储设备，被配置为通过 DMA 直接访问所述硬件卸载卡的内存来获取所述数据访问请求。

可选地，所述可编程片上系统运行的软件处理逻辑包括：存储资源的池化处理、缓存加速处理、访问请求错误处理和/或硬件运维处理的逻辑。

20 根据本说明书实施例的第二方面，提供了一种存储方法，应用于硬件卸载卡，所述硬件卸载卡与存储设备对等地连接在主机上，所述方法包括：从所述主机接收存储任务；执行所述存储任务；将所述存储任务对应的数据访问请求发送给所述存储设备，使所述存储设备基于主机与所述存储设备之间的传输通道，传输所述数据访问请求对应的存储数据。

25 根据本说明书实施例的第三方面，提供了一种硬件卸载卡，包括：存储器和处理器；所述存储器用于存储计算机可执行指令，所述处理器用于执行所述计算机可执行指令，该计算机可执行指令被处理器执行时实现本说明书任意实施例所述存储方法的步骤。

根据本说明书实施例的第四方面，提供了一种计算机可读存储介质，其存储有计算机可执行指令，该计算机可执行指令被处理器执行时实现本说明书任意实施例所述存储方法的步骤。

30 本说明书一个实施例提供了存储系统，该存储系统包括硬件卸载卡以及存储设备，其中，硬件卸载卡与存储设备对等地连接在主机上，所述硬件卸载卡被配置为从所述主机接收存储任务，执行所述存储任务，将所述存储任务对应的数据访问请求发送给所述存储设备，所述存储设备被配置为基于所述主机与所述存储设备之间的传输通道，传输所述数据访问请求对应的存储数据。可见，该存储系统中，存储任务被卸载到硬件卸载卡，利用硬件来加速执行，减少对主机 CPU 资源的占用，任务处理效率更高，并且由于存储设备通过

对等的方式从硬件卸载卡获取数据访问请求，相当于存储数据的传输与硬件卸载卡对存储任务的处理分离，硬件卸载卡负责执行的存储任务是控制相关的任务，并不携带存储数据，存储数据的传输无需经过硬件卸载卡，存储设备直接与主机传输数据访问请求对应的存储数据，实现了数控分离的处理策略，达成接近于物理硬件级别的性能，能够实现更高性能
5 10 和时延。

附图说明

图 1 是本说明书一个实施例提供的存储系统的结构框图；

图 2 是本说明书另一个实施例提供的存储系统的结构框图；

图 3 是本说明书又一个实施例提供的主机的结构框图；

10 图 4 是本说明书一个实施例提供的存储系统的云上多租户应用场景示意图；

图 5 是本说明书一个实施例提供的存储方法的流程示意图；

图 6 是本说明书一个实施例提供的一种计算设备的结构框图。

具体实施方式

在下面的描述中阐述了很多具体细节以便于充分理解本说明书。但是本说明书能够以
15 很多不同于在此描述的其它方式来实施，本领域技术人员可以在不违背本说明书内涵的情况下做类似推广，因此本说明书不受下面公开的具体实施的限制。

在本说明书一个或多个实施例中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本说明书一个或多个实施例。在本说明书一个或多个实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地
20 表示其他含义。还应当理解，本说明书一个或多个实施例中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

应当理解，尽管在本说明书一个或多个实施例中可能采用术语第一、第二等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本说明书一个或多个实施例范围的情况下，第一也可以被称为第二，类似
25 地，第二也可以被称为第一。取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

首先，对本说明书一个或多个实施例涉及的名词术语进行解释。

本地盘：是基于虚拟机所在物理机上的本地磁盘设备，提供高存储 IOPS (Input/Output Operations Per Second，每秒读写次数) 以及低读写时延的存储访问能力。

30 硬件卸载卡：是独立于主机 CPU 的，用于执行任务的处理卡。硬件卸载卡可以基于 GPU (graphics processing unit，图形处理器) /FPGA (Field Programmable Gate Array，可编程阵列逻辑) /ASIC (Application Specific Integrated Circuit，专用集成电路) /SOC (System on a Programmable Chip，片上系统) 等异构硬件实现，将任务转移到硬

件处理。在本申请中，硬件卸载卡可以包括虚拟化卸载、算法加速以及协议栈卸载等能力。

PCIe 总线：是一种高速串行计算机扩展总线，具有支持大带宽、高性能、低 I/O 引脚数和小物理占用空间的传输总线。

SSD：是建立在固态架构上的电子存储驱动器，SSD 内置有 NAND 和 NOR 闪存，用于存储非易失性数据。

HDD：是一种非易失性计算机存储设备，其中包含高速旋转的磁盘或磁盘，用于永久存储数据的辅助存储设备。

随着大数据分析、人工智能等技术的飞速发展，客户需要更加高性能、高可用、可扩展且灵活的存储能力。例如，一种存储任务处理的方案，是基于 virtio/vhost/vhost-user 的纯软件方案。virtio 是半虚拟化虚拟机监测器中位于设备之上的抽象层。Vhost 是虚拟主机。vhost-user 是 virtio 后端。virtio/vhost/vhost-user 的方案本质上是软件定义的虚拟设备，实现上较为灵活，通过实现软件接口的标准化，可以将不同设备类型的软硬件交互进行统一。但因为方案的标准化，而忽略了不同设备类型数据传输的特点，在一些大数据量传输的场景，效率比较低下。同时因为纯软件方案中虚拟化模拟及 IO 处理占用了比较多的 CPU 资源，导致性能无法与物理存储设备相媲美。

由此，在本说明书中，提供了一种存储系统，本说明书同时涉及一种存储方法，一种硬件卸载卡，以及一种计算机可读存储介质，在下面的实施例中逐一进行详细说明。

参见图 1，图 1 示出了根据本说明书一个实施例提供的一种存储系统的结构框图，所述存储系统 100 包括硬件卸载卡 102 以及存储设备 104。所述硬件卸载卡 102 与所述存储设备 104 对等地连接在主机 110 上。

其中，对等地连接表示硬件卸载卡 102 与存储设备 104 处于传输协议的相同层次。对等地连接的实施方式不限。例如，在实际应用中，为了支持硬件卸载卡与存储设备对等地连接在主机上，可以将硬件卸载卡和存储设备作为对等的硬件实体挂在主机的同一个 PCIe 转换器下。

所述硬件卸载卡 102，被配置为从所述主机 110 接收存储任务，执行所述存储任务，将所述存储任务对应的数据访问请求发送给所述存储设备 104。

所述存储任务，是指对存储设备的运行、使用、和/或与存储数据的访问等有关的任务。例如，可以包括对存储设备的健康监测、运维、数据读/写、数据加/解密、数据压缩、循环冗余校验、数据库算子等。其中，所述数据访问请求可以是虚拟机发送所述存储任务时，携带在所述存储任务中的信息，也可以是硬件卸载卡执行存储任务时产生的信息。所述存储任务可以由元数据来描述，该元数据用于携带表示存储任务内容的信息。在存储任务对应的数据访问请求是写请求的情况下，需要写入的数据不会直接放在元数据里一起带过来，而是携带了要写的数据的位置信息，相应地，对应生成的数据访问请求相应携带该位置信息。这样，存储设备可以根据数据访问请求对主机发起请求，以从对应位置获取要

写的数据。

所述数据访问请求，可以包括对存储设备中的存储数据的读和/或写请求。在所述数据访问请求为写请求的情况下，为了实现数控分离，所述数据访问请求并不携带要写入的数据，可以携带要写的数据在主机上的位置信息，从而存储设备能够直接从主机获取在该存储位置保存的存储数据。其中，硬件卸载卡将所述数据访问请求发送给存储设备的实施方式不限，例如，为了加速访问，硬件卸载卡可以将数据访问请求写入硬件卸载卡的内存，由存储设备从该内存中获取数据访问请求。具体地，所述硬件卸载卡 102，可以被配置为将所述数据访问请求保存在所述硬件卸载卡的内存。所述存储设备，被配置为通过 DMA 直接访问所述硬件卸载卡的内存来获取所述数据访问请求。所述数据访问请求携带了要访问的数据的地址。

所述存储设备 104，被配置为基于所述主机 110 与所述存储设备 104 之间的传输通道传输所述数据访问请求对应的存储数据。

其中，所述主机 110 与所述存储设备 104 之间的传输通道可以基于 PCIe (Peripheral Component Interconnect express，高速串行计算机扩展总线标准) 物理链路连接。所述存储设备可以理解为任一种或多种介质类型的一个或多个物理硬盘。

其中，硬件卸载卡 102 与存储设备 104 之间可以通过总线通道通信。例如，在硬件卸载卡执行存储任务时，某些场景下需要读写存储设备的存储数据，则所述硬件卸载卡 102，可以被配置为通过总线通道访问所述存储设备的存储数据。

该存储系统中，存储任务被卸载（该卸载可以理解为转移）到硬件卸载卡，利用硬件来加速执行存储任务，减少对主机 CPU 资源的占用，任务处理效率更高，并且由于存储设备通过对等的方式从硬件卸载卡获取数据访问请求，数据访问请求并不携带存储数据，相当于存储数据的传输与硬件卸载卡对存储任务的处理分离，针对存储设备的存储数据的传输，硬件卸载卡被旁路掉，从而存储设备直接与主机传输数据访问请求对应的存储数据，实现了数控分离的处理策略，数据无需通过硬件卸载卡来转发。由于大量存储数据的传输无需经历主机到硬件卸载卡再到物理存储设备的多次拷贝和流转，避免对硬件卸载卡的任务处理能力和资源处理能力的过高要求，减少硬件卸载卡的总线流量负担，从而达成接近于物理硬件级别的性能，能够实现存储控制面和数据面的加速，达到更高性能 IO 和时延。

在本说明书实施例提供的存储系统中，为了避免存储任务消耗主机 CPU 资源，通过硬件卸载的方式对存储任务处理以实现加速。由于并非所有的存储任务都适合硬件加速，例如控制面任务、面向特定场景需要的处理等，需要更多的灵活性。而对于固定执行的运算指令和访存指令、大批量的数据处理等适合硬件加速，因此，为了提高系统性能、使延迟指标满足系统要求，本说明书实施例中，将硬件卸载卡通过软硬协作的方式实现。

具体地，参见图 2，图 2 示出了根据本说明书另一个实施例提供的一种存储系统的结构框图，所述硬件卸载卡 102 包括可编程片上系统 1022 以及专用硬件 1024。

所述可编程片上系统 1022, 可以被配置为对所述存储任务中的软件子任务进行识别, 调用所述可编程片上系统上运行的软件处理逻辑对所述软件子任务进行处理。

其中, 所述可编程片上系统 1022 (即可编程 SOC), 可以运行控制逻辑来识别存储任务中的软件子任务, 在识别出软件子任务的情况下, 令所述可编程片上系统调用相应的软件处理逻辑来执行所述软件子任务。所述控制逻辑在可编程片上系统内可以表现为程序软件。其中, 所述软件处理逻辑具体可以根据实际应用场景下对软件子任务的处理需要灵活设置。例如, 所述可编程片上系统运行的软件处理逻辑包括: 访问请求错误处理逻辑和/或硬件运维处理逻辑。

另外, 在实际应用中, 由于为了支持硬件卸载卡与存储设备对等地连接在主机上, 硬件卸载卡和存储设备作为对等的硬件实体挂在主机的同一个 PCIe 转换器下。相应地, 如图 2 所示的结构框图, 所述可编程片上系统中可以包括存储设备点对点驱动软件。该存储设备点对点驱动软件也可以理解为 PCIe 点对点驱动软件。通过该存储设备点对点驱动软件, 能够实现存储设备通过 DMA 访问硬件卸载卡内存地址空间的能力。进而, 存储设备可以访问到硬件卸载卡处理后保存在内存中的数据访问请求。存储设备获得数据访问请求后, 通过对数据访问请求的格式解析, 可以从中获得要写的地址在主机内存中的地址, 或者, 从中获得要读的数据在存储设备中的地址, 达成存储设备直接访问数据所在的地址空间, 达到数据面的加速。

所述专用硬件 1024, 可以被配置为执行所述存储任务中的硬件子任务。

其中, 所述专用硬件 1024 的硬件处理逻辑具体可以根据实际应用场景下对硬件子任务的处理需要来设置。所述专用硬件 1024 可以根据场景需要采用任意专用加速硬件来实现。例如, 所述专用硬件 1024 可以表现为 ASIC/FPGA 等专用硬件。

在实际应用中, 所述专用硬件 1024 可以用来提供各种类型的硬件子任务的加速处理能力。例如, 如图 2 所示, 所述专用硬件 1024 可以包括存储加速处理模块, 被配置为对数据读/写任务加速、安全校验 (如密码校验等) 的加速等等。

在该实施例中, 对存储任务中的软件子任务进行识别, 使不适合硬件加速的部分对应识别为软件子任务, 把适合硬件加速的硬件子任务卸载到专用硬件处理, 把不适合硬件加速的部分卸载到可编程片上系统通过软件方式处理, 从而实现了通用的软硬件交互协作框架, 能够灵活配置存储任务为纯软件处理或者专用加速硬件处理。

本说明书实施例中, 可编程片上系统中运行的软件处理逻辑以及专用硬件中运行的硬件处理逻辑的具体实施方式不限, 具体可以根据适合软件或硬件执行的任务来设置。例如, 所述软件处理逻辑可以包括: 存储资源的池化处理、缓存加速处理、访问请求错误处理和/或硬件运维处理的逻辑。相应地, 如图 2 所示的结构框图, 所述可编程片上系统 1022 可以包括 10 错误处理&硬件运维处理模块。

具体地, 访问请求错误处理任务, 可以包括对异常访问请求 (如超时 10、错误 10、无

效 10) 的处理策略。例如，访问请求错误处理任务可以在后端存储设备硬件出现异常或者故障时，对访问该存储设备的访问请求进行拦截，防止访问到无效地址空间导致虚拟机的异常宕机。硬件运维处理任务，可以包括对异常物理硬件（如，存储设备存储故障，存储设备传输故障等异常）的运维策略。例如，硬件运维处理任务该模块可以对接到云运维中心，将异常物理硬件上报到运维中心和机房，进而实现对异常物理硬件的下线维修和上线处理。池化，是指将底层的存储设备的存储资源进行池化融合，形成一个存储的资源池，在池化之后按照用户需要的存储能力进行切分或者组合。缓存加速，是指将高速介质作为低速介质的缓存盘，为低速介质提供加速访问能力。比如支持 SSD 访问速度快，而 HDD 存储空间大、访问速度慢，可以采用 SSD 作为缓存，HDD 作为数据存储盘，解决 HDD 访问速度慢的问题。

由于软件子任务由可编程片上系统执行，因此，软件子任务对应的软件处理逻辑可以在可编程片上系统的软件程序中灵活设置。例如，健康监测、运维策略等软件处理逻辑的参数和策略等均可以在程序中自定义。另外，可编程片上系统还可以通过 PCIe 通道读取到物理盘数据，满足本地盘运维需求。

在实际应用中，存储设备的介质类型可以是多种多样的，为了支持多种介质类型的存储系统，本说明书一个或多个实施例中，结合上述软硬件协作思想，可以通过如下实施例实现支持多介质、数控分离的存储系统，以达到更高效的本地盘硬件加速。具体地，所述可编程片上系统 1022，还可以被配置为识别所述存储设备的介质类型，根据所述介质类型配置对应的交互规则，使所述数据访问请求根据所述交互规则生成。具体地，例如，所述专用硬件如图 2 所示，可以包括存储启动器。所述存储启动器，可以被配置为对接所述可编程片上系统 1022。在硬件卸载卡启动后，所述存储启动器会自动与所述可编程片上系统 1022 内部的存储目标器进行协商确定与后端存储设备的通信协议。所述可编程片上系统 1022 内部的存储目标器，可以识别和管理后端存储设备（包括识别存储设备的介质类型），在存储启动器进行后端存储设备的探测时，与存储启动器通过自动协商确定与存储设备交互需采用的通信协议。在确认好通信协议后，相应创建从存储启动器到存储目标器的传输通道。

例如，所述存储设备的介质类型可以包括 SCM 持久化介质，固态存储介质，机械硬盘存储介质等任一种或多种介质类型。例如，所述交互规则可以包括通信协议、设备驱动程序等任一种或多种与存储设备交互相关的规则。硬件卸载卡可以加载与交互规则对应的协议或者软件，从而硬件卸载卡可以与相应存储设备进行交互。

通过上述实施例提供的存储系统，能够同时支持多种类型本地盘场景，以及未来的本地盘形态创新，同时将协议处理和 10 处理下沉到专用加速硬件芯片实现，实现硬件级别的性能和延时。

需要说明的是，针对不同存储介质类型的存储设备，用于交互的通信协议可能不一样，

根据上述实施例使硬件卸载卡中的可编程片上系统上运行的软件可以对交互规则做灵活的配置，硬件卸载卡能够与不同类型的存储介质交互。但对于要读写的数据本身来说，并不受通信协议的限制，存储设备可以与主机直接传输。

根据上述实施例，虽然不同存储介质遵循的协议可能不同，报文格式不一样，但是本说明书实施例由可编程片上系统运行的软件来识别交互规则，把交互规则写入硬件卸载卡的配置，之后，硬件卸载卡就按照配置的规则去跟存储介质交互。这样，能够支持多种形态的本地盘，不同形态的硬盘都可以挂载在主机的 PCIe 总线上，一方面可以实现硬盘挂载的动态伸缩能力，另一方面可以实现盘的迭代和硬件卸载卡的迭代相互独立。例如，可以将存储虚拟化、存储协议处理以及后台管理任务等存储任务卸载到硬件卸载卡来处理，通过硬件卸载卡的软硬协作来提供处理能力，实现存储设备的线性扩展和可伸缩以及更高级的存储功能，不限于某一种特定存储介质，可以同时支持如 NVMe SSD，HDD 等多种介质类型，通用性高。

为了灵活地适应多租户等应用场景，所述专用硬件 1024，还可以被配置为基于虚拟设备仿真技术建立虚拟设备，所述虚拟设备，用于将所述存储设备的物理存储资源进行抽象，面向主机提供虚拟化的存储资源。具体地，如图 2 所示的存储系统的结构框图，基于所述专用硬件 1024 的虚拟设备仿真，所述硬件卸载卡 102，还可以包括基于虚拟设备仿真技术建立的多台虚拟设备。由于虚拟设备是基于硬件卸载卡的 IO 虚拟化能力仿真出来的虚拟硬件，因此，可以将单个物理的存储设备虚拟成多台虚拟设备，将虚拟设备挂载到主机的多个虚拟机中，可以实现云上多租户对物理设备的共享访问。

在该实施例中，在实现硬件 IO 的性能同时，在硬件卸载卡内部将存储设备的物理存储资源进行抽象，面向主机提供虚拟化的存储资源，使得前端虚拟盘和后端物理盘的形态是分离解耦的，支持后端物理盘形态的多种创新。例如：底层存储设备虽然是 HDD 的盘，但是通过虚拟设备对上层呈现出来的是 NVME 的盘。基于虚拟化的多台虚拟设备，能够满足云计算服务中的多租户场景，给云计算场景下的运维和迁移带来了极大便利。例如，可以基于 SR-IOV (Single Root I/O Virtualization, 单根 I/O 虚拟化) 技术实现虚拟设备仿真。SR-IOV 技术，是一种基于硬件的虚拟化解决方案，可提高性能和可伸缩性。SR-IOV 标准允许在虚拟机之间高效共享 PCIe (Peripheral Component Interconnect Express, 快速外设组件互连) 设备，并且由于是在硬件中实现的，因此可以获得能够与本机性能媲美的 I/O 性能。

考虑到主机在发送存储任务到硬件卸载卡时，若通过主机软件栈发送，则难以避免主机软件栈的一些处理，导致处理效率的低下，因此，本说明书一个或多个实施例中，如图 2 所示的存储系统的结构框图，所述主机 110 中包括虚拟机。其中，所述虚拟设备，被配置为根据与所述主机的虚拟机协商的内存地址，从所述内存地址获取所述存储任务。例如，所述虚拟设备与所述虚拟机可以预先商量好用于保存存储任务的内存地址，由元数据描述的存储任务可以保存在该内存地址对应的区域。由于虚拟机发送存储任务到与硬件卸载卡

预先商量好的内存存储区域，使得硬件卸载卡可以直接去该区域获取存储任务，绕过主机软件栈，提高了处理效率。

对于主机中的虚拟机来说，虚拟机中对存储设备的访问协议与后端物理存储设备是完全分离解耦的。不同虚拟机的 IO 协议格式可以不同，为了使硬件卸载卡中的专用硬件和/或可编程片上系统能够对不同虚拟机的存储任务的协议格式进行识别，如图 2 所示，所述专用硬件 1024，包括存储协议处理模块。所述存储协议处理模块，被配置为解析发送所述存储任务的虚拟机的通信协议格式，将所述存储任务的通信协议格式转换为通用通信协议格式，使进入所述可编程片上系统的任务为通用通信协议格式。

需要说明的是，所述专用硬件中的任意模块如存储协议处理模块可以合并专用硬件的其他模块中，也可以与其他模块分开设置，本说明书对此并不进行限制。例如，所述存储协议处理模块与其他模块分开设置，并从各个虚拟设备接收存储任务进行协议处理。

基于硬件卸载卡内部设置虚拟设备的实施例，在实际应用中为了满足多租户共享单台存储设备的存储资源的需要，可以按需将单台存储设备对应抽象化为多台所述虚拟设备，其中，不同虚拟设备对应主机中的不同虚拟机，多台所述虚拟机共享所述单台存储设备的存储资源。相应地，为了使主机的虚拟机只能访问规定的硬件资源，防止恶意 IO 访问入侵整个存储系统，所述可编程片上系统 1022，包括多租户共享任务处理模块，所述多租户共享任务处理模块，被配置为将共享存储资源的所述不同虚拟机的存储任务对应分配所述单台存储设备的不同存储区域，以及，对所述存储任务进行权限校验和访问地址隔离。

其中，所述权限校验可以是对租户的权限的校验，如访问权限，对数据的处理权限等等。所述访问地址隔离，是指将不同虚拟机对应的不同存储区域的访问地址隔离，避免发生恶意 IO 访问。

另外，为了提高存储设备与主机之间的数据传输效率，本说明书一个或多个实施例中，主机与存储设备之间的数据传输通过 DMA 直接访问虚拟机内存来避免内存拷贝等开销。具体地，所述主机中包括虚拟机，所述虚拟机内设置有内存。主机与存储设备之间传输的数据可以放在虚拟机的内存中。所述主机与所述存储设备之间的传输通道为 DMA 传输通道。所述存储设备，被配置为通过 DMA 直接访问所述虚拟机的内存来传输所述数据访问请求对应的存储数据。

下述结合附图 3，以主机内部组件的应用为例，对存储系统的实施方式进行进一步说明。图 3 示出了根据本说明书又一个实施例提供的主机的结构框图。如图 3 所示，所述主机内部组件可以包括：

主机基础运行环境，被配置为提供基础物理服务器运行的环境。

虚拟机管理器，被配置为对虚拟机的资源分配，生命周期等进行管理。例如，通过该虚拟机管理器，可以设置多个虚拟机共享主机硬件资源。

虚拟机内存映射管理，被配置为管理虚拟机内部的内存地址和主机上内存地址的映射

关系，使得虚拟机内部的应用软件在运行时不需要感知虚拟化层。

虚拟机，包括存储设备驱动软件和存储虚拟化驱动引擎。其中，存储设备驱动软件，是运行在虚拟机内部的硬件设备驱动程序，用于给虚拟机的内部应用提供访问存储硬件的能力。存储虚拟化驱动引擎，用于对接硬件卸载卡提供的虚拟仿真设备，使得虚拟机在访问存储硬件时不需感知底层为虚拟设备硬件或真实硬件。

为了使本说明书实施例提供的存储系统更加易于理解，下面，以图 4 示出的结合本说明书多个实施例提供的存储系统的云上多租户应用场景示意图进行说明。如图 4 所示，主机 110 是云计算平台上的任一云主机。硬件卸载卡 102 中的虚拟设备 C 是从 SCM 持久化介质的硬盘“盘 2”抽象得到的。虚拟设备 C 挂载在虚拟机 A 中，用于支持虚拟机 A 上租户对硬盘“盘 2”的存储资源的使用。硬件卸载卡 102 启动时，硬件卸载卡 102 内的存储启动器探测到“盘 2”，自动与可编程片上系统 1022 内部的存储目标器进行协商确定与“盘 2”的通信协议，通过配置 SCM 持久化介质的通信协议，建立了硬件卸载卡与“盘 2”的传输通道。

租户使用虚拟机 A 发出“向虚拟机的底层存储写入存储数据 B”的存储任务。虚拟机 A 将存储数据 B 在虚拟机内存中的地址携带在存储任务中，并将存储任务写到虚拟机与虚拟设备 C 预先商量好的一段内存空间。虚拟设备 C 从虚拟机的内存获得该存储任务后，将该存储任务发送给存储协议处理模块。由于该存储任务的消息格式遵循虚拟机 A 使用的通信协议，因此，存储协议处理模块将该存储任务转换为通用的通信协议格式。存储协议处理模块将通用的通信协议格式的存储任务发送给可编程片上系统。可编程片上系统识别出该存储任务中存在软件子任务“多租户共享任务”，调用多租户共享任务处理模块对该存储任务进行权限校验和访问地址隔离，确定存储数据 B 需要写入盘 2 的存储区 XX。由此，存储任务对应的数据访问请求为写请求“向盘 2 的存储区 XX 写入存储数据 B”，且该写请求携带了存储数据 B 在主机的主机内存中的地址。该写请求根据“盘 2”的通信协议生成，并保存在硬件卸载卡内存地址空间。盘 2 通过 DMA 访问硬件卸载卡内存地址空间，获得该写请求。盘 2 根据该写请求携带的存储数据 B 的内存地址，通过 DMA 直接访问虚拟机 A 内存获得存储数据 B，将存储数据 B 写入盘 2。

与上述存储系统实施例相对应，本说明书还提供了应用于硬件卸载卡的存储方法实施例，图 5 示出了本说明书一个实施例提供的一种存储方法的流程图。如图 5 所示，该方法包括：

步骤 502：从所述主机接收存储任务。

步骤 504：执行所述存储任务。

步骤 506：将所述存储任务对应的数据访问请求发送给所述存储设备，使所述存储设备基于主机与所述存储设备之间的传输通道，传输所述数据访问请求对应的存储数据。

根据该方法，存储任务被卸载到硬件卸载卡，利用硬件来加速执行，减少对主机 CPU

资源的占用，任务处理效率更高，并且由于存储设备通过对等的方式从硬件卸载卡获取数据访问请求，相当于存储数据的传输与硬件卸载卡对存储任务的控制逻辑的处理分离，存储设备直接与主机传输数据访问请求对应的存储数据，实现了数控分离的处理策略，数据无需通过硬件卸载卡来转发，达成接近于物理硬件级别的性能，能够实现更高性能 10 和 5 时延。

上述为本实施例的一种存储方法的示意性方案。需要说明的是，该存储方法的技术方案与上述的存储系统的技术方案属于同一构思，存储方法的技术方案未详细描述的细节内容，均可以参见上述存储系统的技术方案的描述。

例如，所述存储方法可以包括软件处理部分与硬件处理部分，软件处理部分对应于所述存储系统的可编程片上系统的处理，硬件处理部分对应于所述存储系统的专用硬件的处理，详细可以参见上述存储系统的技术方案的描述，在此不再详细赘述。

图 6 示出了根据本说明书一个实施例提供的一种硬件卸载卡 600 的结构框图。该硬件卸载卡 600 的部件包括但不限于存储器 610 和处理器 620。处理器 620 与存储器 610 通过总线 630 相连接，数据库 650 用于保存数据。

15 硬件卸载卡 600 还包括接入设备 640，接入设备 640 使得硬件卸载卡 600 能够经由一个或多个网络 660 通信。这些网络的示例包括公用交换电话网（PSTN）、局域网（LAN）、广域网（WAN）、个域网（PAN）或诸如因特网的通信网络的组合。接入设备 640 可以包括有线或无线的任何类型的网络接口（例如，网络接口卡（NIC））中的一个或多个，诸如 20 IEEE802.11 无线局域网（WLAN）无线接口、全球微波互联接入（Wi-MAX）接口、以太网接口、通用串行总线（USB）接口、蜂窝网络接口、蓝牙接口、近场通信（NFC）接口，等等。

在本说明书的一个实施例中，硬件卸载卡 600 的上述部件以及图 6 中未示出的其他部件也可以彼此相连接，例如通过总线。应当理解，图 6 所示的硬件卸载卡结构框图仅仅是出于示例的目的，而不是对本说明书范围的限制。本领域技术人员可以根据需要，增添或替换其他部件。

25 其中，处理器 620 用于执行如下计算机可执行指令，该计算机可执行指令被处理器执行时实现上述存储方法的步骤。

上述为本实施例的一种计算设备的示意性方案。需要说明的是，该计算设备的技术方案与上述的存储方法的技术方案属于同一构思，计算设备的技术方案未详细描述的细节内容，均可以参见上述存储方法的技术方案的描述。

30 本说明书一实施例还提供一种计算机可读存储介质，其存储有计算机可执行指令，该计算机可执行指令被处理器执行时实现上述存储方法的步骤。

上述为本实施例的一种计算机可读存储介质的示意性方案。需要说明的是，该存储介质的技术方案与上述的存储方法的技术方案属于同一构思，存储介质的技术方案未详细描述的细节内容，均可以参见上述存储方法的技术方案的描述。

本说明书一实施例还提供一种计算机程序，其中，当所述计算机程序在计算机中执行时，令计算机执行上述存储方法的步骤。

上述为本实施例的一种计算机程序的示意性方案。需要说明的是，该计算机程序的技术方案与上述的存储方法的技术方案属于同一构思，计算机程序的技术方案未详细描述的细节内容，均可以参见上述存储方法的技术方案的描述。

上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

所述计算机指令包括计算机程序代码，所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U 盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器 (ROM, Read-Only Memory)、随机存取存储器 (RAM, Random Access Memory)、电载波信号、电信信号以及软件分发介质等。需要说明的是，所述计算机可读介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减，例如在某些司法管辖区，根据立法和专利实践，计算机可读介质不包括电载波信号和电信信号。

需要说明的是，对于前述的各方法实施例，为了简便描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本说明书实施例并不受所描述的动作顺序的限制，因为依据本说明书实施例，某些步骤可以采用其它顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定都是本说明书实施例所必须的。

在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中未详述的部分，可以参见其它实施例的相关描述。

以上公开的本说明书优选实施例只是用于帮助阐述本说明书。可选实施例并没有详尽叙述所有的细节，也不限制该发明仅为所述的具体实施方式。显然，根据本说明书实施例的内容，可作很多的修改和变化。本说明书选取并具体描述这些实施例，是为了更好地解释本说明书实施例的原理和实际应用，从而使所属技术领域技术人员能很好地理解和利用本说明书。本说明书仅受权利要求书及其全部范围和等效物的限制。

权 利 要 求 书

- 1、一种存储系统，包括：硬件卸载卡以及存储设备，所述硬件卸载卡与所述存储设备对等地连接在主机上；
- 5 所述硬件卸载卡，被配置为从所述主机接收存储任务，执行所述存储任务，将所述存储任务对应的数据访问请求发送给所述存储设备；
- 所述存储设备，被配置为基于所述主机与所述存储设备之间的传输通道，传输所述数据访问请求对应的存储数据。
- 10 2、根据权利要求 1 所述的存储系统，所述硬件卸载卡包括可编程片上系统以及专用硬件；
- 所述可编程片上系统，被配置为对所述存储任务中的软件子任务进行识别，调用所述可编程片上系统上运行的软件处理逻辑对所述软件子任务进行处理；
- 所述专用硬件，被配置为执行所述存储任务中的硬件子任务。
- 15 3、根据权利要求 2 所述的系统，所述可编程片上系统，还被配置为识别所述存储设备的介质类型，根据所述介质类型配置对应的交互规则，使所述数据访问请求根据所述交互规则生成。
- 20 4、根据权利要求 2 所述的系统，所述专用硬件，还被配置为基于虚拟设备仿真技术建立虚拟设备，所述虚拟设备，用于将所述存储设备的物理存储资源进行抽象，面向主机提供虚拟化的存储资源。
- 5、根据权利要求 4 所述的系统，所述虚拟设备，被配置为根据与所述主机的虚拟机协商的内存地址，从所述内存地址获取所述存储任务。
- 25 6、根据权利要求 5 所述的系统，所述专用硬件包括存储协议处理模块，所述存储协议处理模块，被配置为解析发送所述存储任务的虚拟机的通信协议格式，将所述存储任务的通信协议格式转换为通用通信协议格式，使进入所述可编程片上系统的任务为通用通信协议格式。
- 30 7、根据权利要求 4 所述的系统，单台存储设备对应抽象化为多台所述虚拟设备，其中，不同虚拟设备对应主机中的不同虚拟机，多台所述虚拟机共享所述单台存储设备的存

储资源；

所述可编程片上系统包括多租户共享任务处理模块，所述多租户共享任务处理模块，被配置为将所述多台虚拟机的存储任务分别对应分配所述单台存储设备的不同存储区域，以及，对所述存储任务进行权限校验和访问地址隔离。

5

8、根据权利要求 1 所述的系统，所述主机中包括虚拟机，所述虚拟机内设置有内存；所述虚拟机的内存与所述存储设备之间的传输通道为 DMA 传输通道；

所述存储设备，被配置为通过 DMA 直接访问所述虚拟机的内存来传输所述数据访问请求对应的存储数据。

10

9、根据权利要求 1 所述的系统，所述硬件卸载卡，被配置为将所述数据访问请求保存在所述硬件卸载卡的内存；

所述存储设备，被配置为通过 DMA 直接访问所述硬件卸载卡的内存来获取所述数据访问请求。

15

10、根据权利要求 2 所述的系统，所述可编程片上系统运行的软件处理逻辑包括：存储资源的池化处理、缓存加速处理、访问请求错误处理和/或硬件运维处理的逻辑。

20

11、一种存储方法，应用于硬件卸载卡，所述硬件卸载卡与存储设备对等地连接在主机上，所述方法包括：

从所述主机接收存储任务；

执行所述存储任务；

将所述存储任务对应的数据访问请求发送给所述存储设备，使所述存储设备基于主机与所述存储设备之间的传输通道，传输所述数据访问请求对应的存储数据。

25

12、一种硬件卸载卡，包括：

存储器和处理器；

所述存储器用于存储计算机可执行指令，所述处理器用于执行所述计算机可执行指令，该计算机可执行指令被处理器执行时实现权利要求 11 所述存储方法的步骤。

30

13、一种计算机可读存储介质，其存储有计算机可执行指令，该计算机可执行指令被处理器执行时实现权利要求 11 所述存储方法的步骤。