(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
2024 年 7 月 4 日 (04.07.2024)

WIPO | PCT

(10) 国际公布号
WO 2024/140375 A1

(51) 国际专利分类号:
H04J 3/16 (2006.01)

(21) 国际申请号: PCT/CN2023/140344

(22) 国际申请日: 2023 年 12 月 20 日 (20.12.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202211691501.X 2022年12月27日 (27.12.2022) 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: 陈瑜芳 **(CHEN, Yufang)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 刘晓妮 **(LIU, Xiaoni)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 吴俊宏 **(WU, Junhong)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: 北京中博世达专利商标代理有限公司 **(BEIJING ZBSD PATENT & TRADEMARK AGENT LTD.)**; 中国北京市海淀区交大东路31号11号楼8层, Beijing 100044 (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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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, 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, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF,

(54) **Title:** STORAGE DEVICE, AND DATA COMMUNICATION METHOD AND SYSTEM

(54) 发明名称: 一种存储设备、数据通信方法以及系统

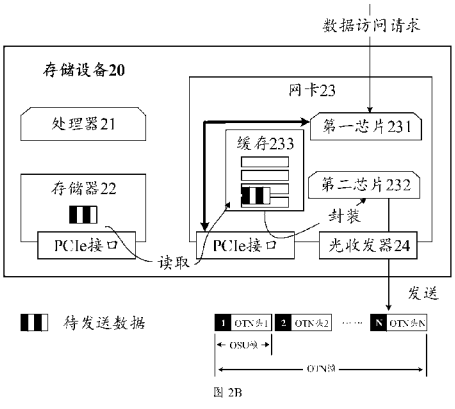

图 2B

20	Storage device
21	Processor
22	Memory
23	Network interface card
24	Optical transceiver
231	First chip
232	Second chip
233	Cache
AA	Data access request
BB	Encapsulate
CC	PCIe interface
DD	Read
EE	Send
FF	Data to be sent
GG	OTN head
HH	OSU frame
II	OTN frame

(57) Abstract: The present application relates to the technical field of optical communications, and provides a storage device, and a data communication method and system. A network interface card in a storage device implements OTN encapsulation of data, so that an OTN frame where the data is mapped can be generated without transmitting and encapsulating the data by a plurality of devices, thereby reducing a transmission delay of the data between a memory and an optical transceiver, and improving the data communication efficiency. Moreover, since the data does not need to be processed by an Ethernet switch, the storage device can encapsulate the data into a protocol stack used by the OTN frame without using an Ethernet protocol, thereby reducing encapsulation procedures required for generation of the OTN frame and reducing the amount of data included in the OTN frame, and facilitating further improvement of the data communication efficiency.

(57) 摘要: 本申请提供一种存储设备、数据通信方法以及系统，涉及光通信技术领域。存储设备中的网卡实现数据的OTN封装，使得数据不需经由多个设备的传输和封装即可生成映射有该数据的OTN帧，减少了数据从存储器到光收发器之间的传输时延，提高了数据通信效率。而且，由于数据无需经由以太交换机进行处理，因此，存储设备将数据封装为OTN帧所采用的协议栈也无需使用以太网协议，减少了生成OTN帧所需的封装流程以及OTN帧中所包含的数据量，有利于进一步提高数据的通信效率。

WO 2024/140375 A1

[见续页]

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))。

一种存储设备、数据通信方法以及系统

本申请要求于2022年12月27日提交国家知识产权局、申请号为202211691501.X、申请名称为“一种存储设备、数据通信方法以及系统”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

本申请涉及光通信技术领域，尤其涉及一种存储设备、数据通信方法以及系统。

背景技术

在数据通信网络（Data Communication Network，DCN）中，光传送网络（Optical Transport Network，OTN）由于其高带宽，大容量，高可靠，低时延等特性，已经成为传送网采用的主流技术，广泛应用于骨干、城域、核心及汇聚等网络。OTN帧用于承载各种业务数据，并提供丰富的管理和监控功能。

目前，服务器中部署有远程直接存储访问（Remote Direct Memory Access，RDMA）应用，RDMA应用调用存储设备中的以太网卡以读取存储设备中的业务数据，并由以太网卡将该业务数据进行封装后发送到以太交换机，该以太交换机将封装后的数据转发到OTN设备，OTN设备为封装后的数据添加OTN帧头生成OTN帧，并将该OTN帧发送到OTN中其他的设备。可知，业务数据需经过多次传输和封装才能确定OTN帧，在业务数据的数据量较大的情况下，OTN中端到端设备之间的通信效率较低。

发明内容

本申请提供了一种存储设备、数据通信方法以及系统，解决了业务数据需经不同的设备进行多次传输和封装才能确定OTN帧的过程，提高了OTN中端到端设备之间的通信效率。

第一方面，本申请实施例提供了一种存储设备。该存储设备包括：处理器、存储器和网卡，网卡包括数据处理芯片、存储介质和光收发器。存储器，用于存储处理器写入的待发送数据。数据处理芯片，用于将待发送数据从存储器写入存储介质。数据处理芯片，还用于读取存储介质中的待发送数据，并将待发送数据映射到数据处理芯片生成的OTN帧的净荷区。光收发器，用于发送OTN帧。

在本实施例中，存储设备中的网卡完成了数据的OTN封装，使得数据不需经由多个设备的传输和封装即可生成映射有该数据的OTN帧，减少了数据从存储器到光收发器之间的传输时延，提高了数据通信效率。而且，由于数据无需经由以太交换机进行处理，因此，存储设备将数据封装为OTN帧所采用的协议栈也无需使用以太网（EtherNet，ETH）协议，减少了生成OTN帧所需的封装流程以及OTN帧中所包含的数据量，有利于进一步提高数据的通信效率。

举例来说，由于网卡可直接将数据不经过以太封装，直接封装成OTN帧，且网卡支持插在端侧存储设备上，因此，存储设备可直接在端侧输出OTN帧，不经过以太交换机的转发，实现了光通信网络中端侧设备到端侧设备的硬管道传输能力，避免了数据通信过程中的丢包，提高了数据的通信效率。

作为一种可选的示例，光收发器，还用于接收目标存储设备的数据写响应。该数据写响应用于指示待发送数据已写入目标存储设备。

举例来说，在光收发器接收到目标存储设备的数据写响应后，存储设备确定数据的本次传输结束，避免了存储设备为该次数据传输预留硬件资源（如计算资源或存储资源）导致的资源消耗，有利于存储设备将有限的硬件资源用于执行其他业务。

在一些可能的情况中，该光收发器还可用于：接收光通信网络中的其他光网络设备发送的OTN帧，并对解析该OTN帧后的数据进行存储。

在一种可选的实现方式中，存储器，还用于维护至少一个远程直接存储访问（remote direct memory access，RDMA）发送队列，该至少一个RDMA发送队列包括第一发送队列，第一发送队列中存储有一个或多个数据的工作队列元素（work queue element，WQE），该一个或多个数据包括前述的待发送数据。光收发器，还用于提供多个OTN通道，其中，一个OTN通道用于传输一个RDMA发送队列对应的数据。前述的数据处理芯片还用于：从第一发送队列中读取待发送数据的WQE，并为该WQE和多个OTN通道中的第一OTN通道建立映射关系。其中，该映射关系用于指示：待发送数据能够通过第一OTN通道进行传输。

在本实施例中，存储设备中的RDMA队列可和光收发器提供的OTN通道建立映射关系，使得不同RDMA队列的数据经由不同的OTN通道进行传输，其中，映射关系是网卡基于RDMA队列中记录

的数据的 WQE 来建立的，避免了光收发器将数据对应的 OTN 帧发送到与数据的 WQE 不匹配的 OTN 通道，提高了数据通信的准确性。而且，在后续的其他数据的通信过程中，若网卡已经建立了该其他数据的 WQE 和第一 OTN 通道的映射关系，则网卡可复用该映射关系，以将映射有该其他数据的 OTN 帧通过第一 OTN 通道进行传输，进一步提高了光通信网络中的数据通信效率。

在另一种可选的实现方式中，数据处理芯片包括：第一芯片和第二芯片。其中，第一芯片，用于从第一发送队列中读取待发送数据的 WQE。以及，第一芯片还用于：根据 WQE 指示的源地址，将待发送数据从存储器写入存储介质。第二芯片，用于：为 WQE 和多个 OTN 通道中的第一 OTN 通道建立映射关系。以及，第二芯片还用于：读取存储介质中的待发送数据，并将待发送数据映射到第二芯片生成的 OTN 帧的净荷区。

作为一种可行的示例，前述的第一芯片可以是数据处理单元（data processing unit，DPU），第二芯片可以是用于执行光线路侧报文处理的光芯片（或称光处理芯片、OTN 芯片等）。

可选的，第一芯片用于读取 RDMA 发送队列中的 WQE，以及从存储器中读取数据可以包括：第一芯片根据 WQE 指示的源地址，将待发送数据从存储器写入多个队列中第一队列对应的存储空间；这里的多个队列是由网卡包括的存储介质所维护的。以及，第二芯片用于为数据的 WQE 与 OTN 通道建立映射关系，可包括：第二芯片为前述第一队列与第一 OTN 通道建立映射关系。

在本实施例中，网卡中不同的芯片用于实现不同的功能，第一芯片实现网卡和应用层的交互，第二芯片实现网卡和光通信网络的交互，因此，由网卡中不同芯片之间进行协调即可实现数据从存储器至光通信网络中的硬管道传输，有利于提高数据的通信效率。

可选的，第二芯片，具体用于：将待发送数据映射到多个光业务单元（Optical Service Unit，OSU）帧。待发送数据承载于这多个 OSU 帧的净荷区。以及，第二芯片，还具体用于将多个 OSU 帧映射到 OTN 帧。

在本实施例中，第二芯片可以将数据映射到不同 OSU 帧的净荷区，使得待发送数据可以以更细的时隙颗粒度来进行数据通信。且 OSU 技术从一开始就考虑无损调整的需求，和 OTN 通信不存在兼容的问题，使得待发送数据的通信过程能够支持更大的无损带宽调整范围，有利于提高数据通信效率。这里的无损带宽调整包括：带宽增加、带宽减少和带宽回退中至少一种。其中的带宽回退用于指示出现问题后恢复原始状态的操作。

可选的，第二芯片，还用于判断第一队列的数据流速率是否大于或等于设定的速率阈值。数据流速率为单位时间内，第一芯片向第一队列中写入的数据量。若数据流速率大于或等于设定的速率阈值，则第二芯片还用于指示第一芯片调低向第一队列写数据的数据流速率。

在本实施例中，在第二芯片处理数据的预期速度过大的情况下，如第二芯片的数据流速度大于或等于设定的速率阈值，则第二芯片可以指示第一芯片调低向第一队列写数据的数据流速率，从而降低第二芯片在单位时间内所要处理（如封装）的数据量，以减小第二芯片的通信负荷，有利于避免第二芯片的网络丢包，提高光通信网络的通信性能。

可选的，处理器用于获取数据访问请求，数据访问请求用于请求待发送数据。处理器还用于判断待发送数据的数据量大于或等于设定的阈值。数据处理芯片具体用于：若待发送数据的数据量大于或等于设定的阈值，将待发送数据从存储器写入存储介质。

在本实施例中，在处理器确定待发送数据的数据量大于或等于设定的阈值的情况下，数据处理芯片才将待发送数据从存储设备的存储器写入网卡中的存储介质。即在大数据量传输的应用场景中，网卡对这些待发送数据进行封装，并通过光收发器来传输 OTN 帧，减少这些数据经由多次传输和以太封装导致的通信时延，有利于提高大数据量通信场景的数据通信效率。

第二方面，本申请实施例提供了一种数据通信方法。该数据通信方法由存储设备执行，该存储设备包括：处理器、存储器和网卡，该存储器用于存储处理器写入的待发送数据，该网卡包括数据处理芯片、存储介质和光收发器。本实施例提供的数据通信方法包括：首先，数据处理芯片将待发送数据从存储器写入存储介质。其次，数据处理芯片读取存储介质中的待发送数据，并将待发送数据映射到数据处理芯片生成的 OTN 帧的净荷区。最后，光收发器发送 OTN 帧。

可选的，存储器还用于维护至少一个 RDMA 发送队列，该至少一个 RDMA 发送队列包括第一发送队列，该第一发送队列中存储有一个或多个数据的工作队列元素 WQE，该一个或多个数据包括前述的

待发送数据。光收发器用于提供多个OTN通道，其中，一个OTN通道用于传输一个RDMA发送队列对应的数据。数据处理芯片读取存储介质中的待发送数据，包括：数据处理芯片从第一发送队列中读取待发送数据的WQE，并根据WQE指示的源地址，将待发送数据从存储器写入存储介质。以及，在光收发器发送OTN帧之前，本实施例提供的数据通信方法还包括：数据处理芯片为前述的WQE和多个OTN通道中的第一OTN通道建立映射关系。其中，映射关系用于指示：待发送数据能够通过第一OTN通道进行传输。

可选的，数据处理芯片包括：第一芯片和第二芯片。数据处理芯片从第一发送队列中读取待发送数据的WQE，并根据WQE指示的源地址，将待发送数据从存储器写入存储介质，包括：第一芯片从第一发送队列中读取待发送数据的WQE，并根据WQE指示的源地址，将待发送数据从存储器写入存储介质。数据处理芯片为WQE和多个OTN通道中的第一OTN通道建立映射关系，包括：第二芯片为WQE和多个OTN通道中的第一OTN通道建立映射关系。数据处理芯片将待发送数据映射到数据处理芯片生成的OTN帧的净荷区，包括：第二芯片读取存储介质中的待发送数据，并将待发送数据映射到第二芯片生成的OTN帧的净荷区。

可选的，存储介质维护有多个队列，多个队列包括第一队列，第一队列对应的存储空间用于存储待发送数据。第二芯片为WQE和多个OTN通道中的第一OTN通道建立映射关系，包括：第二芯片为第一队列与第一OTN通道建立映射关系。

可选的，本实施例提供的数据通信方法还包括：第二芯片判断第一队列的数据流速率是否大于或等于设定的速率阈值。该数据流速率为单位时间内，第一芯片向第一队列中写入的数据量。以及，若数据流速率大于或等于设定的速率阈值，则第二芯片指示第一芯片调低向第一队列写数据的数据流速率。

可选的，数据处理芯片将待发送数据映射到数据处理芯片生成的光传送网络OTN帧的净荷区，包括：第一步，第二芯片将待发送数据映射到多个OSU帧。待发送数据承载于多个OSU帧的净荷区。第二步，第二芯片将多个OSU帧映射到OTN帧。

可选的，在数据处理芯片将待发送数据从存储器写入存储介质之前，本实施例提供的数据通信方法还包括：处理器获取数据访问请求，该数据访问请求用于请求待发送数据。处理器判断待发送数据的数据量大于或等于设定的阈值，若待发送数据的数据量大于或等于设定的阈值，数据处理芯片将待发送数据从存储器写入存储介质。

可选的，本实施例提供的数据通信方法还包括：光收发器接收目标存储设备的数据写响应，数据写响应用于指示待发送数据已写入目标存储设备。

第三方面，本申请实施例提供了一种光通信系统。该光通信系统包括：存储设备和光网络设备。存储设备包括：处理器、存储器和网卡，网卡包括数据处理芯片、存储介质和光收发器。存储器，用于存储处理器写入的待发送数据。数据处理芯片，用于将待发送数据从存储器写入存储介质。数据处理芯片，还用于读取存储介质中的待发送数据，并将待发送数据映射到数据处理芯片生成的OTN帧的净荷区。光收发器，用于向前述的光网络设备发送OTN帧。

本实施例提供的光通信系统可以用于实现第一方面中任一方式的存储设备的功能，也能实现相应的有益效果，在此不予赘述。

第四方面，本申请实施例提供了一种计算机可读存储介质。该计算机可读存储介质包括：计算机软件指令。当计算机软件指令在存储设备中运行时，使得存储设备执行第二方面中任一种可能的实现方式提供的方法。

第五方面，本申请实施例提供了一种计算机程序产品。该计算机程序产品在存储设备上运行时，使得存储设备执行第二方面中任一种可能的实现方式提供的方法。

第二方面至第五方面中任意一种可能的实现方式的有益效果，可以参考第一方面或第一方面任意一种可能的实现方式的描述，在此不予赘述。本申请在上述各方面提供的实现方式的基础上，还可以进行进一步组合以提供更多实现方式。

附图说明

图1为本申请实施例提供的一种光通信系统的结构示意图；

图2A为本申请实施例提供的一种存储设备的结构示意图一；

图2B为本申请实施例提供的一种存储设备的结构示意图二；

图 3 为本申请实施例提供的一种数据通信方法的流程示意图一；

图 4 为本申请实施例提供的协议栈的结构示意图；

图 5 为本申请实施例提供的一种数据通信方法的流程示意图二。

具体实施方式

本申请提供了一种存储设备，该存储设备包括：处理器、存储器和网卡，网卡包括数据处理芯片、存储介质和光收发器。存储器存储处理器写入的待发送数据。数据处理芯片将待发送数据从存储器写入存储介质。以及，数据处理芯片还读取存储介质中的待发送数据，并将待发送数据映射到数据处理芯片生成的 OTN 帧的净荷区。光收发器发送该 OTN 帧。在本实施例中，数据的 OTN 封装从 OTN 中的端侧设备调整到了存储设备中的网卡，使得数据不需经由多个设备的传输和封装即可生成映射有该数据的 OTN 帧，减少了数据从存储器到光收发器之间的传输时延，提高了数据通信效率。而且，由于数据无需经由以太交换机进行处理，因此，存储设备将数据封装为 OTN 帧所采用的协议栈也无需使用以太网协议，减少了生成 OTN 帧所需的封装流程以及 OTN 帧中所包含的数据量，有利于进一步提高数据的通信效率。

举例来说，由于网卡可直接将数据不经过以太封装，直接封装成 OTN 帧，且网卡支持插在端侧存储设备上，因此，存储设备可直接在端侧输出 OTN 帧，不经过以太交换机的转发，实现了光通信网络中端侧设备到端侧设备的硬管道传输能力，避免了数据通信过程中的丢包，提高了数据的通信效率。

为了下述各实施例的描述清楚简洁，首先给出相关技术的简要介绍。

RDMA：通过网络把资料或数据直接传入计算机的存储区，将数据从一个系统快速移动到远程系统存储器中，而不对操作系统造成任何影响，这样就不需要用到计算机的处理功能。它消除了外部存储器复制和上下文切换的开销，因而能解放内存带宽和 CPU 周期，以改进应用系统性能。

基于融合以太网的第二版 RDMA（RDMA over Converged Ethernet version 2，RoCEv2）协议是一种基于用户数据报协议（User Datagram Protocol，UDP）的协议，将无限带宽（InfiniBand，IB）协议报文封装在 UDP 报文中在以太网上传送，存储设备通过网卡将待传输的数据发送给以太交换机汇聚后通过 OTN 网络传输。

图 1 为本申请实施例提供的一种光通信系统的结构示意图，该光通信系统包括：数据中心（data center，DC）1 和数据中心 2，该数据中心 2 与该数据中心 1 通过光纤传输数据并实现通信。

不同的数据中心可以部署在相同或不同的城市。数据中心可以包括：服务器、以太交换机、光网络设备和存储设备。如图 1 所示，数据中心 1 包括：服务器 111、以太交换机 21、光网络设备 31 和存储设备 121，数据中心 2 包括：服务器 112、以太交换机 22、光网络设备 32 和存储设备 122。

以数据中心 1 为例，对数据中心包括的各硬件设备进行说明。以太交换机 21 可以是路由转发设备，例如，路由转发设备可以是路由器或交换机等。光网络设备 31 可以是光通信系统中通过光传输介质（如光纤）等传输光信号的设备。服务器 111 可以是应用服务器或认证授权服务器。服务器 111 可以提供视频服务、游戏服务、消息服务、音乐服务、认证授权服务等。在一种示例中，可以将多个服务的功能集成在服务器 111 上，例如，游戏服务和音乐服务可以部署在服务器 111 上。在另一种示例中，还可以是服务器 111 上集成了部分服务的功能，例如，服务器 111 上部署了游戏服务的部分服务和视频服务的部分服务。服务器 111 还可以利用虚拟化技术提供多个虚拟机，由虚拟机提供各项服务。本申请实施例对服务的部署形态不予限定。

存储设备 121 可以包括：处理器、存储器和网卡等设备，如图 2A 所示，图 2A 为本申请实施例提供的一种存储设备的结构示意图一，图 2A 所示的存储设备 121 可以是一个集中式存储系统。集中式存储系统的特点是有一个统一的入口，所有从外部设备来的数据都要经过这个入口，这个入口就是集中式存储系统的引擎。引擎是集中式存储系统中最为核心的部件，许多存储系统的高级功能都在其中实现。

引擎中可以有一个或多个控制器，图 2A 以引擎包含一个控制器为例予以说明。在一种可能的示例中，若引擎具有多个控制器，任意两个控制器之间可以具有镜像通道，实现任意两个控制器互为备份的功能，从而避免硬件故障导致整个存储设备 121 的不可用。应理解，若引擎包括多个控制器，则引擎也可以称为存储设备 121 的阵列控制器。

引擎还包含前端接口 1211 和后端接口 1214，其中前端接口 1211 用于与计算设备通信，从而为计算设备提供数据访问服务。而后端接口 1214 用于与硬盘通信，以扩充存储设备 121 的容量。通过后端

接口 1214，引擎可以连接更多的硬盘，从而形成一个非常大的存储资源池。

在硬件上，如图 2A 所示，控制器至少包括处理器 1212、内存 1213。处理器 1212 是一个中央处理单元（central processing unit，CPU），用于处理来自存储设备 121 外部（服务器或者其他存储系统）的数据访问请求，也用于处理存储设备 121 内部生成的请求。示例性的，处理器 1212 通过前端接口 1211 接收计算设备发送的写数据请求时，会将这些写数据请求中的数据暂时保存在内存 1213 中。当内存 1213 中的数据总量达到一定阈值时，处理器 1212 通过后端端口将内存 1213 中存储的数据发送给机械硬盘 1221、机械硬盘 1222、固态硬盘（solid state drive，SSD）1223 或其他硬盘 1224 中至少一个硬盘进行持久化存储。

内存 1213 是指与处理器直接交换数据的内部存储器，它可以随时读写数据，而且速度很快，作为操作系统或其他正在运行中的程序的临时数据存储器。内存包括至少两种存储器，例如内存既可以是随机存取存储器，也可以是只读存储器（read only memory，ROM）。举例来说，随机存取存储器是 DRAM，或者 SCM。DRAM 是一种半导体存储器，与大部分随机存取存储器（random access memory，RAM）一样，属于一种易失性存储器（volatile memory）设备。然而，DRAM 和 SCM 在本实施例中只是示例性的说明，内存还可以包括其他随机存取存储器，例如静态随机存取存储器（static random access memory，SRAM）等。而对于只读存储器，举例来说，可以是可编程只读存储器（programmable read only memory，PROM）、可抹除可编程只读存储器（erasable programmable read only memory，EPROM）等。

另外，内存 1213 还可以是双列直插式存储器模块或双线存储器模块（dual in-line memory module，DIMM），即由动态随机存取存储器（DRAM）组成的模块，还可以是 SSD。实际应用中，控制器中可配置多个内存 1213，以及不同类型的内存 1213。本实施例不对内存 1213 的数量和类型进行限定。此外，可对内存 1213 进行配置使其具有保电功能。保电功能是指系统发生掉电又重新上电时，内存 1213 中存储的数据也不会丢失。具有保电功能的内存被称为非易失性存储器。

内存 1213 中存储有软件程序，处理器 1212 运行内存 1213 中的软件程序可实现对硬盘的管理。例如将硬盘抽象化为存储资源池，并将存储资源池以逻辑单元号（logical unit number，LUN）的形式提供给服务器使用等。这里的 LUN 其实就是在服务器上看到的硬盘。当然，一些集中式存储系统本身也是文件服务器，可以为服务器提供共享文件服务。

如图 2A 所示，在该系统中，引擎可以不具有硬盘槽位，硬盘需要放置在硬盘框中，后端接口 1214 与硬盘框通信。后端接口 1214 以适配卡的形态存在于引擎中，一个引擎上可以同时使用两个或两个以上后端接口 1214 来连接多个硬盘框。或者，适配卡也可以集成在主板上，此时适配卡可通过外围部件互连标准（Peripheral Component Interconnect Express，PCIe）总线与处理器 1212 通信。

需要说明的是，图 2A 中只示出了一个引擎，然而在实际应用中，存储系统中可包含两个或两个以上引擎，多个引擎之间做冗余或者负载均衡。

硬盘框包括控制单元 1225 和若干个硬盘。控制单元 1225 可具有多种形态。一种情况下，硬盘框属于智能盘框，如图 2A 所示，控制单元 1225 包括 CPU 和内存。CPU 用于执行地址转换以及读写数据等操作。内存用于临时存储将要写入硬盘的数据，或者从硬盘读取出来将要发送给控制器的数据。另一种情况下，控制单元 1225 是一个可编程的电子部件，例如 DPU。DPU 具有 CPU 的通用性和可编程性，但更具有专用性，可以在网络数据包，存储请求或分析请求上高效运行。DPU 通过较大程度的并行性（需要处理大量请求）与 CPU 区别开来。可选的，这里的 DPU 也可以替换成图形处理单元（graphics processing unit，GPU）、嵌入式神经网络处理器（neural-network processing units，NPU）等处理芯片。通常情况下，控制单元 1225 的数量可以是一个，也可以是两个或两个以上。控制单元 1225 的功能可以卸载到网卡 1226 上。换言之，在该种实施方式中，硬盘框内部不具有控制单元 1225，而是由网卡 1226 来完成数据读写、地址转换以及其他计算功能。此时，网卡 1226 是一个智能网卡。它可以包含 CPU 和内存。CPU 用于执行地址转换以及读写数据等操作。内存用于临时存储将要写入硬盘的数据，或者从硬盘读取出来将要发送给控制器的数据。也可以是一个可编程的电子部件，例如 DPU。硬盘框中的网卡 1226 和硬盘之间没有归属关系，网卡 1226 可访问该硬盘框中任意一个硬盘（如图 2A 所示出的机械硬盘 1221、机械硬盘 1222、固态硬盘 1223 和其他硬盘 1224），因此在存储空间不足时扩展硬盘会较为便捷。

按照引擎与硬盘框之间通信协议的类型，硬盘框可能是串行连接的小型计算机系统接口（serial

attached small computer system interface，SAS）的硬盘框，也可能是 NVMe（Non-Volatile Memory express）硬盘框以及其他类型的硬盘框。SAS 硬盘框，采用 SAS3.0 协议，每个框支持 25 块 SAS 硬盘。引擎通过板载 SAS 接口或者 SAS 接口模块与硬盘框连接。NVMe 硬盘框，更像一个完整的计算机系统，NVMe 硬盘插在 NVMe 硬盘框内。NVMe 硬盘框再通过 RDMA 端口与引擎连接。

在一种可选的实现方式中，存储设备 121 为盘控一体的集中式存储系统，存储设备 121 中不具有上述的硬盘框，引擎用于管理通过硬盘槽连接的多个硬盘。硬盘槽的功能可以由后端接口 1214 实现。示例性的，存储设备 121 可以是指一个存储阵列，如存储介质全部是闪存的全闪存存储阵列。

作为一种可能的示例，网卡 1226 可以包括：数据处理芯片、光收发器和存储介质等。针对于以上的存储设备 121，本示例提供另一种实现方式，如图 2B 所示，图 2B 为本申请实施例提供的一种存储设备的结构示意图二，该存储设备 20 包括：处理器 21、存储器 22 和网卡 23，三者之间可以通过 PCIe 接口进行通信连接。可选的，通信连接的方式也可以采用其他类型的接口或总线等，例如是通用串行总线（universal serial bus，USB）、计算快速互联（compute express link，CXL）或者其他类型的总线或接口等，本申请对此不予限定。

关于处理器 21 和存储器 22 的具体实现方式可以参照图 2A 的相关描述，不予赘述。

如图 2B 所示，网卡 23 包括：第一芯片 231、第二芯片 232、缓存 233 和光收发器 24。

示例性的，该第一芯片 231 可以是 DPU 或者其他具有数据处理功能的处理器等。例如，第一芯片 231 用于将存储在存储器 22 中的数据写入缓存 233。

第二芯片 232 可以是成帧芯片，该第二芯片 232 用于将缓存 233 中存储的数据映射到 OTN 帧的净荷区。

缓存 233 可以用于临时存储第一芯片 231 读取的数据，或者用于临时存储光收发器 24 接收到的数据等。在一种可能的示例中，缓存 233 可以是指 cache。在另一种可能的示例中，缓存 233 也可以替换为其他类型的存储介质，例如 DRAM、SCM、机械硬盘或者 SSD 等等。

光收发器 24 用于发送 OTN 帧到光通信系统中的其他光网络设备，以及，接收来自其他光网络设备发送的 OTN 帧或者其他光信号等。

在一些可能的示例中，第一芯片 231 和第二芯片 232 所包括的逻辑电路可以集成在一个印制电路板（printed circuit boards，PCB）。因此，第一芯片 231 和第二芯片 232 也可合称为数据处理芯片、数据处理模组、数据处理装置或者数据处理单元等。在本申请的后续实施例中，以数据处理芯片为例对第一芯片 231 和第二芯片 232 的功能进行详细说明。

值得注意的是，以上图 2A 和图 2B 仅为本申请实施例提供的两种存储设备的可能实现方式，两种存储设备包括的各设备可以互换，在不同图示中所采用的名字不同，但是均可以实现本申请实施例提供的存储设备的功能，以实现光通信系统中的光信号传输功能。举例来说，网卡 23 可用于实现网卡 1226 的功能，网卡 1226 也可以包括网卡 23 所包括的各个芯片和缓存等，本申请对此不予限定。

下面结合前述实施例提供的存储设备 20，对本申请实施例提供的数据通信方法进行说明。

图 3 为本申请实施例提供的一种数据通信方法的流程示意图一。该数据通信方法可以应用于图 1 所示出的光通信系统，该数据通信方法由存储设备 20 执行，关于该存储设备 20 的硬件实现可参照图 2B 的相关描述，不予赘述。如图 3 所示，本申请实施例提供的数据通信方法包括以下步骤 S310 至 S350。

S310、处理器 21 获取数据访问请求。

其中，该数据访问请求用于请求保存在存储器 22 中的待发送数据。

一种可能的示例中，数据访问请求是处理器 21 根据存储设备 20 所执行的业务所生成的。

另一种可能的示例中，该数据访问请求是处理器 21 从其他设备中接收的，例如该其他设备可以是如图 1 所示的服务器 111，或者与服务器 111 通信的客户端（用户端）等。

S320、处理器 21 判断待发送数据的数据量是否大于或等于设定的阈值。

例如，该设定的阈值为 100 百万字节（million byte，MB）、500MB 或者其他值等。

若待发送数据的数据量大于或等于设定的阈值，则继续执行 S330。

举例来说，RDMA 应用发起数据传输任务时，识别本次任务为长距大数据量传输任务（10GB 数据传输 1000 公里），通知数据处理芯片启动长距数据传输。

S330、数据处理芯片将待发送数据从存储器 22 写入缓存 233。

这里的数据处理芯片可以包括图 2B 所示出的第一芯片 231 和第二芯片 232 等。

举例来说，第一芯片 231 可将待发送数据从存储器 22 写入缓存 233。关于第一芯片将数据从存储器写入缓存的具体实现过程可参照以下图 5 的相关描述，在此不予赘述。

S340、数据处理芯片读取缓存 233 中的待发送数据，并将待发送数据映射到数据处理芯片生成的 OTN 帧的净荷区。

示例性的，第二芯片 232 将待发送数据映射到多个 OSU 帧，该待发送数据承载于多个 OSU 帧的净荷区；以及，第二芯片 232 将多个 OSU 帧映射到 OTN 帧。

在本实施例中，第二芯片可以将数据映射到不同 OSU 帧的净荷区，使得待发送数据可以以更细的时隙颗粒度来进行数据通信，且 OSU 技术从一开始就考虑无损调整的需求，和 OTN 通信不存在兼容的问题，使得待发送数据的通信过程能够支持更大的无损带宽调整范围，有利于提高数据通信效率。

这里的无损带宽调整包括：带宽增加、带宽减少和带宽回退中至少一种。其中的带宽回退用于指示出现问题后恢复原始状态的操作。有关 OSU 技术更多的内容可参照通常技术的描述，在此不予赘述。

S350、光收发器 24 发送 S340 中生成的 OTN 帧。

在本实施例中，存储设备中的网卡实现数据的 OTN 封装功能，使得数据不需经由多个设备的传输和封装即可生成映射有该数据的 OTN 帧，减少了数据从存储器到光收发器之间的传输时延，提高了数据通信效率。

而且，由于数据无需经由以太交换机进行处理，因此，存储设备将数据封装为 OTN 帧所采用的协议栈也无需使用以太网协议，减少了生成 OTN 帧所需的封装流程以及 OTN 帧中所包含的数据量，有利于进一步提高数据的通信效率。

举例来说，由于网卡可直接将数据不经过以太封装，直接封装成 OTN 帧，且网卡支持插在端侧存储设备上，因此，存储设备可直接在端侧输出 OTN 帧，不经过以太交换机的转发，实现了光通信网络中端侧设备到端侧设备的硬管道传输能力，避免了数据通信过程中的丢包，提高了数据的通信效率。

结合协议栈的实现对本申请实施例的有益效果进行说明：通常技术中，数据需经由存储器-网卡-交换机-光网络设备，数据被封装为 OTN 帧的过程中所采用的协议栈包括如下表 1 所示的内容。

表 1

	数据以及校验号		IB 协议	以太网传输协议			光传输协议
格式	FCS	IB payload	IB BTH	UDP header	IP header	ETH header	OTN header

其中，帧校验序列（Frame Check Sequence，FCS）为计算机网络数据链路层的协议数据单元（帧）的尾部字段，是一段 4 个字节的循环冗余校验码。一些示例中，FCS 又称帧尾。

IB payload 用于承载消息负载，如 RDMA 消息或数据等。

IB BTH 为 IB 协议提供的基本传输头（InfiniBand base transport header，IB BTH），该 IB BTH 字段用于指示目的 QP（destination QP）、操作码（operation code）、报文序列号（packet sequence numbers，PSN）和分区（partition）。BTH 字段中的操作码字段（OpCode field）决定了 SEND 消息的起始和结束。

用户数据报协议（user datagram protocol，UDP）字段用于指示报文的载荷是 RDMA 消息。互联网协议（internet protocol，IP）字段用于通过交换机进行三层转发。ETH header 字段用于指示以太网传输过程中的附加字段等。OTN header 字段用于指示光传送网络过程中对光信号进行处理的帧头。

相比之下，在本申请实施例提供的数据通信方法中，数据需经由存储器-网卡-光网络设备，数据被封装为 OTN 帧的过程中所采用的协议栈包括如下表 2 所示的内容。

表 2

	数据以及校验号		IB 协议	光传输协议
格式	FCS	IB payload	IB BTH	OTN header

可知，该协议栈通过传输层 IB 协议直接承载在物理层 OSU 协议上实现极简协议栈，在传输应用上，存储设备直接出 OTN 信号（OTN 帧）对接光网络设备（或光传送设备），做到从端侧到网络侧之间的端到端硬管道传输。在一些可选的示例中，该极简协议栈也可称为 RDMA over OSU，利用 OTN 零丢包、低时延、传输距离长的优势，极大提升不同 DC 之间长距传输的通信效率。

比较表 1 和表 2 的 OTN 帧的格式，可知两者采用的协议栈不同，如图 4 所示，图 4 为本申请实施

例提供的协议栈的结构示意图，本申请提供的技术方案中 OTN 帧所采用的协议栈包括：RDMA 应用层协议（RDMA application protocol）、IB 传输协议、OSU 连接层协议（OSU link layer protocol）、OSU 物理层协议（OSU physical layer protocol，OSU PHY layer）。在本实施例中，RDMA 应用层直接承载在物理层帧格式（OTN 帧），如 RDMA 应用层直接作为 OSU 的服务层封装到 OSU 的净荷中，RDMA 与 OSU 之间相互联动和配合以完成整个业务封装和解封装过程。

在一些可选的实现方式中，若待发送数据的数据量小于设定的阈值，则存储设备可以指示不同 DC 之间的小数据量传输任务仍然走交换机传输路径，如图 1 所示出的存储设备 121-存储设备 121 包括的网卡-以太交换机 21-光网络设备 31。

而批量的大数据量长距传输直接通过存储设备的 OTN 网卡对接 OTN 传输设备进行传输，存储设备可直接在端侧输出 OTN 帧，不经过以太交换机的转发，实现了光通信网络中端侧设备到端侧设备的硬管道传输能力，避免了数据通信过程中的丢包，提高了数据的通信效率。

在另一些可选的实现方式中，以上的光收发器 24 在发送 S340 中生成的 OTN 帧之后，该光收发器 24 还可以接收目标存储设备的数据写响应。该数据写响应用于指示待发送数据已写入目标存储设备。

示例性的，该目标存储设备可以是指图 1 中的存储设备 122。在一些可行的示例中，该数据写响应可以是由存储设备 122 包括的网卡生成并发送的 OTN 帧。

在本实施例中，在光收发器接收到目标存储设备的数据写响应后，存储设备确定数据的本次传输结束，避免了存储设备为该次数据传输预留硬件资源（如计算资源或存储资源）导致的资源消耗，有利于存储设备将有限的硬件资源用于执行其他业务。

针对于以上的 S330 和 S340 的实现过程，本申请实施例提供了一种可能的实现方式，如图 5 所示，图 5 为本申请实施例提供的一种数据通信方法的流程示意图二，网卡接收和发送数据通常采用消息队列的方式，消息队列包括一组队列对（queue pair，QP），QP 包括发送队列和接收队列，如网卡 23 中用于发送数据的消息队列是发送队列（SQ），网卡 23 中用于接收数据的消息队列是接收队列（RQ）。消息队列是多个主机（或存储设备）之间进行通信所采用的连接方式，例如，多个主机之间可以利用 TCP/IP 协议建立多条连接，每条连接都有一个接收队列和发送队列，该接收队列和发送队列用于传输该连接的数据。

示例性的，存储器 22 维护有一个或多个 RDMA 发送队列（send queue，SQ），这一个或多个 SQ 可以包括 SQ 1 至 SQ N 等，每个 SQ 中存储有多个数据的 WQE。示例性的，WQE 包括：数据的源地址、目标地址、存放数据的内存地址、目的存储设备标识、传输完成时长、数据量等等。关于 WQE 的更多内容可参照通常技术的相关描述，在此不予赘述。相应的，存储器 22 还维护有 RDMA 接收队列（receive queue，RQ），RQ 用于接收数据的消息。

本实施例以 SQ 1 为例进行说明，SQ 1 也可称为第一发送队列，该 SQ 1 中存储有一个或多个数据的 WQE，该一个或多个数据包括前述的待发送数据。

如图 5 所示，光收发器 24 还用于提供多个 OTN 通道（如图 5 中的通道 1 和通道 2），其中，一个 OTN 通道用于传输一个 SQ 对应的数据。举例来说，通道 1 用于传输 SQ 1 中 WQE1 和 WQE2 对应的数据。

本申请实施例以第一芯片 231 是 DPU、第二芯片 232 是 OTN 芯片为例进行说明，本实施例提供的数据通信方法包括以下的步骤①至步骤⑩。

步骤①：RDMA 应用发起数据传输任务，并在 SQ 中记录数据传输任务对应的数据的 WQE。如图 5 中的 WQE1 至 WQEm。

步骤②：DPU（第一芯片 231）从 SQ 1 中读取待发送数据的 WQE，并根据 WQE 指示的源地址，将待发送数据从存储器 22 写入缓存 233。

举例来说，RDMA 应用通知 DPU 本次待传输的数据量、期望传输完成时间、存放数据的内存地址、目的地址等 WQE 信息。

步骤③：数据处理芯片为待发送数据的 WQE 和多个 OTN 通道中的第一 OTN 通道建立映射关系。

示例性的，待发送数据的 WQE 可以是例如图 5 所示出的 WQE1，第一 OTN 通道可以是例如图 5 所示出的通道 1。该映射关系用于指示：待发送数据能够通过第一 OTN 通道（通道 1）进行传输。

在一种可能的情形中，数据处理芯片包括的 OTN 芯片与 RDMA QP 队列产生关联关系，QP 队列

生成后（如 SQ 1），OTN 芯片给该 SQ 1 分配对应承载的 OTN 通道。

步骤④：DPU 将待发送数据从存储器 22 写入网卡的缓存。

例如，网卡中的缓存 233 维护有多个队列，这多个队列包括第一队列（QM1），第一队列对应的存储空间用于存储待发送数据。在 OTN 芯片建立 SQ 和 OTN 通道之间的映射关系的过程中，OTN 芯片为 QM1 与通道 1 建立映射关系。如 WQE1 和 WQE2 对应的数据通过 QM1 对应的通道 1 传输，WQEm 对应的数据通过 QM2 对应的通道 2 传输。

在本实施例中，网卡中不同的芯片用于实现不同的功能，DPU 实现网卡和应用层的交互，OTN 芯片实现网卡和光通信网络的交互，因此，由网卡中不同芯片之间进行协调即可实现数据从存储器至光通信网络中的硬管道传输，有利于提高数据的通信效率。

步骤⑤：OTN 芯片读取缓存中的待发送数据，并将待发送数据映射到 OTN 芯片生成的 OTN 帧的净荷区。

如图 5 所示，OTN 芯片包括多个模块：队列管理（queue management，QM）单元、光线路侧报文处理（optical line packet，OLPKT）模块、消费交换（customer exchange，CXC）模块、OTN 低线点（OTN Lite Line Node，OLLN）。其中，QM 用于实现多队列管理、数据反压处理等功能；OLPKT 用于将待发送数据编码转换成 256/257B 格式，并切成 OSU 固定长度信元（OSU 帧）；CXC 模块用于将 OSU 帧进行入通道和出通道映射；OLLN 模块用于完成 OTN 维护信号下插和解析等。

举例来说，OTN 芯片将待发送数据切分为多个 192B 大小的数据单元，并将一个数据单元映射到一个 OSU 帧的净荷区，并在多个数据单元均映射到 OSU 帧的净荷区后，将多个 OSU 帧映射到 OTN 帧。关于 OTN 帧的格式可以参照表 2 的内容，在此不予赘述。

在一种可选的实现方式中，本申请实施例提供的数据通信方法还包括：OTN 芯片判断 QM1 的数据流速率是否大于或等于设定的速率阈值，若数据流速率大于或等于设定的速率阈值，则 OTN 芯片指示 DPU 调低向 QM1 写数据的数据流速率。其中的数据流速率为单位时间内，DPU 向 QM1 中写入的数据量。

值得注意的是，以上设定的速率阈值可以是根据 OTN 芯片和 DPU 的硬件特性来设置的。在一些可选的情形中，设定的速率阈值也可以是用户根据不同 DC 之间的数据通信要求来设置的，本申请对此不予限定，例如该速率阈值可以是 5GB/s、500MB/s 或其他数值等。

在 OTN 芯片处理数据的预期速度过大的情况下，如 OTN 芯片的数据流速度大于或等于设定的速率阈值，则 OTN 芯片可以指示 DPU 调低向第一队列写数据的数据流速率，从而降低 OTN 芯片在单位时间内所要处理（如封装）的数据量，以减小 OTN 芯片的通信负荷，有利于避免 OTN 芯片的网络丢包，提高光通信网络的通信性能。

步骤⑥：光收发器 24 通过通道 1 将待发送数据对应的 OTN 帧发送到目标存储设备。

示例性的，目标存储设备和源存储设备可以是指位于不同 DC 的存储设备，如该目标存储设备是指图 1 中所示出的存储设备 122，源存储设备可以是指图 1 中所示出的存储设备 121。两个存储设备之间可以通过长距离光纤来传输 OTN 帧。

在本申请实施例中，目标存储设备也称为源存储设备的对端存储设备（简称：对端）。

步骤⑦：目标存储设备中的对端 OTN 芯片解析接收到的 OTN 帧，并将前述的待发送数据写入到对端网卡的缓存中。

举例来说，对端 OTN 芯片接收到数据帧（OTN 帧）后，OLLN 模块解析出 OTN 维护信号，判断该 OTN 帧是否有告警，在没有告警的情况下，CXC 模块将数据进行 OTN 通道解映射，解析出 RDMA 数据（待发送数据）后写入对端网卡中缓存维护的 QM 队列。

步骤⑧：对端 DPU 从 QM 队列获取接收到的数据，直接存放入目标存储设备的对端存储器（如内存或硬盘）。

步骤⑨：在对端 DPU 将待发送数据写入对端存储器之后，对端 OTN 芯片通过光收发器向源存储设备的网卡发送数据写响应。

示例性的，该数据写响应可以是采用 OTN 帧的格式来传输的，该数据写响应用于指示：待发送数据已写入目标存储设备的存储器中。

步骤⑩：源存储设备的 DPU 接收到数据写响应后，获知本次数据传输完成，生成传输完成队列信

息 CQ1，并将 CQ1 放置到 CQ 队列，然后通知 RDMA 应用本次数据传输完成。

存储设备中的 RDMA 队列可和光收发器提供的 OTN 通道建立映射关系，使得不同 RDMA 队列的数据经由不同的 OTN 通道进行传输，其中，映射关系是网卡基于 RDMA 队列中记录的数据的 WQE 来建立的，避免了光收发器将数据对应的 OTN 帧发送到与数据的 WQE 不匹配的 OTN 通道，提高了数据通信的准确性。

而且，在后续的其他数据的通信过程中，若网卡已经建立了该其他数据的 WQE 和第一 OTN 通道的映射关系，则网卡可复用该映射关系，以将映射有该其他数据的 OTN 帧通过第一 OTN 通道进行传输，进一步提高了光通信网络中的数据通信效率。

本实施例中的方法步骤可以通过硬件的方式来实现，也可以由处理器执行软件指令的方式来实现。软件指令可以由相应的软件模块组成，软件模块可以被存放于随机存取存储器（random access memory，RAM）、闪存、只读存储器（read-only memory，ROM）、可编程只读存储器（programmable ROM，PROM）、可擦除可编程只读存储器（erasable PROM，EPROM）、电可擦除可编程只读存储器（electrically EPROM，EEPROM）、寄存器、硬盘、移动硬盘、CD-ROM 或者本领域熟知的任何其它形式的存储介质中。一种示例性的存储介质耦合至处理器，从而使处理器能够从该存储介质读取信息，且可向该存储介质写入信息。当然，存储介质也可以是处理器的组成部分。处理器和存储介质可以位于 ASIC 中。另外，该 ASIC 可以位于计算设备中。当然，处理器和存储介质也可以作为分立组件存在于网络设备或终端设备中。

本申请还提供一种芯片系统，该芯片系统包括处理器，用于实现上述方法中数据处理单元的功能。在一种可能的设计中，所述芯片系统还包括存储器，用于保存程序指令和/或数据。该芯片系统，可以由芯片构成，也可以包括芯片和其他分立器件。

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程序或指令。在计算机上加载和执行所述计算机程序或指令时，全部或部分地执行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、网络设备、用户设备或者其它可编程装置。所述计算机程序或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，所述计算机程序或指令可以从一个网站站点、计算机、服务器或数据中心通过有线或无线方式向另一个网站站点、计算机、服务器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集成一个或多个可用介质的服务器、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；也可以是光介质，例如，数字视频光盘（digital video disc，DVD）；还可以是半导体介质，例如，固态硬盘（solid state drive，SSD）。

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，可轻易想到各种等效的修改或替换，这些修改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：处理器、存储器和网卡，所述网卡包括数据处理芯片、存储介质和光收发器；

所述存储器，用于存储所述处理器写入的待发送数据；

所述数据处理芯片，用于将所述待发送数据从所述存储器写入所述存储介质；

所述数据处理芯片，还用于读取所述存储介质中的待发送数据，并将所述待发送数据映射到所述数据处理芯片生成的光传送网络 OTN 帧的净荷区；

所述光收发器，用于发送所述 OTN 帧。

2.根据权利要求 1 所述的存储设备，其特征在于，

所述存储器，还用于维护至少一个远程直接存储访问 RDMA 发送队列，所述至少一个 RDMA 发送队列包括第一发送队列，所述第一发送队列中存储有一个或多个数据的工作队列元素 WQE，所述一个或多个数据包括所述待发送数据；

所述光收发器，还用于提供多个 OTN 通道，其中，一个 OTN 通道用于传输一个 RDMA 发送队列对应的数据；

所述数据处理芯片，还用于：从所述第一发送队列中读取所述待发送数据的 WQE，并为所述 WQE 和所述多个 OTN 通道中的第一 OTN 通道建立映射关系；其中，所述映射关系用于指示：所述待发送数据能够通过所述第一 OTN 通道进行传输。

3.根据权利要求 2 所述的存储设备，其特征在于，所述数据处理芯片包括：第一芯片和第二芯片；

所述第一芯片，用于从所述第一发送队列中读取所述待发送数据的 WQE；

以及，所述第一芯片，还用于：根据所述 WQE 指示的源地址，将所述待发送数据从所述存储器写入所述存储介质；

所述第二芯片，用于：为所述 WQE 和所述多个 OTN 通道中的第一 OTN 通道建立所述映射关系；

以及，所述第二芯片，还用于：读取所述存储介质中的待发送数据，并将所述待发送数据映射到所述第二芯片生成的 OTN 帧的净荷区。

4.根据权利要求 3 所述的存储设备，其特征在于，所述存储介质维护有多个队列；

所述第一芯片，具体用于：根据所述 WQE 指示的源地址，将所述待发送数据从所述存储器写入所述多个队列中第一队列对应的存储空间；

所述第二芯片，具体用于：为所述第一队列与所述第一 OTN 通道建立所述映射关系。

5.根据权利要求 4 所述的存储设备，其特征在于，

所述第二芯片，还用于判断所述第一队列的数据流速率是否大于或等于设定的速率阈值；所述数据流速率为单位时间内，所述第一芯片向所述第一队列中写入的数据量；

若所述数据流速率大于或等于设定的速率阈值，则所述第二芯片还用于指示所述第一芯片调低向所述第一队列写数据的数据流速率。

6.根据权利要求 3-5 中任一项所述的存储设备，其特征在于，

所述第二芯片，具体用于：将所述待发送数据映射到多个光业务单元 OSU 帧；所述待发送数据承载于所述多个 OSU 帧的净荷区；以及，将所述多个 OSU 帧映射到 OTN 帧。

7.根据权利要求 1-6 中任一项所述的存储设备，其特征在于，

所述处理器，用于获取数据访问请求，所述数据访问请求用于请求所述待发送数据；

所述处理器，还用于判断所述待发送数据的数据量大于或等于设定的阈值；

所述数据处理芯片，具体用于：若所述待发送数据的数据量大于或等于设定的阈值，将所述待发送数据从所述存储器写入所述存储介质。

8.根据权利要求 1-6 中任一项所述的存储设备，其特征在于，

所述光收发器，还用于接收目标存储设备的数据写响应，所述数据写响应用于指示所述待发送数据已写入所述目标存储设备。

9.一种数据通信方法，其特征在于，所述方法由存储设备执行，所述存储设备包括：处理器、存储器和网卡，所述存储器用于存储所述处理器写入的待发送数据，所述网卡包括数据处理芯片、存储介质

和光收发器；

所述方法包括：

所述数据处理芯片将所述待发送数据从所述存储器写入所述存储介质；

所述数据处理芯片读取所述存储介质中的待发送数据，并将所述待发送数据映射到所述数据处理芯片生成的光传送网络 OTN 帧的净荷区；

所述光收发器发送所述 OTN 帧。

10.根据权利要求 9 所述的方法，其特征在于，所述存储器还用于维护至少一个远程直接存储访问 RDMA 发送队列，所述至少一个 RDMA 发送队列包括第一发送队列，所述第一发送队列中存储有一个或多个数据的工作队列元素 WQE，所述一个或多个数据包括所述待发送数据；

所述光收发器用于提供多个 OTN 通道，其中，一个 OTN 通道用于传输一个 RDMA 发送队列对应的数据；

所述数据处理芯片读取所述存储介质中的待发送数据，包括：

所述数据处理芯片从所述第一发送队列中读取所述待发送数据的 WQE，并根据所述 WQE 指示的源地址，将所述待发送数据从所述存储器写入所述存储介质；

在所述光收发器发送所述 OTN 帧之前，所述方法还包括：

所述数据处理芯片为所述 WQE 和所述多个 OTN 通道中的第一 OTN 通道建立映射关系；其中，所述映射关系用于指示：所述待发送数据能够通过所述第一 OTN 通道进行传输。

11.根据权利要求 10 所述的方法，其特征在于，所述数据处理芯片包括：第一芯片和第二芯片；

所述数据处理芯片从所述第一发送队列中读取所述待发送数据的 WQE，并根据所述 WQE 指示的源地址，将所述待发送数据从所述存储器写入所述存储介质，包括：

所述第一芯片从所述第一发送队列中读取所述待发送数据的 WQE，并根据所述 WQE 指示的源地址，将所述待发送数据从所述存储器写入所述存储介质；

所述数据处理芯片为所述 WQE 和所述多个 OTN 通道中的第一 OTN 通道建立映射关系，包括：

所述第二芯片为所述 WQE 和所述多个 OTN 通道中的第一 OTN 通道建立所述映射关系；

所述数据处理芯片将所述待发送数据映射到所述数据处理芯片生成的 OTN 帧的净荷区，包括：

所述第二芯片读取所述存储介质中的待发送数据，并将所述待发送数据映射到所述第二芯片生成的 OTN 帧的净荷区。

12.根据权利要求 11 所述的方法，其特征在于，所述存储介质维护有多个队列，所述多个队列包括第一队列，所述第一队列对应的存储空间用于存储所述待发送数据；

所述第二芯片为所述 WQE 和所述多个 OTN 通道中的第一 OTN 通道建立所述映射关系，包括：

所述第二芯片为所述第一队列与所述第一 OTN 通道建立所述映射关系。

13.根据权利要求 12 所述的方法，其特征在于，所述方法还包括：

所述第二芯片判断所述第一队列的数据流速率是否大于或等于设定的速率阈值；所述数据流速率为单位时间内，所述第一芯片向所述第一队列中写入的数据量；

若所述数据流速率大于或等于设定的速率阈值，则所述第二芯片指示所述第一芯片调低向所述第一队列写数据的数据流速率。

14.根据权利要求 11-12 中任一项所述的方法，其特征在于，

所述数据处理芯片将所述待发送数据映射到所述数据处理芯片生成的 OTN 帧的净荷区，包括：

所述第二芯片将所述待发送数据映射到多个光业务单元 OSU 帧；所述待发送数据承载于所述多个 OSU 帧的净荷区；

所述第二芯片将所述多个 OSU 帧映射到 OTN 帧。

15.根据权利要求 9-14 中任一项所述的方法，其特征在于，在所述数据处理芯片将所述待发送数据从所述存储器写入所述存储介质之前，所述方法还包括：

所述处理器获取数据访问请求，所述数据访问请求用于请求所述待发送数据；

所述处理器判断所述待发送数据的数据量大于或等于设定的阈值；

若所述待发送数据的数据量大于或等于设定的阈值，所述数据处理芯片将所述待发送数据从所述存储器写入所述存储介质。

16.根据权利要求 9-15 中任一项所述的方法，其特征在于，所述方法还包括：

所述光收发器接收目标存储设备的数据写响应，所述数据写响应用于指示所述待发送数据已写入所述目标存储设备。

17.一种光通信系统，其特征在于，所述系统包括：存储设备和光网络设备；其中：

所述存储设备包括：处理器、存储器和网卡，所述网卡包括数据处理芯片、存储介质和光收发器；

所述存储器，用于存储所述处理器写入的待发送数据；

所述数据处理芯片，用于将所述待发送数据从所述存储器写入所述存储介质；

所述数据处理芯片，还用于读取所述存储介质中的待发送数据，并将所述待发送数据映射到所述数据处理芯片生成的 OTN 帧的净荷区；

所述光收发器，用于向所述光网络设备发送所述 OTN 帧。

18.一种计算机可读存储介质，其特征在于，包括：计算机软件指令；当所述计算机软件指令在存储设备中运行时，所述存储设备执行权利要求 9-16 中任一项所述的方法。

19.一种计算机程序产品，其特征在于，当所述计算机程序产品在存储设备中运行时，所述存储设备执行权利要求 9-16 中任一项所述的方法。

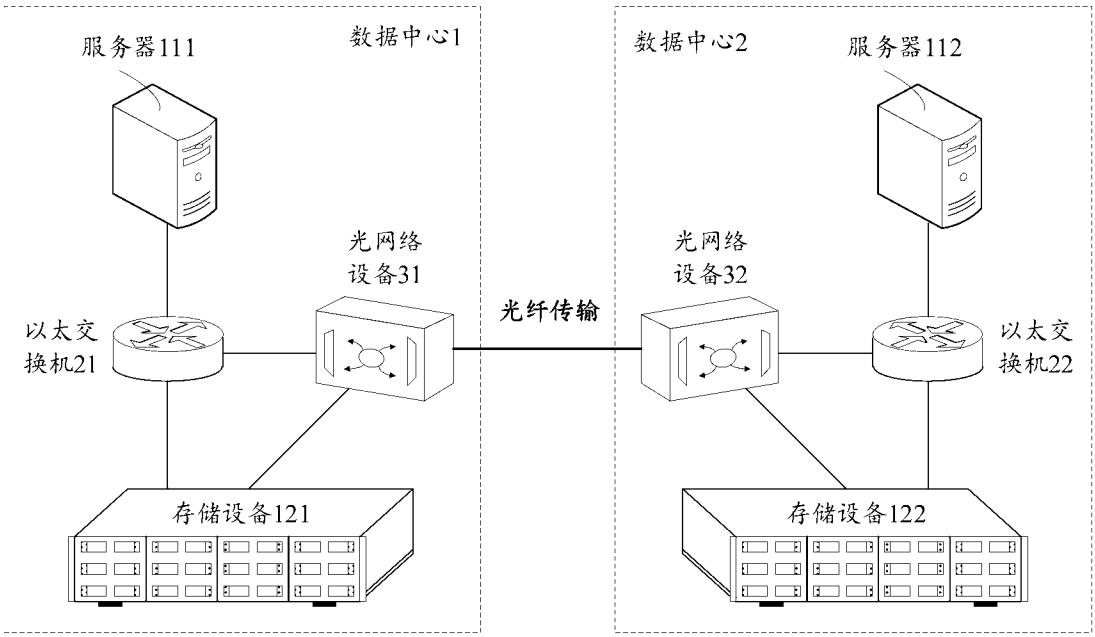
服务器111
数据中心1
数据中心2
服务器112
光网络
设备31
光网络
设备32
光纤传输
以太交
换机21
以太交
换机22
存储设备121
存储设备122

图 1

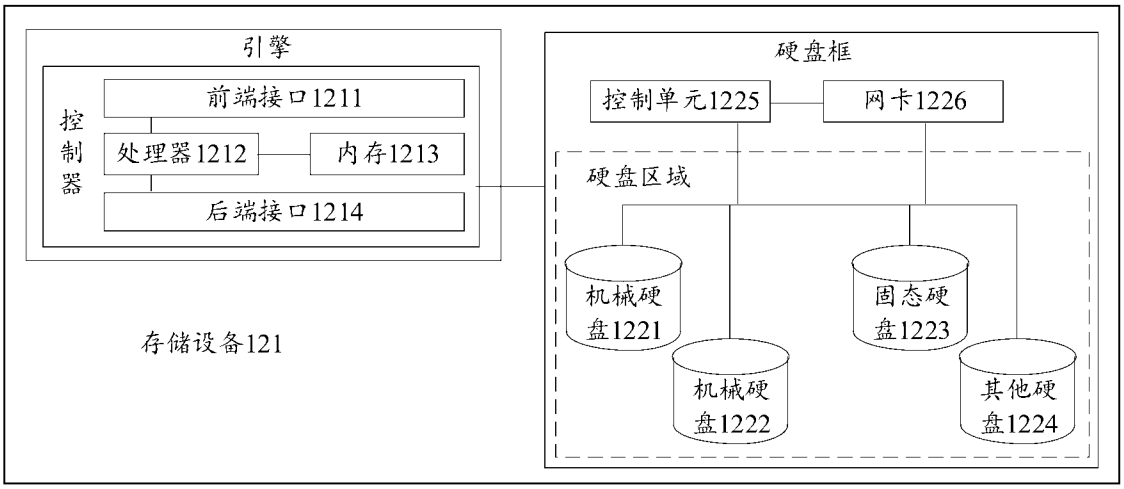
引擎
控制器
前端接口1211
处理器1212
内存1213
后端接口1214
存储设备121
硬盘框
控制单元1225
网卡1226
硬盘区域
机械硬盘1221
机械硬盘1222
固态硬盘1223
其他硬盘1224

图 2A

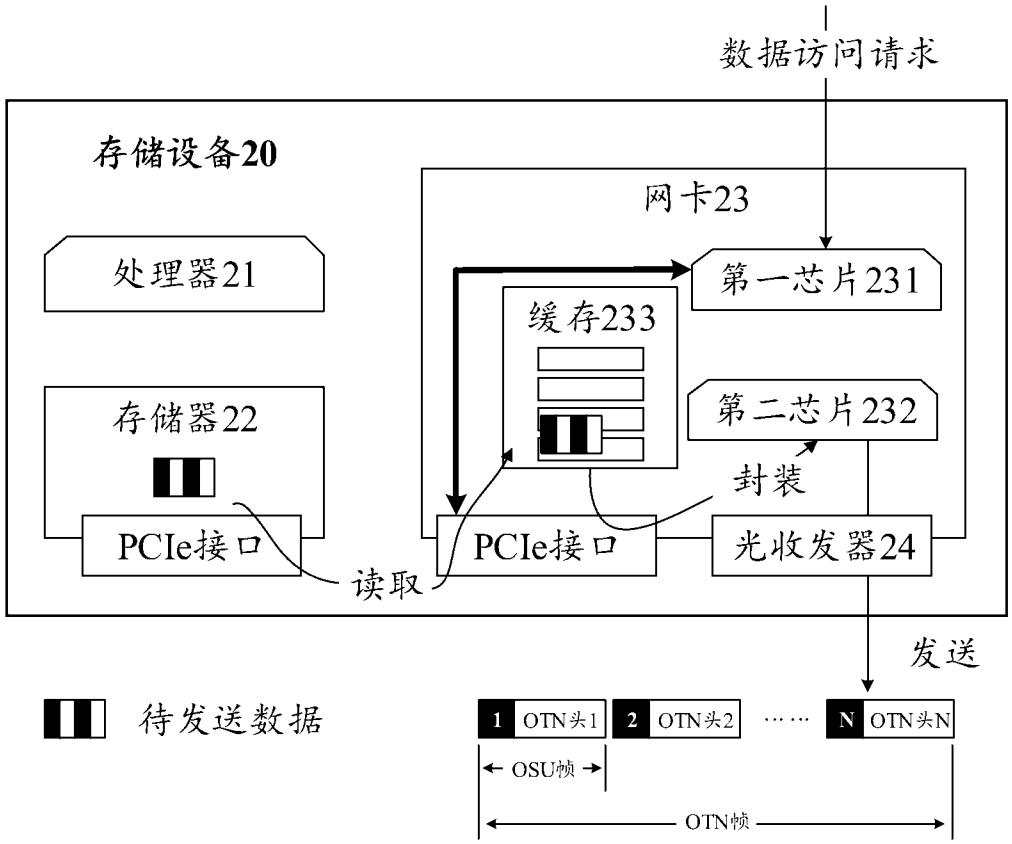
数据访问请求
存储设备20
网卡23
处理器21
第一芯片231
缓存233
存储器22
第二芯片232
封装
PCIe接口
读取
PCIe接口
光收发器24
发送
待发送数据
1 OTN头1
2 OTN头2
…… N OTN头N
OSU帧
OTN帧

图 2B

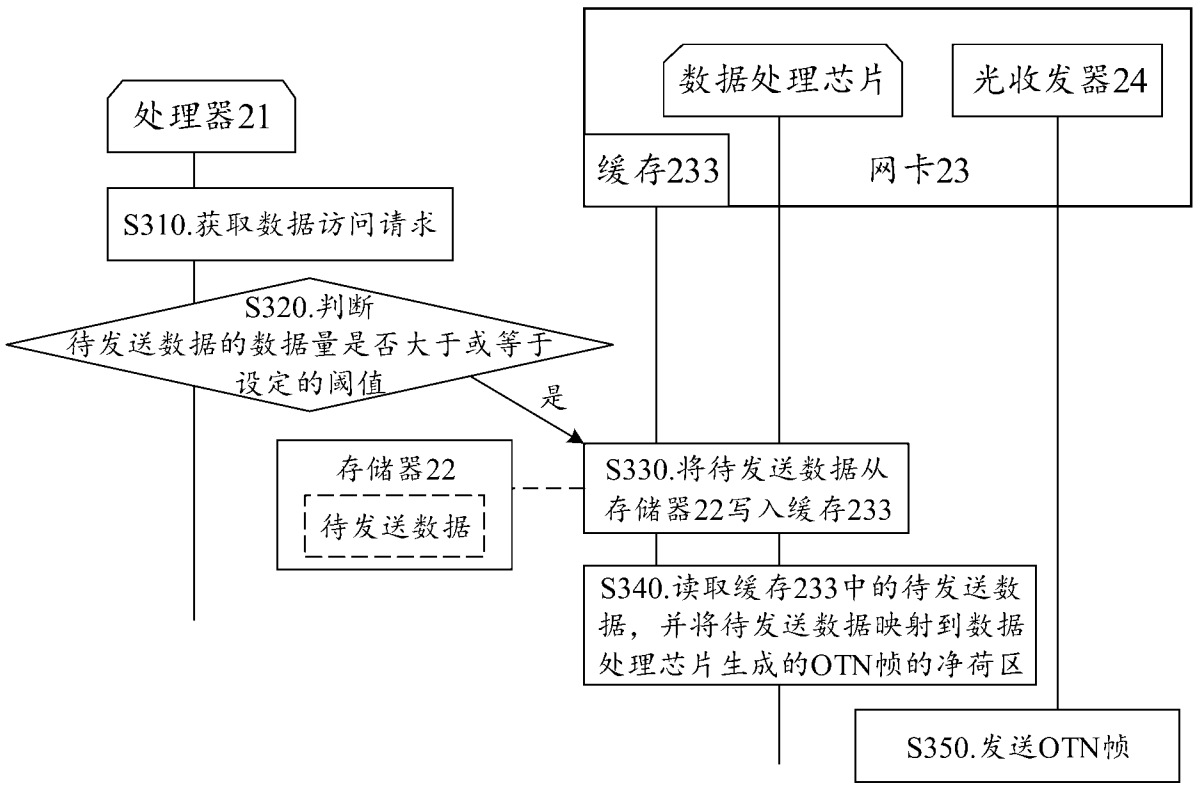
数据处理芯片
光收发器24
处理器21
缓存233
网卡23
S310.获取数据访问请求
S320.判断
待发送数据的数据量是否大于或等于
设定的阈值
是
存储器22
待发送数据
S330.将待发送数据从
存储器22写入缓存233
S340.读取缓存233中的待发送数据，并将待发送数据映射到数据处理芯片生成的OTN帧的净荷区
S350.发送OTN帧

图 3

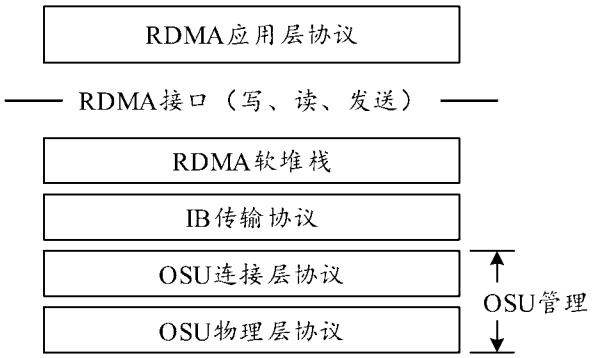
RDMA应用层协议
RDMA接口（写、读、发送）
RDMA软堆栈
IB传输协议
OSU连接层协议
OSU物理层协议
OSU管理

图 4

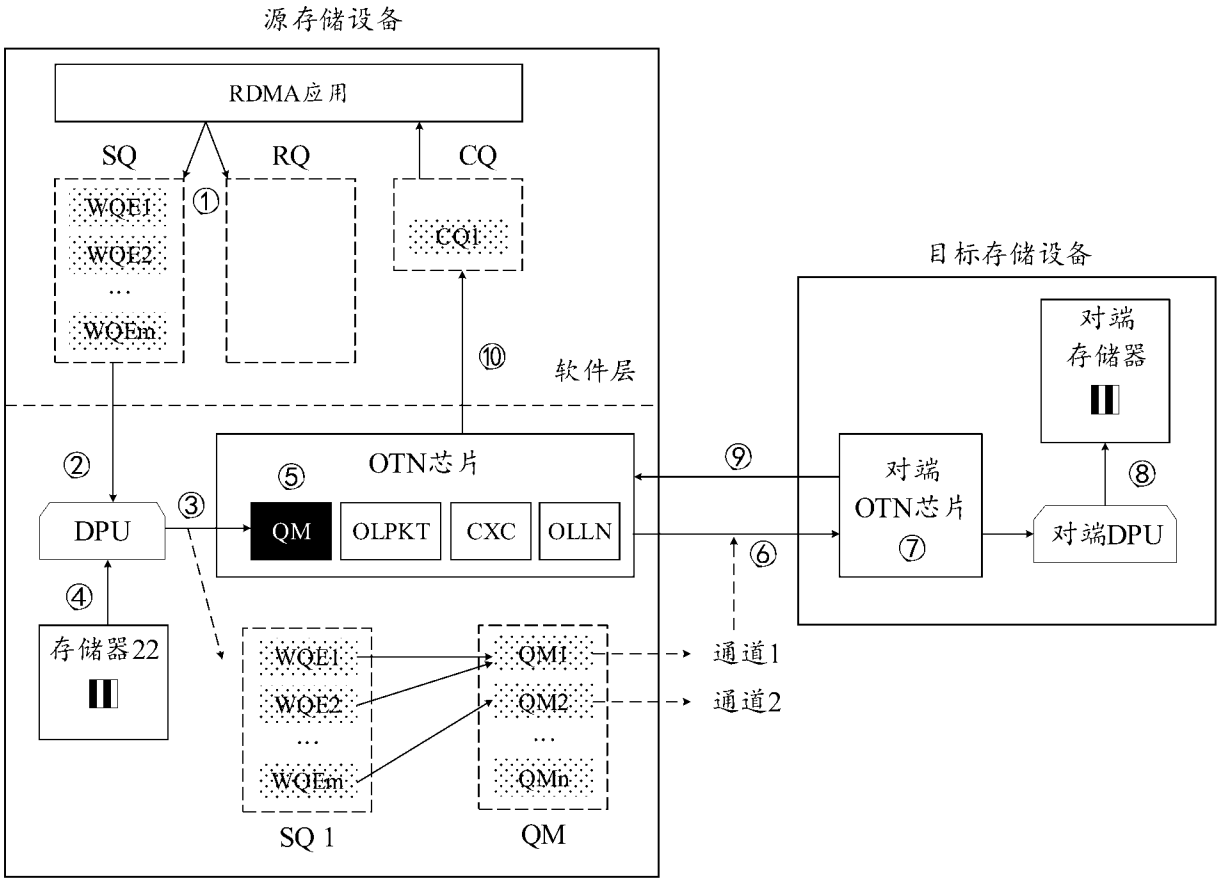
源存储设备
RDMA应用
SQ
RQ
CQ
WQE1
WQE2
…
WQEm
①
CQ1
⑩
软件层
②
DPU
③
⑤
OTN芯片
QM
OLPKT
CXC
OLLN
④
存储器22
WQE1
WQE2
…
WQEm
SQ 1
QM1
QM2
…
QMn
QM
通道1
通道2
⑨
⑥
目标存储设备
对端
OTN芯片
⑦
对端
存储器
⑧
对端DPU

图 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/140344

A. CLASSIFICATION OF SUBJECT MATTER

H04J3/16(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)

IPC:H04J,H04W,H04B,H04L

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)

CNTXT, ENTXTC, ENTXT, CNKI: 避免, 不必, 不需, 无需, 不用, 处理器, 存储, 封装, 光传送网络, 光传输网络, 缓存, 缓冲, 接口控制器, 接口卡, 净荷, 内存, 生成, 适配器, 网卡, 以太网, 映射, 帧, CPU, NIC, HCA, OTN, OSU, payload, PCIE, cache, WQE, RDMA, frame

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 113900972 A (HUAWEI TECHNOLOGIES CO., LTD.) 07 January 2022 (2022-01-07) description, paragraphs [0044]-[0155] and [0205]	1-19
Y	CN 101155006 A (HUAWEI TECHNOLOGIES CO., LTD.) 02 April 2008 (2008-04-02) description, pages 7-11	1-19
A	CN 109491809 A (XI'AN MICROELECTRONICS TECHNOLOGY INSTITUTE) 19 March 2019 (2019-03-19) entire document	1-19
A	CN 115499084 A (CHINA TELECOM CORP., LTD.) 20 December 2022 (2022-12-20) entire document	1-19
A	US 2020026656 A1 (INTERNATIONAL BUSINESS MACHINES CORP.) 23 January 2020 (2020-01-23) entire document	1-19
A	US 2003123493 A1 (NEC CORP.) 03 July 2003 (2003-07-03) entire document	1-19

☐ 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
"D" document cited by the applicant in the international application
"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
13 March 2024	**20 March 2024**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **China No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/140344

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 113900972 A	07 January 2022	WO 2022007470 A1	13 January 2022
		EP 4160425 A1	05 April 2023
		US 2023153264 A1	18 May 2023
CN 101155006 A	02 April 2008	None	
CN 109491809 A	19 March 2019	None	
CN 115499084 A	20 December 2022	None	
US 2020026656 A1	23 January 2020	None	
US 2003123493 A1	03 July 2003	JP 2003188919 A	04 July 2003
		CA 2414346 A1	19 June 2003

国际检索报告

国际申请号
PCT/CN2023/140344

A. 主题的分类

H04J3/16(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)

IPC:H04J,H04W,H04B,H04L

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））

CNTXT,ENTXTC,ENTXT,CNKI:避免,不必,不需,无需,不用,处理器,存储,封装,光传送网络,光传输网络,缓存,缓冲,接口控制器,接口卡,净荷,内存,生成,适配器,网卡,以太网,映射,帧,CPU,NIC,HCA,OTN,OSU,payload,PCIE,cache,WQE,RDMA,frame

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
Y	CN 113900972 A (华为技术有限公司) 2022年1月7日 (2022 - 01 - 07) 说明书第[0044]-[0155]、[0205]段	1-19
Y	CN 101155006 A (华为技术有限公司) 2008年4月2日 (2008 - 04 - 02) 说明书第7-11页	1-19
A	CN 109491809 A (西安微电子技术研究所) 2019年3月19日 (2019 - 03 - 19) 全文	1-19
A	CN 115499084 A (中国电信股份有限公司) 2022年12月20日 (2022 - 12 - 20) 全文	1-19
A	US 2020026656 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2020年1月23日 (2020 - 01 - 23) 全文	1-19
A	US 2003123493 A1 (NEC CORPORATION) 2003年7月3日 (2003 - 07 - 03) 全文	1-19

☐ 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

"D" 申请人在国际申请中引证的文件

"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
2024年3月13日	2024年3月20日

ISA/CN的名称和邮寄地址	受权官员
中国国家知识产权局 中国北京市海淀区蓟门桥西土城路6号 100088	罗啸 电话号码 (+86) 010-53961774

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/140344

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	113900972	A	2022年1月7日	WO	2022007470	A1	2022年1月13日
				EP	4160425	A1	2023年4月5日
				US	2023153264	A1	2023年5月18日
CN	101155006	A	2008年4月2日		无		
CN	109491809	A	2019年3月19日		无		
CN	115499084	A	2022年12月20日		无		
US	2020026656	A1	2020年1月23日		无		
US	2003123493	A1	2003年7月3日	JP	2003188919	A	2003年7月4日
				CA	2414346	A1	2003年6月19日