(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
2018 年 8 月 2 日 (02.08.2018)

WIPO | PCT

(10) 国际公布号
WO 2018/137217 A1

(51) 国际专利分类号:
G06F 3/06 (2006.01) *H04L 29/08* (2006.01)

(21) 国际申请号: PCT/CN2017/072701

(22) 国际申请日: 2017 年 1 月 25 日 (25.01.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(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: 程宏才 (CHENG, Hongcai); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 郭海涛 (GUO, Haitao); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 刘洪广 (LIU, Hongguang); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 陈昊 (CHEN, Hao); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 李思聪 (LI, Sicong); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 谭春毅 (TAN, Chunyi); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 胡瑜 (HU, Yu); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 陈灿 (CHEN, Can); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

WO 2018/137217 A1

(54) **Title:** DATA PROCESSING SYSTEM, METHOD, AND CORRESPONDING DEVICE

(54) **发明名称**: 一种数据处理的系统、方法及对应装置

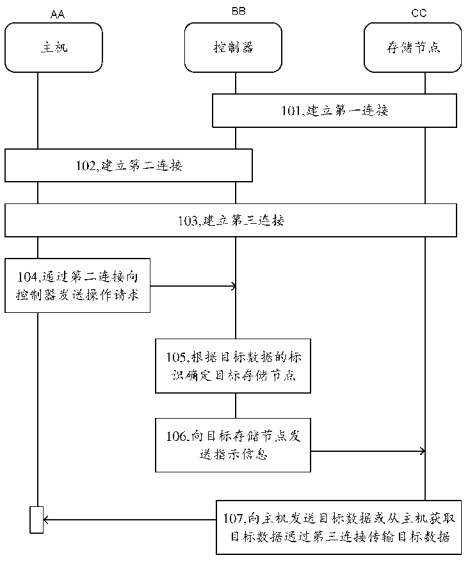

图 4

101 Establish a first connection
102 Establish a second connection
103 Establish a third connection
104 Send an operation request to a controller by means of the second connection
105 Determine a target storage node according to the identifier of target data
106 Send instruction information to the target storage node
107 Send the target data to the host or acquire the target data from the host, and transmit the target data by means of the third connection
AA Host
BB Controller
CC Storage node

(57) **Abstract:** The present application provides a data processing system, method, and corresponding device, for solving the problem that data transmissions between a host and a storage device take a long time. The system comprises a controller and at least two storage nodes, the controller being used for receiving an operation request sent by a host by means of a second connection between the controller and the host, the operation request comprising the identifier of target data to be operated and the operation type thereof; determining, according to the identifier of the target data, at least one target storage node from said at least two storage nodes; sending, by means of a first connection with said at least one target storage node, an instruction message to said at least one target storage node, the instruction message being used for instructing the target storage node to send the target data to the host or acquire the target data from the host; said at least one target storage node being used for sending, according to the instruction message, the target data to the host or acquiring the target data from the host by means of a third connection with the host.

(57) **摘要**: 本申请提供一种数据处理的系统、方法及对应装置，用于解决主机与存储设备间的数据传输耗时较长的问题。该系统包括控制器以及至少两个存储节点，控制器用于接收通过控制器与主机之间的第二连接接述主机发送的操作请求，操作请求包括待操作的目标数据的标识和操作类型；根据目标数据的标识从至少两个存储节点中确定至少一个目标存储节点；通过与至少一个目标存储节点之间的第一连接向至少一个目标存储节点发送指示消息，指示消息用于指示目标存储节点向主机发送目标数据或从主机获取目标数据；至少一个目标存储节点，用于根据指示消息通过与主机之间的第三连接向主机发送目标数据或从主机获取目标数据。

[见续页]

(74) 代理人：北京同达信恒知识产权代理有限公司 (TDIP & PARTNERS)；中国北京市海淀区宝盛南路1号院20号楼8层101-01, Beijing 100192 (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, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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)）。

一种数据处理的系统、方法及对应装置

技术领域

本发明涉及计算机技术领域，特别涉及一种数据处理的系统、方法及对应装置。

背景技术

在计算机存储系统中，主机可以通过管理远端存储设备的控制器访问远端存储设备中的数据。但是，现有技术中的访问方式，主机与存储设备间的数据传输耗时较长，主机读写数据的速率容易受到影响。

发明内容

本申请提供一种数据处理的系统、方法及对应装置，用于解决主机与存储设备间的数据传输需由接收主机操作请求的控制器转发，导致主机与存储设备间的数据传输耗时较长的问题。

本申请的第一方面提供一种数据处理的方法，该方法由数据处理系统中的控制器执行，该系统包括包括控制器以及至少两个存储节点。控制器与每个存储节点建立连接，用于管理该至少两个存储节点。存储节点以及控制器与主机建立连接，主机用于部署应用服务。主机向控制器发送操作请求，该操作请求包括待操作的目标数据的标识和操作类型。控制器接收操作请求，根据目标数据的标识确定目标存储节点。控制器向目标存储节点发送指示消息，该指示消息用于指示目标存储节点通过与主机的连接向主机发送目标数据或从主机获取目标数据。目标存储节点响应该指示消息，通过与主机间的连接向主机发送目标数据或从主机获取目标数据。

由于主机与目标存储节点可以通过二者间的连接直接传输数据，而不用经由控制器转发数据，减短数据传输路径进而减小传输耗时，且避免因控制器的带宽无法满足大量数据传输需求而导致的主机读写数据速率较慢的问题，还可以避免因控制器的运算能力无法满足大量数据处理需求而导致的主机读写数据速率较慢的问题。

在第一方面的一种可能的实现方式中，控制器根据目标数据的标识确定目标存储节点包括存储目标数据的第一数据块的第一存储节点以及存储目标数据的第二数据块的第二存储节点；控制器向第一存储节点发送第一指示消息，第一指示消息包括第一数据块的标识，用于指示第一存储节点与主机传输第一数据块；控制器向第二存储节点发送第二指示消息，第二指示消息包括第二数据块的标识，用于指示第二存储节点与主机传输第二数据块。通过上述方案，控制器能够在目标数据分块存储在多个存储节点时，指示存储目标数据的数据块的存储节点分别通过与主机间的连接向主机发送目标数据或从主机获取目标数据的数据块，提高数据传输的速率，减少传输耗时。

在第一方面的另一种可能的实现方式中，存储节点与主机之间的连接为 RDMA 连接时；主机向控制器发送的操作请求还包括主机在内存中为目标数据指定的目标存储区域的位置参数。控制器根据目标存储区域的位置参数对目标存储区域进行划分，确定目标存储区域中的为每个目标存储节点对应的目标数据的数据块所指定的子存储区域的位置参数。然后，针对每个目标存储节点生成指示消息，该指示消息包括数据块的标识、操作类型以

及该子存储区域的位置参数。目标存储节点接收该指示消息后，在指示消息中的操作类型为读操作时，目标存储节点根据子存储区域的位置参数，通过与主机的 RDMA 连接将目标数据的数据块写入主机内存中子存储区域；在指示消息中的操作类型为写操作时，目标存储节点根据子存储区域的位置参数，通过与主机的 RDMA 连接从主机内存中子存储区域读取目标数据的数据块，并存储数据块。通过上述方案，控制器能够在目标数据分块存储在多个存储节点时，为每个目标存储节点确定其存储数据块对应的主机内存的存储区域，指示该存储节点通过与主机间的 RDMA 连接快速地从主机内存中读取目标数据的数据块或向主机内存写入目标数据的数据块，提高数据传输的速率，减少传输耗时。

在第一方面的另一种可能的实现方式中，存储节点与主机之间的连接为 RDMA 连接时：存储节点创建第一 QP，该第一 QP 包括第一 SQ 以及第一 RQ，然后，存储节点通过第一连接将第一 QP 的参数发送给控制器。控制器通过第二连接将第一 QP 的参数发送给主机。主机创建第二 QP，该第二 QP 包括第二 SQ 以及第二 RQ，通过第二连接将第二 QP 的参数发送给控制器，控制器则通过第一连接将第二 QP 的参数发送给任一存储节点。主机根据接收的第一 QP 的参数将第二 QP 与所述任一存储节点的第一 QP 关联。所述任一存储节点根据接收的第二 QP 的参数将第一 QP 的第一 SQ 与第二 QP 的第二 RQ 绑定，以及将第一 QP 的第一 RQ 与第二 QP 的第二 SQ 绑定。通过上述方案，存储节点与主机能够在控制器的辅助下创建 RDMA 连接，进而能够通过该 RDMA 连接进行数据传输，而无需控制器负责数据的中转，提高数据传输的速率，减少数据传输耗时。

在第一方面的另一种可能的实现方式中，主机与控制器以及至少两个存储节点建立传输控制协议/网际互联协议 TCP/IP 连接；控制器在向目标存储节点发送指示消息后，向主机发送第三指示消息，第三指示消息包括目标存储节点的通信地址，用于指示主机通过与目标存储节点间的 TCP/IP 连接传输目标数据。通过上述方案，控制器不仅向报文发送端发送指示消息，也向报文接收端发送指示消息，使得报文接收端能够从接收的 TCP/IP 报文中获取目标数据，而不是舍弃该报文。

在第一方面的另一种可能的实现方式中，主机与控制节点以及至少两个存储节点建立 TCP/IP 连接，主机向控制器发送的操作请求的操作类型为读操作。控制器在向目标存储节点发送的指示消息中添加该操作类型、主机的通信地址、控制器的通信地址以及目标数据的标识，以指示目标存储节点以控制器的通信地址为源地址以及以主机的通信地址为目的地址发送携带目标数据的 TCP/IP 报文。通过上述方案，目标存储节点修改发送报文的源地址，伪装为控制器向主机发送携带目标数据的报文，进而可以在对现有的主机不作出改变的情况下，实现存储节点通过与主机间的 TCP/IP 连接直接将数据发送给主机，提高主机与存储节点间数据传输的速度，减少数据传输的耗时。

在第一方面的另一种可能的实现方式中，控制器根据主机的 TCP 窗口大小确定主机的数据接收量，根据该数据接收量确定目标存储节点通过每个 TCP/IP 报文携带的目标数据的数据块，该数据块的大小不大于主机的 TCP 窗口大小，然后，控制器生成包括数据块的标识的指示消息，向目标存储节点发送该指示消息。通过上述方案，控制器能够根据主机的 TCP 窗口大小确定每次向主机发送的数据块，指示存储该数据块的存储节点通过 TCP/IP 报文向主机发送该数据块，进而通过存储节点与主机间的 TCP 连接实现将数据发送至主机。

在第一方面的另一种可能的实现方式中，目标存储节点在向主机发送目标数据或从主

机获取目标数据之后，向控制器发送数据传输成功消息；控制器在接收所有目标存储节点发送的数据传输成功消息之后，向主机发送操作成功消息。通过上述方案，控制器能够在目标存储节点与主机间的目标数据传输完成后，告知主机数据读写成功。

本申请的第二方面提供一种数据处理的方法，该方法由数据处理的系统中的存储节点执行。该系统包括包括控制器以及至少两个存储节点。控制器与每个存储节点建立连接，用于管理该至少两个存储节点。存储节点以及控制器与主机建立连接，主机用于部署应用服务。该方法包括：存储节点与主机建立连接，主机用于部署应用服务；存储节点接收控制器发送的指示消息，指示消息包括主机待操作的目标数据的标识和操作类型；存储节点根据指示消息，通过与主机的连接向主机发送目标数据或从主机获取目标数据。

在第二方面的一种可能的实现方式中，存储节点与主机建立 RDMA 连接，建立连接的过程为：存储节点创建第一 QP，通过第一连接将第一 QP 的参数发送给控制器。控制器通过第二连接将第一 QP 的参数发送给主机。主机创建第二 QP 对，通过第二连接将第二 QP 的参数发送给控制器，控制器则通过第一连接将第二 QP 的参数发送给存储节点。主机根据接收的第一 QP 的参数将第二 QP 与存储节点的第一 QP 关联。储节点根据接收的第二 QP 的参数将自身的第一 QP 与第二 QP 关联。通过上述方案，存储节点与主机能够在控制器的辅助下创建 RDMA 连接，进而能够通过该 RDMA 连接进行数据传输，而无需控制器负责数据的中转，提高数据传输的速率，减少数据传输耗时。

在第二方面的另一种可能的实现方式中，控制器发送的指示消息包括目标数据的标识、操作类型以及主机中目标存储区域的位置参数。存储节点响应该指示消息，在指示消息中的操作类型为读操作时，存储节点通过与主机的 RDMA 连接将目标数据写入主机内存中存储区域；在指示消息中的操作类型为写操作时，存储节点通过与主机的 RDMA 连接从主机内存中存储区域读取目标数据，并存储目标数据。通过上述方案，目标存储节点能够根据与主机的 RDMA 从主机内存中读写数据，完成与主机间目标数据的传输。

在第二方面的另一种可能的实现方式中，主机与存储节点建立 TCP/IP 连接，主机向控制器发送的操作请求的操作类型为读操作。控制器在向目标存储节点发送的指示消息中添加该操作类型、主机的通信地址、控制器的通信地址以及目标数据的标识。目标存储节点响应该指示消息，以控制器的通信地址为源地址以及以主机的通信地址为目的地址发送携带目标数据的 TCP/IP 报文。通过上述方案，目标存储节点修改发送报文的源地址，伪装为控制器向主机发送携带目标数据的报文，进而可以在对现有的主机不作出改变的情况下，实现存储节点通过与主机间的 TCP/IP 连接直接将数据发送给主机，提高主机与存储节点间数据传输的速度，减少数据传输的耗时。

本申请的第三方面提供一种数据处理的方法，该方法由处理数据的系统中的第一控制器执行，该系统包括至少两个控制器以及至少两个磁盘，至少两个控制器用于管理至少两个磁盘，至少两个磁盘的每个 LUN 可以归属于一个控制器管理，该至少两个控制器与该至少两个磁盘形成存储阵列。主机可以通过控制器访问该控制器管理的 LUN 中的目标数据。至少两个控制器与主机建立连接，主机用于部署应用服务。该方法包括：第一控制器接收主机发送的操作请求，操作请求包括待操作的目标数据的标识和操作类型；第一控制器根据预设的负载均衡策略或目标数据的 LUN 的归属确定响应操作请求的第二控制器；向第二控制器发送指示消息，指示第二控制器通过与主机的连接向主机发送目标数据或从主机获取目标数据。通过上述方案，主机与负责为主机提供读写目标数据服务的第二控制

器可以通过二者间建立的连接传输目标数据，而不用不经由接收主机的操作请求的第一控制器转发目标数据，缩短了数据传输的路径长度，减小了目标数据传输的耗时。

在第三方面的一种可能的实现方式中，第二控制器与主机之间的连接为 RDMA 连接；主机向第一控制器发送的操作请求还包括主机中目标存储区域的位置参数；第一控制器在向第二控制器发送的指示消息中添加目标数据的标识、操作类型以及该位置参数，以指示第二控制器在操作类型为读操作时从至少两个磁盘中获取目标数据，通过与主机的 RDMA 连接将目标数据写入主机内存中存储区域；以及在操作类型为写操作时通过与主机的 RDMA 连接从主机内存中存储区域读取目标数据，将目标数据写入至少两个磁盘。通过上述方案，主机与负责为主机提供读写目标数据服务的第二控制器可以通过二者间建立的 RDMA 连接快速地传输目标数据，实现高速地读写数据。

在第三方面的另一种可能的实现方式中，第一控制器根据预设负载均衡策略确定由第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据；或者，根据所述目标数据所在的逻辑单元号 LUN 的归属确定由所述第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据。

在第三方面的另一种可能的实现方式中，第一控制器接收管理所述至少两个磁盘的第三控制器的第四指示消息，第四指示消息包括主机待操作的第二目标数据的标识和操作类型；第一控制器响应该第四指示消息，通过与主机的连接向主机发送第二目标数据或从主机获取第二目标数据。本申请的第四方面提供一种数据处理的装置，该装置用于执行上述第一方面或第一方面的任意可能的实现中的方法。具体的，该装置包括用于执行上述第一方面或第一方面的任意可能的实现中的方法的模块。

本申请的第五方面提供一种数据处理的装置，该装置用于执行上述第二方面或第二方面的任意可能的实现中的方法。具体的，该装置包括用于执行上述第二方面或第二方面的任意可能的实现中的方法的模块。

本申请的第六方面提供一种数据处理的装置，该装置用于执行上述第三方面或第三方面的任意可能的实现中的方法。具体的，该装置包括用于执行上述第三方面或第三方面的任意可能的实现中的方法的模块。

本申请的第七方面提供一种数据处理的设备，包括处理器、存储器、通信接口以及总线，处理器、存储器和通信接口之间通过总线连接并完成相互间的通信，存储器中用于存储计算机执行指令，设备运行时，处理器执行存储器中的计算机执行指令以利用设备中的硬件资源执行上述第一方面或第一方面的任意可能的实现中的方法。

本申请的第八方面提供一种数据处理的设备，包括处理器、存储器、通信接口以及总线，处理器、存储器和通信接口之间通过总线连接并完成相互间的通信，存储器中用于存储计算机执行指令，设备运行时，处理器执行存储器中的计算机执行指令以利用设备中的硬件资源执行上述第二方面或第二方面的任意可能的实现中的方法。

本申请的第九方面提供一种数据处理的设备，包括处理器、存储器、通信接口以及总线，处理器、存储器和通信接口之间通过总线连接并完成相互间的通信，存储器中用于存储计算机执行指令，设备运行时，处理器执行存储器中的计算机执行指令以利用设备中的硬件资源执行上述第三方面或第三方面的任意可能的实现中的方法。

本申请的第十方面提供一种数据处理的系统，该系统包括第七方面所述的设备以及至少两个第八方面所述的设备，用于实现第八方面所述的设备与主机通过二者间的连接直接

传输数据，而不用经由第七方面所述的设备转发数据。

本申请的第十一方面提供一种数据处理的系统，该系统包括至少两个第九方面所述的设备以及至少两个磁盘，用于实现响应主机的操作请求的第九方面所述的设备与主机通过二者间的连接直接传输数据，而不用经由接收主机操作请求的设备转发数据。

本申请的第十二方面提供了一种计算机可读介质，所述计算机可读存储介质中存储有指令，当其在计算机上运行时，使得计算机执行第一方面或第一方面的任意可能的实现中的方法的指令。

本申请的第十三方面提供了一种计算机可读介质，所述计算机可读存储介质中存储有指令，当其在计算机上运行时，使得计算机执行第二方面或第二方面的任意可能的实现中的方法的指令。

本申请的第十四方面提供了一种计算机可读介质，所述计算机可读存储介质中存储有指令，当其在计算机上运行时，使得计算机执行第三方面或第三方面的任意可能的实现中的方法的指令。

本申请的在上述各方面提供的实现的基础上，还可以进行进一步组合以提供更多实现。

附图说明

为了更清楚地说明本发明实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简要介绍。

图 1 为现有技术中 SAN 存储系统的示意图；

图 2 为离散聚合列表 SGL 的示意图；

图 3 为本发明实施例中 SAN 系统的示意图；

图 4 为本发明实施例中 SAN 系统中传输数据方法的流程示意图；

图 5 为本发明实施例中主机与存储节点建立 RDMA 连接的流程示意图；

图 6 为本发明实施例中存储节点与主机的结构示意图；

图 7 为本发明实施例中主机与存储节点通过 RDMA 连接传输目标数据的流程示意图；

图 8a-图 8b 为本发明实施例中主机内存中存储目标数据的存储区域的示意图；

图 9 为本发明实施例中主机与存储节点通过 iSER 连接传输目标数据的流程示意图；

图 10 为本发明实施例中主机与存储节点通过 TCP/IP 连接传输数据方法的流程示意图；

图 11 为本发明实施例中主机与存储节点通过 TCP/IP 连接传输数据的另一方法的流程示意图；

图 12 为本发明实施例中存储节点向主机发送的报文的帧结构示意图；

图 13 为本发明实施例中装置 50 的结构示意图；

图 14 为本发明实施例中装置 60 的结构示意图；

图 15 为本发明实施例中设备 70 的结构示意图；

图 16 为现有技术中存储阵列的系统框图；

图 17 为本发明实施例中存储阵列的系统框图；

图 18 为本发明实施例中数据处理方法的流程示意图；

图 19 为本发明实施例中装置 80 的结构示意图；

图 20 为本发明实施例中设备 90 的结构示意图。

具体实施方式

为了便于理解，下面先介绍现有技术中主机访问远端存储设备中数据的技术方案。

图 1 为一种分布式的存储区域网络（storage area network，SAN）系统的示意图，该系统包括控制器 12 以及至少两个存储节点，如存储节点 13-1、存储节点 13-2、...、存储节点 13-n。该系统用于处理主机 11 的请求消息，主机 11 与控制器 12 建立网络连接，如基于光纤通道（fibre channel，FC）的连接、或基于以太网的连接。主机 11 通过与控制器 12 之间的连接向控制器 12 发送操作请求。控制器 12 与该至少两个存储节点建立网络连接，SAN 系统中采用分布式方式存储数据，如有数据需要写入 SAN 系统时，控制器按照预置算法将数据拆分成多个数据块，分别存储在不同的存储节点中，控制器 12 中记录有每个存储节点所存储数据的信息，该信息也称为存储节点的分区视图。示例地，表 1 为存储节点中分区视图的一种示例，如表所示，表 1 中包括数据块标识、关联原始数据标识、存储介质和校验值，其中，校验值为控制节点根据预置算法计算的该数据块的校验信息,用于在读取或写入数据块时确定数据块的完整性；存储介质用于标识数据块存储在存储节点中的目标存储介质信息。

表 1 存储节点中分区视图的示例

数据块标识	关联原始数据标识	存储介质	校验值
数据块 11	数据 1	存储介质 1	校验值 1
数据块 21	数据 2	存储介质 2	校验值 2
数据块 31	数据 3	存储介质 3	校验值 3

可选的，当 SAN 系统中采用分布式方式存储数据时，在控制器中保存有 SAN 系统中存储节点的全局视图，该全局视图记录有每个存储节点所存储数据的信息。示例地，表 2 为全区视图的一种示例。

表 2 控制节点中全局分区视图的示例

数据块标识	关联原始数据标识	校验值	存储节点	存储介质
数据块 11	数据 1	校验值 1	存储节点 1	存储介质 1
数据块 12	数据 1	校验值 2	存储节点 2	存储介质 1
数据块 13	数据 1	校验值 3	存储节点 3	存储介质 1

当控制器 12 接收主机 11 发送的读请求时，根据读请求中待操作的目标数据的标识以及该分区视图确定存储目标数据的目标存储节点（比如目标存储节点为存储节点 13-1）。然后，控制器 12 向目标存储节点 13-1 发送读请求。针对控制器 12 发送的读请求，目标存储节点 13-1 向控制器 12 返回目标数据；控制器 12 接收目标存储节点 13-1 返回的目标数据，将目标数据返回给主机 11。

当控制器 12 接收主机 11 发送的写请求时，根据写请求中待操作的目标数据的标识以及该分区视图确定存储目标数据的目标存储节点 13-1。然后，控制器 12 向目标存储节点发送写请求。针对控制器 12 发送的写请求，目标存储节点 13-1 将目标数据写入存储介质，并向控制器 12 返回写数据响应消息。当控制器 12 接收的响应消息指示写数据成功时，向主机 11 返回写操作成功消息。

通过上面的流程可以看出，无论是主机从存储节点读数据的场景，还是主机向存储节点写数据的场景，均需要由控制器来转发待操作的数据，而且，控制器在转发该待操作的数据时，还需要对数据进行封包、解包等处理。实际情况中，SAN 系统可用于处理多个主机的操作请求，此时，控制器可能会同时处理多个主机对存储节点的操作请求，使得控制器的数据传输负担以及运算负担过重，制约主机与存储节点间读写数据的速度。

为了解决上述接收主机操作请求的控制器制约主机与存储节点间读写数据速度，增大主机与存储节点间读写数据耗时的问题，本发明的实施例提供一种数据处理的系统，下面通过附图以及具体实施例对该数据处理的系统做详细的说明。

首先介绍本发明各实施例中涉及的部分概念。

（1）远程直接数据存取（remote direct memory access，RDMA），为一种直接进行远程内存存取的技术，可以直接将数据从一个设备的存储器快速迁移到另一个远程设备的存储器中，减少了设备的中央处理器（central processing unit，CPU）参与数据传输过程的消耗，进而提升了系统处理业务的性能，具有高带宽、低时延及低 CPU 占用率的特点。

（2）队列对（queue pair，QP），包括接收队列（send queue，SQ）以及发送队列（receive queue，RQ）。建立 RDMA 连接的两端均需分别建立 QP，然后将自身的 QP 中的 SQ 与对端的 QP 的 RQ 关联，以及将自身的 QP 中的 RQ 与对端的 QP 中的 SQ 关联，进而实现自身的 QP 与对端的 QP 关联，两端建立 RDMA 连接。

（3）具有 RDMA 功能的网卡（RDMA enabled network interface card，RNIC），用于实现基于 RDMA 连接的数据传输。在设备 A 与设备 B 的基于 RDMA 连接的数据传输中，设备 A 的 RNIC 可以直接从设备 A 的内存读取数据，将读取的数据发送给设备 B 的 RNIC，设备 B 的 RNIC 将从设备 A 的 RNIC 接收的该数据写入设备 B 的内存中。本发明的实施例中，RNIC 可以为支持 RDMA 的主机总线适配器（host bus adapter，HBA）。

（4）RDMA 操作类型，包括用于传输命令的 RDMA send、RDMA receive，以及用于传输数据的 RDMA read、RDMA write。上述四种操作类型的具体实施方式请参照现有技术，在此不予详述。

（5）iSER 连接，指的是基于远程内存直接访问方式扩展的网络小型计算机系统接口（iSCSI extensions for RDMA，iSER）协议的连接，iSCSI 指的是互联网小型计算机系统接口（internet small computer system interface,iscsi），iSER 协议支持 RDMA 传输。

（6）基于网络的非易失性存储总线（NVMe over fabric，NOF）连接，NVMe 指非易失性存储总线（non-volatile memory express ，NVMe），NOF 协议支持 RDMA 传输。

（7）离散聚合列表（scatter gather list，SGL），指一种数据封装形式。在 RDMA 传输数据的过程中，要求源物理地址以及目标物理地址均必须是连续的。但实际情况中，数据的存储地址在物理空间上不一定是连续的，在这种情况下，将不连续的物理存储地址以 SGL 的形式封装，设备在传输完一块存储在物理连续的存储空间的数据后，根据该 SGL 传输下一块存储在物理连续的存储空间的数据。

图 2 所示为 SGL 的示意图，SGL 包括多个离散聚合实体（scatter gather entries，SGE），每个 SGE 包括地址字段（address）、长度（length）字段以及 flag，其中，address 字段表征存储区域的起始位置，length 字段表征存储区域的长度，flag 字段表征该 SGE 是否为该 SGL 中最后一个，flag 字段还可以包括其他辅助信息，例如数据块描述符。每个 SGE 根据自身包含的地址字段以及长度字段表征一段连续的存储区域，若干个表征的存储区域顺次相连的

SGE 组成一组，SGL 中的不同组的 SGE 所表征的存储区域之间不相邻，每组 SGE 的最后一个 SGE 指向下一组 SGE 的起始地址。

（8）逻辑单元号（logical unit number，LUN），LUN 只是个号码标识，不代表任何实体属性。在存储阵列中，将多个磁盘按照预置算法（如独立冗余磁盘阵列（redundant array of independent disks，RAID）配置关系）组成一个逻辑磁盘，再将该逻辑磁盘按照预置规则切分成不同的条带，每个条带称为一个 LUN，其中，一个 LUN 可以用于描述存储阵列的一个磁盘的一个连续的存储区域，或者描述一个磁盘的多个不连续的存储区域的集合，或者描述多个磁盘中的存储区域的集合。

图 3 为本发明实施例中的一种分布式存储系统的示意图，该系统包括主机 40、控制器 20 以及控制器 20 管理的至少两个存储节点，如存储节点 31、存储节点 32、存储节点 33。控制器 20 与存储节点 31、存储节点 32、存储节点 33 分别建立连接，控制器 20 与主机 40 建立连接，存储节点 31、存储节点 32、存储节点 33 分别与主机 40 建立连接。主机 40 用于部署应用服务。

为了便于区分，本发明的实施例中，将控制器 20 与存储节点间的连接称为第一连接，将控制器 20 与主机 40 间的连接称为第二连接，将存储节点与主机 40 间的连接称为第三连接。上述第一连接、第二连接以及第三连接可以为基于有线通信的连接，如基于 FC 的连接；上述第一连接、第二连接以及第三连接还可以为基于无线通信的连接，如基于蜂窝（cellular communication）通信的连接，又如无线保真（wireless fidelity，WIFI）连接。

图 4 所示为根据图 3 所示的系统进行数据处理的方法，该方法包括：

步骤 101：控制器与存储节点建立第一连接。

存储节点的数量为两个或两个以上，控制器分别与每个存储节点建立第一连接，控制器用于通过与存储节点之间的第一连接对存储节点进行管理。例如，根据主机的请求指示存储节点存储/返回/删除数据。

步骤 102：控制器与主机建立第二连接。

步骤 103：主机与存储节点建立第三连接。

步骤 104：主机通过第二连接向控制器发送操作请求，该操作请求包括待操作的目标数据的标识和操作类型。

本发明的实施例中，将读请求、写请求统称为操作请求。操作请求包括操作类型，用于表明请求的操作为读操作或写操作。目标数据的标识用于唯一标识目标数据，例如，在采用键值对（key-value）的方式存储数据时，key 值就是数据的标识。

步骤 105：控制器接收操作请求，根据目标数据的标识以及全局视图或分区视图确定目标存储节点。

控制器从该操作请求中获取目标数据的标识，根据该目标数据的标识以及上述全局视图或分区视图确定目标存储节点，当所述操作请求的操作类型为写操作时，所述目标存储节点为待写入所述目标数据的存储节点，当所述操作请求的操作类型为读操作时，所述目标存储节点为存储有所述目标数据的存储节点。

例如，在操作请求中的操作类型为读操作时，控制器在上述全局视图或分区视图中查找目标数据的标识，确定存储目标数据的标识的目标存储节点。该目标存储节点可以为一个或一个以上，比如，目标数据被切分为多个数据块，多个数据块被分别存储在多个存储节点中。

又例如，在操作请求中的操作类型为写操作时，控制器会根据预置算法将待写入数据切分成多个数据块，分别存储在多个存储节点中，此时，存储数据块的每个存储节点均为用于存储目标数据的目标存储节点。

步骤 106：控制器向目标存储节点发送指示消息，该指示消息用于指示目标存储节点通过与主机的连接向主机发送目标数据或从主机获取目标数据。

作为一种可选的方式，在目标存储节点为多个时，控制器可以向多个目标存储节点发送指示消息，指示每个目标存储节点通过与主机间的第三连接传输目标数据的数据块。由于多个目标存储节点可以根据指示消息同时与主机进行目标数据的数据块的传输，能够提高目标数据传输的效率，减少目标数据传输的耗时。

作为另一种可选的方式，控制器向多个目标存储节点逐一发送该指示消息，在前一个目标存储节点与主机间的数据传输结束之后，再向后一个目标存储节点发送指示消息。通过上述方案，控制器可以控制目标数据传输的有序进行，保证传输数据的正确性。

步骤 107：目标存储节点根据控制器发送的指示消息向主机发送目标数据或从主机获取目标数据。

通过上述方案，主机通过控制器指示目标存储节点向主机发送目标数据或从主机获取目标数据，且目标存储节点与主机直接通过二者间的连接传输目标数据，而不用经由控制器转发数据，减短数据传输路径进而减小传输耗时，且避免因控制器的带宽无法满足大量数据传输需求而导致的主机读写数据速率较慢的问题，还可以避免因控制器的运算能力无法满足大量数据处理需求而导致的主机读写数据速率较慢的问题。

进一步地，本发明实施例中，主机与存储节点之间建立的第三连接可以有多种实现方式，下面分别进行介绍。

（一）第三连接为 RDMA 连接。

参照图 5，主机与存储节点建立 RDMA 连接的过程如下：

步骤 201：存储节点创建第一 QP。

为了便于区分，本发明实施例中将存储节点建立的 QP 称为第一 QP。第一 QP 包括发送队列 SQ 和接收队列 RQ，SQ 用于发送数据，RQ 用于接收数据。可选的，第一 QP 还包括完成队列，该完成队列用于检测第一 QP 的 SQ 的发送数据任务以及第一 QP 的 RQ 的接收数据任务是否完成。

步骤 202：存储节点通过第一连接将第一 QP 的参数发送给控制器。

该第一 QP 的参数可以包括第一 QP 的标识，第一 QP 的标识可以是用于标识该 QP 的数字、字母或其他形式的组合。除此之外，也可以包括为第一 QP 配置的保护域（protect domain，PD）的标识，PD 用于表征存储节点为 RDMA 连接所分配的具有 RDMA 功能的网卡（如 RNIC）；第一 QP 的参数还可以包括存储节点分配的协助建立 RDMA 连接以及管理 RDMA 连接的连接管理器（connection manager，CM）的标识。

步骤 203：控制器通过第二连接将第一 QP 的参数发送给主机。

步骤 204：主机创建第二 QP。

为了便于区分，本发明实施例中将主机创建的 QP 称为第二 QP。第二 QP 包括发送队列 SQ 和接收队列 RQ。可选的，第二 QP 还包括完成队列，该完成队列用于检测第二 QP 的 SQ 的发送数据任务以及第二 QP 的 RQ 的接收数据任务是否完成。

步骤 205：主机通过第二连接将第二 QP 的参数发送给控制器。

该第二 QP 的参数可以包括第二 QP 的标识，除此之外，也可以包括为第二 QP 配置的保护域 PD 的标识、为第二 QP 配置的连接管理器 CM 的标识等。

步骤 206：控制器通过第一连接将第二 QP 的参数发送给存储节点。

步骤 207：主机根据接收的第一 QP 的参数将第二 QP 与存储节点的第一 QP 关联。

步骤 208：存储节点根据接收的第二 QP 的参数将自身的第一 QP 与第二 QP 关联。

上述第一 QP 与第二 QP 的关联指的是，根据第一 QP 的标识以及第二 QP 的标识将第一 QP 的 SQ 与第二 QP 的 RQ 绑定，进而创建从存储节点向主机发送数据的通路；以及，将第一 QP 的 RQ 与第二 QP 的 SQ 绑定，进而创建从主机向存储节点发送数据的通路。

需要说明的是，上述步骤 201 存储节点创建第一 QP 与步骤 204 主机创建第二 QP 可以同时进行，也可以为先执行步骤 201 后执行步骤 204，还可以为先执行步骤 204 后执行步骤 201。

通过上述方案，主机能够与存储节点建立 RDMA 连接，由于该 RDMA 连接为主机与存储节点之间直接建立的连接，主机与存储节点通过该 RDMA 连接进行目标数据的传输无需中转，可以减少传输耗时。而且，RDMA 连接本身的传输速率也很快，能够进一步减小传输耗时。

接下来，结合图 6 进一步介绍存储节点与主机基于 RDMA 连接进行数据传输的实现方式，图 6 为主机与存储节点通过 RDMA 连接传输数据的示意图，主机 40 包括 RNIC41（如 HBA 卡）以及内存 42，存储节点 31 包括 RNIC311 以及内存 312。

存储节点 31 的 RNIC311 可以向主机 41 的 RNIC41 发送读取内存 42 指定位置的数据的请求，RNIC41 从内存 42 指定位置读取数据，将该数据发送至 RNIC311，RNIC311 将接收的数据写入内存 312，上述过程称为存储节点以 RDMA read 方式从主机 40 读取数据。

存储节点 31 的 RNIC311 还可以向主机 40 的 RNIC41 发送向内存 42 指定位置写入数据的请求，RNIC41 缓存该请求携带的数据，并将数据写入内存 42 中该请求所指定的位置，上述过程称为存储节点以 RDMA write 方式向主机 40 写入数据。

进一步地，图 7 为基于 RDMA 连接的数据处理方法的流程示意图，该方法包括如下步骤：

步骤 301，主机通过第二连接向控制器发送操作请求，该操作请求包括操作类型、目标数据的标识以及主机在内存中为目标数据指定的目标存储区域的位置参数。

该目标存储区域的位置参数用于标识目标数据在主机内存中的存储位置，其表现形式可以为内存中的偏移量。该操作请求还可以包括目标数据的长度、远端密钥（remote key、Rkey）等信息。

步骤 302，控制器接收操作请求，根据该操作请求中的目标数据的标识确定目标存储节点。

当所述操作请求为写请求时，所述目标存储节点为待写入所述目标数据的存储节点，当所述操作请求为读请求时，所述目标存储节点为存储有所述目标数据的存储节点。

步骤 303，控制器向目标存储节点发送指示消息，该指示消息包括该目标存储节点对应的目标数据的数据块的标识、操作类型以及目标存储区域的位置参数。

步骤 304，目标存储节点响应指示消息，根据目标存储区域的位置参数通过与主机的 RDMA 连接向主机发送目标数据或从主机获取目标数据。

具体的，当指示消息中的操作类型为写操作时，目标存储节点以前述 RDMA read 方式，

从主机内存中目标存储区域读取目标数据。然后，目标存储节点将读取的目标数据写入该目标存储节点的磁盘中。

当指示消息中的操作类型为读操作时，目标存储节点以前述 RDMA write 方式，向主机内存中该目标存储区域写入目标数据。然后，主机将主机的内存中目标存储区域存储的该目标数据写入主机的磁盘。

步骤 305，目标存储节点在完成通过 RDMA 连接向主机发送目标数据或从主机获取目标数据后，向控制器发送数据传输成功消息。

在操作类型为写操作时，目标存储节点在以 RDMA read 方式读取主机内存中该目标存储区域的数据后，向控制器发送数据传输成功消息。

在操作类型为读操作时，目标存储节点在以 RDMA write 方式将存储的目标数据写入主机内存中该目标存储区域后，向控制器发送数据传输成功消息。

步骤 306，控制器在接收到目标存储节点发送的数据传输成功消息后，向主机发送操作成功消息。

通过上述方案，待操作的数据经由主机与存储节点之间的 RDMA 连接传输，不用经由控制器，不仅减轻控制器的带宽负担以及运算负担，而且通过高速的 RDMA 连接实现数据传输，数据传输速率更快，整个写数据的过程耗时更短。

作为一种可选的方式，目标数据被划分为多个数据块，不同数据块对应的目标存储节点的磁盘中的存储区域不连续。具体又可以有两种场景：

场景 1，目标数据对应的目标存储节点的个数大于 1，目标数据的不同数据块可以对应不同的目标存储节点。例如，目标数据被划分为第一数据块以及第二数据块，其中，第一数据块对应第一存储节点，第二数据块对应第二存储节点。

场景 2，目标数据对应的目标存储节点的个数为 1，但是，目标数据的不同数据块对应的目标存储节点的磁盘中的存储区域不连续。例如，目标数据被划分为第一数据块以及第二数据块，其中，第一数据块对应目标存储节点中的第一存储区域，第二数据块对应目标存储节点中的第二存储区域，第一存储区域与第二存储区域不连续。

在上述两种场景中，控制器还要为目标数据的每个数据块确定在主机内存中目标存储区域中对应的子存储区域，进而指示目标存储节点从该子存储区域获取该目标数据的数据块，或者将该目标数据的数据块写入主机内存中该子存储区域。下面以场景 1 为例进行说明。

请参照图 8a，主机在内存中为目标数据指定的目标存储区域为连续存储区域时，控制器确定目标数据的第一数据块由第一存储节点存储，从目标数据在内存中的存储区域中划分出用于存储该第一数据块的第一子存储区域，以及确定目标数据的第二数据块由第二存储节点存储，从目标数据在内存中的存储区域中划分出用于存储该第二数据块的第二子存储区域，该第一子存储区域以及第二子存储区域为连续存储区域。

请参照图 8b，该目标存储区域为非连续存储区域时，控制器确定目标数据的第一数据块由第一存储节点存储，从目标数据在内存中的存储区域中划分出用于存储该第一数据块的第一存储区域，该第一存储区域为由多个非连续存储区域所组成的存储区域。控制器确定目标数据的第二数据块由第二存储节点存储，从目标数据在内存中的存储区域中划分出用于存储该第二数据块，该第一存储区域为由多个非连续存储区域所组成的存储区域。

针对场景 2，控制器为目标数据的数据块确定在主机的内存中对应的子存储区域的实

现方式与场景 1 相一致，在此不再重复。

在确定出目标数据的数据块在主机的内存中对应的子存储区域后，控制器向目标存储节点发送指示消息，该指示消息包括该目标存储节点对应的目标数据的数据块的标识、操作类型以及为该数据块确定的子存储区域的位置参数。然后，目标存储节点响应指示消息，根据子存储区域的位置参数通过与主机的 RDMA 连接向主机发送目标数据的数据块或者从主机获取目标数据的数据块。再然后，目标存储节点在完成通过 RDMA 连接向主机发送目标数据或从主机获取目标数据的数据块后，向控制器发送数据传输成功消息。最后，控制器在接收到所有目标存储节点发送的数据传输成功消息后，向主机发送操作成功消息。

在场景 2 中，在确定出目标数据的数据块在主机的内存中对应的子存储区域后，控制器向目标存储节点发送指示消息，该指示消息包括该目标存储节点对应的目标数据的数据块的标识、操作类型以及为该数据块确定的子存储区域的位置参数。然后，目标存储节点响应指示消息，根据子存储区域的位置参数通过与主机的 RDMA 连接向主机发送目标数据的数据块或者从主机获取目标数据的数据块。然后，目标存储节点在完成通过 RDMA 连接向主机发送目标数据或从主机获取目标数据的数据块后，向控制器发送数据传输成功消息。然后，控制器在接收到所有目标存储节点发送的数据传输成功消息后，向主机发送操作成功消息。

通过上述方案，控制器能够为目标数据的数据块确定主机的内存中对应的字存储区域，进而使得该数据块对应的目标存储节点能够通过 RDMA 连接从主机获取该数据块或者向主机发送该数据块。

作为一种可能的实现方式,下面结合图 9 介绍 iSER 协议下数据处理方法的流程，其中，iSER 协议支持 RDMA 连接，包括如下步骤：

步骤 401,主机向控制器发送建立 iSER 连接请求。

步骤 402,控制器向主机返回建立连接响应，与主机建立 iSER 连接。

步骤 403,主机创建第二 QP，并通过与控制器的 iSER 连接将第二 QP 的参数发送给控制器。

步骤 404,控制器通过与存储节点间的第一连接将第二 QP 的参数发送给每个存储节点。

步骤 405,存储节点创建第一 QP，并通过第一连接将第一 QP 的参数发送给控制器。

步骤 406,控制器通过 iSER 连接将第一 QP 的参数发送给主机。

步骤 407,主机根据第一 QP 的参数将第一 QP 与第二 QP 关联。

步骤 408,存储节点根据第二 QP 的参数将第二 QP 与第一 QP 关联。

步骤 409,主机通过 iSER 连接向控制器发送控制请求。

该控制请求用于请求控制器准许主机以 RDMA 的方式向控制器发送命令请求，如后文中的操作请求。该控制请求的具体实现方式请参照现有技术。

步骤 410,控制器通过 iSER 连接向主机返回控制响应。

该控制响应表征控制器准许主机以 RDMA 的方式向控制器发送命令请求。

步骤 411,主机通过 iSER 连接向控制器发送操作请求。

该操作请求包括操作类型、待操作的目标数据的标识以及主机在内存中为该目标数据指定的目标存储区域的位置参数。

在一种可能的实现方式中，该操作请求可以不包括该目标存储区域的位置参数，主机

在发送操作请求之后再向控制器发送目标存储区域的位置参数。

步骤 412,控制器根据该操作请求确定目标存储节点，并根据目标存储区域的位置参数确定主机在内存中为每个目标存储节点对应的数据块所指定的子存储区域的位置参数。

控制器根据本地存放的分区视图或全局视图确定目标数据存放或者应该被存放在哪些存储节点上，确定出的存储节点即为目标存储节点，其中，每个目标存储节点存储该目标数据的一个或多个数据块。然后，控制器对主机内存中用于存放目标数据的存储区域进行划分，确定主机内存中为每个目标存储节点所对应的数据块指定的存储区域。

步骤 413,控制器向每个目标存储节点发送指示消息，该指示消息包括该目标存储节点对应的数据块的标识以及为该目标存储节点确定的子存储区域的位置参数。

步骤 414,目标存储节点响应该指示消息，通过与主机的 RDMA 连接向主机发送目标数据的数据块或从主机获取目标数据的数据块。

在指示消息中的操作类型为写操作时，目标存储节点根据指示消息中的子存储区域的位置参数，以 RDMA read 的方式从该子存储区域读取数据。然后将读取的数据写入目标存储节点的内存，再将该数据从内存写入磁盘，记录该数据在磁盘中的存储位置。

在指示消息中的操作类型为读操作时，目标存储节点根据指示消息中的子存储区域的位置参数，以 RDMA write 的方式将该目标存储节点所存储的目标数据的数据块写入该子存储区域。

步骤 415,目标存储节点在完成通过 RDMA 连接向主机发送目标数据或从主机获取目标数据的数据块后，向控制器发送数据传输成功消息。

步骤 416,控制器在接收到每个目标存储节点发送的数据传输成功响应后，通过 iSER 连接向主机发送操作成功消息。

在上述步骤 401 至步骤 416 的流程中，主机与存储节点之间的数据通过二者之间的 RDMA 连接，而不是经由控制器转发，不仅降低了控制器的负载，避免现有技术中因控制器转发数据带来的负载过高导致主机与存储节点数据传输速度较慢的问题。而且，在目标存储节点为两个或两个以上时，不同存储节点可以同时与主机通过 RDMA 连接传输数据，进一步提高主机与存储节点间传输数据的速率。

作为另一种可能的实现方式,NOF 协议也支持 RDMA 连接，NOF 协议下数据处理方法的流程与上述步骤 401 至步骤 416 的流程相一致。

继续参照图 8b，与 iSER 协议不同的是,NOF 协议还支持主机将内存中的该目标存储区域配置为非连续，主机将目标存储区域的位置参数以 SGL 的形式封装得到 SGL 包，并以 RDMA write 方式将该 SGL 包写入控制器的内存。控制器对该 SGL 包解析，并确定主机的目标存储区域中为每个目标存储节点对应的数据块所指定的子存储区域的位置参数，以 SGL 的形式封装子存储区域的位置参数，并将 SGL 形式封装的子存储区域的位置参数发送给存储节点，存储节点对该 SGL 封包进行解析，获得该子存储区域的位置参数。通过上述方式，在主机将数据存储在内存的离散存储区域的情况下实现存储节点与主机之间通过 RDMA 连接进行数据传输，使得主机能够充分利用内存存储区域，提高内存利用率。

（二）、第三连接为 TCP/IP 连接。

主机与存储节点均包括支持 TCP/IP 协议的通信端口，主机的该通信端口与存储节点的该通信端口能够建立通信链路，即为第三连接，本发明实施例中，主机与存储节点能够通过该第三连接传输数据。

图 10 为在第三连接为 TCP/IP 连接时，主机与存储节点之间传输数据方法的流程示意图，该方法包括如下步骤：

步骤 501,主机通过第二连接向控制器发送操作请求，该操作请求包括目标数据的标识以及操作类型；

步骤 502,控制器响应该操作请求，确定存储目标数据的目标存储节点，以及确定每个目标存储节点所对应的数据块。

步骤 503,控制器通过第一连接向每个目标存储节点发送指示消息，该指示消息包括主机的通信地址，以及该目标存储节点所对应的数据块的标识。

具体的，数据块的标识用于使控制器在分布式系统中确定目标数据在存储节点的存储位置。可选的，该标识中还可以包括验证信息，如身份验证、数据访问许可信息。

步骤 504,控制器通过第二连接向主机发送每个目标存储节点的通信地址以及每个目标存储节点所对应的数据块的标识。

上述通信地址可以为 IP 地址或者媒体访问控制（media access control，MAC）地址。控制器可以从主机发送的操作请求中获得主机的通信地址，而控制器本地可以存储其连接的每个存储节点的通信地址。

步骤 505,每个目标存储节点与主机基于对方的通信地址以 TCP/IP 报文的方式传输目标数据中该目标存储节点所对应的数据块。

具体地，步骤 505 可以有多种实现方式，包括但不限于以下方式：

方式 1，控制器指示目标存储节点发起与主机间的数据传输。

例如，在操作请求中的操作类型为读操作时，控制器在确定出每个目标存储节点以及每个目标存储节点所对应的数据块后，向每个目标存储节点发送指示消息，该指示消息包括主机的通信地址以及该存储节点需返回的数据块的标识。存储节点响应该指示消息，发送以主机为目的地的 TCP/IP 报文，该 TCP/IP 报文包括指示消息所指示的数据块。由于控制器已将目标存储节点的通信地址以及该目标存储节点所对应的数据块的标识发送给主机，所以，主机在接收到目标存储节点发送的 TCP/IP 报文时，可以确认该报文为合法报文，且确定该 TCP/IP 报文中包含的数据块为目标数据中的数据块，并从 TCP/IP 报文中获取该数据块。主机在接收每个目标存储节点发送的 TCP/IP 报文后，可以获得目标数据，实现从存储节点读取该目标数据。

又例如，操作请求中的操作类型为写操作时，存储节点响应控制器发送的指示消息，向主机发送 TCP/IP 读请求报文，该 TCP/IP 读请求报文包括该目标存储节点负责存储的数据块的标识。主机响应该 TCP/IP 读请求报文，向该存储节点发送携带有该目标存储节点所对应的数据块的报文。

方式 2，控制器指示主机发起与目标存储节点间的数据传输。

一方面，控制器向主机返回每个目标存储节点的通信地址以及该目标存储节点负责传输的数据块的标识，主机接收到上述信息后，主动向目标存储节点发送数据传输请求报文。

另一方面，控制器向每个目标存储节点发送指示消息，该指示消息包括主机的通信地址以及待操作的数据块的标识，该指示消息的作用不是指示目标存储节点主动向主机返回数据，而是告知目标存储节点合法的数据传输需求，以便在目标存储节点接收到主机的数据传输请求报文时，将其识别为合法报文，响应该报文，与主机进行数据传输。

通过上述方案，主机与存储节点之间能够通过 TCP/IP 连接传输待操作的数据，而不

用经由控制器转发数据，避免因控制器的带宽无法满足大量数据传输需求而导致的主机读写数据速率较慢的问题，而且，也可以避免因控制器的运算能力无法满足大量数据处理需求而导致的主机读写数据速率较慢的问题。

作为一种可能的实现方式，在步骤 505 之后，还包括如下步骤：

步骤 506,目标存储节点在完成与主机通过 TCP/IP 连接传输目标数据的数据块之后，向控制器发送数据传输成功响应；

步骤 507：控制器在接收到所有目标存储节点发送的数据传输成功响应后，向主机发送操作成功消息。

通过上述方案，控制器可以在每个目标存储节点的数据传输任务完成后，向主机发送操作成功消息，告知主机数据读取成功，便于主机及时确认读写数据成功。

接下来，结合图 11 具体介绍当第三连接为 TCP/IP 连接时本发明提供的另一种数据处理的方法，该方法包括如下步骤：

步骤 601,主机通过第二连接向控制器发送操作请求，该操作请求包括目标数据的标识以及操作类型，操作类型为读操作；

步骤 602,控制器确定存储目标数据的目标存储节点，以及确定每个目标存储节点所存储的目标数据的数据块。

步骤 603,控制器通过第一连接向每个目标存储节点发送指示消息，该指示消息包括主机的通信地址、控制器的通信地址以及该目标存储节点所存储的数据块的标识。

作为一种可能的方式，控制器根据主机的 TCP 窗口大小确定主机的数据接收量，根据该数据接收量确定目标存储节点通过每个 TCP/IP 报文携带的目标数据的数据块，该数据块的大小不大于主机的 TCP 窗口大小，然后，控制器生成包括所述数据块的标识的所述指示消息，向目标存储节点发送该指示消息。

其中，控制器获得主机的 TCP 窗口可以有多种实现方式，例如，主机与控制器之间的第二连接为支持 TCP/IP 的连接（如 iSCSI 连接），主机在于控制器建立第二连接时会协商 TCP 窗口的大小,控制器进而获知主机的 TCP 窗口。又例如,主机在向控制器发送的 TCP/IP 报文中携带有主机当前可用的 TCP 窗口大小，控制器从该 TCP/IP 报文中确定主机的 TCP 窗口。

例如，控制器根据该 TCP 窗口确定每次均向主机发送 1000 字节的数据块，然后，控制器确定目标数据的第一位起长度为 1000 字节的数据块所在的目标存储节点，向该目标存储节点发送指示消息,指示该目标存储节点向主机发送该 1000 字节的数据块,在该 1000 字节的数据块发送成功后，再指示存储目标数据的第二个长度为 1000 字节的数据块所在的目标存储节点向主机发送该第二个长度为 1000 字节的数据，以此类推，直至目标存储节点将目标数据全部发送至主机。

在一种可能的实现方式中，控制器确定每次向主机发送的数据块的长度可以不同，这是因为主机的可用 TCP 窗口是动态变化的。例如，控制器确定第一次向主机发送目标数据的第一个长度为 1000 字节的数据块，指示该目标存储节点向主机发送该 1000 字节的数据块，然后，再确定第二次向主机发送长度为 800 字节的数据块，指示存储该第一个长度为 1000 字节的数据块之后的长度为 800 字节的数据块的目标存储节点向主机发送该长度为 800 字节的数据块。

在一种可能的实施方式中，控制器可以将主机的 TCP 窗口同时分配给多个目标存储节

点，例如，主机的 TCP 窗口大小为 1000 字节，目标数据的大小为 450 字节的第一数据块存储在第一目标存储节点，目标数据的大小为 500 字节的第二数据块存储在第二目标存储节点，控制器可以同时向第一目标存储节点以及第二目标存储节发送指示消息，指示第一目标存储节点向主机发送第一数据块，指示第二目标存储节点向主机发送第二数据块，由于第一数据块与第二数据块的大小之和不大于主机的 TCP 窗口大小，因而主机能够成功接收第一数据块以及第二数据块。通过上述方式，可以充分利用主机的 TCP 窗口，提高数据传输的效率。

步骤 604,目标存储节点响应该指示消息，以控制器的通信地址为源地址以及以主机的通信地址为目的地址发送携带目标数据的数据块的 TCP/IP 报文。

图 12 所示为目标存储节点发送的 TCP/IP 报文的网络层报文头以及传输层报文头的示意图，目标存储节点通过将网络层报文头中源 IP 地址设置为控制器的 IP 地址，将传输层报文头中源端口号、TCP 序列号、TCP 响应序列号、发送窗口大小等参数均设置为控制器的参数。控制器可以在给目标存储节点发送的指示消息中添加上述参数，以使目标存储节点获得上述参数。

步骤 605,主机接收目标存储节点发送的源地址为控制器的通信地址的该 TCP/IP 报文，从该 TCP/IP 报文中获取目标数据的数据块。

具体的，控制器一次指示所有目标存储节点中的一个或多个向主机发送目标数据的数据块，这些数据块的大小之和不大于主机的 TCP 窗口大小。主机在接收上述数据块之后，向控制器返回接收数据响应。控制器在接收该接收数据响应之后，确定向目标存储节点发送的指示消息均被成功响应，继续向所有目标存储节点中的一个或多个发送指示消息，指示该一个或多个目标存储节点向主机分别发送目标数据的数据块，所有数据块的大小之和不大于主机的 TCP 窗口大小。通过上述方案，控制器可以根据主机返回的接收数据响应消息控制目标存储节点有序地向主机发送目标数据的数据块，保证数据传输的准确性。

结合上一段所述的实施方式，如果控制器在向目标存储节点发送指示消息后预设时长内未能从主机接收该接收数据响应消息，控制器向该存储节点发送重传指令，指示存储节点重新向主机发送数据。本实现方式中，通过上述重传机制保证数据传输的顺利进行。

通过步骤 601 至步骤 606 所述方案，存储节点修改发送报文的源地址，伪装为控制器向主机发送携带目标数据中的数据块的 TCP/IP 报文，主机将该 TCP/IP 报文识别为由控制器返回的目标数据的数据块。由于主机能够将存储节点发送的 TCP/IP 报文识别为控制器所发送的报文，进而可以在对现有的主机不作出改变的情况下，实现存储节点通过与主机间的 TCP/IP 连接直接将数据发送给主机，而不用经由控制器转发数据，减小将数据由目标存储节点发送至主机的耗时。除此之外，还可以避免因控制器的带宽无法满足大量数据传输需求而导致的主机读写数据速率较慢的问题，而且，也可以避免因控制器的运算能力无法满足大量数据处理需求而导致的主机读写数据速率较慢的问题。再者，由于可以不对现有计算机的硬件以及其工作协议进行改变，成本较低。

图 13 所示为本发明实施例的一种数据处理的装置 50，该装置 50 对应图 3 所示数据处理系统中的控制器 20,用于实现图 4 至图 12 所述数据处理方法中控制器的功能。所述装置用于管理所述至少两个存储节点，所述装置 50 包括：

第一接收模块 51，用于接收所述主机发送的操作请求，所述操作请求包括待操作的目标数据的标识和操作类型；

确定模块 52，用于根据所述目标数据的标识从所述至少两个存储节点中确定至少一个目标存储节点；

第一发送模块 53，用于向所述至少一个目标存储节点发送指示消息，所述指示消息用于指示所述至少一个目标存储节点通过与所述主机的连接向所述主机发送所述目标数据或从所述主机获取所述目标数据。

应理解的是，本发明实施例的装置 50 可以通过专用集成电路（application-specific integrated circuit，ASIC）实现，或可编程逻辑器件（programmable logic device，PLD）实现，上述 PLD 可以是复杂程序逻辑器件（complex programmable logic device，CPLD），现场可编程门阵列（field-programmable gate array，FPGA），通用阵列逻辑（generic array logic，GAL）或其任意组合。也可以通过软件实现图 4 至图 12 中所示的数据处理方法时，装置 50 及其各个模块也可以为软件模块。

可选的，所述确定模块 52，用于根据所述目标数据的标识确定所述目标存储节点包括存储所述目标数据的第一数据块的第一存储节点以及存储所述目标数据的第二数据块的第二存储节点；

所述第一发送模块 53，具体用于向所述第一存储节点发送第一指示消息，所述第一指示消息包括所述第一数据块的标识，用于指示所述第一存储节点向所述主机发送所述第一数据块或从所述主机获取所述第一数据块；向所述第二存储节点发送第二指示消息，所述第二指示消息包括所述第二数据块的标识，用于指示所述第二存储节点向所述主机发送所述第二数据块或从所述主机获取所述第二数据块。

可选的，所述操作请求还包括所述主机在内存中为所述目标数据指定的目标存储区域的位置参数；

所述确定模块 52 还用于根据所述目标存储区域的位置参数确定所述目标存储区域中的为所述至少一个目标存储节点中每个目标存储节点对应的所述目标数据的数据块所指定的子存储区域的位置参数；生成包括所述数据块的标识、所述操作类型以及所述子存储区域的位置参数的所述指示消息，所述指示消息用于指示接收到所述指示消息的目标存储节点在所述操作类型为读操作时，根据所述子存储区域的位置参数通过与所述主机的 RDMA 连接将所述目标数据的数据块写入所述主机内存中所述子存储区域，或者，在所述操作类型为写操作时，根据所述子存储区域的位置参数，通过与所述主机的 RDMA 连接从所述主机内存中所述子存储区域读取所述目标数据的数据块，并存储所述数据块。

可选的，所述装置 50 还包括第二接收模块 54 和第二发送模块 55，

第二接收模块 54，用于从所述存储节点中任一存储节点接收所述任一存储节点创建的第一 QP 的参数；

第二发送模块 55，用于将所述第一 QP 的参数发送给所述主机；

所述第一接收模块 51，还用于从所述主机接收所述主机创建的第二 QP 的参数；

所述第一发送模块 53，还用于将所述第二 QP 的参数发送给所述任一存储节点。

可选的，所述装置 50 与所述主机之间的连接为 TCP/IP 连接；所述指示消息包括所述操作类型、所述主机的通信地址以及所述目标数据的标识；所述装置 50 还包括：

第二发送模块 55，用于向所述主机发送第三指示消息，所述第三指示消息包括一个所述目标存储节点的通信地址，用于指示所述主机通过与所述目标存储节点间的 TCP/IP 向所述目标存储节点发送所述目标数据或从所述目标存储节点获取所述目标数据。

可选的，所述装置 50 与所述主机之间的连接为 TCP/IP 连接，所述操作类型为读操作；所述指示消息包括所述操作类型、所述主机的通信地址、所述装置 50 的通信地址以及所述目标数据的标识，所述指示消息用于指示所述目标存储节点以所述装置 50 的通信地址为源地址以及以所述主机的通信地址为目的地址发送携带所述目标数据的 TCP/IP 报文。

可选的，所述确定模块 52,还用于根据所述主机的 TCP 窗口大小确定所述主机的数据接收量，根据所述数据接收量确定所述至少一个目标存储节点中每个目标存储节点通过每个 TCP/IP 报文携带的所述目标数据的数据块；生成包括所述数据块的标识的所述指示消息，所述指示消息用于指示接收到所述指示信息的目标存储节点向所述主机发送所述数据块。

上述装置 50 的各模块的实现方式，与图 4 至图 12 对应的数据处理的方法中由控制器执行的步骤的实施方式相同，在此不予重复。

图 14 所示为本发明实施例的一种数据处理的装置 60，用于实现图 4 至图 12 所述数据处理方法中存储节点的功能。所述装置 60 与控制器通信连接，所述控制器用于管理所述装置 60；所述装置 60 包括：

接收模块 61，用于接收所述控制器发送的指示消息，所述指示消息包括主机待操作的目标数据的标识和操作类型；

传输模块 62，用于根据所述指示消息，通过与所述主机的连接向所述主机发送所述目标数据或从所述主机获取所述目标数据。

可选的，装置 60 还包括连接模块 63，用于创建第一 QP，所述第一 QP 中包括第一发送队列 SQ 以及第二接收队列 RQ；将所述第一 QP 的参数发送给所述控制器；从所述控制器接收所述主机创建的第二 QP 的参数，所述第二 QP 中包括第二发送队列 SQ 以及第二接收队列 RQ；根据所述第一 QP 的参数以及所述第二 QP 的参数将所述第一 SQ 与所述第二 QP 的第二 RQ 绑定，以及将所述第一 RQ 与所述第二 SQ 绑定，进而与所述主机建立 RDMA 连接。

可选的，所述指示消息包括所述目标数据的标识、所述操作类型以及所述主机的内存中目标存储区域的位置参数；

所述传输模块 62，用于在所述操作类型为读操作时，通过与所述主机的 RDMA 连接将所述目标数据写入所述主机内存中所述目标存储区域；在所述操作类型为写操作时，通过与所述主机的 RDMA 连接从所述主机内存中所述目标存储区域读取所述目标数据，并存储所述目标数据。

可选的，所述指示消息包括所述操作类型、所述主机的通信地址、所述控制器的通信地址以及所述目标数据的标识；

所述传输模块 62，用于以所述控制器的通信地址为源地址以及以所述主机的通信地址为目的地址发送携带所述目标数据的 TCP/IP 报文。

上述装置 60 的各模块的实现方式，与图 4 至图 12 对应的数据处理的方法中由存储节点执行的步骤的实施方式相同，在此不摘赘述。

图 15 所示为本发明实施例的一种数据处理的设备 70，用于实现图 4 至图 12 所述数据处理方法中控制器的功能。设备 70 包括处理器 71、存储器 72、通信接口 73 以及总线 74，所述处理器 71、所述存储器 72 和所述通信接口 73 之间通过所述总线 74 连接并完成相互间的通信，所述存储器 72 中用于存储计算机执行指令，所述设备 70 运行时，所述处理器

71 执行所述存储器 72 中的计算机执行指令以利用所述设备中的硬件资源执行图 4 至图 12 对应的数据处理的方法中由控制器执行的步骤。

上述处理器 71 可以包括一个处理单元，也可以包括多个处理单元。例如，该处理器 71 可以是中央处理器 CPU，也可以是特定集成电路 ASIC，或者是被配置成实施本发明实施例的一个或多个集成电路，例如：一个或多个微处理器（digital signal processor，DSP），或，一个或者多个现场可编程门阵列 FPGA。

上述存储器 72 可以包括一个存储单元，也可以包括多个存储单元，且用于存储可执行程序代码、设备 70 运行所需要参数、数据等。且存储器 72 可以包括随机存储器（random-access memory，RAM），也可以包括非易失性存储器（non-volatile memory，NVM），例如磁盘存储器，闪存（flash）等。上述通信接口 73，在一些实施例中可以为支持 TCP/IP 协议的接口，在另一些实施例中可以为支持 RDMA 协议的接口。

上述总线可 74 以是工业标准体系结构（industry standard architecture，ISA）总线、外部设备互连（peripheral component，PCI）总线或扩展工业标准体系结构（extended industry standard architecture，EISA）总线等。该总线 74 可以分为地址总线、数据总线、控制总线等。为便于表示，图中仅用一条线表示，但并不表示仅有一根总线或一种类型的总线。

应理解 ，根据本发明实施例一种数据处理的设备 700 可对应于本发明实施例中的装置 500，并可以对应于执行根据本发明实施例中图 4 至图 12 所述数据处理方法中控制器为执行主体的操作步骤的相应流程，为了简洁，在此不再赘述。

上述实施例，可以全部或部分地通过软件、硬件、固件或其他任意组合来实现。当使用软件实现时，上述实施例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。在计算机上加载或执行所述计算机程序指令时，全部或部分地产生按照本发明实施例所述的流程或功能。所述计算机可以为通用计算机、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、服务器或数据中心通过有线（例如红外、无线、微波等）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包含一个或多个可用介质集合的服务器、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）、光介质（例如，DVD）、或者半导体介质（例如固态硬盘 SSD）等。

本发明实施例还提供一种存储设备，用于实现图 4 至图 12 所述数据处理方法中存储节点的功能。该存储设备的结构可以继续参照图 15，该设备包括处理器、存储器、通信接口以及总线，所述处理器、所述存储器和所述通信接口之间通过所述总线连接并完成相互间的通信，所述存储器中用于存储计算机执行指令，所述设备运行时，所述处理器执行所述存储器中的计算机执行指令以利用所述设备中的硬件资源执行图 4 至图 12 对应的数据处理的方法中由存储节点执行的步骤。

本发明实施例提供的数据处理方法除了可以应用于分布式存储系统之外，还可以用于非分布式的存储阵列。下面先介绍现有技术中主机向存储阵列读写目标数据的技术方案。

图 16 为存储阵列的系统框图，如图所示，该系统包括：主机 810、第一控制器 821、第二控制器 822 以及多个磁盘，如磁盘 830-1、830-2、...830-n。其中，所述第一控制器 821

以及第二控制器 822 用于管理多个磁盘，且为主机 810 提供访问磁盘上数据的服务，第一控制器 821 和第二控制器 822 的工作方式，可以为主备模式，即同一时刻仅有一个控制器为主用状态，负责接收主机的操作请求，并对该操作请求进行处理，包括直接与主机传输目标数据；或者，将该操作请求转发给备控制器，由备控制器进行数据的处理。第一控制器 821 和第二控制器 822 也可以没有主备之分，均能够接收主机发送的操作请求，对该操作请求进行处理。

现有技术中主机 810 访问磁盘中数据的过程如下：

在管理多个磁盘的多个控制器有主备之分时，主机 810 向主控制器发送操作请求，在操作请求的操作类型为读操作时，主控制器根据预置算法从磁盘读取操作请求中请求操作的目标数据，向主机返回目标数据；在操作请求的操作类型为写操作时，主控制器根据预置算法确定操作请求中携带的目标数据在磁盘中的目标存储位置，将该目标数据写入确定出的目标存储位置。

在管理多个磁盘的多个控制器没有主备之分时，主机 810 可以将操作请求发送给任一控制器，或者，主机 810 确定管理目标数据所归属的逻辑单元号 LUN 的控制器，向该控制器发送操作请求。上述两种情形中，控制器与主机间的数据传输与上一段中介绍的主机与主控制器之间的数据传输方式相同。

然而，当接收主机 810 发送的操作请求的控制器（如第一控制器 821）无法为主机进行目标数据的处理时，例如第一控制器与磁盘间的链路发生故障，第一控制器 821 需要将主机 810 的操作请求转发给管理多个磁盘的其他控制器处理，例如第二控制器 822，第二控制器 822 接收第一控制器转发的操作请求，为主机 810 进行目标数据的处理。

第二控制器 822 为主机 810 进行目标数据的处理的过程包括：当操作请求的操作类型为读操作时，第二控制器 822 从至少一个磁盘（如磁盘 830-1）中读取该目标数据，将目标数据发送给第一控制器 821，由第一控制器 821 将目标数据返回给主机 810。当该操作请求的操作类型为写操作时，第一控制器 821 从主机 810 获得目标数据，将目标数据发送给第二控制器 822，第二控制器 822 将目标数据写入至少一个磁盘中。

可见，接收到主机的操作请求的控制器无法为主机进行目标数据的处理时，虽然可以通过其他控制器为主机进行目标数据的处理，但是，负责目标数据的处理的控制器与主机之间需要经由接收主机的操作请求的控制器进行目标数据的传输，导致数据传输的路径较长，数据传输耗时较长。

本发明还提供多个实施例用于解决上述存储阵列中接收主机操作请求的控制器需要为进行目标数据处理的控制器转发目标数据，导致数据传输耗时较长的问题。

图 17 所示为本发明实施例的存储阵列的系统框图，包括至少两个控制器以及至少两个磁盘。至少两个控制器可以为如图 17 所示的第一控制器 911、第二控制器 912，至少两个磁盘可以为如图 17 所示的磁盘 920-1、磁盘 920-2、…、磁盘 920-n。

至少两个磁盘中的每个磁盘可以为机械磁盘、固态硬盘 SSD、固态混合硬盘（solid state hybrid drive，SSHD）等。至少两个磁盘的每个 LUN 可以归属于一个控制器管理，该至少两个控制器与该至少两个磁盘形成存储阵列。主机 930 可以通过控制器访问该控制器管理的 LUN 中的目标数据。

至少两个控制器中每个控制器与主机 930 建立 RDMA 连接，该 RDMA 连接的建立过程与步骤 201 至步骤 208 中所述的方法类似。不同之处在于，步骤 201 至步骤 208 所述方

法中，由控制器为建立RDMA连接的主机与存储节点同步二者创建的QP的参数，而在本实施例中，主机930与控制器可以通过二者间的非RDMA连接（如TCP/IP连接）向对方发送各自创建的QP的参数，进而根据QP的参数关联对方创建的QP，建立RDMA连接。

结合图17所示的系统，本发明实施例提供一种数据处理的方法，参照图18，该方法包括如下步骤：

步骤701：主机向第一控制器发送操作请求，该操作请求包括待操作的目标数据的标识和操作类型。

具体的，主机中记录有数据与管理该数据所归属的LUN的控制器的映射关系，主机根据该映射关系确定第一控制器管理目标数据所归属的LUN，因此，主机向第一控制器发送所述操作请求。

可选的，第一控制器为存储阵列中任一控制器,主机向存储阵列中的任一控制器发送该操作请求。

可选的，第一控制器为存储阵列的主控制器,主机向存储阵列中的主控制器发送该操作请求。

步骤702，第一控制器接收主机发送的操作请求，根据目标数据的标识确定由至少两个控制器中的第二控制器向主机发送目标数据或从主机获取目标数据。

第一控制器确定由第二控制器向主机发送目标数据或从主机获取目标数据可以有多种实现方式，包括但不限于以下方式：

方式1，第一控制器根据预设的负载均衡策略确定向主机发送目标数据或从主机获取目标数据的第二控制器。

具体的,方式1可以包括如下实现方式：

其一，第一控制器为主控制器，主控制器负责接收主机的操作请求，并根据备控制器的预设负载均衡策略（如轮询、分配至负载最小的备控制器等）分配响应该操作请求的备控制器，该主控制器自身并不直接响应主机的操作请求。

其二，结合所述其一，与之不同之处在于，作为主控制器的第一控制器在自身的负载不大于阈值时响应主机的操作请求，在自身的负载大于该阈值时根据备控制器的负载分配响应该操作请求的备控制器。

其三，存储阵列中控制器没有主备之分，每个控制器均保存有其他控制器的负载，接收主机操作请求的第一控制器在自身的负载超过阈值后，将操作请求发送给当前负载最小的其他控制器。

其四，存储阵列中控制器没有主备之分，每个控制器不知道其他控制器的负载，接收主机操作请求的第一控制器在自身的负载超过阈值后，将操作请求发送给任一其他的控制器，若接收的第二控制器自身的负载没有超过阈值，则响应该操作请求，若超过，则第二控制器将该操作请求转发给处第一控制器之外的其他控制器。

方式2，第一控制器根据目标数据所在的LUN的归属确定响应该操作请求的第二控制器。

第一控制器接收操作请求后，确定目标数据所在的LUN实际不由自身管理，第一控制器确定实际管理该LUN的第二控制器，确定由实际管理该LUN的第二控制器响应主机的操作请求。由于管理目标数据所在LUN的第二控制器对目标数据的访问速度较快，通过第二控制器向主机发送目标数据或从主机获取目标数据，能够减少目标数据的传输耗

时。

方式 3，第一控制器在自身无法访问目标数据时，根据上述方式 1 或方式 2 确定响应主机的操作请求的主控制器。

第一控制器与管理的磁盘间的链路发生故障，第一控制器无法从磁盘读写目标数据。第一控制器通过上述方式 1 或方式 2，确定其他控制器响应该操作请求保证主机的操作请求继续被正确响应。

步骤 703：第一控制器向第二控制器发送指示消息。

指示消息包括目标数据的标识、操作类型以及主机的标识，用于指示第二控制器通过与主机间的连接向主机发送目标数据或从主机获取目标数据。

步骤 704：第二控制器根据指示消息，通过与主机间的连接向主机发送目标数据或从主机获取目标数据。

主机与第二控制器之间的连接可以有多种实现方式，例如 RDMA 连接，TCP/IP 连接、快速外围组件互连 PCIe 连接等。

通过上述方案，主机与负责为主机提供读写目标数据服务的第二控制器可以通过二者间建立的连接传输目标数据，而不用不经由接收主机的操作请求的第一控制器转发目标数据，缩短了数据传输的路径长度，减小了目标数据传输的耗时。

作为一种可能的实现方式，第一控制器与第二控制器分别与主机建立 RDMA 连接。主机向第一控制器发送的操作请求包括目标数据的标识、操作类型、主机的标识以及主机在内存中为目标数据指定的目标存储区域的位置参数。第一控制器在向第二控制器发送的操作请求中同样包含目标数据的标识、操作类型、主机的标识以及该位置参数。

第二控制器根据该指示消息向主机发送目标数据或从主机获取目标数据的过程为：

在主机的操作请求的操作类型为读操作时，第二控制器确定目标数据所在的磁盘，从该磁盘读取目标数据，根据主机的标识确定与主机间的 RDMA 连接，通过该 RDMA 连接以前述 RDMA write 的方式将目标数据写入主机内存的目标存储区域。

在主机的操作请求的操作类型为写操作时，第二控制器根据主机的标识确定与主机间的 RDMA 连接，通过该 RDMA 连接以前述 RDMA read 的方式从主机内存的目标存储区域读取数据，根据预置算法确定目标数据在磁盘中的存储位置，将目标数据写入磁盘。

通过上述方案，主机与负责为主机提供读写目标数据服务的第二控制器可以通过二者间建立的 RDMA 连接快速地传输目标数据，实现高速地读写数据。

需要说明的是，第二控制器与主机间的 RDMA 连接，可以在第二控制器接收第一控制器发送的所述指示消息之前建立，也可以在接收所述指示消息之后建立。

作为一种可能的实现方式，存储阵列的控制器之间可以对主机的操作请求进行多次转发，由最后接收操作请求的控制器与主机进行目标数据的传输。

例如，至少两个磁盘的第三控制器接收主机发送的第二操作请求，该第二操作请求包括主机待操作的第三目标数据的标识和操作类型。第三控制器根据步骤 702 所述方式确定由第一控制器与主机传输该第三目标数据。第三控制器向第一控制器发送指示消息，该指示消息包括主机待操作的第三目标数据的标识和操作类型。

第一控制器接收第三控制器发送的该指示消息后，根据步骤 702 所述方式确定由第二控制器向主机发送第三目标数据或从主机获取第三目标数据。第一控制器向第二控制器发送指示消息，该指示消息用于指示第二控制器与主机传输第三目标数据。

第二控制器根据第一控制器发送的该指示消息，通过与主机间的连接与主机传输第三目标数据。

通过上述方案，在控制器对主机的操作请求进行两次或两次以上的转发时，主机仍然可以与负责为主机提供读写目标数据服务的第二控制器可以通过二者间建立的连接传输目标数据，缩短了数据传输的路径长度，减小了目标数据传输的耗时。

请参照图 17，本发明实施例还提供一种数据处理的系统，该系统包括图 17 所述的第一控制器 911、第二控制器 912 以及多个磁盘，如磁盘 920-1、磁盘 920-2、...磁盘 920-n，第一控制器 911、第二控制器 912 用于执行步骤 701 至步骤 704 所述的数据处理的方法，用于实现响应主机操作请求的控制器与主机通过二者间的直接链路传输目标数据。

图 19 所示为本发明实施例的一种处理数据的装置 80，用于实现图 18 所述方法中第一控制器的功能，所述装置 80 与第二控制器以及至少两个磁盘连接，所述装置 80 以及所述第二控制器用于管理所述至少两个磁盘；所述装置包括：

接收模块 81，用于接收主机发送的操作请求，所述操作请求包括待操作的目标数据的标识和操作类型；

确定模块 82，用于确定由所述至少两个控制器中的第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据；

发送模块 83，用于向所述第二控制器发送指示消息，所述指示消息用于指示所述第二控制器通过与所述主机间的连接向所述主机发送所述目标数据或从所述主机获取所述目标数据。

可选的，所述第二控制器与所述主机之间的连接为 RDMA 连接；所述操作请求还包括所述主机中目标存储区域的位置参数；所述指示消息包括所述目标数据的标识、所述操作类型以及所述位置参数，所述指示消息用于指示所述第二控制器在所述操作类型为读操作时，从所述至少两个磁盘中获取所述目标数据；通过与所述主机的所述 RDMA 连接将所述目标数据写入所述主机内存中所述目标存储区域；以及在所述操作类型为写操作时，通过与所述主机的所述 RDMA 连接从所述主机内存中所述目标存储区域读取所述目标数据；将所述目标数据写入所述至少两个磁盘。

可选的，所述确定模块 82，用于根据预设负载均衡策略确定由所述第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据；或根据所述目标数据所在的逻辑单元号 LUN 的归属确定由所述第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据。

可选的，所述接收模块 81，还用于接收管理所述至少两个磁盘的第三控制器的指示消息，所述第三控制器发送的指示消息包括主机待操作的第二目标数据的标识和操作类型；

所述确定模块 82，还用于根据所述第三控制器发送的指示消息确定所述第二目标数据；

所述装置还包括：

传输模块 84，用于通过与所述主机的连接与所述主机传输所述第二目标数据。

可选的，所述接收模块 81，还用于接收管理所述至少两个磁盘的第四控制器的第五指示消息，所述第五指示消息包括主机待操作的第三目标数据的标识和操作类型；

所述确定模块 82，还用于根据所述第三目标数据的标识确定由所述第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据；

所述发送模块 83，还用于向所述第二控制器发送第六指示消息，所述第五指示消息用于指示所述第二控制器与所述主机传输所述第三目标数据。

上述装置 80 的各模块的实现方式，可以参照图 18 所示方法中由第一控制器执行的步骤的实现方式，在此不予详述。

图 20 所示为本发明实施例提供的一种数据处理的设备 90，设备 90 包括处理器 91、存储器 92、第一通信接口 93、第二通信接口 94 以及总线 95，所述处理器 91、所述存储器 92 和所述第一通信接口 93、所述第二通信接口 94 之间通过所述总线 95 连接并完成相互间的通信，所述存储器 92 中用于存储计算机执行指令，所述第一通信接口用于与主机进行通信，所述第二通信接口用于与至少两个磁盘通信。所述设备 90 运行时，所述处理器 91 执行所述存储器 92 中的计算机执行指令以利用所述设备中的硬件资源执行图 18 所示的数据处理的方法中由第一控制器执行的步骤。

处理器 91 的物理实现可以参照前述处理器 71，存储器 92 的物理实现可以参照前述存储器 72，总线 95 的物理实现可以参照前述总线 74。

上述第一通信接口 93，可以为支持 RDMA 协议的接口，也可以为支持 TCP/IP 协议的接口，在另一些实施例中可以为支持 RDMA 协议的接口。

上述第一通信接口 94，可以为支持 PCIe 协议的接口，也可以为现有技术中的各种用于访问磁盘的接口。

上述实施例，可以全部或部分地通过软件、硬件、固件或其他任意组合来实现。当使用软件实现时，上述实施例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。在计算机上加载或执行所述计算机程序指令时，全部或部分地产生按照本发明实施例所述的流程或功能。所述计算机可以为通用计算机、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、服务器或数据中心通过有线（例如红外、无线、微波等）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包含一个或多个可用介质集合的服务器、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）、光介质（例如，DVD）、或者半导体介质（例如固态硬盘 SSD）等。

显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变型在内。

权 利 要 求

1、一种数据处理的系统，其特征在于，包括控制器以及至少两个存储节点，所述控制器用于管理所述至少两个存储节点；

所述控制器，用于接收通过所述控制器与主机之间的第二连接接收所述主机发送的操作请求，所述操作请求包括待操作的目标数据的标识和操作类型；根据所述目标数据的标识从所述至少两个存储节点中确定至少一个目标存储节点；通过与所述至少一个目标存储节点之间的第一连接向所述至少一个目标存储节点发送指示消息，所述指示消息用于指示所述目标存储节点向所述主机发送所述目标数据或从所述主机获取所述目标数据；

所述至少一个目标存储节点，用于根据所述指示消息，通过与所述主机之间的第三连接向所述主机发送所述目标数据或从所述主机获取所述目标数据。

2、根据权利要求 1 所述的系统，其特征在于，当根据所述目标数据的标识确定目标存储节点，向所述目标存储节点发送指示消息时，所述控制器用于：

根据所述目标数据的标识确定所述至少一个目标存储节点包括存储所述目标数据的第一数据块的第一存储节点以及存储所述目标数据的第二数据块的第二存储节点；

向所述第一存储节点发送第一指示消息，所述第一指示消息包括所述第一数据块的标识，用于指示所述第一存储节点向所述主机发送所述第一数据块或从所述主机获取所述第一数据块；

向所述第二存储节点发送第二指示消息，所述第二指示消息包括所述第二数据块的标识，用于指示所述第二存储节点向所述主机发送所述第二数据块或从所述主机获取所述第二数据块。

3、根据权利要求 1 或 2 所述的系统，其特征在于，所述第三连接为远程直接数据存取 RDMA 连接；所述操作请求还包括所述主机在内存中为所述目标数据指定的目标存储区域的位置参数；

所述控制器，还用于根据所述目标存储区域的位置参数确定所述目标存储区域中的为所述目标存储节点对应的所述目标数据的数据块所指定的子存储区域的位置参数；生成包括所述数据块的标识、所述操作类型以及所述子存储区域的位置参数的所述指示消息；

当通过与所述主机之间的第三连接向所述主机发送所述目标数据或从所述主机获取所述目标数据时，所述目标存储节点用于：

在所述操作类型为读操作时，根据所述子存储区域的位置参数，通过与所述主机的 RDMA 连接将所述目标数据的数据块写入所述主机的内存中的所述子存储区域；

在所述操作类型为写操作时，根据所述子存储区域的位置参数，通过与所述主机的 RDMA 连接从所述主机的内存中的所述子存储区域读取所述目标数据的数据块，并存储所述数据块。

4、根据权利要求 1 至 3 中任一所述的系统，其特征在于，所述至少两个存储节点中的任一存储节点还用于：

创建第一队列对 QP，所述第一 QP 包括第一发送队列 SQ 以及第一接收队列 RQ；将所述第一 QP 的参数发送给所述控制器，以及从所述控制器接收所述主机创建的第二 QP 的参数，所述第二 QP 包括第二 SQ 以及第二 RQ；根据所述第一 QP 的参数以及所述第二 QP 的参数将所述第一 SQ 与所述第二 QP 的第二 RQ 绑定，以及将所述第一 RQ 与所述第二 SQ 绑定，进而与所述主机建立所述 RDMA 连接；

所述控制器，还用于从所述主机接收所述第二 QP 的参数以及从所述任一存储节点接收所述第一 QP 的参数；将所述第一 QP 的参数发送给所述主机，将所述第二 QP 的参数发送给所述任一存储节点。

5、根据权利要求 1 至 2 中任一项所述的系统，其特征在于，所述第二连接与所述第三连接为 TCP/IP 连接；

所述控制器，还用于向所述主机发送第三指示消息，所述第三指示消息包括所述目标存储节点的通信地址，用于指示所述主机通过与所述目标存储节点间的 TCP/IP 连接向所述目标存储节点发送所述目标数据或从所述目标存储节点获取所述目标数据。

6、根据权利要求 1 至 2 中任一项所述的系统，其特征在于，所述第二连接与所述第三连接为 TCP/IP 连接，所述操作类型为读操作；所述指示消息包括所述操作类型、所述主机的通信地址、所述控制器的通信地址以及所述目标数据的标识；

在向所述主机发送所述目标数据或从所述主机获取所述目标数据时，所述目标存储节点用于：

以所述控制器的通信地址为源地址、以所述主机的通信地址为目的地址发送携带所述目标数据的 TCP/IP 报文。

7、根据权利要求 6 所述的系统，其特征在于，所述控制器还用于：根据所述主机的 TCP 窗口大小确定所述主机的数据接收量，根据所述数据接收量确定所述目标存储节点通过每个 TCP/IP 报文携带的所述目标数据的数据块，生成包括所述数据块的标识的所述指示消息。

8、根据权利要求 1 至 7 任一项所述的系统，其特征在于，所述至少一个目标存储节点中的每个目标存储节点，还用于在向所述主机发送所述目标数据或从所述主机获取所述目标数据之后，向所述控制器发送数据传输成功消息；

所述控制器，还用于在接收所述数据传输成功消息之后，向所述主机发送操作成功消息。

9、一种数据处理的方法，其特征在于，所述方法由控制器执行，所述控制器用于管理至少两个存储节点，所述方法包括：

所述控制器接收主机发送的操作请求，所述操作请求包括待操作的目标数据的标识和操作类型；

所述控制器根据所述目标数据的标识从所述至少两个存储节点中确定至少一个目标存储节点；

所述控制器向所述至少一个目标存储节点发送指示消息，所述指示消息用于指示所述至少一个目标存储节点通过与所述主机的连接向所述主机发送所述目标数据或从所述主机获取所述目标数据。

10、根据权利要求 9 所述的方法，其特征在于，所述控制器根据所述目标数据的标识从所述至少两个存储节点中确定至少一个目标存储节点，包括：

所述控制器根据所述目标数据的标识确定所述至少一个目标存储节点包括存储所述目标数据的第一数据块的第一存储节点以及存储所述目标数据的第二数据块的第二存储节点；

所述控制器向所述至少一个目标存储节点发送指示消息，包括：

所述控制器向所述第一存储节点发送第一指示消息，所述第一指示消息包括所述第一

数据块的标识，用于指示所述第一存储节点向所述主机发送所述第一数据块或从所述主机获取所述第一数据块；

所述控制器向所述第二存储节点发送第二指示消息，所述第二指示消息包括所述第二数据块的标识，用于指示所述第二存储节点向所述主机发送所述第二数据块或从所述主机获取所述第二数据块。

11、根据权利要求 9 至 10 中任一项所述的方法，其特征在于，所述操作请求还包括所述主机在内存中为所述目标数据指定的目标存储区域的位置参数；在所述控制器向所述至少一个目标存储节点发送指示消息之前，还包括：

所述控制器根据所述目标存储区域的位置参数确定所述目标存储区域中的为所述至少一个目标存储节点中每个目标存储节点对应的所述目标数据的数据块所指定的子存储区域的位置参数；

所述控制器生成包括所述数据块的标识、所述操作类型以及所述子存储区域的位置参数的所述指示消息，所述指示消息用于指示接收到所述指示消息的目标存储节点在所述操作类型为读操作时，根据所述子存储区域的位置参数通过与所述主机的 RDMA 连接将所述目标数据的数据块写入所述主机内存中所述子存储区域，或者，在所述操作类型为写操作时，根据所述子存储区域的位置参数，通过与所述主机的 RDMA 连接从所述主机内存中所述子存储区域读取所述目标数据的数据块，并存储所述数据块。

12、根据权利要求 9 至 11 中任一项所述的方法，其特征在于，所述方法还包括：

所述控制器从所述主机接收所述主机创建的第二 QP 的参数以及从所述至少两个存储节点中的任一存储节点接收所述任一存储节点创建的第一 QP 的参数；

所述控制器将所述第一 QP 的参数发送给所述主机，以及将所述第二 QP 的参数发送给所述任一存储节点。

13、根据权利要求 9 或 10 所述的方法，其特征在于，所述控制器与所述主机之间的连接为 TCP/IP 连接；所述指示消息包括所述操作类型、所述主机的通信地址以及所述目标数据的标识；所述方法还包括：

所述控制器向所述主机发送第三指示消息，所述第三指示消息包括一个目标存储节点的通信地址，用于指示所述主机通过与所述目标存储节点间的 TCP/IP 连接向所述目标存储节点发送所述目标数据或从所述目标存储节点获取所述目标数据。

14、根据权利要求 9 或 10 所述的方法，其特征在于，所述控制器与所述主机之间的连接为 TCP/IP 连接，所述操作类型为读操作；所述指示消息包括所述操作类型、所述主机的通信地址、所述控制器的通信地址以及所述目标数据的标识，所述指示消息用于指示所述目标存储节点以所述控制器的通信地址为源地址以及以所述主机的通信地址为目的地址发送携带所述目标数据的 TCP/IP 报文。

15、根据权利要求 14 所述的方法，其特征在于，在所述向所述至少一个目标存储节点发送指示消息之前，所述方法还包括：

所述控制器根据所述主机的 TCP 窗口大小确定所述主机的数据接收量，根据所述数据接收量确定所述至少一个目标存储节点中每个目标存储节点通过每个 TCP/IP 报文携带的所述目标数据的数据块；

所述控制器生成包括所述数据块的标识的所述指示消息，所述指示消息用于指示接收到所述指示消息的目标存储节点向所述主机发送所述数据块。

16、一种数据处理的方法，其特征在于，所述方法由存储节点执行，所述存储节点与控制器通信连接，所述控制器用于管理所述存储节点；所述方法包括：

所述存储节点接收所述控制器发送的指示消息，所述指示消息包括主机待操作的目标数据的标识和操作类型；

所述存储节点根据所述指示消息，通过与所述主机的连接向所述主机发送所述目标数据或从所述主机获取所述目标数据。

17、根据权利要求 16 所述的方法，其特征在于，所述方法还包括所述存储节点与所述主机建立连接的步骤，所述步骤包括：

所述存储节点创建第一 QP，所述第一 QP 中包括第一发送队列 SQ 以及第二接收 RQ；

所述存储节点将所述第一 QP 的参数发送给所述控制器；

所述存储节点从所述控制器接收所述主机创建的第二 QP 的参数，所述第二 QP 中包括第二 SQ 以及第二 RQ；

所述存储节点根据所述第一 QP 的参数以及所述第二 QP 的参数将所述第一 SQ 与所述第二 QP 的第二 RQ 绑定，以及将所述第一 RQ 与所述第二 SQ 绑定，进而与所述主机建立所述 RDMA 连接。

18、根据权利要求 16 或 17 所述的方法，其特征在于，所述指示消息包括所述目标数据的标识、所述操作类型以及所述主机的内存中的目标存储区域的位置参数；

所述存储节点通过与所述主机的连接向所述主机发送所述目标数据或从所述主机获取所述目标数据，包括：

在所述操作类型为读操作时，所述存储节点通过与所述主机的 RDMA 连接将所述目标数据写入所述主机内存中所述目标存储区域；

在所述操作类型为写操作时，所述存储节点通过与所述主机的 RDMA 连接从所述主机内存中所述目标存储区域读取所述目标数据，并存储所述目标数据。

19、根据权利要求 16 所述的方法，其特征在于，所述指示消息包括所述操作类型、所述主机的通信地址、所述控制器的通信地址以及所述目标数据的标识；

所述所述存储节点通过与所述主机的连接向所述主机发送所述目标数据或从所述主机获取所述目标数据，包括：

所述存储节点以所述控制器的通信地址为源地址、以所述主机的通信地址为目的地址发送携带所述目标数据的 TCP/IP 报文。

20、一种处理数据的系统，其特征在于，包括至少两个控制器以及至少两个磁盘，所述至少两个控制器用于管理所述至少两个磁盘；

所述至少两个控制器中的第一控制器，用于接收主机发送的操作请求，所述操作请求包括待操作的目标数据的标识和操作类型；确定由所述至少两个控制器中的第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据；向所述第二控制器发送指示消息，所述指示消息用于指示所述第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据；

所述第二控制器，用于接收所述指示消息；根据所述指示消息，通过与所述主机的连接向所述主机发送所述目标数据或从所述主机获取所述目标数据。

21、根据权利要求 20 所述的系统，其特征在于，所述第二控制器与所述主机之间的连接为 RDMA 连接；所述操作请求还包括所述主机的内存中目标存储区域的位置参数；

所述指示消息包括所述目标数据的标识、所述操作类型以及所述位置参数；

所述第二控制器，用于根据所述指示消息，通过与所述主机的连接向所述主机发送所述目标数据或从所述主机获取所述目标数据，包括：

在所述操作类型为读操作时，从所述至少两个磁盘中获取所述目标数据；通过与所述主机的 RDMA 连接将所述目标数据写入所述主机内存中所述目标存储区域；

在所述操作类型为写操作时，通过与所述主机的 RDMA 连接从所述主机内存中所述目标存储区域读取所述目标数据；将所述目标数据写入所述至少两个磁盘。

22、根据权利要求 20 或 21 所述的系统，其特征在于，所述第一控制器用于：

根据预设负载均衡策略确定由所述第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据；或

根据所述目标数据所在的逻辑单元号 LUN 的归属确定由所述第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据。

23、一种处理数据的方法，其特征在于，所述方法由第一控制器执行，所述第一控制器与第二控制器以及至少两个磁盘连接，所述第一控制器以及所述第二控制器用于管理所述至少两个磁盘；所述方法包括：

所述第一控制器接收主机发送的操作请求，所述操作请求包括待操作的目标数据的标识和操作类型；

所述第一控制器确定由所述至少两个控制器中的第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据；

所述第一控制器向所述第二控制器发送指示消息，所述指示消息用于指示所述第二控制器通过与所述第二控制器与所述主机间的连接向所述主机发送所述目标数据或从所述主机获取所述目标数据。

24、根据权利要求 23 所述的方法，其特征在于，所述第二控制器与所述主机之间的连接为 RDMA 连接，所述操作请求还包括所述主机的内存中中目标存储区域的位置参数；所述指示消息包括所述目标数据的标识、所述操作类型以及所述位置参数，所述指示消息用于指示所述第二控制器在所述操作类型为读操作时，从所述至少两个磁盘中获取所述目标数据；通过与所述主机的所述 RDMA 连接将所述目标数据写入所述主机内存中所述目标存储区域；以及在所述操作类型为写操作时，通过与所述主机的所述 RDMA 连接从所述主机内存中所述目标存储区域读取所述目标数据；将所述目标数据写入所述至少两个磁盘。

25、根据权利要求 23 或 24 所述的方法，其特征在于，所述第一控制器确定由所述至少两个控制器中的第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据，包括：

根据预设负载均衡策略确定由所述第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据；或

根据所述目标数据所在的逻辑单元号 LUN 的归属确定由所述第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据。

26、根据权利要求 23 至 25 任一项所述的方法，其特征在于，还包括：

所述第一控制器接收管理所述至少两个磁盘的第三控制器发送的指示消息，所述第三控制器发送的指示消息包括所述主机待操作的第二目标数据的标识和操作类型；

所述第一控制器响应所述第三控制器发送的指示消息，通过与所述主机间的连接向所述主机发送所述第二目标数据或从所述主机获取所述第二目标数据。

27、一种数据处理的装置，其特征在于，所述装置用于管理至少两个存储节点，所述装置包括：

第一接收模块，用于接收主机发送的操作请求，所述操作请求包括待操作的目标数据的标识和操作类型；

确定模块，用于根据所述目标数据的标识从所述至少两个存储节点中确定至少一个目标存储节点；

第一发送模块，用于向所述至少一个目标存储节点发送指示消息，所述指示消息用于指示所述至少一个目标存储节点通过与所述主机的连接向所述主机发送所述目标数据或从所述主机获取所述目标数据。

28、根据权利要求 27 所述的装置，其特征在于，所述确定模块，用于： 根据所述目标数据的标识确定所述至少一个目标存储节点包括存储所述目标数据的第一数据块的第一存储节点以及存储所述目标数据的第二数据块的第二存储节点；

所述第一发送模块，具体用于向所述第一存储节点发送第一指示消息，所述第一指示消息包括所述第一数据块的标识，用于指示所述第一存储节点向所述主机发送所述第一数据块或从所述主机获取所述第一数据块；

向所述第二存储节点发送第二指示消息，所述第二指示消息包括所述第二数据块的标识，用于指示所述第二存储节点向所述主机发送所述第二数据块或从所述主机获取所述第二数据块。

29、根据权利要求 27 或 28 所述的装置，其特征在于，所述操作请求还包括所述主机在内存中为所述目标数据指定的目标存储区域的位置参数；

所述确定模块，还用于根据所述目标存储区域的位置参数确定所述目标存储区域中的为所述至少一个目标存储节点中每个目标存储节点对应的所述目标数据的数据块所指定的子存储区域的位置参数；

生成包括所述数据块的标识、所述操作类型以及所述子存储区域的位置参数的所述指示消息，所述指示消息用于指示接收到所述指示消息的目标存储节点在所述操作类型为读操作时，根据所述子存储区域的位置参数通过与所述主机的 RDMA 连接将所述目标数据的数据块写入所述主机内存中所述子存储区域，或者，在所述操作类型为写操作时，根据所述子存储区域的位置参数，通过与所述主机的 RDMA 连接从所述主机内存中所述子存储区域读取所述目标数据的数据块，并存储所述数据块。

30、根据权利要求 27 至 29 中任一项所述的装置，其特征在于，还包括：

第二接收模块，用于从所述存储节点中的任一存储节点接收所述任一存储节点创建的第一 QP 的参数；

第二发送模块，用于将所述第一 QP 的参数发送给所述主机；

所述第一接收模块，还用于从所述主机接收所述主机创建的第二 QP 的参数；

所述第一发送模块，还用于将所述第二 QP 的参数发送给所述任一存储节点。

31、根据权利要求 27 或 28 所述的装置，其特征在于，所述装置与所述主机之间的连接为 TCP/IP 连接；所述指示消息包括所述操作类型、所述主机的通信地址以及所述目标数据的标识；所述装置还包括：

第二发送模块，用于向所述主机发送第三指示消息，所述第三指示消息包括一个目标存储节点的通信地址，用于指示所述主机通过与所述目标存储节点间的 TCP/IP 连接向所述目标存储节点发送所述目标数据或从所述目标存储节点获取所述目标数据。

32、根据权利要求 27 或 28 所述的装置，其特征在于，所述装置与所述主机之间的连接为 TCP/IP 连接，所述操作类型为读操作；所述指示消息包括所述操作类型、所述主机的通信地址、所述装置的通信地址以及所述目标数据的标识，所述指示消息用于指示所述目标存储节点以所述装置的通信地址为源地址以及以所述主机的通信地址为目的地址发送携带所述目标数据的 TCP/IP 报文。

33、根据权利要求 32 所述的装置，其特征在于，所述确定模块还用于：

根据所述主机的 TCP 窗口大小确定所述主机的数据接收量，根据所述数据接收量确定所述至少一个目标存储节点中每个目标存储节点通过每个 TCP/IP 报文携带的所述目标数据的数据块；

生成包括所述数据块的标识的所述指示消息，所述指示消息用于指示接收到所述指示信息的目标存储节点向所述主机发送所述数据块。

34、一种数据处理的装置，其特征在于，所述装置与控制器通信连接，所述控制器用于管理所述装置；所述装置包括：

接收模块，用于接收所述控制器发送的指示消息，所述指示消息包括主机待操作的目标数据的标识和操作类型；

传输模块，用于根据所述指示消息，通过与所述主机的连接向所述主机发送所述目标数据或从所述主机获取所述目标数据。

35、根据权利要求 34 所述的装置，其特征在于，所述装置还包括连接模块，用于：

创建第一 QP，所述第一 QP 中包括第一发送队列 SQ 以及第二接收队列 RQ；

将所述第一 QP 的参数发送给所述控制器；

从所述控制器接收所述主机创建的第二 QP 的参数，所述第二 QP 中包括第二 SQ 以及第二 RQ；

根据所述第一 QP 的参数以及所述第二 QP 的参数将所述第一 SQ 与所述第二 QP 的第二 RQ 绑定，以及将所述第一 RQ 与所述第二 SQ 绑定，进而与所述主机建立所述 RDMA 连接。

36、根据权利要求 34 或 35 所述的装置，其特征在于，所述指示消息包括所述目标数据的标识、所述操作类型以及所述主机的内存中目标存储区域的位置参数；

所述传输模块，用于在所述操作类型为读操作时，通过与所述主机的 RDMA 连接将所述目标数据写入所述主机内存中所述目标存储区域；

在所述操作类型为写操作时，通过与所述主机的 RDMA 连接从所述主机内存中所述目标存储区域读取所述目标数据，并存储所述目标数据。

37、根据权利要求 34 所述的装置，其特征在于，所述指示消息包括所述操作类型、所述主机的通信地址、所述控制器的通信地址以及所述目标数据的标识；

所述传输模块，用于以所述控制器的通信地址为源地址、以所述主机的通信地址为目的地址发送携带所述目标数据的 TCP/IP 报文。

38、一种处理数据的装置，其特征在于，所述装置与第二控制器以及至少两个磁盘连接，所述装置以及所述第二控制器用于管理所述至少两个磁盘；所述装置包括：

接收模块，用于接收主机发送的操作请求，所述操作请求包括待操作的目标数据的标识和操作类型；

确定模块，用于确定由所述至少两个控制器中的第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据；

发送模块，用于向所述第二控制器发送指示消息，所述指示消息用于指示所述第二控制器通过与所述主机间的连接向所述主机发送所述目标数据或从所述主机获取所述目标数据。

39、根据权利要求 38 所述的装置，其特征在于，所述第二控制器与所述主机之间的连接为 RDMA 连接所述操作请求还包括所述主机的内存中目标存储区域的位置参数；所述指示消息包括所述目标数据的标识、所述操作类型以及所述位置参数，所述指示消息用于指示所述第二控制器在所述操作类型为读操作时，从所述至少两个磁盘中获取所述目标数据；通过与所述主机的所述 RDMA 连接将所述目标数据写入所述主机内存中所述目标存储区域；以及在所述操作类型为写操作时，通过与所述主机的所述 RDMA 连接从所述主机内存中所述目标存储区域读取所述目标数据；将所述目标数据写入所述至少两个磁盘。

40、根据权利要求 38 或 39 所述的装置，其特征在于，所述确定模块，用于：

根据预设负载均衡策略确定由所述第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据；或

根据所述目标数据所在的逻辑单元号 LUN 的归属确定由所述第二控制器向所述主机发送所述目标数据或从所述主机获取所述目标数据。

41、根据权利要求 38 至 40 任一项所述的装置，其特征在于，所述接收模块，还用于：接收管理所述至少两个磁盘的第三控制器的指示消息，所述第三控制器发送的指示消息包括主机待操作的第二目标数据的标识和操作类型；

所述装置还包括：传输模块，用于响应所述第三控制器发送的指示消息，通过与所述主机间的连接向所述主机发送所述第二目标数据或从所述主机获取所述第二目标数据。

42、一种数据处理的设备，其特征在于，包括处理器、存储器、通信接口以及总线，所述处理器、所述存储器和所述通信接口之间通过所述总线连接并完成相互间的通信，所述存储器中用于存储计算机执行指令，所述设备运行时，所述处理器执行所述存储器中的计算机执行指令以利用所述设备中的硬件资源执行权利要求 9 至 15 中任一所述的方法。

43、一种数据处理的设备，其特征在于，包括处理器、存储器、通信接口以及总线，所述处理器、所述存储器和所述通信接口之间通过所述总线连接并完成相互间的通信，所述存储器中用于存储计算机执行指令，所述设备运行时，所述处理器执行所述存储器中的计算机执行指令以利用所述设备中的硬件资源执行权利要求 16 至 19 中任一所述的方法。

44、一种数据处理的设备，其特征在于，包括处理器、存储器、通信接口以及总线，所述处理器、所述存储器和所述通信接口之间通过所述总线连接并完成相互间的通信，所述存储器中用于存储计算机执行指令，所述设备运行时，所述处理器执行所述存储器中的计算机执行指令以利用所述设备中的硬件资源执行权利要求 23 至 26 中任一所述的方法。

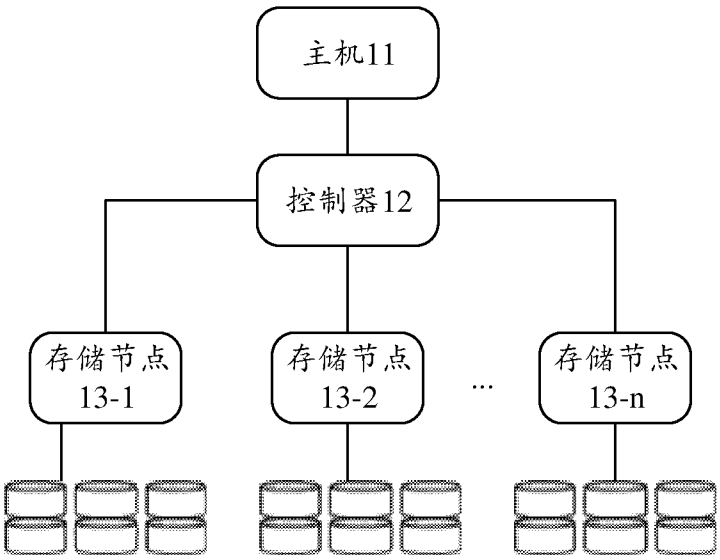
主机11
控制器12
存储节点
13-1
存储节点
13-2
...
存储节点
13-n

图 1

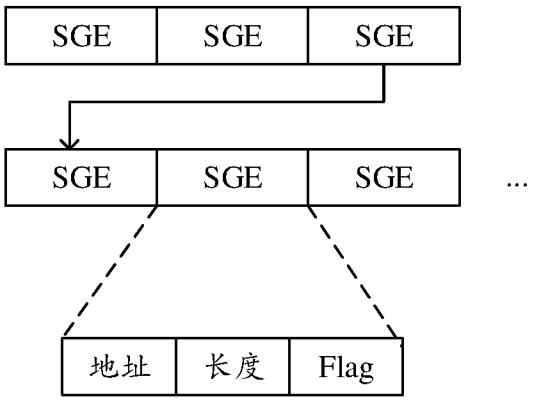
SGE
SGE
SGE
SGE
SGE
SGE
...
地址
长度
Flag

图 2

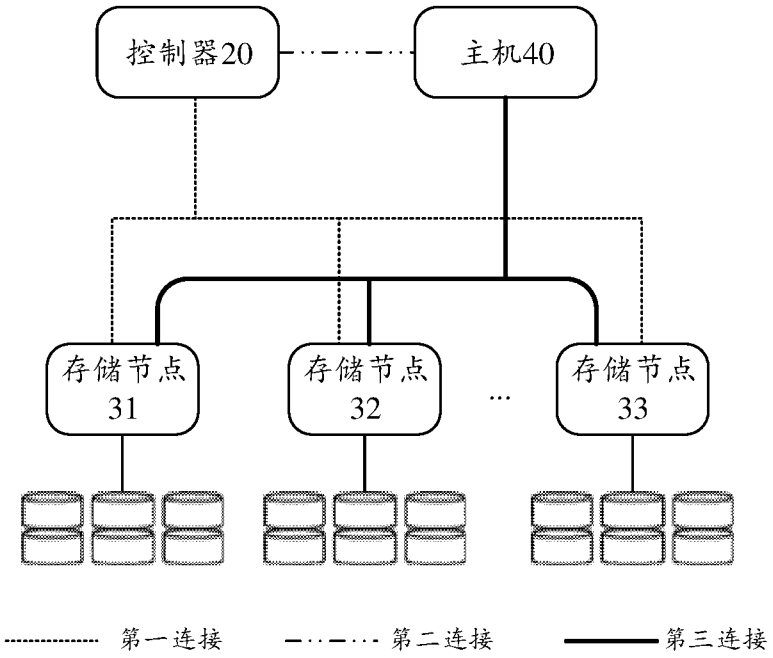

图 3

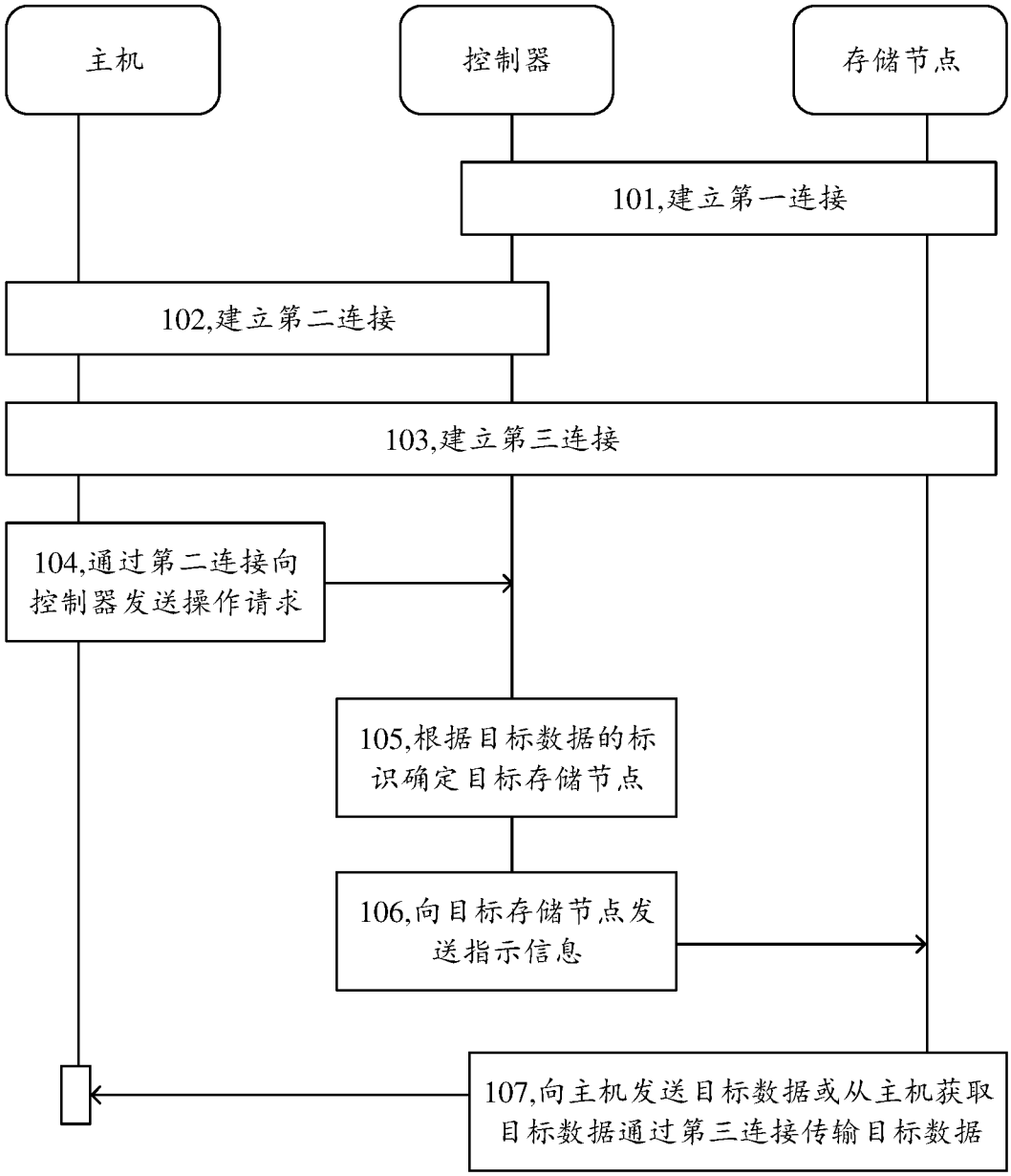
主机
控制器
存储节点
101,建立第一连接
102,建立第二连接
103,建立第三连接
104,通过第二连接向控制器发送操作请求
105,根据目标数据的标识确定目标存储节点
106,向目标存储节点发送指示信息
107,向主机发送目标数据或从主机获取目标数据通过第三连接传输目标数据

图 4

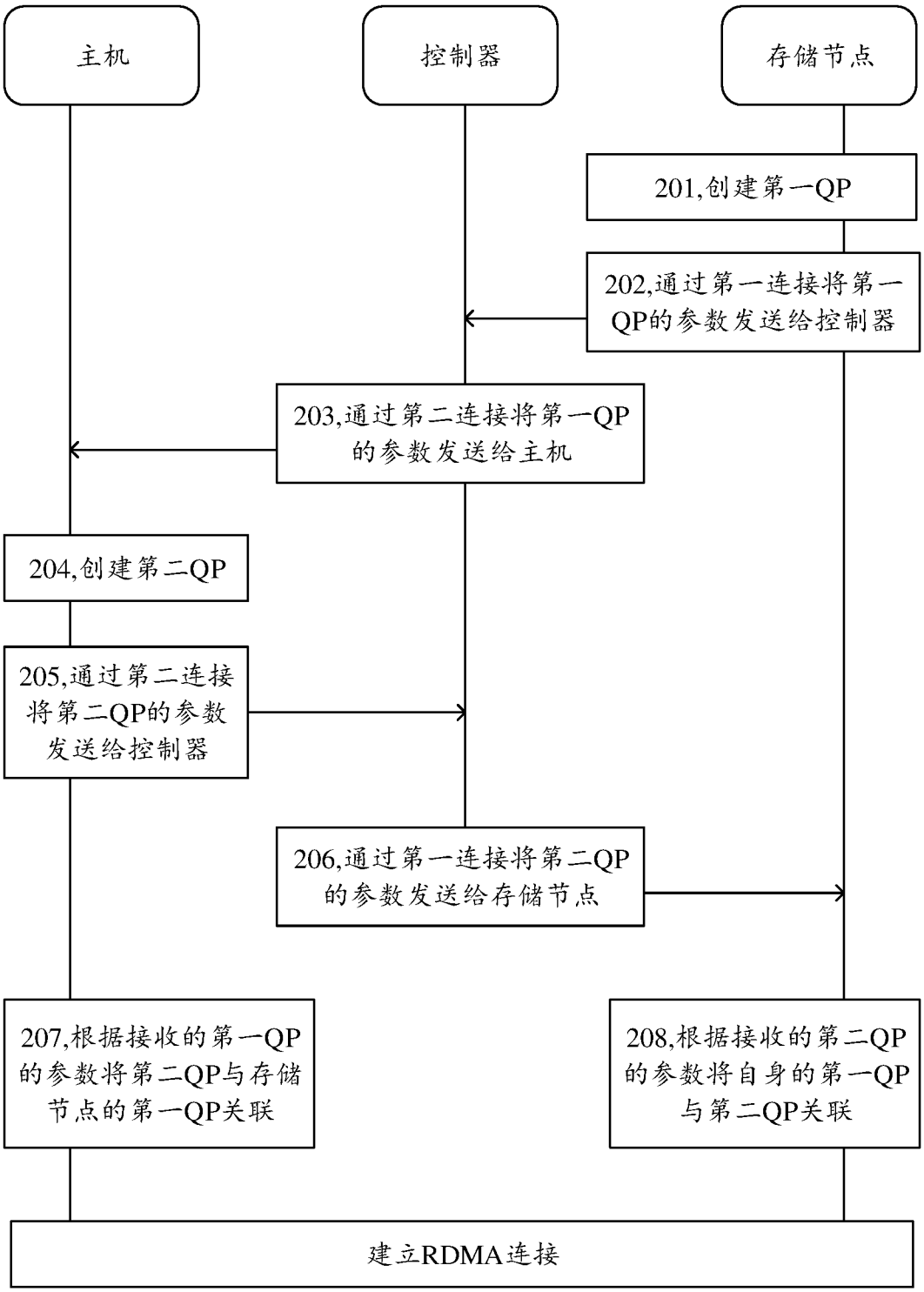
主机
控制器
存储节点
201,创建第一QP
202,通过第一连接将第一QP的参数发送给控制器
203,通过第二连接将第一QP的参数发送给主机
204,创建第二QP
205,通过第二连接将第二QP的参数发送给控制器
206,通过第一连接将第二QP的参数发送给存储节点
207,根据接收的第一QP的参数将第二QP与存储节点的第一QP关联
208,根据接收的第二QP的参数将自身的第一QP与第二QP关联
建立RDMA连接

图 5

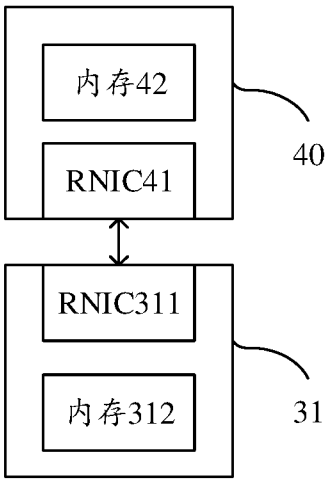
内存42
RNIC41
40
RNIC311
内存312
31

图 6

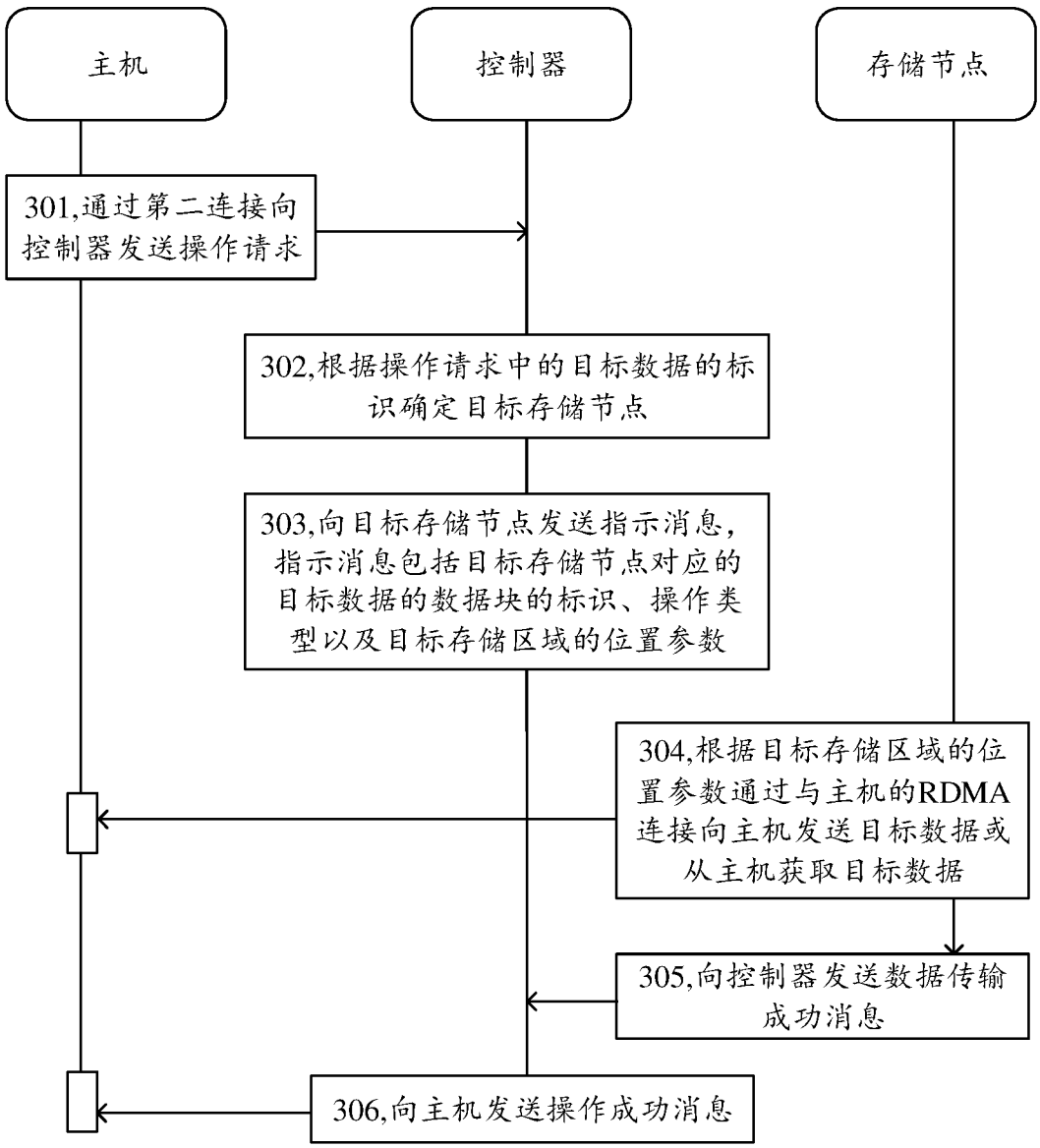
主机
控制器
存储节点
301,通过第二连接向控制器发送操作请求
302,根据操作请求中的目标数据的标识确定目标存储节点
303,向目标存储节点发送指示消息,指示消息包括目标存储节点对应的目标数据的数据块的标识、操作类型以及目标存储区域的位置参数
304,根据目标存储区域的位置参数通过与主机的RDMA连接向主机发送目标数据或从主机获取目标数据
305,向控制器发送数据传输成功消息
306,向主机发送操作成功消息

图 7

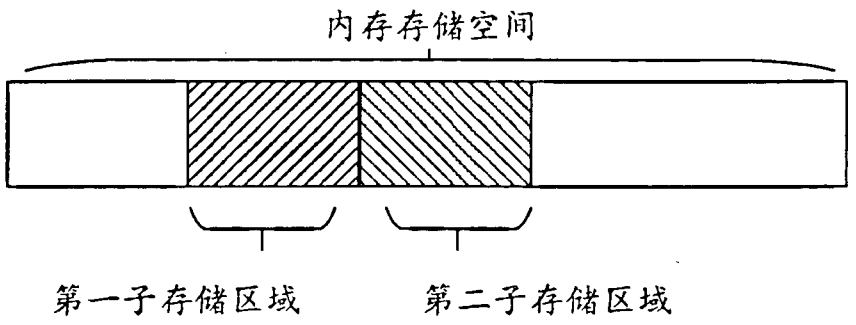

图 8a

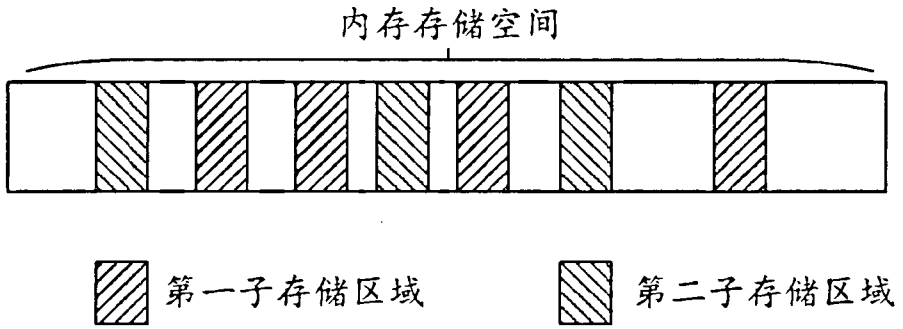

图 8b

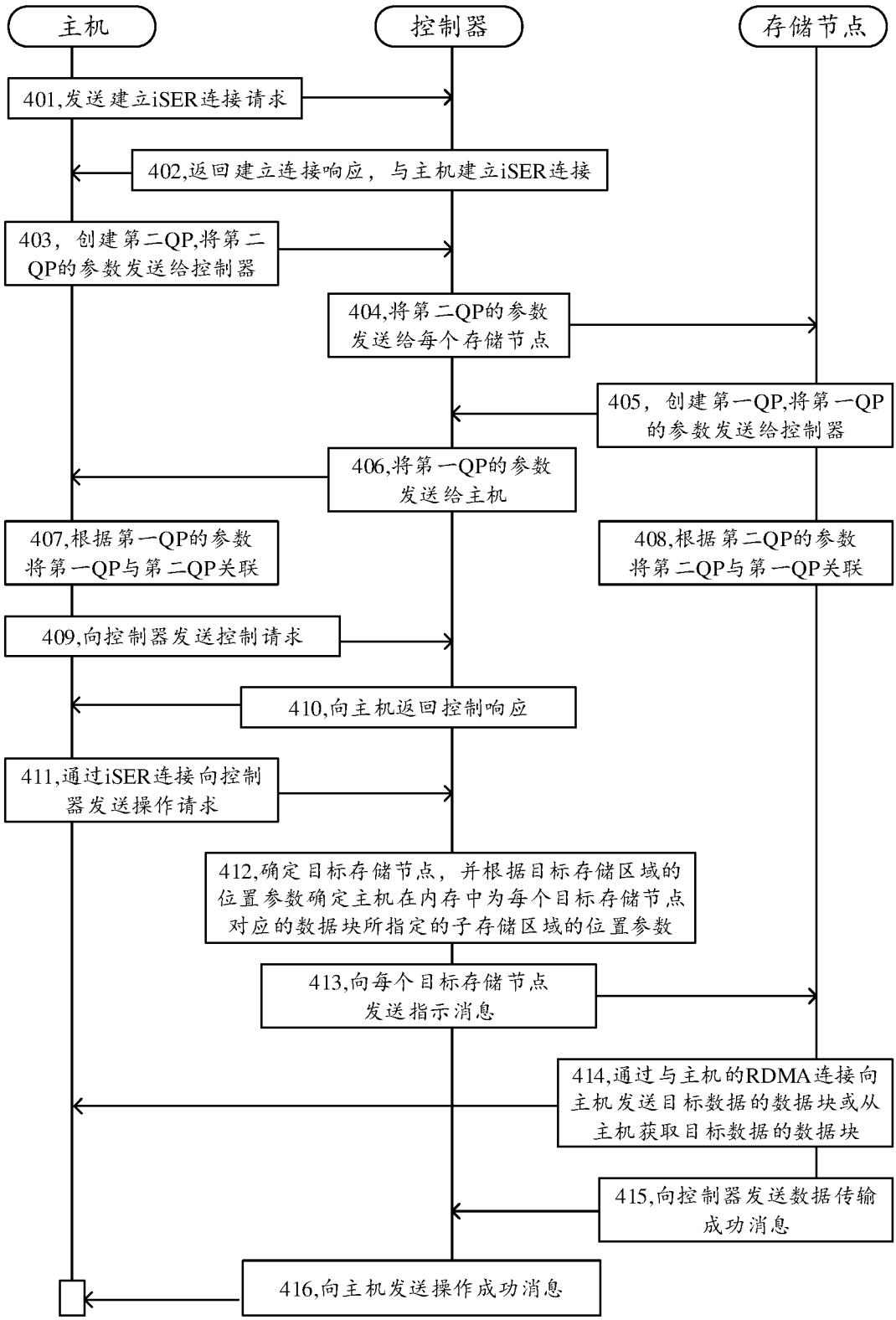
主机
控制器
存储节点
401,发送建立iSER连接请求
402,返回建立连接响应，与主机建立iSER连接
403，创建第二QP,将第二QP的参数发送给控制器
404,将第二QP的参数发送给每个存储节点
405，创建第一QP,将第一QP的参数发送给控制器
406,将第一QP的参数发送给主机
407,根据第一QP的参数将第一QP与第二QP关联
408,根据第二QP的参数将第二QP与第一QP关联
409,向控制器发送控制请求
410,向主机返回控制响应
411,通过iSER连接向控制器发送操作请求
412,确定目标存储节点，并根据目标存储区域的位置参数确定主机在内存中为每个目标存储节点对应的数据块所指定的子存储区域的位置参数
413,向每个目标存储节点发送指示消息
414,通过与主机的RDMA连接向主机发送目标数据的数据块或从主机获取目标数据的数据块
415,向控制器发送数据传输成功消息
416,向主机发送操作成功消息

图 9

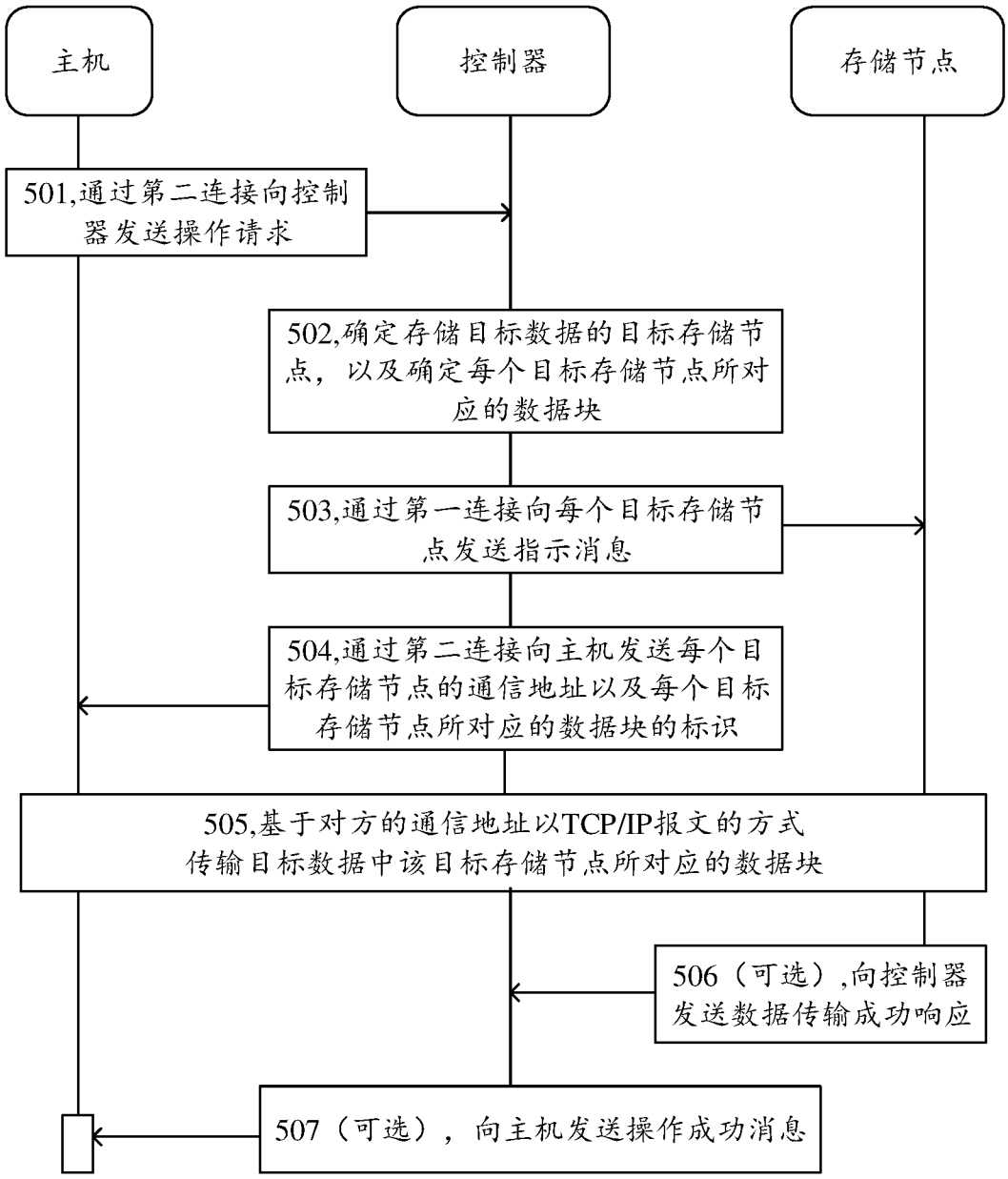
主机
控制器
存储节点
501,通过第二连接向控制器发送操作请求
502,确定存储目标数据的目标存储节点，以及确定每个目标存储节点所对应的数据块
503,通过第一连接向每个目标存储节点发送指示消息
504,通过第二连接向主机发送每个目标存储节点的通信地址以及每个目标存储节点所对应的数据块的标识
505,基于对方的通信地址以TCP/IP报文的方式传输目标数据中该目标存储节点所对应的数据块
506（可选）,向控制器发送数据传输成功响应
507（可选），向主机发送操作成功消息

图 10

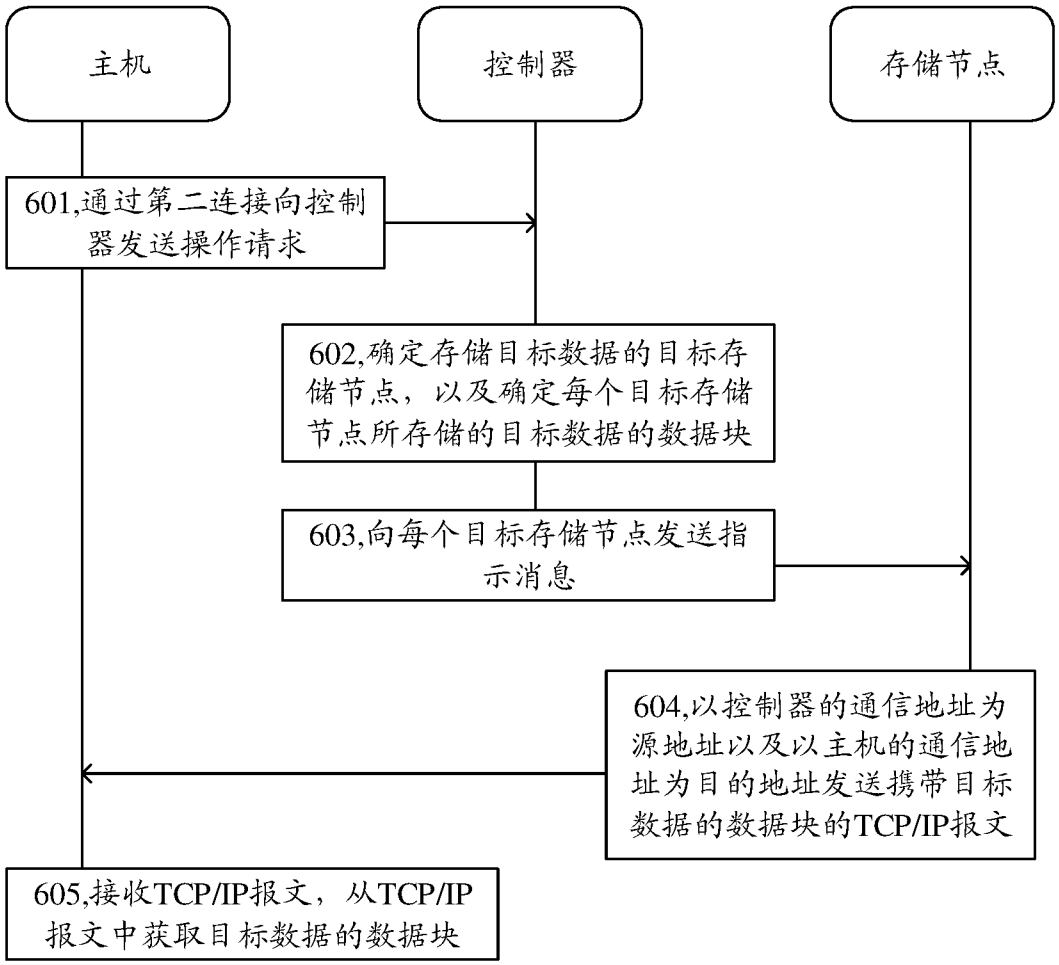
主机
控制器
存储节点
601,通过第二连接向控制器发送操作请求
602,确定存储目标数据的目标存储节点，以及确定每个目标存储节点所存储的目标数据的数据块
603,向每个目标存储节点发送指示消息
604,以控制器的通信地址为源地址以及以主机的通信地址为目的地址发送携带目标数据的数据块的TCP/IP报文
605,接收TCP/IP报文，从TCP/IP报文中获取目标数据的数据块

图 11

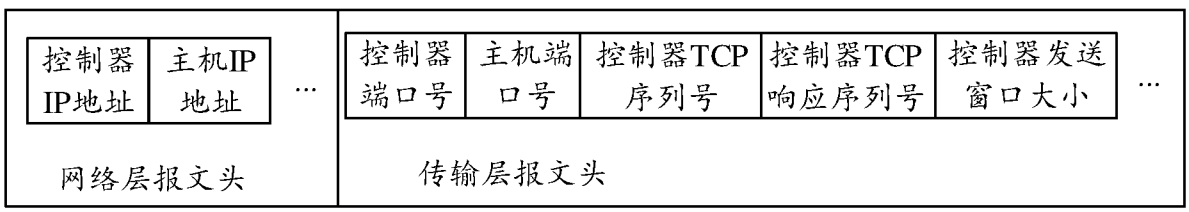
控制器IP地址
主机IP地址
…
网络层报文头
控制器端口号
主机端口号
控制器TCP序列号
控制器TCP响应序列号
控制器发送窗口大小
…
传输层报文头

图 12

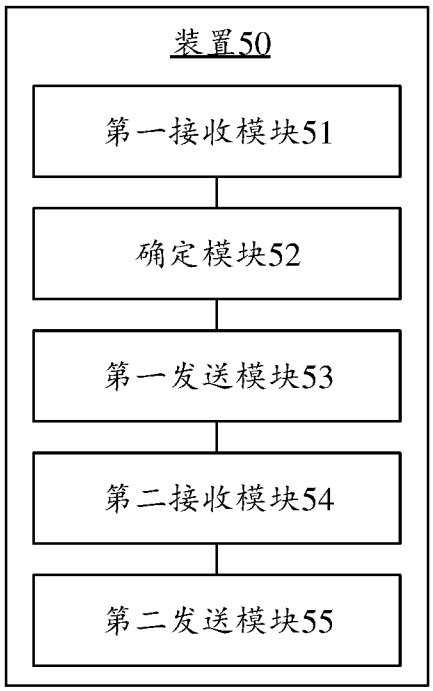
装置50
第一接收模块51
确定模块52
第一发送模块53
第二接收模块54
第二发送模块55

图 13

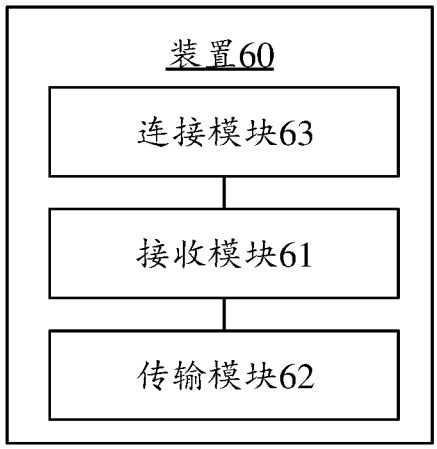
装置60
连接模块63
接收模块61
传输模块62

图 14

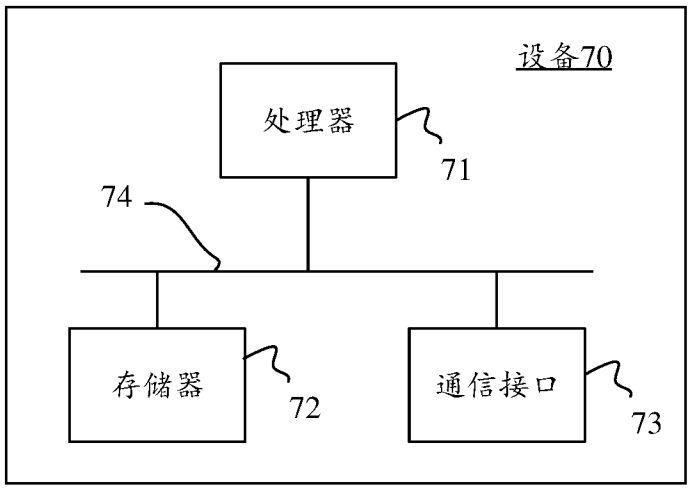
设备70
处理器
71
74
存储器
72
通信接口
73

图 15

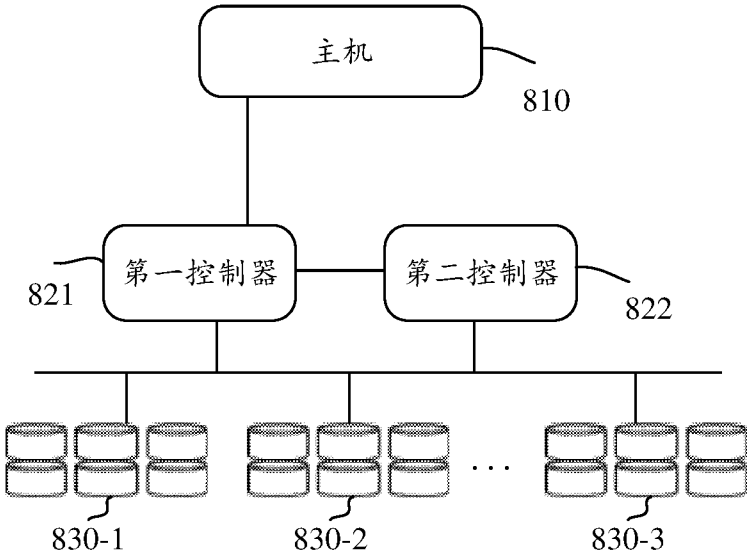
主机
810
821
第一控制器
第二控制器
822
830-1
830-2
830-3

图 16

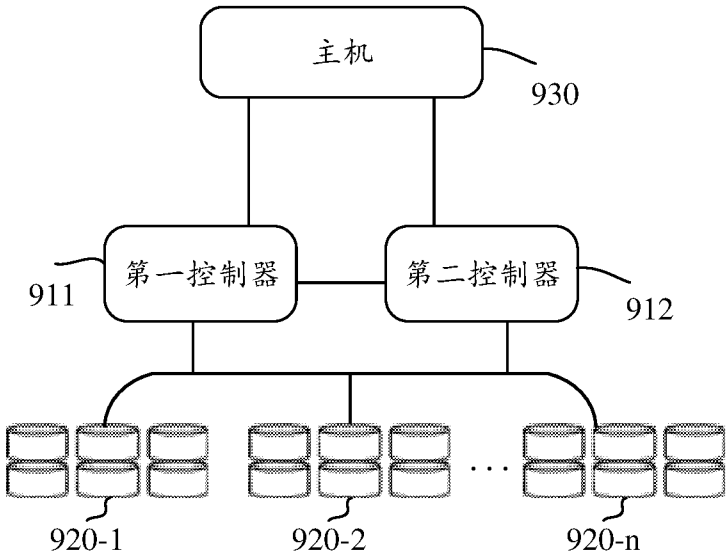
主机
930
第一控制器
911
第二控制器
912
920-1
920-2
920-n

图 17

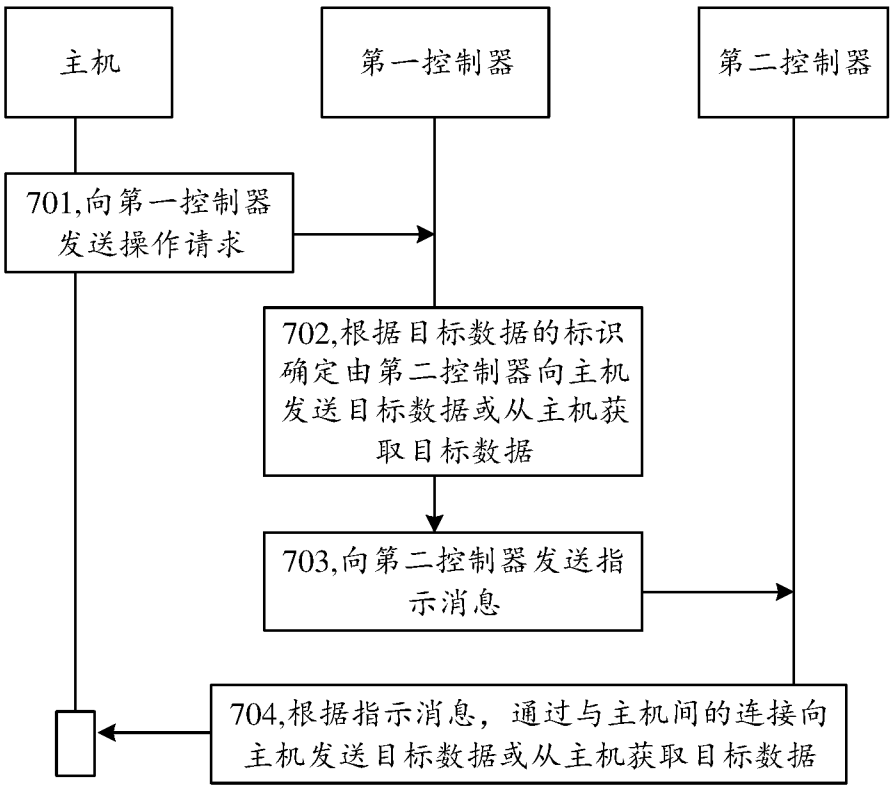
主机
第一控制器
第二控制器
701,向第一控制器发送操作请求
702,根据目标数据的标识确定由第二控制器向主机发送目标数据或从主机获取目标数据
703,向第二控制器发送指示消息
704,根据指示消息，通过与主机间的连接向主机发送目标数据或从主机获取目标数据

图 18

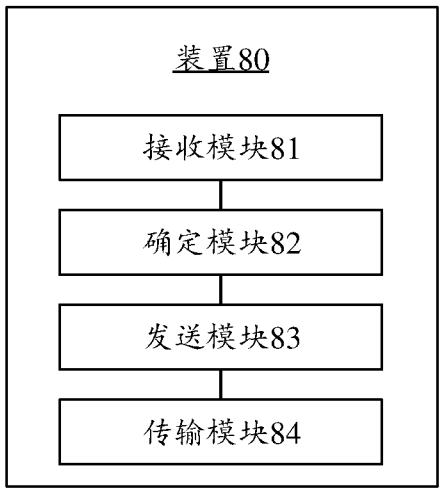
装置80
接收模块81
确定模块82
发送模块83
传输模块84

图 19

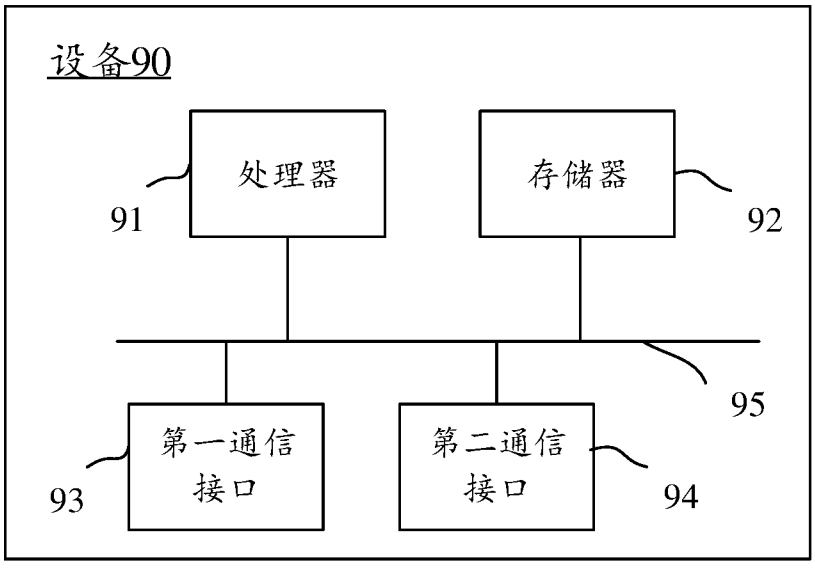
设备90
处理器
91
存储器
92
95
第一通信接口
93
第二通信接口
94

图 20

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/072701

A. CLASSIFICATION OF SUBJECT MATTER

G06F 3/06 (2006.01) i; H04L 29/08 (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; 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)

CNPAT, WPI, EPODOC, CNKI: 存储, 直接, 控制器, 标识, 目标, 指示, 主机, 远程直接数据存取, 存取, 主, 备, 读, 写, store, save, buffer, direct, controller, identifier, target, ID, indicate, master, host, DMA, RDMA, access, slave, read, write

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 103530066 A (HUAWEI TECHNOLOGIES CO., LTD.), 22 January 2014 (22.01.2014), description, paragraphs [0067]-[0095]	1-44
X	CN 103516755 A (HUAWEI TECHNOLOGIES CO., LTD.), 15 January 2014 (15.01.2014), description, paragraphs [0035]-[0087]	1-44
X	CN 103984662 A (HUAWEI TECHNOLOGIES CO., LTD.), 13 August 2014 (13.08.2014), description, paragraphs [0180]-[0188]	1-44
A	CN 103905526 A (SHENZHEN COSHIP ELECTRONICS CO., LTD.), 02 July 2014 (02.07.2014), entire document	1-44
A	CN 104580346 A (JIDIAN XINYUAN INTERNATIONAL TECHNOLOGY DEVELOPMENT BEIJING CO., LTD.), 29 April 2015 (29.04.2015), entire document	1-44
A	CN 103440202 A (HUAWEI TECHNOLOGIES CO., LTD.), 11 December 2013 (11.12.2013), entire document	1-44
A	US 2013198311 A1 (TAMIR, E. et al.), 01 August 2013 (01.08.2013), entire document	1-44

☐ 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 September 2017	19 October 2017
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer WANG, Jian Telephone No. (86-10) 62413911

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/072701

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103530066 A	22 January 2014	WO 2015035887 A1	19 March 2015
CN 103516755 A	15 January 2014	None	
CN 103984662 A	13 August 2014	WO 2015180538 A1	03 December 2015
CN 103905526 A	02 July 2014	None	
CN 104580346 A	29 April 2015	None	
CN 103440202 A	11 December 2013	None	
US 2013198311 A1	01 August 2013	WO 2013109640 A1	25 July 2013
		DE 112013000601 T5	18 December 2014
		US 2013198312 A1	01 August 2013
		US 2017249281 A1	31 August 2017
		CN 104246742 A	24 December 2014
		US 2014325013 A1	30 October 2014

国际检索报告

国际申请号
PCT/CN2017/072701

A. 主题的分类

G06F 3/06(2006.01)i; H04L 29/08(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; 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））

CNPAT, WPI, EPODOC, CNKI：存储,直接,控制器,标识,目标,指示,主机,远程直接数据存取,存取,主,备,读,写,store, save, buffer, direct, controller, identifier, target, ID, indicate, master, host, DMA, RDMA, access, slave, read, write

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 103530066 A (华为技术有限公司) 2014年 1月 22日 (2014-01-22) 说明书第[0067]-[0095]段	1-44
X	CN 103516755 A (华为技术有限公司) 2014年 1月 15日 (2014-01-15) 说明书第[0035]-[0087]段	1-44
X	CN 103984662 A (华为技术有限公司) 2014年 8月 13日 (2014-08-13) 说明书第[0180]-[0188]段	1-44
A	CN 103905526 A (深圳市同洲电子股份有限公司) 2014年 7月 2日 (2014-07-02) 全文	1-44
A	CN 104580346 A (奇点新源国际技术开发北京有限公司) 2015年 4月 29日 (2015-04-29) 全文	1-44
A	CN 103440202 A (华为技术有限公司) 2013年 12月 11日 (2013-12-11) 全文	1-44
A	US 2013198311 A1 (TAMIR, ELIEZER 等) 2013年 8月 1日 (2013-08-01) 全文	1-44

☐ 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
2017年 9月 22日	2017年 10月 19日

ISA/CN的名称和邮寄地址	受权官员
中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088	王健
传真号 (86-10)62019451	电话号码 (86-10)62413911

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/072701

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	103530066	A	2014年 1月 22日	WO	2015035887	A1	2015年 3月 19日
CN	103516755	A	2014年 1月 15日		无		
CN	103984662	A	2014年 8月 13日	WO	2015180538	A1	2015年 12月 3日
CN	103905526	A	2014年 7月 2日		无		
CN	104580346	A	2015年 4月 29日		无		
CN	103440202	A	2013年 12月 11日		无		
US	2013198311	A1	2013年 8月 1日	WO	2013109640	A1	2013年 7月 25日
				DE	112013000601	T5	2014年 12月 18日
				US	2013198312	A1	2013年 8月 1日
				US	2017249281	A1	2017年 8月 31日
				CN	104246742	A	2014年 12月 24日
				US	2014325013	A1	2014年 10月 30日