(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 1 月 4 日 (04.01.2024)

WIPO | PCT

(10) 国际公布号
WO 2024/001850 A1

(51) 国际专利分类号:
G06F 15/163 (2006.01)

(21) 国际申请号: PCT/CN2023/101171

(22) 国际申请日: 2023 年 6 月 19 日 (19.06.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202210733448.9 2022年6月27日 (27.06.2022) CN
202211260921.2 2022年10月14日 (14.10.2022) 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: 周铁刚 (**ZHOU, Yigang**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

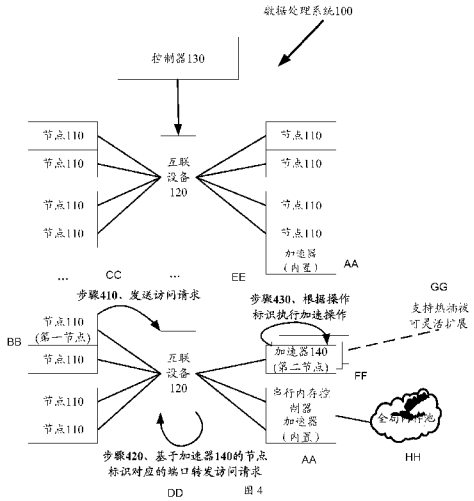
(74) 代理人: 北京中博世达专利商标代理有限公司 (**BEIJING ZBSD PATENT & TRADEMARK AGENT LTD.**); 中国北京市海淀区交大东路31号11号楼8层, Beijing 100044 (CN)。

(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,

(54) **Title:** DATA PROCESSING SYSTEM, METHOD AND APPARATUS, AND CONTROLLER

(54) 发明名称: 数据处理系统、方法、装置和控制器



100 Data processing system
110 Node
120 Interconnection device
130 Controller
140 Accelerator
AA Accelerator (built-in)
BB First node
CC Step 410, send an access request
DD Step 420, forward the access request on the basis of a port corresponding to a node identifier of the accelerator 140
EE Step 430, execute an acceleration operation according to an operation identifier
FF Second node
GG Support hot plug and be able to be flexibly expanded
HH Global memory pool

(57) **Abstract:** Disclosed are a data processing system, method and apparatus, and a controller, which relate to the field of data processing. The data processing system comprises a controller and a plurality of nodes, which are connected by means of an interconnection device, and the plurality of nodes comprise a first node and a second node. The method comprises: when a second node requests access to a data processing system, a controller allocating a second node identifier to the second node, and an interconnection device forwarding, on the basis of the second node identifier, a message indicating that a first node accesses the second node, wherein the second node identifier is used for uniquely indicating the second node; and the second node acquiring, according to a source address indicated by the first node, data to be processed, and storing and processing, according to a destination address indicated by the first node, processed data of the data to be processed. In this way, the scale of a data processing system is elastically expanded according to requirements, and a data processing request of a node is extended to different nodes in the system, such that accelerator resources in the system can be shared by a plurality of nodes, thereby adapting to computing requirements in different application scenarios.

WO 2024/001850 A1

[见续页]

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, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:
— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 公开了数据处理系统、方法、装置和控制器，涉及数据处理领域。数据处理系统包括通过互联设备连接的控制器和多个节点，多个节点包括第一节点和第二节点。当第二节点请求接入数据处理系统时，控制器为第二节点分配第二节点标识，互联设备基于第二节点标识转发第一节点访问第二节点的消息，第二节点标识用于唯一指示第二节点。第二节点根据第一节点指示的源地址获取待处理数据，根据第一节点指示的目的地址存储处理待处理数据的处理后数据。如此，按需弹性扩展数据处理系统的规模，将节点的数据处理请求扩充到系统内不同节点间，使系统内的加速器资源可以为多个节点共享，适应不同的应用场景下的计算需求。

数据处理系统、方法、装置和控制器

本申请要求于2022年06月27日提交国家知识产权局、申请号为202210733448.9申请名称为“一种内存数据访问的方法和计算系统”的中国专利申请的优先权，本申请要求于2022年10月14日提交国家知识产权局、申请号为202211260921.2申请名称为“数据处理系统、方法、装置和控制器”的中国专利申请的优先权，这些全部内容通过引用结合在本申请中。

技术领域

本申请涉及数据处理领域，尤其涉及一种数据处理系统、方法、装置和控制器。

背景技术

目前，通过高带宽、低时延的片间互联总线和交换机将多个节点互联成一个高性能集群，俗称超节点（Super Node）。相对异构计算服务器架构（如：中央处理器（central processing unit，CPU）+领域特定结构（Domain Specific Architecture，DSA）），超节点能够提供更高的计算能力；相对以太网络互联带宽，超节点能够提供更大的带宽。但是，超节点通常以静态模式进行配置，无法灵活扩展，无法适应不同的应用场景下的计算需求。

发明内容

本申请提供了数据处理系统、方法、装置和控制器，由此实现灵活扩展超节点的规模，适应不同的应用场景下的计算需求。

第一方面，提供了一种数据处理系统。数据处理系统包括多个节点和控制器，多个节点包括第一节点和第二节点，多个节点和控制器通过高速互联链路连接。控制器，用于当第二节点请求接入数据处理系统时，为第二节点分配第二节点标识，其中，第二节点在数据处理系统中的全局物理地址为第二节点的节点标识和第二节点的物理地址；控制器，还用于将第二节点的全局物理地址发送给第一节点。

相对于在数据处理系统的启动阶段以静态模式进行配置，本申请通过控制器控制接入新节点，可以按需弹性扩展数据处理系统的规模，例如，支持接入加速器、增大全局内存的存储空间等，从而满足不同的应用场景下的计算需求。

结合第一方面，在一种可能的实现方式中，数据处理系统还包括互联设备，互联设备基于高速互联链路连接多个节点；控制器，还用于向互联设备发送第二节点标识与端口的对应关系，第二节点标识对应的端口用于向第二节点转发消息。

其中，互联设备也可以称为互联芯片或交换芯片。互联设备用于基于第二节点标识与端口的对应关系转发第一节点访问第二节点的消息。例如，互联设备存储有第一节点标识与第一端口的对应关系，互联设备基于第一节点标识与第一端口的对应关系转发第二节点访问第一节点的消息。互联设备存储有第二节点标识与第二端口的对应关系，互联设备基于第二节点标识与第二端口的对应关系转发第一节点访问第二节点的消息。

从而，使互联设备基于节点标识转发节点间进行通信的数据，使系统内的加速器资源可以为多个节点共享，适应不同的应用场景下的计算需求。

需要说明的是，全局物理地址是指由数据处理系统包含的多个节点中任意一个节点可以访问的地址。数据处理系统包含的多个节点中任意一个节点存储有其他节点的全局物理地址，以便于任意一个节点根据其他节点的全局物理地址访问其他节点的存储空间。全局物理地址用于唯一指示数据处理系统中的一个节点的存储空间。可理解地，全局物理地址包含节点标识和节点的物理地址。由于节点标识用于唯一指示数据处理系统中的一个节点，则全局物理地址中的物理地址用于唯一指示数据处理系统中的一个节点的存储空间。节点的物理地址是指节点内存储空间的地址。虽然不同的节点内存储空间的物理地址可以是相同的，但是数据处理系统中的任一节点根据全局物理地址中的节点标识区分不同节点内的存储空间。

例如，第一节点的全局物理地址包括第一节点标识和第一节点的物理地址。由于第一节点标识用于唯一指示第一节点，第一节点的物理地址用于唯一指示第一节点的存储空间，则第一节点的全局物理地址可以用于指示第一节点的存储空间。

第二节点的全局物理地址包括第二节点标识和第二节点的物理地址。由于第二节点标识用于唯一指示第二节点，第二节点的物理地址用于唯一指示第二节点的存储空间，则第二节点的全局物理地址可以用于指示第二节点的存储空间。

由此，第一节点可以根据第二节点的全局物理地址访问第二节点的存储空间。第二节点可以根据第一节点的全局物理地址访问第一节点的存储空间。

结合第一方面，在另一种可能的实现方式中，第二节点用于根据第一节点指示的源地址获取待处理数据，源地址用于指示存储待处理数据的节点的节点标识和物理地址；第二节点还用于处理待处理数据，以及根据第一节点指示的目的地址存储处理后数据，目的地址用于指示存储处理后数据的节点的节点标识和物理地址。

如此，将节点的数据处理请求扩充到系统内不同节点间，根据全局物理地址中节点标识唯一指示的节点的物理地址，获取待处理数据或存储处理后数据，从而使系统内的加速器资源可以为多个节点共享，适应不同的应用场景下的计算需求。

示例地，源地址用于指示第一节点的全局物理地址。目的地址用于指示第二节点的全局物理地址。

结合第一方面，在另一种可能的实现方式中，第二节点具体用于根据第一节点指示的操作标识对待处理数据执行加速操作得到处理后数据。其中，第二节点包括处理器、加速器和内存控制器中任一个。例如将通用处理器（如：CPU）的作业卸载到加速器，由加速器处理计算需求较高的作业（如：HPC、大数据作业、数据库作业等），解决由于通用处理器浮点算力不足，无法满足HPC、AI等场景的重浮点计算需求的问题，从而，缩短数据处理时长以及降低系统能耗，提升系统性能。节点内部也可以集成加速器。独立部署的加速器和集成加速器的节点支持灵活插拔，可以按需弹性扩展数据处理系统的规模，从而满足不同的应用场景下的计算需求。

结合第一方面，在另一种可能的实现方式中，多个节点的存储介质经过统一编址构成全局内存池。例如，全局内存池包括源地址指示的存储空间或/和目的地址指示的存储空间。从而，节点执行数据处理的过程中从全局内存池读取数据或对全局内存池写入数据，以提升数据处理的速度。

结合第一方面，在另一种可能的实现方式中，第一节点还用于根据第一节点的物理地址访问第一节点的存储空间。第二节点还用于根据第二节点的物理地址访问第一节点的存储空间。

结合第一方面，在另一种可能的实现方式中，控制器还用于当第二节点退出数据处理系统时，控制第一节点老化第二节点的全局物理地址，以及控制互联设备老化第二节点标识与端口的对应关系。

如此，在数据处理系统中设置控制器和互联设备，基于节点接入机制和退出机制，可弹性增加及减少节点，实现了可弹性扩展的超节点架构，既解决了传统超节点架构无法动态扩展的问题，又避免了传统IO总线架构规模受限和带宽低问题， 并支持在节点或者互联设备故障情况下的动态容错机制。

第二方面，提供了一种数据处理方法，数据处理系统包括多个节点，多个节点包括第一节点和第二节点，多个节点和控制器通过高速互联链路连接。方法包括：当第二节点请求接入数据处理系统时，控制器为第二节点分配第二节点标识，其中，第二节点在数据处理系统中的全局物理地址为第二节点的节点标识和第二节点的物理地址；控制器将第二节点的全局物理地址发送给第一节点。

结合第二方面，在一种可能的实现方式中，数据处理系统还包括互联设备，互联设备基于高速互联链路连接多个节点。方法还包括：控制器向互联设备发送第二节点标识与端口的对应关系，第二节点标识对应的端口用于向第二节点转发消息。

其中，互联设备基于第二节点标识与端口的对应关系转发第一节点访问第二节点的消息，第二节点标识用于唯一指示第二节点。

结合第二方面，在另一种可能的实现方式中，方法还包括：当第二节点退出数据处理系统时，

控制器控制第一节点老化第二节点的全局物理地址，以及控制互联设备老化第二节点标识与端口的对应关系。

结合第二方面，在另一种可能的实现方式中，方法还包括：第二节点根据第一节点指示的源地址获取待处理数据，进而，处理待处理数据，以及根据第一节点指示的目的地址存储处理后数据。其中，目的地址用于指示存储处理后数据的节点的节点标识和物理地址。源地址用于指示存储待处理数据的节点的节点标识和物理地址。

第三方面，提供了一种控制装置，所述装置包括用于执行第二方面或第二方面任一种可能设计中的控制器执行的方法的各个模块。

第四方面，提供了一种数据传输装置，所述装置包括用于执行第二方面或第二方面任一种可能设计中的互联设备执行的方法的各个模块。

第五方面，提供了一种数据处理节点，所述节点包括用于执行第二方面或第二方面任一种可能设计中的节点执行的方法的各个模块。

第六方面，提供一种控制器，该控制器包括至少一个处理器和存储器，存储器用于存储一组计算机指令；当处理器作为第二方面或第二方面任一种可能实现方式中的控制器执行所述一组计算机指令时，执行第二方面或第二方面任一种可能实现方式中的数据处理方法的操作步骤。

第七方面，提供一种芯片，包括：处理器和供电电路；其中，所述供电电路用于为所述处理器供电；所述处理器用于执行第二方面或第二方面任一种可能实现方式中的数据处理方法的操作步骤。

第八方面，提供一种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计算设备中运行时，使得计算设备执行如第二方面或第二方面任意一种可能的实现方式中所述方法的操作步骤。

第九方面，提供一种计算机程序产品，当计算机程序产品在计算机上运行时，使得计算设备执行如第二方面或第二方面任意一种可能的实现方式中所述方法的操作步骤。

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 为本申请提供的一种数据处理系统的架构示意图；

图 2 为本申请提供的一种全局内存池的部署场景示意图；

图 3 为本申请提供的一种节点接入数据处理系统方法的流程示意图；

图 4 为本申请提供的一种数据处理方法的流程示意图；

图 5 为本申请提供的一种描述符的结构示意图；

图 6 为本申请提供的一种节点退出数据处理系统方法的流程示意图；

图 7 为本申请提供的一种控制装置的结构示意图；

图 8 为本申请提供的一种数据处理节点的结构示意图；

图 9 为本申请提供的一种计算设备的结构示意图。

具体实施方式

为了便于描述，首先对本申请涉及的术语进行简单介绍。

超节点（Super Node），指通过高带宽、低时延的片间互联总线和交换机将多个节点互联成一个高性能集群。超节点的规模大于缓存一致非统一内存寻址（Cache-Coherent Non Uniform Memory Access，CC-NUMA）架构下的节点规模，超节点内节点的互联带宽大于以太网络互联带宽。

高性能计算（High Performance Computing，HPC）集群，指一个计算机集群系统。HPC 集群包含利用各种互联技术连接在一起的多个计算机。互联技术例如可以是 InfiniBand、基于聚合以太网的远程直接内存访问（Remote Direct Memory Access over Converged Ethernet，RoCE）或传输控制协议（Transmission Control Protocol，TCP）。HPC 提供了超高浮点计算能力，可用于解决计算密集型和海量数据处理等业务的计算需求。连接在一起的多个计算机的综合计算能力可以来处理大型计算问题。例如，科学研究、气象预报、金融、仿真实验、生物制药、基因测序和图像处

理等行业涉及的利用 HPC 集群来解决的大型计算问题和计算需求。利用 HPC 集群处理大型计算问题可以有效地缩短处理数据的计算时间，以及提高计算精度。

内存操作指令，可以称为内存语义或内存操作函数。内存操作指令包括内存分配（malloc）、内存设置（memset）、内存复制（memcpy）、内存移动（memmove）、内存释放（memory release）和内存比较（memcmp）中至少一种。

内存分配用于支持应用程序运行分配一段内存。

内存设置用于设置全局内存池的数据模式，例如初始化。

内存复制用于将源地址（source）指示的存储空间存储的数据复制到目的地址（destination）指示的存储空间。

内存移动用于将源地址（source）指示的存储空间存储的数据复制到目的地址（destination）指示的存储空间，并删除源地址（source）指示的存储空间存储的数据。

内存比较用于比较两个存储空间存储的数据是否相等。

内存释放用于释放内存中存储的数据，以提高系统内存资源的利用率，进而提升系统性能。

广播（broadcast）通信，指在计算机网络中传输数据包时，目的地址指示计算机网络中广播域的设备的一种传输方式。以广播通信方式发送的数据包可以称为广播消息。

为了解决超节点的规模无法灵活扩展，无法适应不同的应用场景下的计算需求的问题，本申请提供一种数据处理系统包括通过互联设备连接的控制器和多个节点，多个节点包括第一节点和第二节点。当第二节点请求接入数据处理系统时，控制器为第二节点分配第二节点标识，互联设备基于第二节点标识转发第一节点访问第二节点的消息，第二节点标识用于唯一指示第二节点。第二节点根据第一节点指示的源地址获取待处理数据，根据第一节点指示的目的地址存储处理待处理数据的处理后数据。如此，按需弹性扩展数据处理系统的规模，将节点的数据处理请求扩充到系统内不同节点间，使系统内的加速器资源可以为多个节点共享，适应不同的应用场景下的计算需求。

图 1 为本申请提供的一种数据处理系统的架构示意图。如图 1 所示，数据处理系统 100 是一种提供高性能计算的实体。数据处理系统 100 包括多个节点 110。

节点 110 可以是处理器、服务器、台式计算机、存储阵列的控制器和存储器等。处理器可以是中央处理器（central processing unit，CPU）、图形处理器（graphics processing unit，GPU）、数据处理单元（data processing unit，DPU）、神经处理单元（neural processing unit，NPU)和嵌入式神经网络处理器（neural-network processing unit，NPU）等用于数据处理的 XPU。例如，节点 110 可以包括计算节点和存储节点。

当节点 110 是计算能力（Computing Power）较高的 GPU、DPU、NPU 等数据处理的 XPU 时，节点 110 可以作为加速器，将通用处理器（如：CPU）的作业卸载到加速器，由加速器处理计算需求较高的作业（如：HPC、大数据作业、数据库作业等），解决由于通用处理器浮点算力不足，无法满足 HPC、AI 等场景的重浮点计算需求的问题，从而，缩短数据处理时长以及降低系统能耗，提升系统性能。节点的计算能力也可以称为节点的计算算力。节点 110 内部也可以集成加速器。独立部署的加速器和集成加速器的节点支持灵活插拔，可以按需弹性扩展数据处理系统的规模，从而满足不同的应用场景下的计算需求。

多个节点 110 基于具有高带宽、低时延的高速互联链路连接。示例地，如图 1 所示，互联设备 120（如：交换机）基于高速互联链路连接多个节点 110。例如，互联设备 120 通过光纤、铜缆或铜线连接多个节点 110。互联设备可称为交换芯片或互联芯片。

互联设备 120 基于高速互联链路连接的多个节点 110 组成的数据处理系统 100 也可以称为超节点。多个超节点通过数据中心网络进行连接。数据中心网络包括多个核心交换机和多个汇聚交换机。数据中心网络可以组成一个规模域。多个超节点可以组成一个性能域。两个以上超节点可以组成宏机柜。宏机柜之间也可以基于数据中心网络连接。

如图 1 所示，数据处理系统 100 还包括连接互联设备 120 的控制器 130。控制器 130 基于控制平面链路与互联设备 120 进行通信，基于数据平面链路与多个节点 110 进行通信。控制器 130 用于控制节点接入或退出数据处理系统 100。

例如，控制器 130 为请求接入的节点分配节点标识，向数据处理系统 100 中已接入的活跃节点（如：多个节点 110）配置接入节点的节点标识和接入节点的物理地址，向互联设备配置节点标识与端口的对应关系。请求接入的节点可以简称为接入节点。可理解地，互联设备基于接入节点的节点标识与端口的对应关系向接入节点转发数据。每个活跃节点存储接入节点的全局物理地址，全局物理地址包括接入节点的节点标识和接入节点的物理地址。由于接入节点的节点标识用于唯一指示接入节点，则全局物理地址中的物理地址用于唯一指示接入节点的存储空间，即接入节点的物理地址是指接入节点内存储空间的地址，使活跃节点基于接入节点的全局物理地址访问接入节点。

需要说明的是，全局物理地址是指由数据处理系统 100 包含的多个活跃节点中任意一个活跃节点可以访问的地址。数据处理系统 100 包含的多个活跃节点中任意一个活跃节点存储有其他活跃节点的全局物理地址，以便于任意一个活跃节点根据其他活跃节点的全局物理地址访问其他活跃节点的存储空间。

又如，当活跃节点退出数据处理系统 100 时，控制器 130 老化请求退出的节点的全局物理地址。请求退出的节点可以简称为退出节点。控制器 130 向互联设备 120 发送第一老化消息，以及广播第二老化消息。第一老化消息用于指示互联设备老化退出节点的节点标识与端口的对应关系。第二老化消息用于指示活跃节点老化退出节点的全局物理地址，即数据处理系统 100 中每个活跃节点接收到第二老化消息，老化请求退出节点的节点标识和节点的物理地址。

示例地，当第一节点请求接入数据处理系统 100 时，控制器 130 为第一节点分配用于唯一指示第一节点的第一节点标识，并对互联设备配置第一节点标识对应的第一端口的对应关系，以及对多个节点 110 配置第一节点的全局物理地址，第一节点的全局物理地址包括第一节点标识和第一节点的物理地址，使互联设备基于第一节点标识与第一端口的对应关系向第一节点转发数据，每个节点 110 存储有第一节点的全局物理地址，使每个节点 110 根据第一节点的全局物理地址访问第一节点的存储空间，即根据第一节点标识确定访问第一节点，根据第一节点的物理地址访问第一节点的存储空间。当第一节点退出数据处理系统 100 时，控制互联设备老化第一节点标识与第一端口的对应关系；控制多个节点 110 老化第一节点标识和第一节点的物理地址。

在一种可能的示例中，节点根据 None-Posted Write 指令进行内存访问。当第一节点访问第一节点的内存时，None-Posted Write 指令用于指示第一节点的内存的物理地址，使第一节点访问第一节点的内存。当第一节点访问系统中的远端节点的内存时，None-Posted Write 指令用于指示远端节点的节点标识和远端节点的物理地址，使第一节点访问远端节点的内存。

其中，数据处理系统 100 支持运行大数据、数据库、高性能计算、人工智能、分布式存储和云原生等应用。本申请实施例中数据包括大数据、数据库、高性能计算、人工智能（Artificial Intelligence，AI）、分布式存储和云原生等应用的业务数据。在一些实施例中，控制器 130 可以接收用户操作客户端发送的处理请求，对处理请求指示的作业进行控制。客户端可以是指计算机，也可称为工作站（workstation）。

控制器和节点可以是独立的物理设备。控制器也可称为控制节点、控制设备或命名节点。节点可以称为计算设备或数据节点或存储节点。

在一些实施例中，数据处理系统 100 中节点 110 的存储介质经过统一编址构成全局内存池，实现跨超节点内节点（简称：跨节点）的内存语义访问。全局内存池为由节点的存储介质经过统一编址构成的节点共享的资源。

本申请提供的全局内存池可以包括超节点中计算节点的存储介质和存储节点的存储介质。计算节点的存储介质包括计算节点内的本地存储介质和计算节点连接的扩展存储介质中至少一种。存储节点的存储介质包括存储节点内的本地存储介质和存储节点连接的扩展存储介质中至少一种。

例如，全局内存池包括计算节点内的本地存储介质和存储节点内的本地存储介质。

又如，全局内存池包括计算节点内的本地存储介质、计算节点连接的扩展存储介质，以及存储节点内的本地存储介质和存储节点连接的扩展存储介质中任意一种。

又如，全局内存池包括计算节点内的本地存储介质、计算节点连接的扩展存储介质、存储节点内的本地存储介质和存储节点连接的扩展存储介质。

示例地，如图 2 所示，为本申请提供的一种全局内存池的部署场景示意图。全局内存池 200 包括 N 个计算节点中每个计算节点内的存储介质 210、N 个计算节点中每个计算节点连接的扩展存储介质 220、M 个存储节点中每个存储节点内的存储介质 230 和 M 个存储节点中每个存储节点连接的扩展存储介质 240。

应理解，全局内存池的存储容量可以包括计算节点的存储介质中的部分存储容量和存储节点的存储介质中的部分存储容量。全局内存池是经过统一编址的超节点内计算节点和存储节点均可以访问的存储介质。全局内存池的存储容量可以通过大内存、分布式数据结构、数据缓存、元数据等内存接口供计算节点或存储节点使用。计算节点运行应用程序可以使用这些内存接口对全局内存池进行内存操作。如此，基于计算节点的存储介质的存储容量和存储节点的存储介质构建的全局内存池北向提供了统一的内存接口供计算节点使用，使计算节点使用统一的内存接口将数据写入全局内存池的计算节点提供的存储空间或存储节点提供的存储空间，实现基于内存操作指令的数据的计算和存储，以及降低数据处理的时延，提升数据处理的速度。

上述是以计算节点内的存储介质和存储节点内的存储介质构建全局内存池为例进行说明。全局内存池的部署方式可以灵活多变，本申请实施例不予限定。例如，全局内存池由存储节点的存储介质构建。又如，全局内存池由计算节点的存储介质构建。使用单独的存储节点的存储介质或计算节点的存储介质构建全局内存池可以减少存储侧的存储资源的占用，以及提供更灵活的扩展方案。

依据存储介质的类型划分，本申请实施例提供的全局内存池的存储介质包括动态随机存取存储器（Dynamic Random Access Memory，DRAM）、固态驱动器（Solid State Disk 或 Solid State Drive，SSD）和存储级内存（storage-class-memory，SCM）。

在一些实施例中，可以根据存储介质的类型设置全局内存池，即利用一种类型的存储介质构建一种内存池，不同类型的存储介质构建不同类型的全局内存池，使全局内存池应用于不同的场景，计算节点根据应用的访问特征选择存储介质，增强了用户对系统控制权限，提升了用户的系统体验又扩展了系统适用的应用场景。例如，将计算节点中的 DRAM 和存储节点中的 DRAM 进行统一编址构成 DRAM 内存池。DRAM 内存池用于对访问性能要求高，数据容量适中，无数据持久化诉求的应用场景。又如，将计算节点中的 SCM 和存储节点中的 SCM 进行统一编址构成 SCM 内存池。SCM 内存池则用于对访问性能不敏感，数据容量大，对数据持久化有诉求的应用场景。

接下来，结合图 3 至图 6 对本申请提供的数据处理方法的实施方式进行详细描述。

图 3 为本申请提供的一种节点接入数据处理系统方法的流程示意图。在这里以扩展数据处理系统 100 的规模，加速器 140 请求接入数据处理系统 100 为例进行说明。如图 3 所示，该方法包括以下步骤。

步骤 310、加速器 140 发送广播消息，广播消息用于指示对加速器 140 进行认证。

加速器 140 与数据处理系统 100 中的互联设备 120 建立物理连接并上电后，加速器 140 请求接入数据处理系统 100。例如加速器 140 发送广播消息，即向数据处理系统 100 中的多个节点 110 和互联设备 120 发送广播消息，请求接入数据处理系统 100。广播消息包括加速器 140 的设备标识（Device_ID）和加速器 140 的存储空间的物理地址。加速器 140 存储有设备标识，设备标识可以是加速器 140 出厂时预先配置的。其中，节点 110 接收到广播消息丢弃，控制器 130 接收到广播消息，根据广播消息的指示对加速器 140 进行认证，即执行步骤 320 至步骤 340。

步骤 320、控制器 130 为加速器 140 分配节点标识。

控制器 130 存储有设备标识表，设备标识表包括数据处理系统 100 中已认证的活跃节点的设备标识。控制器 130 接收到广播消息，根据加速器 140 的设备标识查询设备标识表。如果控制器 130 确定设备标识表包括加速器 140 的设备标识，表示加速器 140 已接入数据处理系统 100，加速器 140 为活跃节点。如果控制器 130 确定设备标识表未包括与加速器 140 的设备标识相同的标识，表示加速器 140 为请求认证的节点，则控制器 130 更新设备标识表，即将加速器 140 的设备标识写入设备标识表。进而，控制器 130 可以为加速器 140 分配节点标识，加速器 140 的节点标识用于唯一指示加速器 140，互联设备 120 基于加速器 140 的节点标识向加速器 140 发送数据，使节点 110 与加速器 140 进行通信。

步骤 330、控制器 130 向互联设备 120 发送加速器 140 的节点标识与端口的对应关系。

控制器 130 通过连接互联设备 120 的光纤、铜缆或铜线等物理介质，向互联设备 120 发送加速器 140 的节点标识与端口的对应关系。

步骤 340、互联设备 120 根据节点标识与端口的对应关系更新转发表。

转发表用于指示互联设备 120 根据节点标识与端口的对应关系将通信流量转发到节点标识指示的节点。转发表包括节点标识与端口的对应关系。

例如，加速器 140 的节点标识对应的端口可以用于指示互联设备 120 向加速器 140 转发数据。互联设备 120 接收到加速器 140 的节点标识与端口的对应关系后，更新转发表，即将加速器 140 的节点标识与端口的对应关系写入转发表。

在一种示例中，节点标识与端口的对应关系可以以表格的形式呈现，如表 1 所示。

表 1

节点标识	端口
节点标识 1	端口 1
节点标识 2	端口 2

如表 1 所示，假设加速器 140 的节点标识为节点标识 1，互联设备 120 采用端口 1 连接加速器 140。互联设备 120 接收到节点标识 1，根据节点标识 1 查询表 1，确定节点标识 1 对应端口 1，通过端口 1 向加速器 140 发送数据。

需要说明的是，表 1 只是以表格的形式示意对应关系在存储设备中的存储形式，并不是对该对应关系在存储设备中的存储形式的限定，当然，该对应关系在存储设备中的存储形式还可以以其他的形式存储，本实施例对此不做限定。

步骤 350、控制器 130 向多个节点 110 发送加速器 140 的信息。

控制器 130 向多个节点 110 发送加速器 140 的节点标识、加速器 140 的设备标识和加速器 140 的存储空间的物理地址。多个节点 110 是指数据处理系统 100 中的已认证的活跃节点（Active Nodes）。

加速器 140 的节点标识和加速器 140 的物理地址可以作为加速器 140 在数据处理系统 100 中的全局物理地址。加速器 140 的节点标识用于唯一指示加速器 140。加速器 140 的物理地址是指加速器 140 内存储空间的地址。加速器 140 的全局物理地址用于唯一指示数据处理系统 100 中加速器 140 的存储空间。每个活跃节点存储加速器 140 的节点标识和加速器 140 的存储空间的物理地址，以便于活跃节点根据加速器 140 的节点标识和加速器 140 的物理地址对加速器 140 的存储空间进行读操作或写操作，即向加速器 140 的物理地址指示的存储空间写入数据，或者，从加速器 140 的物理地址指示的存储空间读取数据。

每个活跃节点根据加速器 140 的设备标识查询设备列表，确定加速器 140 的设备标识对应的软件驱动，运行加速器 140 的设备标识对应的软件驱动，以便于多个节点 110 与加速器 140 进行通信，实现访问加速器 140 的功能。

如此，相对于在数据处理系统的启动阶段以静态模式进行配置，可以按需弹性扩展数据处理系统的规模，支持接入加速器、增大全局内存的存储空间等，从而满足不同的应用场景下的计算需求。

在加速器 140 接入数据处理系统 100 后，数据处理系统 100 中活跃节点可以基于加速器 140 的节点标识和加速器 140 的物理地址与加速器 140 进行通信。假设第一节点为节点 110，第二节点为加速器 140，节点 110 请求加速器 140 执行加速操作。图 4 为本申请提供的一种数据处理方法的流程示意图。如图 4 所示，该方法包括以下步骤。

步骤 410、节点 110 发送访问请求，访问请求用于指示请求加速器 140 执行加速操作。

访问请求可以包括加速器 140 的节点标识，以便于互联设备 120 基于加速器 140 的节点标识转发访问请求。

访问请求还可以包括源地址、目的地址和操作标识，以便于加速器 140 根据源地址获取待处理数据，处理待处理数据，以及根据目的地址存储处理后数据。其中，源地址用于指示存储待处

理数据的节点的节点标识和物理地址。目的地址用于指示存储处理后数据的节点的节点标识和物理地址。

需要说明的是，本申请实施例所述的包括节点标识和物理地址的源地址和包括节点标识和物理地址的目的地址可以是一种全局物理地址，将节点的数据处理请求扩充到系统内不同节点间，根据节点标识唯一指示的节点的物理地址，获取待处理数据或存储处理后数据，从而使系统内的加速器资源可以为多个节点共享，适应不同的应用场景下的计算需求。

另外，本申请实施例对存储待处理数据的节点和存储处理后数据的节点不予限定。例如，源地址包含请求执行加速操作的节点的节点标识和物理地址。目的地址包含执行加速操作的节点的节点标识和物理地址。节点 110 请求加速器 140 执行加速操作，源地址包括节点 110 的节点标识和节点 110 的物理地址，目的地址包括加速器 140 的节点标识和加速器 140 的物理地址。又如，源地址包含的物理地址可以指示任意一个节点 110 或加速器 140。目的地址包含的物理地址可以指示任意一个节点 110 或加速器 140。

在一些实施例中，节点 110 可以复用领域特定结构（Domain Specific Architecture，DSA）架构中的描述符（Descriptor）指示加速操作。示例地，如图 5 所示，为本申请提供的一种描述符的结构示意图。其中，描述符包括操作标识、源地址和目的地址。

源地址用于指示加速操作所使用的待处理数据的存储位置。目的地址用于指示加速操作的结果的存储位置，即处理后数据的存储位置。

例如，描述符可以是一个 64 字节的描述符。源地址的长度为 64 比特。目的地址的长度为 64 比特。节点标识的长度可以是 12 比特。物理地址的长度可以是 52 比特。根据节点中不同的内存配置，自适应配置描述符中节点标识的长度和物理地址的长度。内存配置包括内存的存储容量大小和内存类型。内存类型包括 DRAM、SSD 和 SCM。

描述符还可以包括操作自定义域，操作自定义域用于指示按照不同的操作标识可自定义的操作。

节点 110 通过加速器驱动生成 64 字节描述符，并运行 None-Posted Write 指令，通过节点 110 与互联设备 120 的互联总线将描述符写入加速器 140 的请求队列。节点 110 以内存映射输入输出（Memory Mapped IO，MMIO）方式将读写节点 110 的请求队列的寄存器地址映射给加速器 140，而加速器 140 通过 None-Posted Write 指令基于互联总线访问该寄存器。

另外，操作标识也可以称为操作算子。操作标识用于指示所执行的加速操作。加速操作包括以下任一项。

内存交互（SWAP）：指示节点的内存之间的数据块交换。

归约（Reduce）：指针对多份本地数据执行算术或逻辑计算。

广播（Broadcast）：指将本地数据块广播给系统中的节点。

查询（Search）：指数据库的查询操作， 返回匹配结果。

加密/解密（Encryption/Decryption）：指数据块加解密操作。

压缩/解压缩（Compress/Decompress）：指数据块压缩/解压缩操作。

步骤 420、互联设备 120 基于加速器 140 的节点标识对应的端口转发访问请求。

互联设备 120 存储有节点标识与端口的对应关系。互联设备 120 接收到访问请求后，获取到加速器 140 的节点标识，查询转发表，根据转发表所确定的加速器 140 的节点标识对应的端口向加速器 140 转发访问请求。

步骤 430、加速器 140 根据操作标识执行加速操作。

加速器 140 从请求队列中读取描述符，并解析描述符的各字段，执行操作标识指示的加速操作。

例如，加速器 140 根据描述符需求在加速器 140 的本地内存中分配一段存储空间。该段存储空间可以用于存储加速器 140 执行加速操作的中间数据。

又如，如果操作标识用于指示对远端节点的数据处理，加速器 140 驱动本地的 SDMA 引擎，将远端节点（如：发出访问请求的节点 110）的内存的数据写入到加速器 140 的本地内存。

又如，如果描述符指示对加速器 140 的本地数据的请求，驱动数据读引擎，将网络或者本地

内存中的数据读取到加速器 140 的本地缓存中。

可理解地，加速器 140 根据源地址指示的节点标识和物理地址读取待处理数据。源地址指示的物理地址包括发出访问请求的节点 110 中存储空间的物理地址、系统中其他的节点 110 中存储空间的物理地址和加速器 140 中存储空间的物理地址中任意一个。即，加速器 140 读取待处理数据的位置包括发出访问请求的节点 110 的存储空间、系统中其他的节点 110 的存储空间和加速器 140 的存储空间中任意一个。

加速器 140 根据目的地址指示的节点标识和物理地址存储加速操作的处理后数据，以便于发出访问请求的节点 110 根据目的地址指示的物理地址读取加速操作的处理后数据。目的地址指示的物理地址包括发出访问请求的节点 110 中存储空间的物理地址、系统中其他的节点 110 中存储空间的物理地址和加速器 140 中存储空间的物理地址中任意一个。即，加速器 140 将处理后数据存储到以下任意一个位置，包括发出访问请求的节点 110 的存储空间、系统中其他的节点 110 的存储空间、加速器 140 的存储空间。

可选地，源地址指示的物理地址和目的地址指示的物理地址可以指示全局内存池中存储介质的存储空间。即加速器 140 可以从全局内存池中读取待处理数据或/和将处理后数据存储到全局内存池，以提升数据处理的速度。

加速器 140 按照描述符中描述的操作标识对本地缓存的待处理数据执行加速操作。例如，操作标识指示压缩操作，加速器 140 对本地缓存的待处理数据执行压缩操作。又如，操作标识指示加密操作，加速器 140 对本地缓存的待处理数据执行加密操作。

加速器 140 执行完加速操作后，释放本地缓存。加速器 140 还可以通过请求完成中断（Request Completion Interrupt）消息触发请求描述符节点的中断，由发出访问请求的节点 110 取回加速操作的结果。

从而，将通用处理器（如：CPU）的作业卸载到加速器，由加速器处理计算需求较高的作业（如：HPC、大数据作业、数据库作业等），解决由于通用处理器浮点算力不足，无法满足 HPC、AI 等场景的重浮点计算需求的问题，缩短数据处理时长以及降低系统能耗，提升系统性能。独立部署的加速器和集成加速器的节点支持灵活插拔，可以按需弹性扩展数据处理系统的规模，从而满足不同的应用场景下的计算需求。

在另一些实施例中，数据处理系统中的任一个活跃节点可以退出系统。图 6 为本申请提供的一种节点退出数据处理系统方法的流程示意图。在这里以加速器 140 请求退出数据处理系统 100 为例进行说明。如图 6 所示，该方法包括以下步骤。

在一些实施例中，加速器 140 可以主动退出数据处理系统，即加速器 140 执行步骤 610。

步骤 610、加速器 140 广播老化消息，老化消息用于指示加速器 140 退出数据处理系统。执行步骤 630 和步骤 640。

在另一些实施例中，控制器 130 接收到互联设备 120 发送的链路故障信息，或者加速器 140 与其他节点 110 的心跳消息超时，执行步骤 620。

步骤 620、控制器 130 广播老化消息，老化消息用于指示加速器 140 退出数据处理系统。执行步骤 630 和步骤 640。

老化消息可以包括加速器 140 的节点标识、加速器 140 的物理地址和加速器 140 的设备标识。

步骤 630、互联设备 120 老化加速器 140 的节点标识与端口的对应关系。

例如，互联设备 120 接收到第一老化消息，删除转发表中加速器 140 的节点标识与端口对应的转发表项。

步骤 640、节点 110 老化加速器 140 的节点标识和加速器 140 的物理地址。

例如，节点 110 接收到第二老化消息后，删除存储的加速器 140 的节点标识和物理地址和设备列表中加速器 140 的设备标识对应的软件驱动。

如此，在数据处理系统中设置控制器和互联设备，基于节点接入机制和退出机制，可弹性增加及减少节点，实现了可弹性扩展的超节点架构，既解决了传统超节点架构无法动态扩展的问题，又避免了传统 IO 总线架构规模受限和带宽低问题， 并支持在节点或者互联设备故障情况下的动态容错机制。

可以理解的是，为了实现上述实施例中的功能，控制器包括了执行各个功能相应的硬件结构和/或软件模块。本领域技术人员应该很容易意识到，结合本申请中所公开的实施例描述的各示例的单元及方法步骤，本申请能够以硬件或硬件和计算机软件相结合的形式来实现。某个功能究竟以硬件还是计算机软件驱动硬件的方式来执行，取决于技术方案的特定应用场景和设计约束条件。

上文中结合图 1 至图 6，详细描述了根据本实施例所提供的数据处理方法，下面将结合图 7，描述根据本实施例所提供的控制装置和数据处理节点。

图 7 为本实施例提供的可能的控制装置的结构示意图。这些控制装置可以用于实现上述方法实施例中控制器的功能，因此也能实现上述方法实施例所具备的有益效果。在本实施例中，该控制装置可以是如图 3、图 4、图 6 所示的控制器 130，还可以是应用于服务器的模块（如芯片）。

如图 7 所示，控制装置 700 包括通信模块 710、控制模块 720 和存储模块 730。控制装置 700 用于实现上述图 3、图 4、图 6 中所示的方法实施例中控制器 130 的功能。

通信模块 710 用于接收第二节点发送的接入请求，接入请求用于指示对第二节点进行认证。

控制模块 720，用于当第二节点请求接入数据处理系统时，为第二节点分配第二节点标识，第二节点标识用于唯一指示第二节点。例如，控制模块 720 用于执行图 3 中步骤 320。

通信模块 710，还用于向互联设备发送第二节点标识与第二端口的对应关系，第二节点标识对应的第二端口用于向第二节点转发消息。例如，控制模块 720 用于执行图 3 中步骤 330。

通信模块 710，还用于将第二节点的全局物理地址发送给数据处理系统中的其他节点，例如，向数据处理系统中的其他节点广播第二节点的节点标识和物理地址。例如，控制模块 720 用于执行图 3 中步骤 350。

通信模块 710，还用于向互联设备发送第一老化消息，第一老化消息用于指示互联设备老化第二节点标识与第二端口的对应关系。例如，控制模块 720 用于执行图 6 中步骤 620。

通信模块 710，还用于广播第二老化消息，第二老化消息用于指示老化第二节点标识和第二节点的物理地址。

存储模块 730 用于存储节点的节点标识，以便于控制模块 720 控制节点接入数据处理系统和退出数据处理系统。

图 8 为本实施例提供的可能的数据处理节点的结构示意图。这些数据处理节点可以用于实现上述方法实施例中数据处理系统中节点的功能，因此也能实现上述方法实施例所具备的有益效果。在本实施例中，该数据处理节点可以是如图 3、图 4、图 6 所示的节点 110 或加速器 140，还可以是应用于服务器的模块（如芯片）。

如图 8 所示，数据处理节点 800 包括通信模块 810、数据处理模块 820 和存储模块 830。数据处理节点 800 用于实现上述图 3、图 4、图 6 中所示的方法实施例中节点 110 或加速器 140 的功能。

通信模块 810 用于接收所述第一节点发送的访问请求，所述访问请求包括源地址、目的地址和操作标识，所述源地址用于指示存储待处理数据的节点的节点标识和物理地址，所述目的地址用于指示存储处理后数据的节点的节点标识和物理地址。

数据处理模块 820，用于根据所述操作标识对所述待处理数据执行加速操作得到处理后数据，以及根据所述第一节点指示的目的地址存储所述处理后数据。例如，数据处理模块 820 用于执行图 4 中步骤 430。

存储模块 830 用于存储内存操作指令、待处理数据或处理后数据，以便于数据处理模块 820 执行加速操作。

应理解的是，本申请实施例的控制装置 700 和数据处理节点 800 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 logical 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 3、图 4、图 6 所示的数据处理方法时，及其各个模块也可以为软件模块，控制装置 700 和数据处理节点 800 及其各个模块也可以为软件模块。

根据本申请实施例的控制装置 700 和数据处理节点 800 可对应于执行本申请实施例中描述的方法，并且控制装置 700 和数据处理节点 800 中的各个单元的上述和其它操作和/或功能分别为了

实现图 3、图 4、图 6 中的各个方法的相应流程，为了简洁，在此不再赘述。

图 9 为本实施例提供的一种计算设备 900 的结构示意图。如图所示，计算设备 900 包括处理器 910、总线 920、存储器 930、通信接口 940 和内存单元 950（也可以称为主存（main memory）单元）。处理器 910、存储器 930、内存单元 950 和通信接口 940 通过总线 920 相连。

应理解，在本实施例中，处理器 910 可以是 CPU，该处理器 910 还可以是其他通用处理器、数字信号处理器（digital signal processing，DSP）、ASIC、FPGA 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者是任何常规的处理器等。

处理器还可以是图形处理器（graphics processing unit，GPU）、神经网络处理器（neural network processing unit，NPU）、微处理器、ASIC、或一个或多个用于控制本申请方案程序执行的集成电路。

通信接口 940 用于实现计算设备 900 与外部设备或器件的通信。在本实施例中，计算设备 900 用于实现图 3 所示的控制器 130 的功能时，通信接口 940 用于获取广播消息，处理器 910 为节点 110 分配节点标识，广播节点标识和节点的物理地址。计算设备 900 用于实现图 6 所示的控制器 130 的功能时，通信接口 940 用于广播老化消息。

总线 920 可以包括一通路，用于在上述组件（如处理器 910、内存单元 950 和存储器 930）之间传送信息。总线 920 除包括数据总线之外，还可以包括电源总线、控制总线和状态信号总线等。但是为了清楚说明起见，在图中将各种总线都标为总线 920。总线 920 可以是快捷外围部件互连标准（Peripheral Component Interconnect Express，PCIe)总线，或扩展工业标准结构（extended industry standard architecture，EISA）总线、统一总线（unified bus，Ubus 或 UB）、计算机快速链接（compute express link，CXL）、缓存一致互联协议（cache coherent interconnect for accelerators，CCIX）等。总线 920 可以分为地址总线、数据总线、控制总线等。

作为一个示例，计算设备 900 可以包括多个处理器。处理器可以是一个多核（multi-CPU）处理器。这里的处理器可以指一个或多个设备、电路、和/或用于处理数据（例如计算机程序指令）的计算单元。在本实施例中，计算设备 900 用于实现图 4 所示的加速器 140 的功能时，处理器 910 还用于根据第一节点指示的源地址获取待处理数据，根据第一节点指示的操作标识对待处理数据执行加速操作得到处理后数据，以及根据第一节点指示的目的地址存储处理后数据，源地址用于指示存储待处理数据的节点的节点标识和物理地址；第二节点所述目的地址用于指示存储处理后数据的节点的节点标识和物理地址。

计算设备 900 用于实现图 4 所示的控制器 130 的功能时，处理器 910 还用于在节点请求接入数据处理系统时，为所述节点分配节点标识，将所述节点的全局物理地址发送给数据处理系统中的节点，向互联设备发送所述节点标识与端口的对应关系，所述节点标识对应的端口用于向所述节点转发消息，以及在所述节点退出所述数据处理系统时，老化所述节点的全局物理地址和节点标识与端口的对应关系。

计算设备 900 用于实现图 4 所示的互联设备 120 的功能时，处理器 910 还用于基于节点标识与端口的对应关系转发节点间进行通信的数据。

值得说明的是，图 9 中仅以计算设备 900 包括 1 个处理器 910 和 1 个存储器 930 为例，此处，处理器 910 和存储器 930 分别用于指示一类器件或设备，具体实施例中，可以根据业务需求确定每种类型的器件或设备的数量。

内存单元 950 可以对应上述方法实施例中用于存储待处理数据和处理后数据等信息的全局内存池。内存单元 950 可以是易失性存储器池或非易失性存储器池，或可包括易失性和非易失性存储器两者。其中，非易失性存储器可以是只读存储器（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）或闪存。易失性存储器可以是随机存取存储器（random access memory，RAM），其用作外部高速缓存。通过示例性但不是限制性说明，许多形式的 RAM 可用，例如静态随机存取存储器（static RAM，SRAM）、动态随机存取存储器（DRAM）、同步动态随机存取存储器（synchronous DRAM，SDRAM）、双倍数据速率

同步动态随机存取存储器（double data date SDRAM，DDR SDRAM）、增强型同步动态随机存取存储器（enhanced SDRAM，ESDRAM）、同步连接动态随机存取存储器（synchlink DRAM，SLDRAM）和直接内存总线随机存取存储器（direct rambus RAM，DR RAM）。

存储器 930 可以对应上述方法实施例中用于存储计算机指令、内存操作指令、节点标识等信息的存储介质，例如，磁盘，如机械硬盘或固态硬盘。

上述计算设备 900 可以是一个通用设备或者是一个专用设备。例如，计算设备 900 可以是边缘设备（例如，携带具有处理能力芯片的盒子）等。可选地，计算设备 900 也可以是服务器或其他具有计算能力的设备。

应理解，根据本实施例的计算设备 900 可对应于本实施例中的控制装置 700 和数据处理节点 800，并可以对应于执行根据图 3、图 4、图 6 中任一方法中的相应主体，并且控制装置 700 和数据处理节点 800 中的各个模块的上述和其它操作和/或功能分别为了实现图 3、图 4、图 6 中的各个方法的相应流程，为了简洁，在此不再赘述。

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计算设备中。

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. 一种数据处理系统，其特征在于，所述数据处理系统包括多个节点和控制器，所述多个节点包括第一节点和第二节点，所述多个节点和所述控制器通过高速互联链路连接；

所述控制器，用于当所述第二节点请求接入所述数据处理系统时，为所述第二节点分配第二节点标识，其中，所述第二节点在所述数据处理系统中的全局物理地址为所述第二节点的节点标识和所述第二节点的物理地址；

所述控制器，还用于将所述第二节点的全局物理地址发送给所述第一节点。

2. 根据权利要求 1 所述的系统，其特征在于，所述数据处理系统还包括互联设备，所述互联设备基于高速互联链路连接所述多个节点；

所述控制器，还用于向所述互联设备发送所述第二节点标识与端口的对应关系，所述第二节点标识对应的端口用于向所述第二节点转发消息。

3. 根据权利要求 2 所述的系统，其特征在于，

所述互联设备用于基于所述第二节点标识与端口的对应关系转发所述第一节点访问所述第二节点的消息，所述第二节点标识用于唯一指示所述第二节点。

4. 根据权利要求 1-3 中任一项所述的系统，其特征在于，所述多个节点的存储介质经过统一编址构成全局内存池。

5. 根据权利要求 1-4 中任一项所述的系统，其特征在于，

所述第二节点用于根据所述第一节点指示的源地址获取待处理数据，所述源地址用于指示存储所述待处理数据的节点的节点标识和物理地址；

所述第二节点还用于处理所述待处理数据，以及根据所述第一节点指示的目的地址存储处理后数据，所述目的地址用于指示存储所述处理后数据的节点的节点标识和物理地址。

6. 根据权利要求 5 所述的系统，其特征在于，所述多个节点的存储介质构成的全局内存池包括所述源地址指示的存储空间或/和所述目的地址指示的存储空间。

7. 根据权利要求 5 或 6 所述的系统，其特征在于，所述第二节点具体用于根据所述第一节点指示的操作标识对所述待处理数据执行加速操作得到所述处理后数据。

8. 根据权利要求 5-7 中任一项所述的系统，其特征在于，

所述第一节点还用于根据所述第一节点的物理地址访问所述第一节点的存储空间。

9. 根据权利要求 1-8 中任一项所述的系统，其特征在于，所述第二节点包括处理器、加速器和内存控制器中任一个。

10. 根据权利要求 1-9 所述的系统，其特征在于，

所述控制器还用于当所述第二节点退出所述数据处理系统时，控制所述第一节点老化所述第二节点的全局物理地址，以及控制互联设备老化所述第二节点标识与端口的对应关系。

11. 一种数据处理方法，其特征在于，数据处理系统包括多个节点，所述多个节点包括第一节点和第二节点，所述多个节点和所述控制器通过高速互联链路连接；所述方法包括：

当所述第二节点请求接入所述数据处理系统时，所述控制器为所述第二节点分配第二节点标识，其中，所述第二节点在所述数据处理系统中的全局物理地址为所述第二节点的节点标识和所述第二节点的物理地址；

所述控制器将所述第二节点的全局物理地址发送给所述第一节点。

12. 根据权利要求 11 所述的方法，其特征在于，所述数据处理系统还包括互联设备，所述互联设备基于高速互联链路连接所述多个节点；所述方法还包括：

所述控制器向所述互联设备发送所述第二节点标识与端口的对应关系，所述第二节点标识对应的端口用于向所述第二节点转发消息。

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：

所述互联设备基于所述第二节点标识与端口的对应关系转发所述第一节点访问所述第二节点的消息，所述第二节点标识用于唯一指示所述第二节点。

14. 根据权利要求 11-13 中任一项所述的方法，其特征在于，所述多个节点的存储介质经过统一编址构成全局内存池。

15. 根据权利要求 11-14 中任一项所述的方法，其特征在于，所述方法还包括：

所述第二节点根据所述第一节点指示的源地址获取待处理数据，所述源地址用于指示存储所述待处理数据的节点的节点标识和物理地址；

所述第二节点处理所述待处理数据，以及根据所述第一节点指示的目的地址存储处理后数据，所述目的地址用于指示存储所述处理后数据的节点的节点标识和物理地址。

16. 根据权利要求 15 所述的方法，其特征在于，所述多个节点的存储介质构成的全局内存池包括所述源地址指示的存储空间或/和所述目的地址指示的存储空间。

17. 根据权利要求 15 或 16 所述的方法，其特征在于，所述第二节点处理所述待处理数据包括：

所述第二节点根据所述第一节点指示的操作标识对所述待处理数据执行加速操作得到所述处理后数据。

18. 根据权利要求 15-17 中任一项所述的方法，其特征在于，所述方法还包括：

所述第一节点根据所述第一节点的物理地址访问所述第一节点的存储空间。

19. 根据权利要求 11-18 中任一项所述的方法，其特征在于，所述第二节点包括处理器、加速器和内存控制器中任一个。

20. 根据权利要求 11-19 所述的方法，其特征在于，所述方法还包括：

当所述第二节点退出所述数据处理系统时，所述控制器控制所述第一节点老化所述第二节点的全局物理地址，以及控制互联设备老化所述第二节点标识与端口的对应关系。

21. 一种控制装置，其特征在于，所述控制装置应用于数据处理系统，所述数据处理系统包括基于高速互联技术连接的多个节点，所述多个节点包括第一节点和第二节点，所述装置包括：

控制模块，用于当所述第二节点请求接入所述数据处理系统时，为所述第二节点分配第二节点标识，其中，所述第二节点标识用于唯一指示所述第二节点，所述第二节点在所述数据处理系统中的全局物理地址为所述第二节点的节点标识和所述第二节点的物理地址；

所述控制模块，还用于将所述第二节点的全局物理地址发送给所述第一节点。

22. 根据权利要求 21 所述的装置，其特征在于，所述数据处理系统还包括互联设备，所述互联设备基于高速互联链路连接所述多个节点；所述装置还包括通信模块；

所述通信模块，还用于向所述互联设备发送所述第二节点标识与端口的对应关系，所述第二节点标识对应的端口用于向所述第二节点转发消息。

23. 根据权利要求 22 所述的装置，其特征在于，

所述通信模块，还用于当所述第二节点退出所述数据处理系统时，向所述互联设备发送第一老化消息，所述第一老化消息用于指示所述互联设备老化所述第二节点标识与端口的对应关系；

所述通信模块，还用于向所述第一节点发送第二老化消息，所述第二老化消息用于指示所述第二节点的全局物理地址。

24. 一种数据处理节点，其特征在于，所述数据处理节点为数据处理系统中基于高速互联链路连接的多个节点中的其中一个节点，所述装置包括：

通信模块，用于接收所述多个节点中的其他节点发送的访问请求，所述访问请求包括源地址、目的地址和操作标识，所述源地址用于指示存储待处理数据的节点的节点标识和物理地址，所述目的地址用于指示存储处理后数据的节点的节点标识和物理地址；

数据处理模块，用于根据所述操作标识对所述待处理数据执行加速操作得到处理后数据，以及根据所述目的地址存储所述处理后数据。

25. 一种控制器，其特征在于，所述控制器包括存储器和至少一个处理器，所述存储器用于存储一组计算机指令；当所述处理器执行所述一组计算机指令时，控制器执行如权利要求 11、12 和 20 中任一所述的方法。

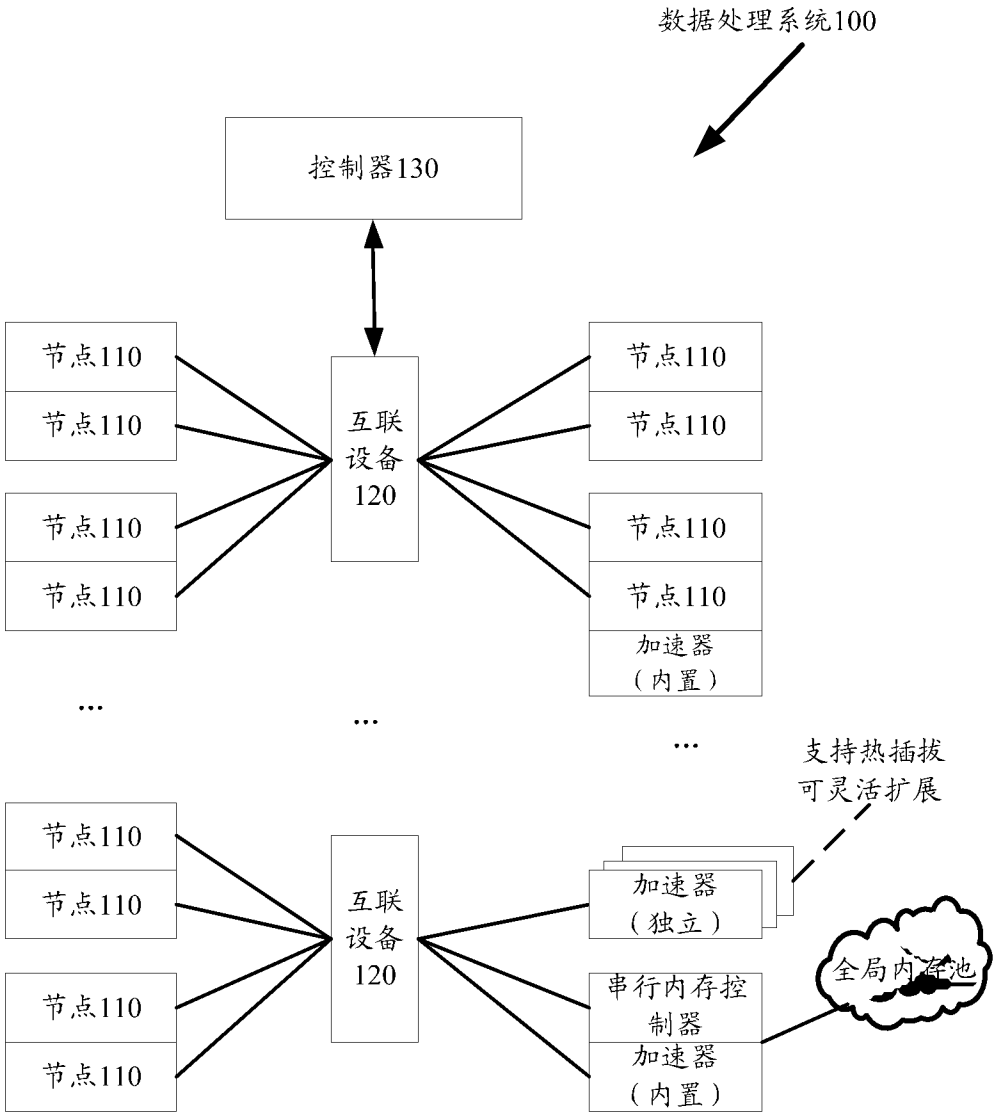

数据处理系统100
控制器130
节点110
节点110
节点110
节点110
互联设备120
节点110
节点110
节点110
节点110
加速器（内置）
支持热插拔
可灵活扩展
节点110
节点110
节点110
节点110
互联设备120
加速器（独立）
串行内存控制器
加速器（内置）
全局内存池

图 1

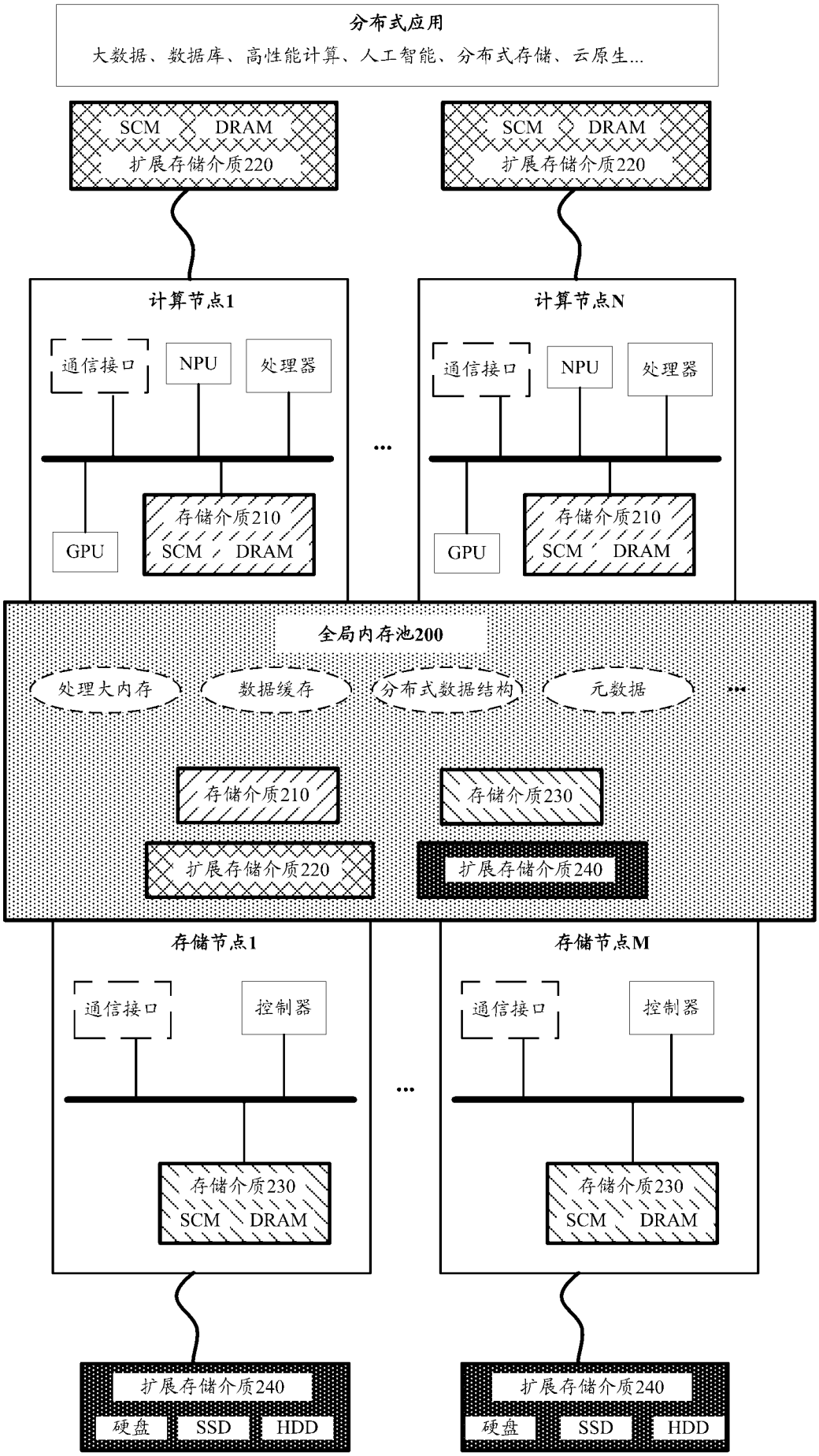

分布式应用
大数据、数据库、高性能计算、人工智能、分布式存储、云原生...
SCM DRAM
扩展存储介质220
SCM DRAM
扩展存储介质220
计算节点1
通信接口
NPU
处理器
GPU
存储介质210
SCM DRAM
计算节点N
通信接口
NPU
处理器
GPU
存储介质210
SCM DRAM
全局内存池200
处理大内存
数据缓存
分布式数据结构
元数据
存储介质210
存储介质230
扩展存储介质220
扩展存储介质240
存储节点1
通信接口
控制器
存储介质230
SCM DRAM
存储节点M
通信接口
控制器
存储介质230
SCM DRAM
扩展存储介质240
硬盘
SSD
HDD
扩展存储介质240
硬盘
SSD
HDD

图 2

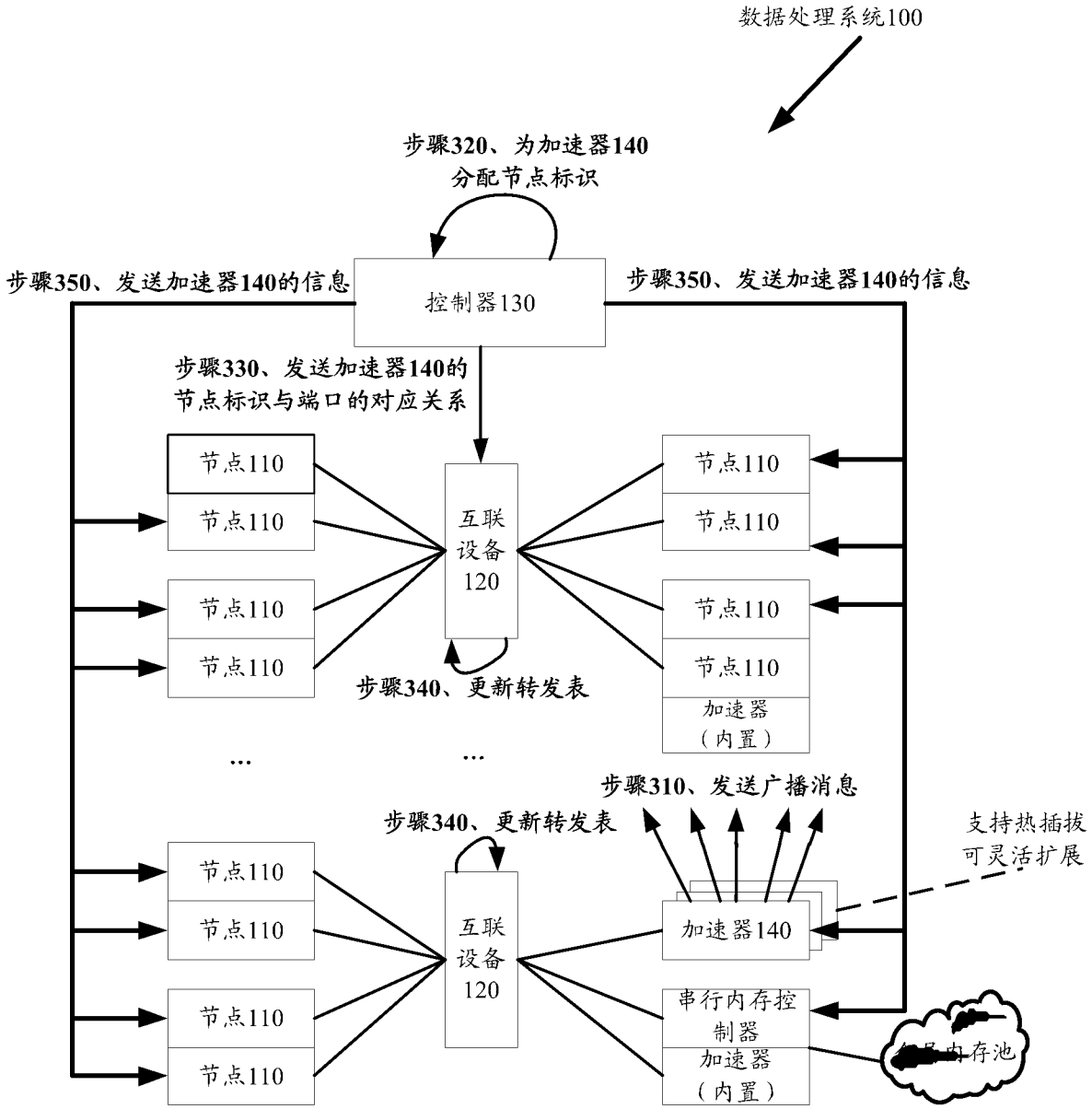

数据处理系统100
步骤320、为加速器140
分配节点标识
步骤350、发送加速器140的信息
控制器130
步骤350、发送加速器140的信息
步骤330、发送加速器140的
节点标识与端口的对应关系
节点110
节点110
节点110
节点110
互联
设备
120
节点110
节点110
节点110
节点110
加速器
（内置）
步骤340、更新转发表
步骤310、发送广播消息
步骤340、更新转发表
支持热插拔
可灵活扩展
节点110
节点110
节点110
节点110
互联
设备
120
加速器140
串行内存控
制器
加速器
（内置）
内存池

图 3

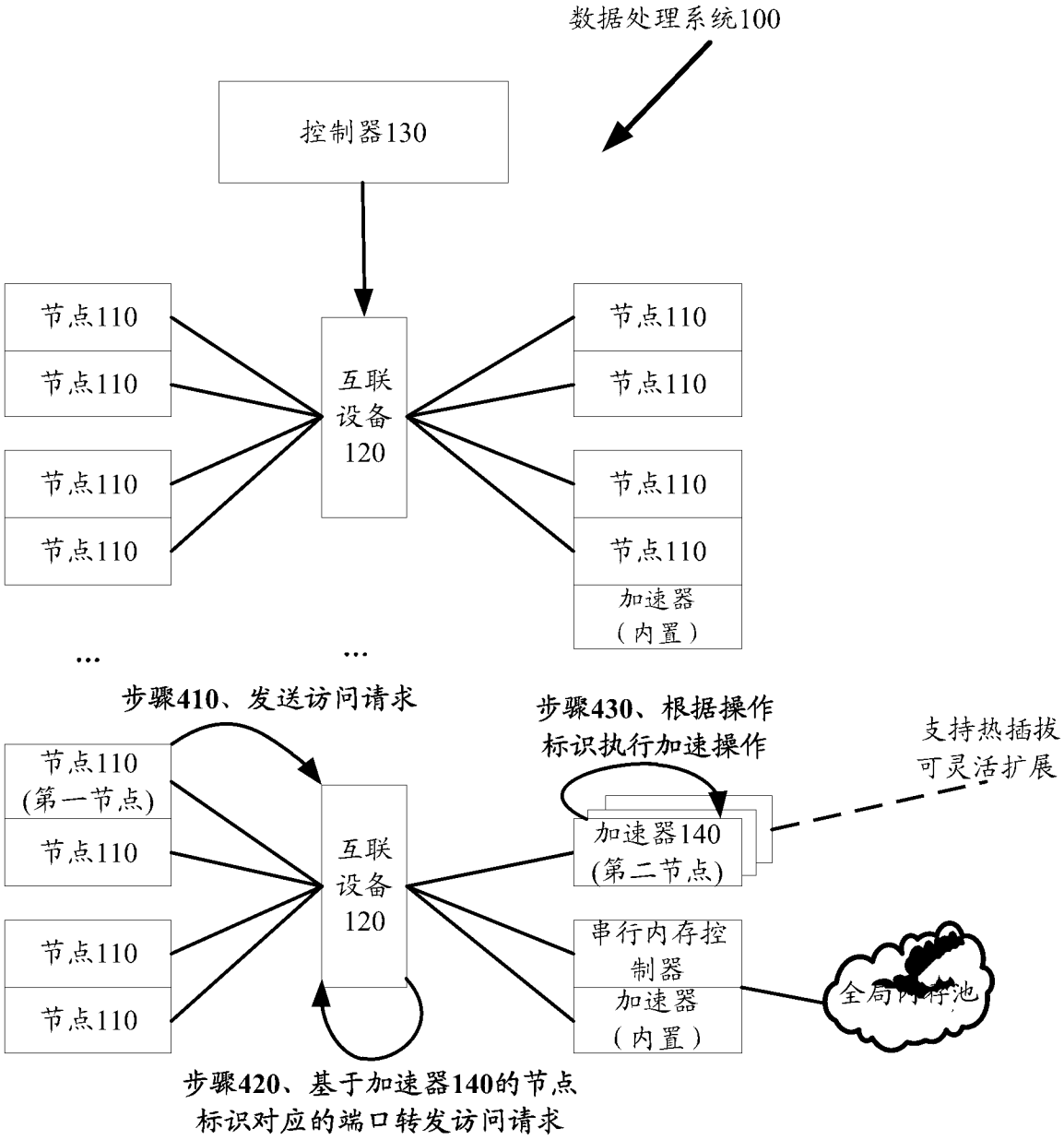

数据处理系统100
控制器130
节点110
节点110
节点110
节点110
互联设备120
节点110
节点110
节点110
节点110
加速器（内置）
步骤410、发送访问请求
步骤430、根据操作标识执行加速操作
支持热插拔
可灵活扩展
节点110（第一节点）
节点110
节点110
节点110
互联设备120
加速器140（第二节点）
串行内存控制器
加速器（内置）
全局内存池
步骤420、基于加速器140的节点标识对应的端口转发访问请求

图 4

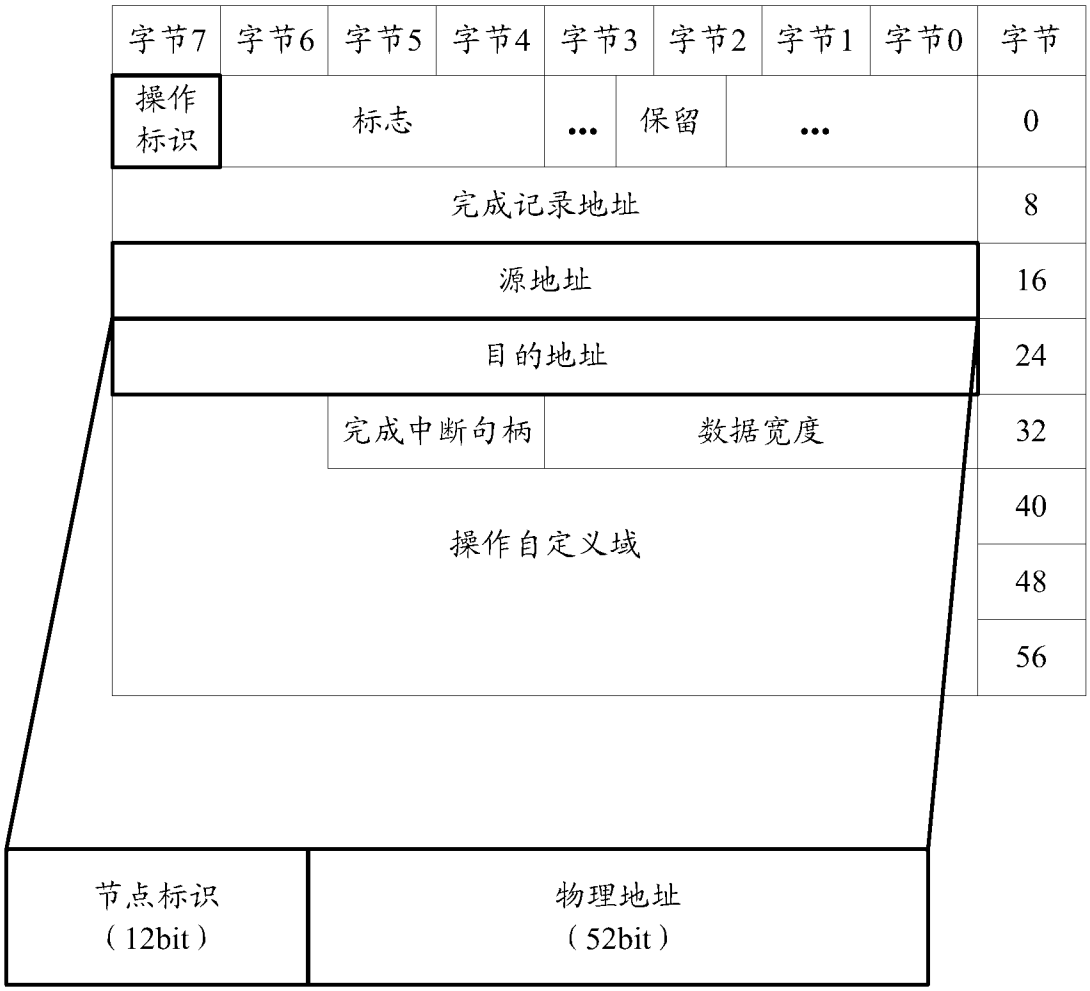

字节7
字节6
字节5
字节4
字节3
字节2
字节1
字节0
字节
操作标识
标志
...
保留
...
0
完成记录地址
8
源地址
16
目的地址
24
完成中断句柄
数据宽度
32
操作自定义域
40
48
56
节点标识（12bit）
物理地址（52bit）

图 5

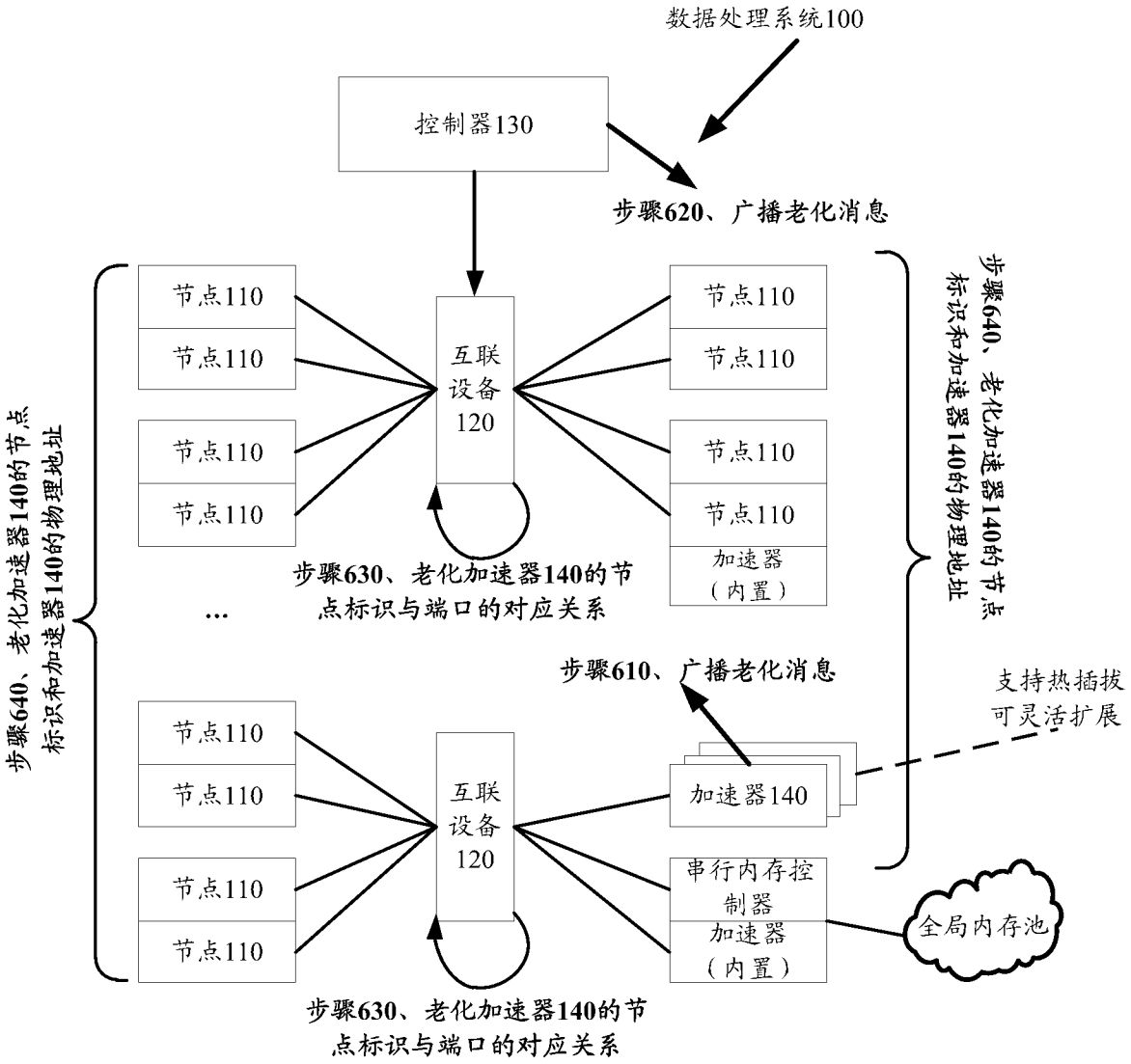

数据处理系统100
控制器130
步骤620、广播老化消息
节点110
节点110
节点110
节点110
互联设备120
节点110
节点110
节点110
节点110
加速器（内置）
步骤630、老化加速器140的节点标识与端口的对应关系
...
步骤640、老化加速器140的节点标识和加速器140的物理地址
步骤640、老化加速器140的节点标识和加速器140的物理地址
步骤610、广播老化消息
支持热插拔
可灵活扩展
节点110
节点110
节点110
节点110
互联设备120
加速器140
串行内存控制器
加速器（内置）
全局内存池
步骤630、老化加速器140的节点标识与端口的对应关系

图 6

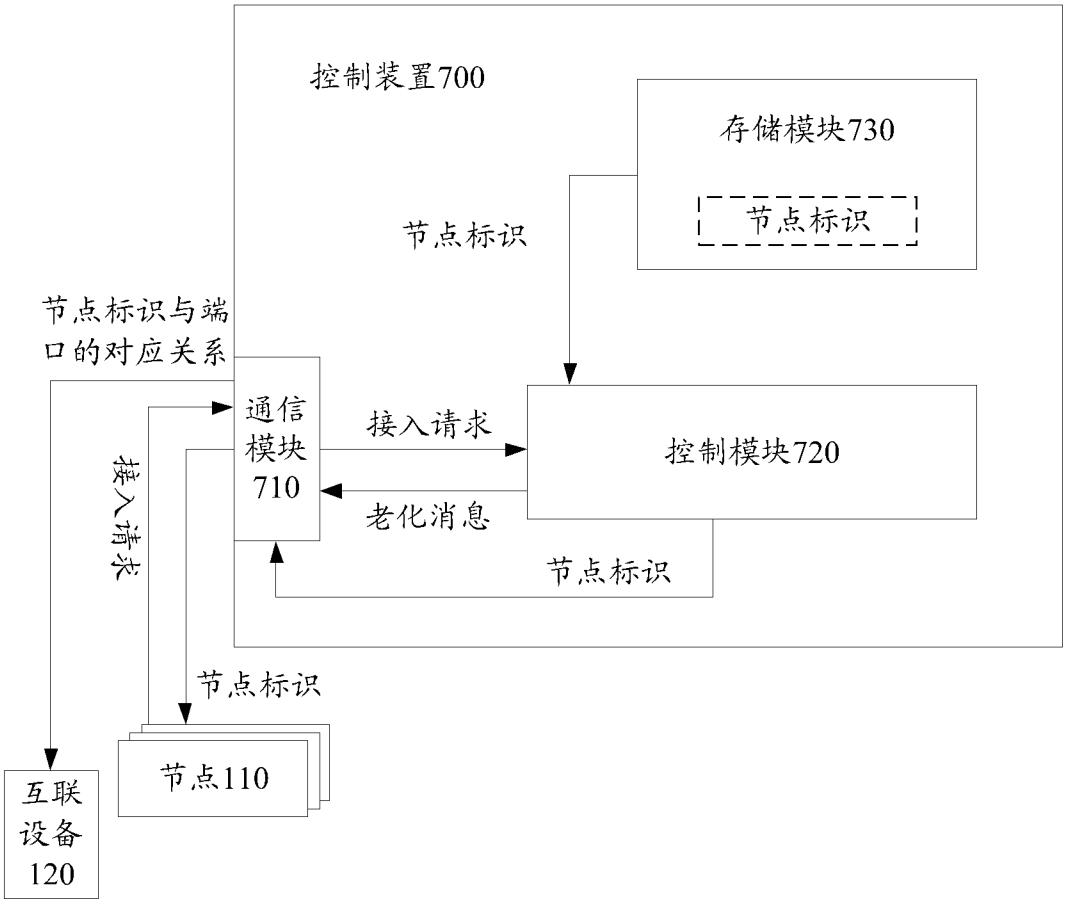

控制装置700
存储模块730
节点标识
节点标识
节点标识与端口的对应关系
通信模块710
接入请求
控制模块720
老化消息
接入请求
节点标识
节点标识
节点110
互联设备120

图 7

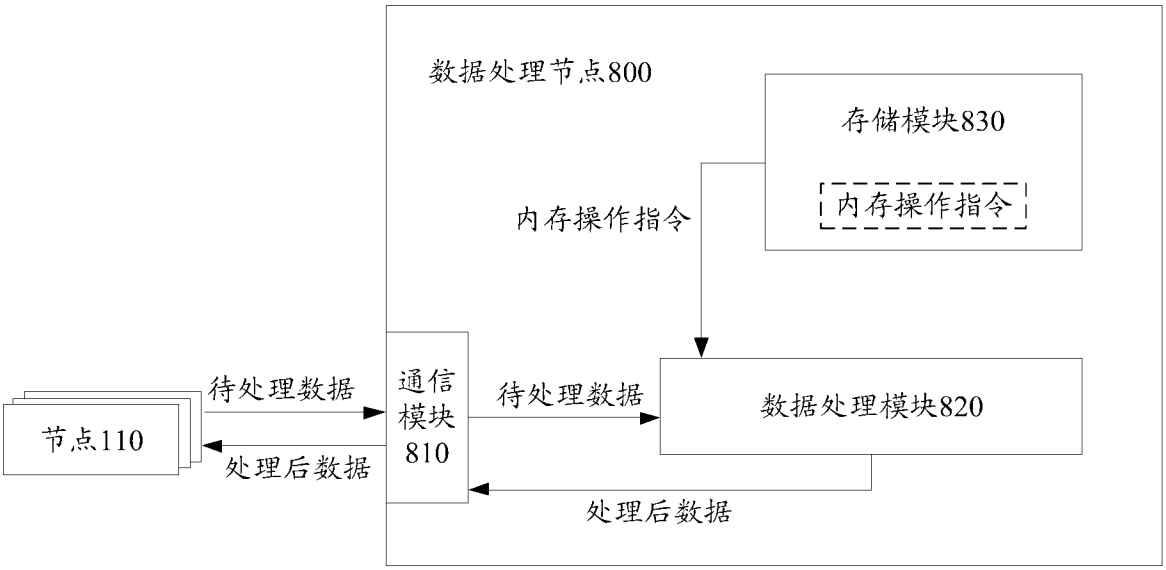

数据处理节点800
存储模块830
内存操作指令
内存操作指令
待处理数据
通信模块810
待处理数据
数据处理模块820
节点110
处理后数据
处理后数据

图 8

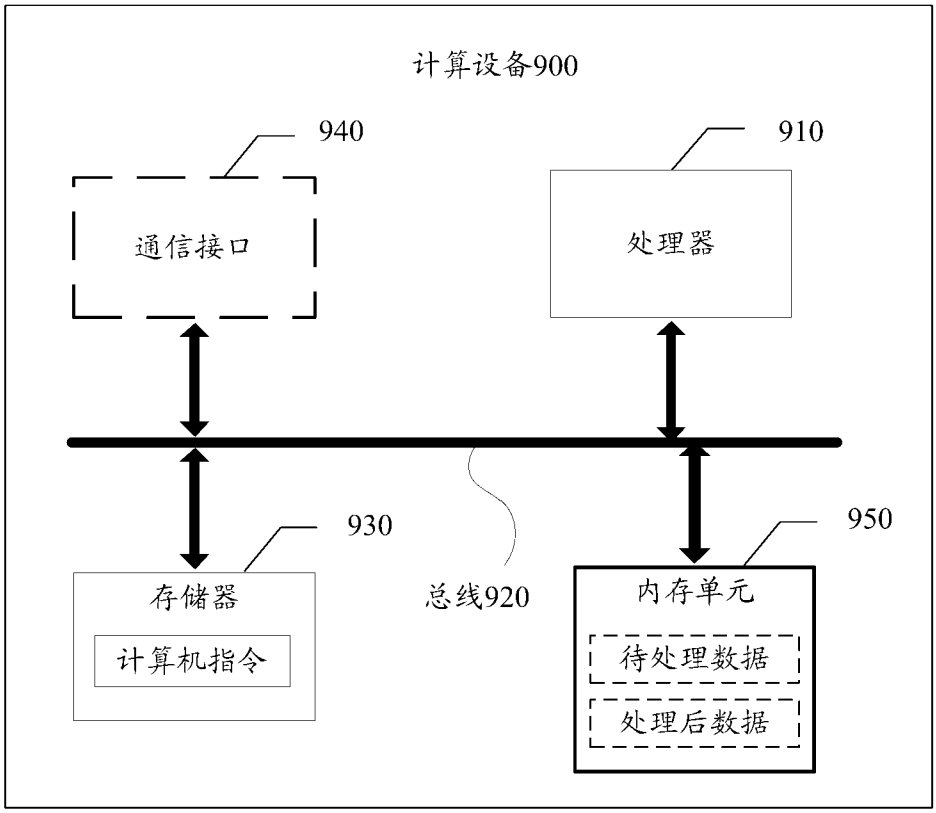

计算设备900
940
通信接口
910
处理器
930
存储器
计算机指令
总线920
950
内存单元
待处理数据
处理后数据

图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/101171

A. CLASSIFICATION OF SUBJECT MATTER

G06F 15/163(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:G06F

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

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

CNABS; CNTXT; EPTXT; VEN; WOTXT; USTXT; CNKI; IEEE: 节点, 多个, 其他, 第二, 互联, 加入, 接入, 插入, 全局, 地址, 分配, 分发, 对应, 端口, 转发, 标识, 老化, 退出, 删除, node, multiple, other, another, second, interconnection, add, access, insert, global, address, allocation, distribute, corresponding, port, retransmission, ID, identification, aging, exit, delete

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 113568562 A (HUAWEI TECHNOLOGIES CO., LTD.) 29 October 2021 (2021-10-29) description, paragraphs 55-120	1-4, 9-14, 19-25
Y	CN 113568562 A (HUAWEI TECHNOLOGIES CO., LTD.) 29 October 2021 (2021-10-29) description, paragraphs 55-120	5-10, 15-20, 25
Y	CN 112165505 A (DBAPPSECURITY CO., LTD.) 01 January 2021 (2021-01-01) description, paragraphs 33-72	5-10, 15-20, 25
Y	CN 113157611 A (SHANDONG YINGXIN COMPUTER TECHNOLOGY CO., LTD.) 23 July 2021 (2021-07-23) description, paragraphs 51-116	5-10, 15-20, 25
A	CN 111654519 A (HUAWEI TECHNOLOGIES CO., LTD.) 11 September 2020 (2020-09-11) entire document	1-25
A	US 2020183854 A1 (IBM) 11 June 2020 (2020-06-11) entire document	1-25

☐ 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
01 September 2023	**13 September 2023**

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/CN2023/101171

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	113568562	A	29 October 2021	WO	2021218038	A1	04 November 2021
				EP	3958107	A1	23 February 2022
				US	2022057954	A1	24 February 2022
				CN	114610232	A	10 June 2022
				CN	114860163	A	05 August 2022
				EP	3958107	A4	17 August 2022
				JP	2022539950	A	14 September 2022
CN	112165505	A	01 January 2021	CN	112165505	B	19 July 2022
CN	113157611	A	23 July 2021	CN	113157611	B	18 April 2023
CN	111654519	A	11 September 2020	None			
US	2020183854	A1	11 June 2020	US	11734192	B2	22 August 2023

国际检索报告

国际申请号
PCT/CN2023/101171

A. 主题的分类

G06F 15/163(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:G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNABS;CNTXT;EPTXT;VEN;WOTXT;USTXT;CNKI;IEEE:节点，多个，其他，第二，互联，加入，接入，插入，全局，地址，分配，分发，对应，端口，转发，标识，老化，退出，删除，node，multiple，other，another，second，interconnection，add，access，insert，global，address，allocation，distribute，corresponding，port，retransmission，ID，identification，aging，exit，delete

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 113568562 A (华为技术有限公司) 2021年10月29日 (2021-10-29) 说明书第55-120段	1-4、9-14、19-25
Y	CN 113568562 A (华为技术有限公司) 2021年10月29日 (2021-10-29) 说明书第55-120段	5-10、15-20、25
Y	CN 112165505 A (杭州安恒信息技术股份有限公司) 2021年1月1日 (2021-01-01) 说明书第33-72段	5-10、15-20、25
Y	CN 113157611 A (山东英信计算机技术有限公司) 2021年7月23日 (2021-07-23) 说明书第51-116段	5-10、15-20、25
A	CN 111654519 A (华为技术有限公司) 2020年9月11日 (2020-09-11) 全文	1-25
A	US 2020183854 A1 (IBM) 2020年6月11日 (2020-06-11) 全文	1-25

☐ 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
2023年9月1日	2023年9月13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	赵静 电话号码 (+86) 020-28958139

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/101171

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	113568562	A	2021年10月29日	WO	2021218038	A1	2021年11月4日
				EP	3958107	A1	2022年2月23日
				US	2022057954	A1	2022年2月24日
				CN	114610232	A	2022年6月10日
				CN	114860163	A	2022年8月5日
				EP	3958107	A4	2022年8月17日
				JP	2022539950	A	2022年9月14日
CN	112165505	A	2021年1月1日	CN	112165505	B	2022年7月19日
CN	113157611	A	2021年7月23日	CN	113157611	B	2023年4月18日
CN	111654519	A	2020年9月11日		无		
US	2020183854	A1	2020年6月11日	US	11734192	B2	2023年8月22日