



(12) 发明专利申请

(10) 申请公布号 CN 102906726 A

(43) 申请公布日 2013.01.30

(21) 申请号 201180005166.7

(22) 申请日 2011.12.09

(85) PCT申请进入国家阶段日

2012.07.12

(86) PCT申请的申请数据

PCT/CN2011/083770 2011.12.09

(71) 申请人 华为技术有限公司

地址 518129 中国广东省深圳市龙岗区坂田  
华为总部办公楼

(72) 发明人 章晓峰 方帆 秦岭

(51) Int. Cl.

G06F 15/16 (2006.01)

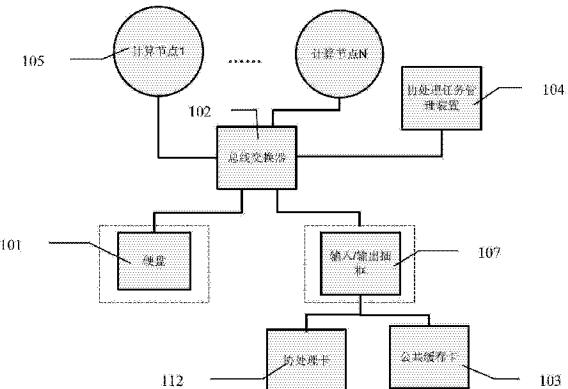
权利要求书 3 页 说明书 13 页 附图 4 页

(54) 发明名称

协处理加速方法、装置及系统

(57) 摘要

本发明实施例公开一种协处理加速方法，包括：接收计算机系统中计算节点发出的携带有待处理数据地址信息的协处理请求消息，根据该协处理请求消息获得待处理数据，并将该待处理数据存储到公共缓存卡中；将存储在公共缓存卡中的待处理数据分配到计算机系统中的空闲的协处理卡进行处理。相应地，本发明实施例还公开了一种协处理任务管理装置、加速管理板卡及计算机系统，通过以上技术方案，使用添加的公共缓存卡来作为计算机系统硬盘和各个协处理卡之间的公共数据缓冲通道，待处理数据不必通过计算节点的内存来中转，避免了数据从计算节点内存传输的开销，从而突破了内存延迟、带宽的瓶颈，提高了协处理速度。



1. 一种计算机系统,包括:至少一个计算节点、总线交换器和至少一个协处理卡,其特征在于,所述系统还包括:公共缓存卡和协处理任务管理装置,所述公共缓存卡为所述计算机系统中各个计算节点和各个协处理卡之间的数据传输提供临时存储;所述公共缓存卡和所述至少一个协处理卡通过所述总线交换器互连;

所述计算节点用于,发送协处理请求消息,所述协处理请求消息携带有待处理数据的地址信息;所述待处理数据为所述计算节点请求处理的数据;

所述协处理任务管理装置用于:

接收所述协处理请求消息;

根据所述协处理请求消息携带的待处理数据的地址信息获得所述待处理数据,并将所述待处理数据存储到所述公共缓存卡中;

将存储到所述公共缓存卡中的待处理数据分配到所述计算机系统中的空闲的协处理卡进行处理。

2. 如权利要求1所述的计算机系统,其特征在于,所述协处理任务管理装置还用于,在将所述待处理数据存储到所述公共缓存卡之前,在所述公共缓存卡内申请存储空间,所述存储空间用于存储所述待处理数据。

3. 如权利要求1或2所述的计算机系统,其特征在于,所述协处理任务管理装置还用于,将所述空闲的协处理卡处理完成的数据存储到所述协处理请求消息指定的目的地址。

4. 如权利要求3所述的计算机系统,其特征在于,所述计算节点还用于,从所述协处理请求消息指定的目的地址中获取所述空闲的协处理卡处理完成的数据。

5. 如权利要求1-4任一项所述的计算机系统,其特征在于,所述协处理任务管理装置还用于,在将存储到所述公共缓存卡中的待处理数据分配到所述计算机系统中的空闲的协处理卡进行处理之后,将所述待处理数据从所述公共缓存卡中擦除。

6. 如权利要求1-5任一项所述的计算机系统,其特征在于,所述协处理任务管理装置具体用于:

接收多个所述协处理请求消息;

根据各个协处理请求消息获得对应的待处理数据,并将各个待处理数据存储到所述公共缓存卡中;

从各个协处理请求消息中获得所述各个协处理请求消息的请求优先级和请求类型;

根据所述各个协处理请求消息的请求优先级和请求类型,确定各个协处理请求消息对应的待处理数据的处理顺序;

将所述各个待处理数据,按照所述处理顺序依次从所述公共缓存卡分配到所述计算机系统中的空闲的协处理卡进行处理。

7. 如权利要求1-6任一项所述的计算机系统,其特征在于:所述公共缓存卡为快速外设组件互连PCIE缓存卡,其存储介质为闪存固态硬盘flash SSD、相变换存储固态硬盘PCM SSD或动态存储器DRAM。

8. 如权利要求1-7任一项所述的计算机系统,其特征在于,所述协处理卡为PCIE协处理卡。

9. 如权利要求8所述的计算机系统,其特征在于,所述PCIE协处理卡为图形处理器GPU加速卡。

10. 如权利要求 1-9 任一项所述的计算机系统,其特征在于,所述公共缓存卡和所述至少一个协处理卡通过 PCIE 接口与所述总线交换器互连。

11. 如权利要求 1-10 任一项所述的计算机系统,其特征在于,所述协处理任务管理装置根据所述协处理请求消息,从所述计算机系统的硬盘中获得待处理数据。

12. 如权利要求 1-10 任一项所述的计算机系统,其特征在于,所述协处理任务管理装置采用直接内存存取 DMA 方式将所述待处理数据存储到所述公共缓存卡中。

13. 一种协处理加速方法,包括 :

接收计算机系统中计算节点发出的至少一个协处理请求消息,所述协处理请求消息携带有待处理数据的地址信息;所述待处理数据为所述计算节点请求处理的数据;

根据所述协处理请求消息携带的待处理数据的地址信息,获得所述待处理数据,并将所述待处理数据存储到公共缓存卡中;所述待处理数据为所述协处理请求消息请求处理的数据;

将存储在所述公共缓存卡中的待处理数据分配到所述计算机系统中的空闲的协处理卡进行处理。

14. 如权利要求 13 所述的方法,其特征在于,在将所述待处理数据存储到公共缓存卡之前,所述方法还包括:在所述公共缓存卡内申请存储空间,所述存储空间用于存储所述待处理数据。

15. 如权利要求 13 或 14 所述的方法,其特征在于,所述协处理请求消息为多个;所述将存储在所述公共缓存卡中的待处理数据分配到所述计算机系统中的空闲的协处理卡进行处理,具体包括 :

从各个协处理请求消息中获得所述各个协处理请求消息的请求优先级和请求类型;

根据所述各个协处理请求消息的请求优先级和请求类型,确定各个协处理请求消息对应的待处理数据的处理顺序;

将所述各个协处理请求消息对应的待处理数据,按照所述处理顺序依次从所述公共缓存卡分配到所述计算机系统中的空闲的协处理卡进行处理。

16. 如权利要求 13-15 任一项所述的方法,其特征在于,所述方法还包括:

将所述空闲的协处理卡处理完成的数据存储到到所述协处理请求消息指定的目的地址。

17. 如权利要求 13-16 任一项所述的方法,其特征在于,在将存储到所述公共缓存卡中的待处理数据分配到所述计算机系统中的空闲的协处理卡进行处理之后,所述方法还包括,将所述待处理数据从所述公共缓存卡中擦除。

18. 如权利要求 13-17 任一项所述的方法,其特征在于,所述将所述待处理数据存储到公共缓存卡中,具体包括 :

将所述待处理数据采用 DMA 方式存储到公共缓存卡中。

19. 如权利要求 13-18 任一项所述的方法,其特征在于,所述公共缓存卡为 PCIE 缓存卡。

20. 如权利要求 13-19 任一项所述的方法,其特征在于,所述协处理卡为 PCIE 协处理卡。

21. 如权利要求 20 所述的方法,其特征在于,所述 PCIE 协处理卡为 GPU 加速卡。

22. 一种协处理任务管理装置,其特征在于,包括:

消息接收模块,用于接收计算机系统中计算节点发出的至少一个协处理请求消息,所述协处理请求消息携带有待处理数据的地址信息;所述待处理数据为所述计算节点请求处理的数据;

第一数据传送模块,用于根据所述协处理请求消息携带的待处理数据的地址信息,获得所述待处理数据,并将所述待处理数据存储到公共缓存卡中;所述待处理数据为所述协处理请求消息请求处理的数据;

第二数据传送模块,用于将存储在所述公共缓存卡中的待处理数据分配到所述计算机系统中的空闲的协处理卡进行处理。

23. 如权利要求 22 所述的装置,其特征在于,所述第二数据传送模块还用于,

将所述空闲的协处理卡处理完成的数据存储到所述协处理请求消息指定的目的地址。

24. 如权利要求 22 或 23 所述的装置,其特征在于,所述装置还包括:

缓存管理模块,用于在将所述待处理数据存储到公共缓存卡之前,在所述公共缓存卡中申请存储空间,所述存储空间用于缓存所述待处理数据。

25. 如权利要求 22-24 任一项所述的装置,其特征在于,所述第二数据传送模块具体包括:

获取单元,用于在所述协处理请求消息为多个时,从各个协处理请求消息中获得所述各个协处理请求消息的请求优先级和请求类型;

请求顺序确定单元,用于根据所述各个协处理请求消息的请求优先级和请求类型,确定各个协处理请求消息对应的待处理数据处理顺序;

数据处理单元,用于将所述各个协处理请求消息对应的待处理数据,按照所述处理顺序依次从所述公共缓存卡分配到所述计算机系统中的空闲的协处理卡进行处理。

26. 如权利要求 22-25 任一项所述的装置,其特征在于,所述第一数据传送模块通过 DMA 方式将所述待处理数据存储到公共缓存卡中。

27. 一种加速管理板卡,其特征在于,包括:控制器和 PCIE 接口单元;所述控制器通过所述 PCIE 接口单元与计算机系统的总线交换器数据连接;所述控制器用于接收所述计算机系统的中央处理器 CPU 发出的至少一个协处理请求消息,所述协处理请求消息携带有待处理数据的地址信息;所述待处理数据为所述 CPU 请求处理的数据;并根据所述协处理请求消息携带的待处理数据的地址信息,从所述计算机系统中的硬盘获取所述待处理数据,将所述待处理数据存储到公共缓存单元中;

所述控制器还用于,将存储在所述公共缓存单元中的待处理数据分配到所述计算机系统中的空闲的 GPU 加速卡进行处理,所述 GPU 加速卡通过自身的第一 PCIE 接口和所述计算机系统的总线交换器连接。

28. 如权利要求 27 所述的加速管理板卡,其特征在于,所述公共缓存单元位于所述加速管理板卡内部。

29. 如权利要求 27 所述的加速管理板卡,其特征在于,所述公共缓存单元位于所述加速管理板卡外部,所述公共缓存单元为 PCIE 缓存卡,所述 PCIE 缓存卡包含第二 PCIE 接口,所述 PCIE 缓存卡通过所述第二 PCIE 接口与所述连 PICE 接口单元连接。

## 协处理加速方法、装置及系统

### 技术领域

[0001] 本发明涉及计算机领域,尤其涉及一种协处理加速方法、装置及系统。

### 背景技术

[0002] 随着计算机技术的发展,计算机被应用到越来越广阔的领域。除了日常生活中常见的计算办公应用之外,计算机还被应用到一些非常复杂的领域,例如大型科学计算、海量数据处理等,它们通常对计算机的处理能力有更高的要求。然而单个计算机的处理能力毕竟有限,在上述这些大规模的运算场景下,容易成为系统性能提升的瓶颈,集群系统的出现有效解决了这一问题,所谓集群系统,就是通过高速互联网络连接起来的多台自治计算机和相关资源组成的高性能系统,其中每一台自治计算机称为一个计算节点。在集群中,由于每个计算节点的CPU(central processing unit,中央处理器)是作为通用计算设备来设计的,在某些特定应用领域,例如图像处理、音频处理等,处理效率往往不高,所以出现了很多协处理器,例如网络协处理器,GPU(Graphics processing unit,图形处理器),压缩协处理器等,这些协处理器可以辅助计算节点进行任务处理,即协处理;协处理器辅助计算节点处理的任务称为协处理任务。在大型计算机系统海量计算场景下,如何利用协处理器辅助计算节点进行协处理,将直接关系到计算机系统的工作效率。

[0003] 现有技术中,协处理器大都以PCIE(Peripheral Component Interconnect Express,快速外设组件互连)协处理卡的方式添加在计算机系统当中,并由计算机系统的计算节点控制协处理器进行协处理任务的处理,同时利用计算节点的内存作为协处理器和计算节点数据传输的通道,以中转待处理数据和经协处理器处理完毕的数据。

[0004] 采用现有技术中的这种架构,所有待处理数据必须通过计算节点的内存来中转,增加了内存开销,而由于内存带宽、延迟等因素的限制,协处理速度不高。

### 发明内容

[0005] 本发明的实施例提供一种计算机系统、协处理加速方法、协处理任务管理装置及加速管理板卡,用于减少计算机系统的内存开销和提高计算机系统中协处理器的协处理速度。

[0006] 本发明实施例提供一种计算机系统,包括:至少一个计算节点、总线交换器和至少一个协处理卡,所述计算机系统还包括:公共缓存卡和协处理任务管理装置,所述公共缓存卡为所述计算机系统中各个计算节点和各个协处理卡之间的数据传输提供临时存储;所述公共缓存卡和所述至少一个协处理卡通过所述总线交换器互连;

[0007] 所述计算节点用于,发送协处理请求消息;

[0008] 所述协处理任务管理装置用于:接收所述协处理请求消息,所述协处理请求消息携带有待处理数据的地址信息;所述待处理数据为所述计算节点请求处理的数据;根据所述协处理请求消息携带的待处理数据的地址信息,获得所述待处理数据,并将所述待处理数据存储到所述公共缓存卡中;将存储到所述公共缓存卡中的待处理数据分配到所述计算

机系统中的空闲的协处理卡进行处理。

[0009] 本发明实施例提供一种协处理加速方法，包括：

[0010] 接收计算机系统中计算节点发出的至少一个协处理请求消息，所述协处理请求消息携带有待处理数据的地址信息；所述待处理数据为所述计算节点请求处理的数据；

[0011] 根据所述协处理请求消息携带有待处理数据的地址信息，获得所述待处理数据，并将所述待处理数据存储到公共缓存卡中；所述待处理数据为所述协处理请求消息请求处理的数据；

[0012] 将存储在所述公共缓存卡中的待处理数据分配到所述计算机系统中的空闲的协处理卡进行处理。

[0013] 本发明的实施例提供一种协处理任务管理装置，包括：

[0014] 消息接收模块，用于接收计算机系统中计算节点发出的至少一个协处理请求消息，所述协处理请求消息携带有待处理数据的地址信息；所述待处理数据为所述计算节点请求处理的数据；

[0015] 第一数据传送模块，用于根据所述协处理请求消息携带有待处理数据的地址信息，获得所述待处理数据，并将所述待处理数据存储到公共缓存卡中；所述待处理数据为所述协处理请求消息请求处理的数据；

[0016] 第二数据传送模块，用于将存储在所述公共缓存卡中的待处理数据分配到所述计算机系统中的空闲的协处理卡进行处理。

[0017] 本发明实施例还提供一种加速管理板卡，包括：控制器和 PCIE 接口单元；所述控制器通过所述 PCIE 接口单元与计算机系统的总线交换器数据连接；所述控制器用于接收所述计算机系统的中央处理器 CPU 发出的至少一个协处理请求消息，所述协处理请求消息携带有待处理数据的地址信息，所述待处理数据为所述 CPU 请求处理的数据；并根据所述协处理请求消息携带有待处理数据的地址信息，从所述计算机系统中的硬盘获取所述待处理数据，将所述待处理数据存储到公共缓存单元中；

[0018] 所述控制器还用于，将存储在所述公共缓存单元中的待处理数据分配到所述计算机系统中的空闲的 GPU 加速卡进行处理，所述 GPU 加速卡通过自身的第一 PCIE 接口和所述计算机系统的总线交换器连接。

[0019] 本发明实施例通过以上技术方案，使用公共缓存卡来作为计算机系统的各个计算节点和各个协处理卡之间的公共数据缓冲通道，待处理数据不必通过计算节点的内存来中转，避免了待处理数据从计算节点内存传输的开销，突破了内存延迟、带宽的瓶颈，提高了待处理数据的协处理速度。

## 附图说明

[0020] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0021] 图 1 为现有技术中一种协处理系统的架构图；

[0022] 图 2 为本发明实施例一提供的一种协处理加速方法的流程图；

- [0023] 图 3 为本发明实施例二提供的一种协处理加速方法的流程图；
- [0024] 图 4 为本发明实施例三提供的一种协处理任务管理装置示意图；
- [0025] 图 5 为本发明实施例三的第二数据传送模块示意图；
- [0026] 图 6 为本发明实施例四提供的一种计算机系统结构图；
- [0027] 图 7 为本发明实施例五提供的一种加速管理板卡示意图。

## 具体实施方式

[0028] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0029] 为使本领域一般技术人员更好的了解本发明实施例提供的技术方案，对现有技术中的协处理系统及协处理方案做一个介绍。

[0030] 如图 1 所示，现有技术方案中，协处理卡通过 PCIE 接口安置在输入 / 输出插框上，帮助计算节点完成协处理任务，输入 / 输出插框通过 PCIE 总线交换器与计算节点数据连接。步骤 1，计算节点 1 将数据从硬盘复制到计算节点 1 的内存中；步骤 2，计算节点 1 利用 DMA(Direct Memory Access, 直接内存存取) 技术将数据从计算节点 1 的内存复制到协处理卡的内存进行处理；步骤 3，计算节点 1 利用 DMA 将处理完的数据从协处理卡内存复制到计算节点 1 的内存；步骤 4，计算节点 1 在该数据上进一步处理或将该数据重新保存到硬盘中。

[0031] 本发明实施例提供的技术方案可以应用于多处理器架构的大型计算设备、云计算、CRAN(cloud radio access net, 云无线接入网) 业务等多种海量计算场景。如图 2 所示，本发明实施例一提供一种协处理加速方法，用于提高计算机系统中协处理的速度，根据图 2，该方法包括：

[0032] S101，接收计算机系统中计算节点发出的至少一个协处理请求消息，所述协处理请求消息携带有待处理数据的地址信息；

[0033] 需要说明的是，所述待处理数据为所述计算节点通过所述协处理消息请求处理的数据；本发明所有实施例中关于待处理数据的解释均如此。

[0034] 具体地，在计算机系统中，存在至少一个计算节点和至少一个协处理卡，协处理卡可以辅助计算节点进行任务处理，即协处理；当计算节点需要协处理卡辅助进行任务处理时，就发出协处理请求消息。在一个实施例中，协处理请求消息可以为一个包含若干字段的数据报文。

[0035] 在一个实施例中，协处理请求消息具体包括但不限于以下信息：

[0036] 1. 请求计算节点标志；

[0037] 在一个计算机系统中，存在至少一个计算节点，请求计算节点标志用于标示和区分发起服务请求的计算节点；具体地，可以为计算机系统中的每个计算节点分配一个唯一的 ID 号，当某个计算节点发出协处理请求消息时，就将该计算节点的 ID 号作为请求计算节点标志。

[0038] 2. 请求类型；

[0039] 请求类型用于表示计算节点请求的协处理类型,常见的协处理类型有:图形处理类、浮点运算类、网络类、Hash 运算类等,具体地,可以用协处理请求消息中的一个字段来表示请求类型,例如,请求类型字段为 graphic 表示图形处理类、请求类型字段为 float 表示浮点运算类、请求类型字段为 net 表示网络类、请求类型字段为 Hash 表示 Hash 运算类。需要说明的是,在计算机系统中,可以配置一种或多种类型的协处理卡,因此可以允许的请求类型需要根据当前计算机系统中配置的协处理卡的种类来确定。例如,在一个实施例中,系统可以只配置一种类型的协处理卡,比如 GPU 加速卡,这时请求类型就只有图形处理类;在另一个实施例中,系统可以同时配置多种类型的协处理卡,比如浮点运算协处理卡、Hash 运算协处理卡、网络协处理卡以及 GPU 加速卡等,这时请求类型就对应有浮点运算类、Hash 运算类、网络类和图形处理类等,本发明实施例不做特别地限定。

[0040] 3. 待处理数据的地址信息;

[0041] 一个实施例中,待处理数据的地址信息可以包括源地址和待处理数据的长度;

[0042] 其中,源地址表示等待协处理卡处理的数据(即,待处理数据)所在存储空间的首地址。在一个实施例中,源地址可以是计算机系统的非易失性存储器中的某一地址,进一步地,所述非易失性存储器可以为硬盘或 flash(闪存),需要说明的是,所述硬盘具体可以包括磁盘类硬盘及固态类硬盘(如 flash SSD、PCMSSD)。

[0043] 待处理数据长度表示待处理数据所需的存储空间的大小;

[0044] 4. 目的地址;

[0045] 目的地址为经过协处理卡处理完成的数据最终的存储地址。在一个实施例中,目的地址可以是计算机系统的硬盘中的某一地址,例如硬盘中的某一地址。需要说明的是,所述硬盘具体可以包括磁盘类硬盘及固态类硬盘(如 flash SSD、PCMSSD)

[0046] 5. 请求优先级;

[0047] 请求优先级是由计算节点根据协处理任务的性质、紧迫程度或来源等来指定的,在一个实施例中,请求优先级可以分成高、中、低三个级别,当然可以理解的是,在另一个实施例中,优先级还可以细分成更多的级别,比如极高、高、普通、正常、较低、极低等,也可以是用阿拉伯数字 1、2、3.... 代表的优先级别,本实施例不做特别地限定。

[0048] 在一个实施例中,请求计算节点标志、请求类型、源地址、待处理数据长度、目的地址和请求优先级这些信息可以分别以字段的形式添加到协处理请求消息中,上述字段共同组成一个协处理请求消息。

[0049] S102,根据所述协处理请求消息携带的待处理数据的地址信息,获得待处理数据,并将所述待处理数据存储到公共缓存卡中;

[0050] 需要说明的是,所述公共缓存卡为所述计算机系统中各个计算节点和各协处理卡之间的数据传输提供临时缓存;

[0051] 具体地,在一个实施例中,可以根据协处理请求消息携带的地址信息,从计算机系统的硬盘中获取待处理数据。

[0052] 在一个实施例中,协处理请求消息中的地址信息包括:源地址和待处理数据长度。具体地,根据协处理请求消息中的源地址和待处理数据长度这两个字段的信息来获取待处理数据,所述待处理数据具体指存放在计算机系统的硬盘中等待协处理卡进行处理的原始数据。由于协处理请求消息中的源地址字段表示待处理数据在计算机系统的硬盘中的首地

址,因此,计算机系统的硬盘中从源地址开始,大小为待处理数据长度的连续地址空间内的数据即为待处理数据;将该待处理数据存储到公共缓存卡中。

[0053] 在一个实施例中,将待处理数据存储到公共缓存卡中,可以采用复制或迁移的方式。

[0054] 具体地,可以通过 DMA 方式实现上述复制或迁移操作。具体地,在进行数据复制或迁移之前,待处理数据所在存储器的 I/O 接口先向发出 DMA 请求指令,向计算机系统的总线逻辑控制器提出总线请求,当计算机系统中的计算节点执行完当前总线周期内的指令并释放总线控制权后,总线逻辑控制器输出总线应答,表示 DMA 已经响应,并将总线控制权交给 DMA 控制器,DMA 控制器获得总线控制权后,通知待复制数据所在存储器的 I/O 接口开始 DMA 传输,然后输出读写命令,直接控制数据的传输,整个数据传输过程不需要计算机系统中计算节点的参与,有效节省了系统的资源。

[0055] 需要说明的是,所述硬盘具体可以包括磁盘类硬盘及固态类硬盘(如 flash SSD、PCMSSD)。

[0056] 需要说明的是,公共缓存卡是添加在计算机系统中,作为各个计算节点和各个协处理卡进行数据传输的公共临时存储,它不同于协处理卡的缓存,如 GPU 加速卡的缓存,公共缓存卡是供计算机系统中所有协处理卡共用的缓存区,作为计算机系统硬盘和所有协处理卡传输数据的缓冲通道,公共缓存卡可以是具有快速存取能力的任何存储介质。在一个实施例中,公共缓存卡可以为 PCIE 公共缓存卡,其存储介质为 Flash SSD(Solid State Storage, 固态硬盘)、PCM SSD 或 DRAM(动态存储器)等。

[0057] S103,将存储在所述公共缓存卡中的待处理数据分配到所述计算机系统中的空闲的协处理卡进行处理。

[0058] 需要说明的是,所述空闲的协处理卡,可以是当前没有协处理任务的协处理卡;也可以是根据负载均衡策略,选择出的负载较小或相对空闲的协处理卡,例如,可以将当前 CPU 利用率最低的协处理卡作为空闲的协处理卡。

[0059] 具体地,在一个实施例中,根据协处理请求消息中的请求类型以及与该请求类型匹配的各协处理卡的利用率,来判断是否有与协处理请求消息中的请求类型匹配的空闲的协处理卡,若有匹配的空闲的协处理卡,则将公共缓存卡中的待处理数据分配给该空闲处理器进行处理。例如,在一个实施例中,若某一计算节点请求图形协处理服务,则通过系统函数调用来获取当前计算机系统中所有 GPU 加速卡的 CPU 利用率,若某个 GPU 加速卡的 CPU 的利用率小于 5%,即可判定该 GPU 加速卡处于空闲状态,然后将待处理数据从公共缓存卡复制或迁移到该 GPU 加速卡的存储器进行处理;当然可以理解的是,在另一个实施例中,若某一计算节点请求其他类型的协处理服务,如浮点运算类,则应判断是否有浮点运算协处理卡空闲,这里不再赘述。

[0060] 进一步地,为了根据优先级对多个协处理请求排序,使优先级高的协处理请求最先得到处理,使协处理卡得到更加合理的利用,在另一个实施例中,S103 可以具体包括如下步骤:

[0061] (1) 从各个协处理请求消息中获得所述各个协处理请求消息的请求优先级和请求类型;

[0062] (2) 根据所述各个协处理请求消息的请求优先级和请求类型,确定各个协处理请

求消息的处理顺序；

[0063] 具体地，确定各协处理请求消息的处理顺序的方法是：将请求类型不同的协处理请求消息放入不同的消息队列，请求类型相同的协处理请求消息在对应的消息队列中按照请求优先级从高到低的顺序排队，请求优先级相同且请求类型相同的协处理请求消息，在对应的消息队列中按照请求的先后顺序排队；与请求类型匹配的空闲的协处理卡将按照对应任务队列中的先后顺序处理待处理数据。

[0064] (3) 将所述各个协处理请求消息对应的待处理数据，按照所述处理顺序依次从所述公共缓存卡分配到所述计算机系统中的空闲的协处理卡进行处理。

[0065] 需要说明的是，将待处理数据从公共缓存卡分配到空闲的协处理卡进行处理的具体方法之前已经作了详细说明，此处不再赘述。

[0066] 本发明实施例一通过以上技术方案，根据计算机系统中各个计算节点的发送的协处理请求消息，将各个计算节点请求处理的待处理数据分配给系统中的空闲的协处理卡进行处理，计算节点不需要消耗自身的资源来进行待处理数据的分配，减少了各个计算节点自身的资源开销；并使用公共缓存卡来作为计算机系统的各个计算节点和各个协处理卡之间的公共数据缓冲通道，待处理数据不必通过计算节点的内存来中转，避免了待处理数据从计算节点内存传输的开销，突破了内存延迟、带宽的瓶颈，提高了待处理数据的协处理速度。

[0067] 实施例二，本发明实施例提供一种协处理加速方法，用于提高计算机系统中协处理的速度，如图 3 所示，该方法包括：

[0068] S201，接收计算机系统中计算节点发出的至少一个协处理请求消息；

[0069] 在一个实施例中，每个协处理消息携带有该协处理消息对应的待处理数据（即，计算节点通过该协处理消息请求处理的待处理数据）的地址信息；

[0070] 具体地，在计算机系统中，存在至少一个计算节点和至少一个协处理卡，协处理卡可以辅助计算节点进行任务处理，即协处理；当计算节点需要协处理卡辅助进行任务处理时，就发出协处理请求消息。在一个实施例中，协处理请求消息可以为一个包含若干字段的数据报文。

[0071] 在一个实施例中，协处理请求消息具体包括但不限于以下信息：

[0072] 1. 请求计算节点标志；

[0073] 在一个计算机系统中，存在至少一个计算节点，请求计算节点标志用于标示和区分发起服务请求的计算节点；具体地，可以为计算机系统中的每个计算节点分配一个唯一的 ID 号，当某个计算节点发出协处理请求消息时，就将该计算节点的 ID 号作为请求计算节点标志。

[0074] 2. 请求类型；

[0075] 请求类型用于表示计算节点请求的协处理类型，常见的协处理类型有：图形处理类、浮点运算类、网络类、Hash 运算类等，具体地，可以用协处理请求消息中的一个字段来表示请求类型，例如，请求类型字段为 graphic 表示图形处理类、请求类型字段为 float 表示浮点运算类、请求类型字段为 net 表示网络类、请求类型字段为 Hash 表示 Hash 运算类。需要说明的是，在计算机系统中，可以配置一种或多种类型的协处理卡，因此可以允许的请求类型需要根据当前计算机系统中配置的协处理卡的种类来确定。例如，在一个实施例中，系

统可以只配置一种类型的协处理卡,比如 GPU 加速卡,这时请求类型就只有图形处理类;在另一个实施例中,系统可以同时配置多种类型的协处理卡,比如浮点运算协处理卡、Hash 运算协处理卡、网络协处理卡以及 GPU 加速卡等,这时请求类型就对应有浮点运算类、Hash 运算类、网络类和图形处理类等,本发明实施例不做特别地限定。

[0076] 3. 待处理数据的地址信息;

[0077] 一个实施例中,待处理数据的地址信息可以包括源地址和待处理数据的长度;

[0078] 其中,源地址表示等待协处理卡处理的数据(即,待处理数据)所在存储空间的首地址。在一个实施例中,源地址可以是计算机系统的非易失性存储器中的某一地址,进一步地,所述非易失性存储器可以为硬盘或 flash(闪存),需要说明的是,所述硬盘具体可以包括磁盘类硬盘及固态类硬盘(如 flash SSD、PCMSSD)。

[0079] 待处理数据长度表示待处理数据所需的存储空间的大小;

[0080] 4. 目的地址;

[0081] 目的地址为经过协处理卡处理完成的数据最终的存储地址。在一个实施例中,目的地址可以是计算机系统的硬盘中的某一地址,例如硬盘中的某一地址。需要说明的是,所述硬盘具体可以包括磁盘类硬盘及固态类硬盘(如 flash SSD、PCMSSD)

[0082] 5. 请求优先级;

[0083] 请求优先级是由计算节点根据协处理任务的性质、紧迫程度或来源等来指定的,在一个实施例中,请求优先级可以分成高、中、低三个级别,当然可以理解的是,在另一个实施例中,优先级还可以细分成更多的级别,比如极高、高、普通、正常、较低、极低等,也可以是用阿拉伯数字 1、2、3.... 代表的优先级别,本实施例不做特别地限定。

[0084] 在一个实施例中,请求计算节点标志、请求类型、源地址、待处理数据长度、目的地址和请求优先级这些信息可以分别以字段的形式添加到协处理请求消息中,上述字段共同组成一个协处理请求消息。

[0085] S202,在公共缓存卡内申请存储空间,以缓存待处理数据;所述公共缓存卡设置在所述计算机系统中,为所述计算机系统中各个计算节点和协处理卡之间的数据传输提供临时存储;

[0086] 具体地,根据协处理请求消息携带的待处理数据的地址信息中的待处理数据长度字段,向公共缓存卡申请与待处理数据长度对应大小的内存空间,用于缓存待处理数据。

[0087] S203,根据所述协处理请求消息携带的待处理数据的地址信息,获得待处理数据,并将所述待处理数据存储到上述公共缓存卡中申请的存储空间;

[0088] 具体地,在一个实施例中,可以根据协处理请求消息携带的地址信息,从计算机系统的硬盘中获取待处理数据;

[0089] 在一个实施例中,协处理请求消息中的地址信息包括:源地址和待处理数据长度。具体地,根据协处理请求消息中的源地址和待处理数据长度这两个字段的信息来获取待处理数据,所述待处理数据具体指存放在计算机系统的硬盘中等待协处理卡进行处理的原始数据。由于协处理请求消息中的源地址字段表示待处理数据在计算机系统的硬盘中的首地址,因此,计算机系统的硬盘中从源地址开始,大小为待处理数据长度的连续地址空间内的数据即为待处理数据。需要说明的是,所述硬盘具体可以包括磁盘类硬盘及固态类硬盘(如 flash SSD、PCMSSD)

[0090] 在一个实施例中,将待处理数据存储到公共缓存卡,可以采用复制或迁移的方式。

[0091] S204,将存储在所述公共缓存卡中的待处理数据分配到所述计算机系统中的空闲的协处理卡进行处理;

[0092] 需要说明的是,所述空闲的协处理卡,可以是当前没有协处理任务的协处理卡;也可以是根据负载均衡策略,选择出的负载较小、相对空闲的的协处理卡,例如,可以将当前CPU利用率最低的协处理卡作为空闲的协处理卡。

[0093] 具体地,在一个实施例中,根据协处理请求消息中的请求类型以及与该请求类型匹配的各协处理卡的利用率,来判断是否有与协处理请求消息中的请求类型匹配的空闲的协处理卡,若有匹配的空闲的协处理卡,则将公共缓存卡中的待处理数据分配给该空闲处理器进行处理。例如,在一个实施例中,若某一计算节点请求图形协处理服务,则通过系统函数调用来获取当前计算机系统中所有GPU加速卡的CPU的利用率,若某个GPU加速卡的CPU的利用率小于5%,即可判定该GPU加速卡处于空闲状态,然后将待处理数据从公共缓存卡复制或迁移到该GPU加速卡的存储器进行处理;当然可以理解的是,在另一个实施例中,若某一计算节点请求其他类型的协处理服务,如浮点运算类,则应判断是否有浮点运算协处理卡空闲,这里不再赘述。

[0094] 进一步地,为了根据优先级对多个协处理请求排序,使优先级高的协处理请求最先得到处理,使协处理卡得到更加合理的利用,在另一个实施例中,S204可以具体包括如下步骤:

[0095] (1) 从各个协处理请求消息中获得所述各个协处理请求消息的请求优先级和请求类型;

[0096] (2) 根据所述各个协处理请求消息的请求优先级和请求类型,确定各个协处理请求消息的处理顺序;

[0097] 具体地,确定各协处理请求消息的处理顺序的方法是:将请求类型不同的协处理请求消息放入不同的消息队列,请求类型相同的协处理请求消息在对应的消息队列中按照请求优先级从高到低的顺序排队,请求优先级相同且请求类型相同的协处理请求消息,在对应的消息队列中按照请求的先后顺序排队;与请求类型匹配的空闲的协处理卡将按照对应任务队列中的先后顺序处理待处理数据。

[0098] (3) 将所述各个协处理请求消息对应的待处理数据,按照所述处理顺序依次从所述公共缓存卡分配到所述计算机系统中的空闲的协处理卡进行处理。

[0099] 进一步的,在将所述待处理数据从所述公共缓存卡分配到所述计算机系统中的空闲的协处理卡进行处理之后,本发明实施例二提供的协处理加速方法还包括:

[0100] S205,将所述待处理数据从所述公共缓存卡中擦除;

[0101] S206,将所述空闲的协处理卡处理完成的数据存储到到所述协处理请求消息指定的目的地址;

[0102] 需要说明的是,所述目的地址,为协处理请求消息中携带的目的地址,它表示经过协处理卡处理完成的数据最终的存储地址。

[0103] 进一步的,将空闲的协处理卡处理完成的数据存储到到所述协处理请求消息指定的目的地址之后,本发明实施例二提供的协处理加速方法还包括:

[0104] S207,根据协处理请求消息中的请求计算节点标志向发起协处理请求的计算节点

发送服务请求完成消息。

[0105] 在一个实施例中，服务请求完成消息可以为一个包含特定含义字段的数据报文，所述报文包含的特定字段可以为“finish”、“ok”或“yes”，用于表示当前的协处理任务已经完成。

[0106] 本发明实施例二通过以上技术方案，根据计算机系统中各个计算节点的发送的协处理请求消息，将各个计算节点请求处理的待处理数据分配给系统中的空闲的协处理卡进行处理，计算节点不需要消耗自身的资源来进行待处理数据的分配，减少了各个计算节点自身的资源开销；并使用公共缓存卡来作为计算机系统的各个计算节点和各个协处理卡之间的公共数据缓冲通道，待处理数据不必通过计算节点的内存来中转，避免了待处理数据从计算节点内存传输的开销，突破了内存延迟、带宽的瓶颈，提高了待处理数据的协处理速度。

[0107] 实施例三，本发明实施例提供一种协处理任务管理装置，用于统一管理计算机系统中的协处理任务，如图 4 所示，该协处理任务管理装置包括：

[0108] 消息接收模块 420，用于接收计算机系统中计算节点发出的至少一个协处理请求消息；所述协处理请求消息携带有待处理数据的地址信息；

[0109] 具体地，在计算机系统中，如果计算节点需要协处理卡处理待处理数据，就发送协处理请求消息给消息接收模块 420，消息接收模块 420 接收计算节点发出的协处理请求消息，该协处理请求消息包含的内容与本发明实施例一的 S101 所述的协处理请求消息内容完全一致，本实施例不再赘述。

[0110] 在另一个实施例中，消息接收模块 420 还用于，在协处理卡将数据处理完成之后，根据协处理请求消息中的请求计算节点标志向发起协处理请求的计算节点发送服务请求完成消息。

[0111] 具体地，在协处理卡将数据处理完成之后，消息接收模块 420 根据该协处理请求消息中的请求计算节点标志向发起该协处理请求的计算节点发送服务请求完成消息。在一个实施例中，服务请求完成消息可以为一个包含特定含义字段的数据报文，所述报文包含的特定字段可以为“finish”、“OK”或“yes”，用于表示当前的协处理任务已经完成。

[0112] 第一数据传送模块 430，用于根据所述协处理请求消息携带的待处理数据的地址信息，获得待处理数据，并将所述待处理数据存储到公共缓存卡中；

[0113] 具体地，在一个实施例中，第一数据传送模块 430 可以根据协处理请求消息携带的地址信息，从计算机系统的硬盘中获取待处理数据；在一个实施例中，协处理请求消息中的地址信息包括：源地址和待处理数据长度。具体地，第一数据传送模块 430 根据协处理请求消息中的源地址和待处理数据长度这两个字段信息来获取待处理数据，所述待处理数据具体指存放在计算机系统的硬盘中等待协处理卡进行处理的原始数据；由于协处理请求消息中的源地址字段表示待处理数据在计算机系统的硬盘中的首地址，因此，计算机系统的硬盘中从源地址开始，大小为待处理数据长度的连续地址空间内的数据即为待处理数据。

[0114] 需要说明的是，所述硬盘具体可以包括磁盘类硬盘及固态类硬盘（如 flashSSD、PCMSSD）。

[0115] 需要说明的是，公共缓存卡是添加在计算机系统中，作为各个计算节点和协处理卡进行数据传输的临时存储，它不同于协处理卡的缓存，如 GPU 加速卡的缓存，公共缓存卡

是供计算机系统中所有协处理卡共用的缓存区，作为计算机系统的硬盘和所有协处理卡传输数据的缓冲通道，公共缓存卡可以是具有快速存取能力的任何存储介质。在一个实施例中，公共缓存卡可以为 PCIE 公共缓存卡，其存储介质为 Flash SSD、PCM SSD 或 DRAM 等。

[0116] 第二数据传送模块 440，用于将存储在所述公共缓存卡中的待处理数据分配到所述计算机系统中的空闲的协处理卡进行处理。

[0117] 需要说明的是，所述空闲的协处理卡，可以是当前没有协处理任务的协处理卡；也可以是根据负载均衡策略，选择出的负载较小、相对空闲的的协处理卡，例如，可以将当前 CPU 利用率最低的协处理卡作为空闲的协处理卡。

[0118] 具体地，在一个实施例中，第二数据传送模块 440 根据协处理请求消息中的请求类型以及与该请求类型匹配的各协处理卡的利用率，来判断是否有与协处理请求消息中的请求类型匹配的空闲的协处理卡，若有匹配的空闲的协处理卡，则第二数据传送模块 440 将公共缓存卡中的待处理数据分配给该空闲处理器进行处理。例如，在一个实施例中，若某一计算节点请求图形协处理服务，则第二数据传送模块 440 通过系统函数调用来获取当前计算机系统中所有 GPU 加速卡的 CPU 的利用率，若某个 GPU 加速卡的 CPU 的利用率小于 5%，即可判定该 GPU 加速卡处于空闲状态，然后将待处理数据从公共缓存卡复制或迁移到该 GPU 加速卡的存储器进行处理；当然可以理解的是，在另一个实施例中，若某一计算节点请求其他类型的协处理服务，如浮点运算类，则应判断是否有浮点运算协处理卡空闲，这里不再赘述。

[0119] 进一步地，在另一个实施例中，第二数据传送模块 440 还可以用于将协处理卡处理完成的数据存储到协处理请求消息指定的目的地址。

[0120] 在一个实施例中，如图 5 所示，当所述协处理请求消息为多个时，为了根据优先级对多个协处理请求排序，使优先级高的协处理请求最先得到处理，使协处理卡得到更加合理的利用，第二数据传送模块可以具体包括：

[0121] 获取单元 4401，用于从各个协处理请求消息中获得所述各个协处理请求消息的请求优先级和请求类型；

[0122] 请求顺序确定单元 4402，用于根据所述各个协处理请求消息的请求优先级和请求类型，确定各协处理请求消息处理顺序；

[0123] 在一个实施例中，请求顺序确定单元 4402 确定各协处理请求消息的处理顺序的方法是：将请求类型不同的协处理请求消息放入不同的消息队列，请求类型相同的协处理请求消息在对应的消息队列中按照请求优先级从高到低的顺序排队，请求优先级相同且请求类型相同的协处理请求消息，在对应的消息队列中按照请求的先后顺序排队；与请求类型匹配的空闲的协处理卡将按照对应任务队列中的先后顺序处理待处理数据。

[0124] 数据处理单元 4403，用于将所述各个协处理请求消息对应的待处理数据，按照所述处理顺序依次从所述公共缓存卡分配到所述计算机系统中的空闲的协处理卡进行处理。

[0125] 在一个实施例中，第一数据传送模块 430 可以采用复制或迁移的方式将待处理数据存储到公共缓存卡中，第二数据传送模块 440 可以采用复制或迁移的方式将协处理卡处理完成的数据存储到协处理请求消息指定的目的地址。进一步地，第一数据传送模块 430 和第二数据传送模块 440 可以通过 DMA 方式实现数据在计算节点硬盘、公共缓存卡以及协处理卡之间的复制或迁移。具体地，以第一数据传送模块 430 为例，在进行数据复制或迁移

之前,待复制数据所在存储器的 I/O 接口先向第一数据传送模块 430 发出 DMA 请求指令,第一数据传送模块 430 根据 DMA 请求指令,向计算机系统的总线逻辑控制器提出总线请求,当计算机系统中的计算节点执行完当前总线周期内的指令并释放总线控制权后,总线逻辑控制器输出总线应答,表示 DMA 已经响应,并将总线控制权交给第一数据传送模块 430,第一数据传送模块 430 获得总线控制权后,通知待复制数据所在存储器的 I/O 接口开始 DMA 传输,然后输出读写命令,直接控制数据的传输,整个数据传输过程不需要计算机系统中计算节点的参与,有效节省了系统的资源。

[0126] 第二数据传送模块 440 的具体工作可以参照本发明实施例一 S103。

[0127] 进一步的,为了便于公共缓存卡存储空间的管理,本发明实施例三提供的协处理任务管理装置还包括:

[0128] 缓存管理模块 450,用于在第一数据传送模块 430 将所述待处理数据存储到公共缓存卡之前,在所述公共缓存卡中申请存储空间,所述存储空间用于缓存所述待处理数据。

[0129] 本发明实施例三通过以上技术方案,通过协处理任务管理装置将计算机系统中各个计算节点的协处理任务以协处理请求消息进行统一处理,计算节点不需要消耗自身的资源来进行待处理数据的分配,减少了各个计算节点的资源开销;同时,以添加的公共缓存卡来作为计算机系统硬盘和各个协处理卡之间公共的数据缓冲通道,实现数据的复制或迁移,避免了待处理数据从计算节点内存传输的开销,从而突破了内存延迟、带宽的瓶颈,提高了待处理数据的协处理速度;进一步地,在向公共缓存卡复制数据之前,利用缓存管理模块在公共缓存卡中申请空间,使公共缓存卡空间的管理更加方便;进一步地,通过任务优先级管理模块,使优先级高的协处理请求最先得到处理,使协处理卡得到更加合理的利用,提高了协处理的效率。

[0130] 实施例四,如图 6 所示,本发明实施例四提供一种计算机系统,包括:

[0131] 硬盘 101、总线交换器 102、公共缓存卡 103、协处理任务管理装置 104、至少一个计算节点(例如,图 6 中的计算节点 105)和至少一个协处理卡(例如,图 6 中的协处理卡 112);协处理卡 112、硬盘 101 和公共缓存卡 103 与总线交换器 102 数据连接,总线交换器 102 使协处理卡 112、硬盘 101 和公共缓存卡 103 互连;至少一个计算节点 105 用于发送协处理请求消息,所述协处理请求消息携带有待处理数据的地址信息,所述待处理数据为所述计算节点 105 请求处理的数据;

[0132] 协处理任务管理装置 104 用于:接收所述协处理请求消息;根据所述协处理请求消息携带的待处理数据的地址信息获得待处理数据,并将所述待处理数据存储到公共缓存卡 103 中;所述待处理数据为所述协处理请求消息求处理的数据;将存储到公共缓存卡 103 中的待处理数据分配到所述计算机系统中至少一个协处理卡中空闲的的一个协处理卡(假设图 6 中的协处理卡 112 空闲)进行处理。

[0133] 在一个实施例中,所述计算机系统还包括硬盘 101,协处理任务管理装置 104 根据所述协处理请求消息,从硬盘 101 中获得待处理数据。需要说明的是,硬盘 101 具体可以为磁盘类硬盘或固态类硬盘(如 flash SSD、PCMSSD)。

[0134] 进一步的,为了便于缓存卡存储空间的管理,在一个实施例中,协处理任务管理装置 104 还用于,在将待处理数据存储到公共缓存卡 103 之前,在公共缓存卡 103 内申请存储空间,该存储空间用于存储所述待处理数据。在另一个实施例中,协处理任务管理装置 104

还用于将公共缓存卡 103 中的待处理数据分配到协处理卡 112 进行处理之后, 将所述待处理数据从公共缓存卡 103 中擦除。

[0135] 在另一个实施例中, 协处理任务管理装置 104 还用于, 将协处理卡 112 处理完成的数据存储到所述协处理请求消息指定的目的地址; 相应地, 至少一个计算节点 105 还用于, 从所述目的地址中获取协处理卡 112 处理完成的数据。

[0136] 在一个实施例中, 协处理任务管理装置 104 可以采用复制或迁移的方式将待处理数据存储到公共缓存卡 103 中, 也可以采用复制或迁移的方式将协处理卡 112 处理完成的数据存储到所述协处理请求消息指定的目的地址。进一步地, 可以通过 DMA 方式实现数据上述复制或迁移操作。

[0137] 在一个实施例中, 上述公共缓存卡 103 可以为 PCIE 缓存卡, 其存储介质可以为 flash SSD、PCM SSD 或 DRAM。

[0138] 在一个实施例中, 协处理卡 112、硬盘 101 和公共缓存卡 103 均可以通过 PCIE 总线直接与总线交换器 102 连接。

[0139] 在另一个实施例中, 如图 6 所示, 协处理卡 112 和公共缓存卡 103 通过输入 / 输出插框与总线交换器 102 连接。具体地, 协处理卡 112 和公共缓存卡 103 插置于输入 / 输出插框 107 的 PCIE 插槽中, 输入 / 输出插框 107 通过 PCIE 总线和总线交换器 102 连接。

[0140] PCIE 作为最新的总线接口标准, 与其他总线接口标准相比, 具有较高数据传输速率, 因此采用 PCIE 总线进行数据连接, 可以提高数据在硬盘、协处理卡和公共缓存卡之间的传输速度, 进一步提高计算机系统的协处理速度。

[0141] 当然可以理解的是, 实际应用中的另一个实施例中, 协处理卡 112、硬盘 101 和公共缓存卡 103 也可以通过 AGP 总线与总线交换器 102 连接。本发明实施例并不做特别地限定。

[0142] 需要说明的是, 本发明实施例四提供的计算机系统包括一个协处理卡 112 和一个计算节点 105 仅仅是一个举例, 故不应理解为对本发明实施例四提供的计算机系统的计算节点和协处理卡数量的限定。可以理解的是, 在一个实施例中, 计算节点和协处理卡的数量可以是大于零的任意整数值, 但是实际应用中出于节约成本的考虑, 协处理卡的数量应该不多于计算机系统中计算节点的个数, 比如说, 当前协处理装置中包含有 20 个计算节点, 则协处理卡的数量可以为 1、5、10、15 或者 20 等。

[0143] 进一步地, 在一个实施例中, 协处理卡的类型可以只有一种, 比如 GPU 加速卡; 也可以有多种, 例如浮点运算协处理卡、Hash 运算协处理卡、网络协处理卡以及 GPU 加速卡等。当然可以理解的是, 计算机系统中包含的协处理卡类型越多, 整个系统能够支持的协处理任务类型就会越多, 协处理功能就会越强大。

[0144] 本发明实施例四通过以上技术方案, 通过协处理任务管理装置对计算机系统中的协处理任务进行统一管理, 减少了各个计算节点的资源开销; 同时计算机系统中的多个协处理卡可以共用公共缓存卡来作为硬盘和协处理卡之间的数据缓冲通道, 并通过协处理任务管理装置实现数据的复制或迁移, 避免了数据从计算节点内存传输的开销, 从而突破了内存延迟、带宽的瓶颈, 提高了协处理的速度; 进一步地, 通过 PCIE 总线来连接计算机系统中的协处理卡、硬盘、公共缓存卡和总线交换器, 可以有效提高数据的传输速率, 进而提高了协处理的速度。

[0145] 实施例五,如图 7 所示,本发明实施例五提供一种加速管理板卡,用于提高计算机系统的协处理速度,包括:控制器 710 和 PCIE 接口单元 720;控制器 710 和 PCIE 接口单元 720 数据连接;控制器 710 接收计算机系统中计算节点的 CPU 发出的至少一个协处理请求消息,该协处理请求消息携带有待处理数据的地址信息,并根据所述待处理数据的地址信息,从所述计算机系统中的硬盘获取待处理数据,将所述待处理数据存储到公共缓存单元中;所述待处理数据为所述 CPU 请求处理的数据;

[0146] 控制器 710 还用于,将存储在所述公共缓存单元中的待处理数据分配到计算机系统中的空闲的 GPU 加速卡进行处理;具体地点,如图 7 所示,GPU 加速卡 80 通过自身的第一 PCIE 接口 810 和加速管理板卡 70 的 PCIE 接口单元 720 数据连接。

[0147] 在一个实施例中,公共缓存单元可以集成在所述加速管理板卡内部,如图 7 所示,公共缓存单元 730 通过加速管理板卡 710 上的总线与控制器 710 连接,具体地,所述加速板卡上的总线可以为 PCIE 总线。

[0148] 在另一个实施例中,公共缓存单元也可以设置于所述加速管理板卡外部,作为一个单独的物理实体;进一步的公共缓存单元可以为 PCIE 缓存卡。具体地,如图 7 所示,PCIE 缓存卡 90 包含有第二 PCIE 接口 910,PCIE 缓存卡 90 通过其第二 PCIE 接口 910 与加速管理板卡 70 的 PCIE 接口单元 720 连接。

[0149] 需要说明的是,PCIE 作为最新的总线接口标准,与其他总线接口标准相比,具有较高数据传输速率,因此上述实施例中采用 PCIE 接口作为 GPU 加速卡和控制器,以及控制器和公共缓存单元数据连接的接口,只是为取得最佳技术效果的一种举例,故不应理解为对本发明实施例的限制。

[0150] 本发明实施例通过以上技术方案,通过单独的控制器对计算机系统中的协处理任务进行统一管理,减少了各个计算节点的资源开销;同时计算机系统中的多个协处理卡可以共用公共缓存单元来作为硬盘和协处理卡之间的数据缓冲通道,避免了数据从计算节点内存传输的开销,从而突破了内存延迟、带宽的瓶颈,提高了协处理的速度。

[0151] 以上所述仅为本发明的几个实施例,本领域的技术人员依据申请文件公开的可以对本发明进行各种改动或变型而不脱离本发明的精神和范围。

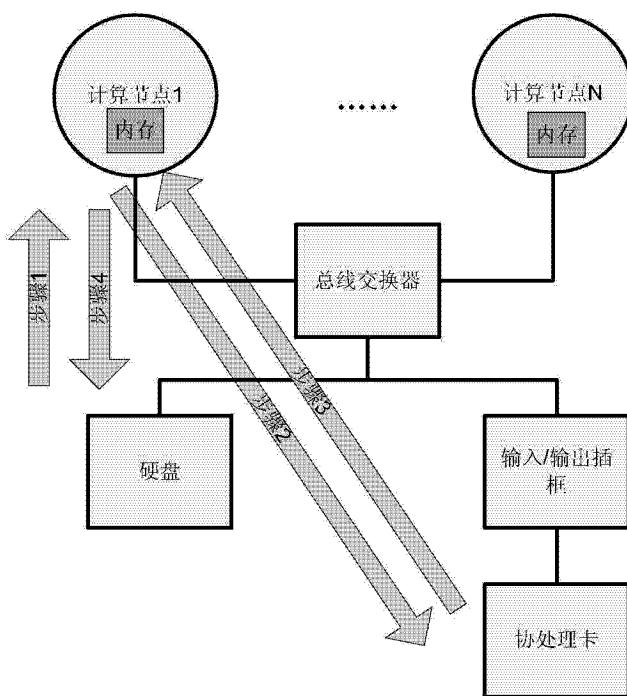


图 1

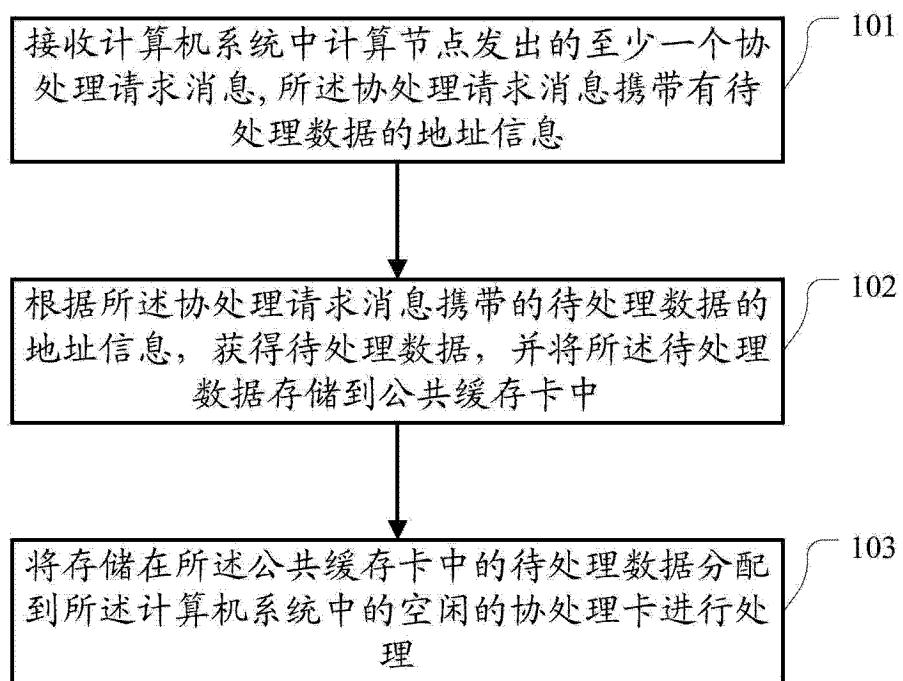


图 2

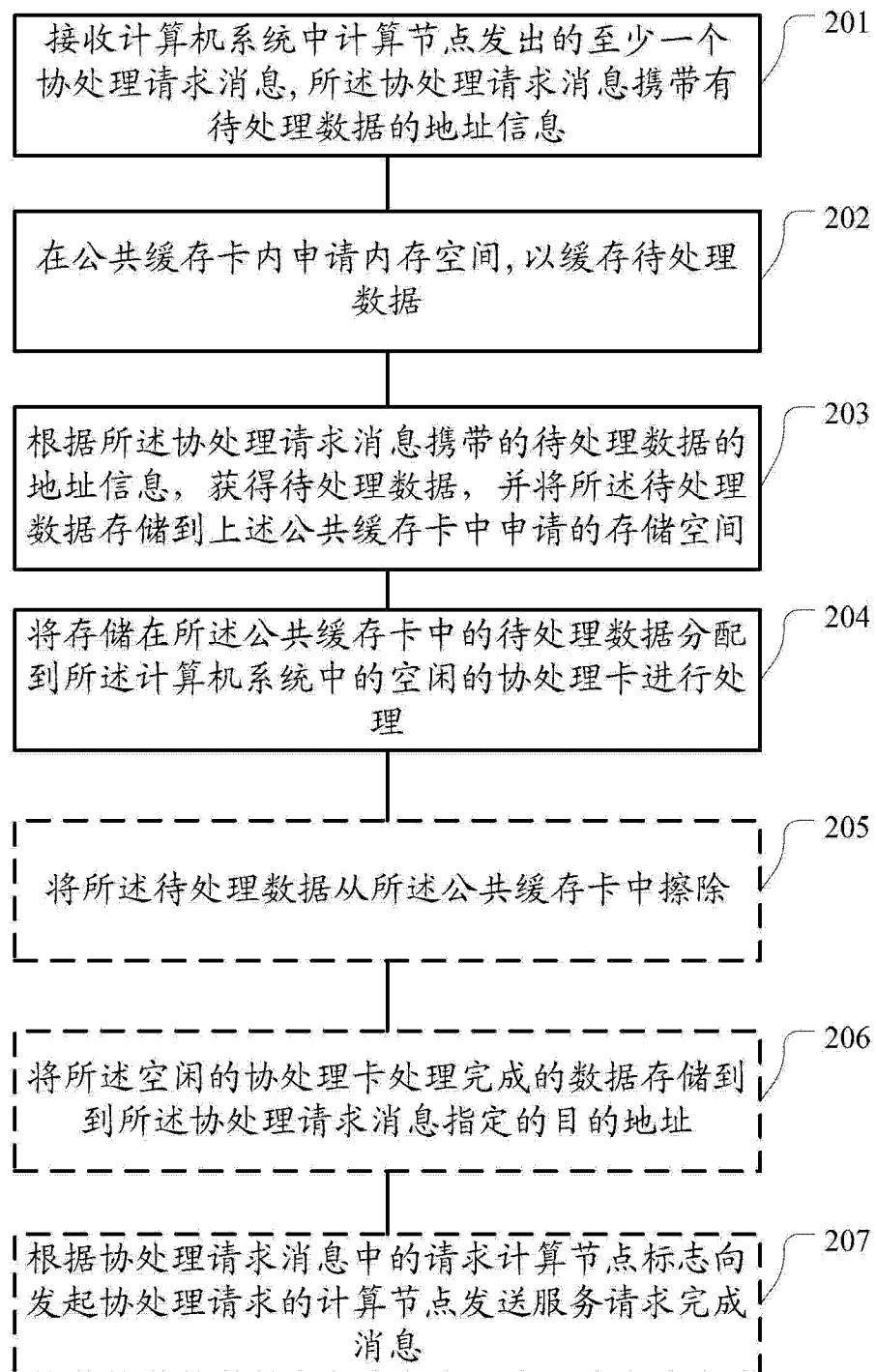


图 3

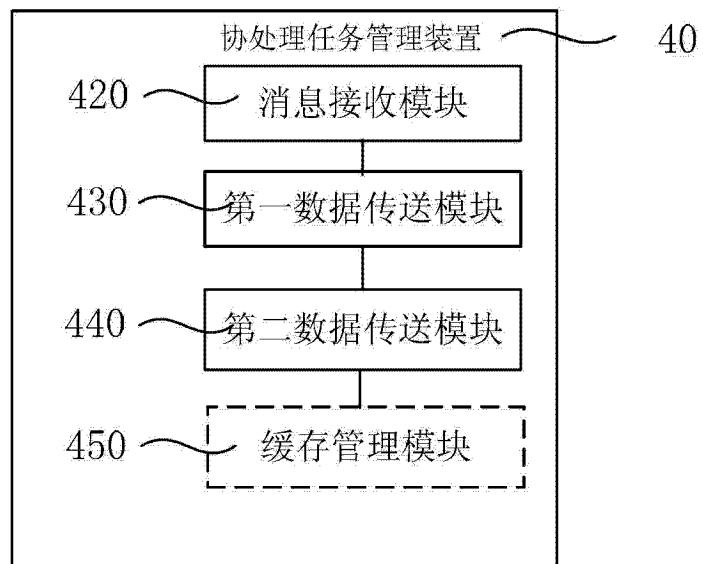


图 4

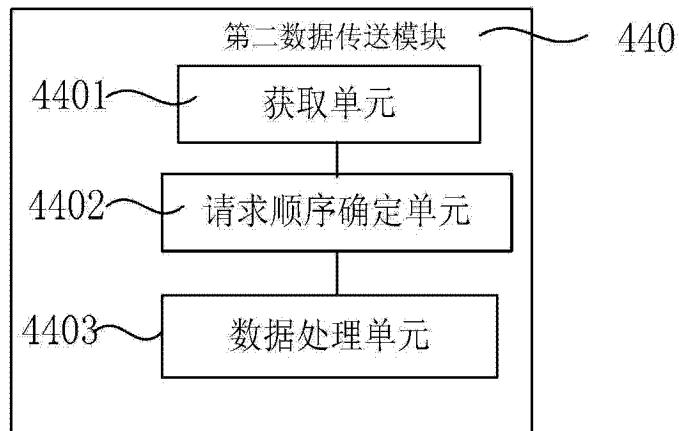


图 5

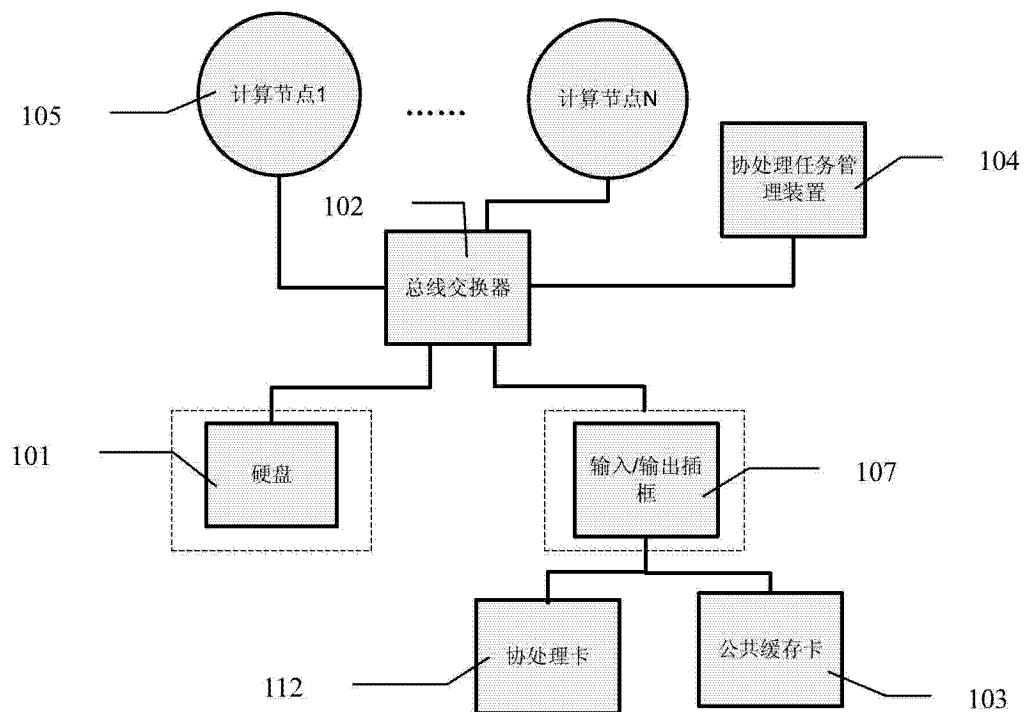


图 6

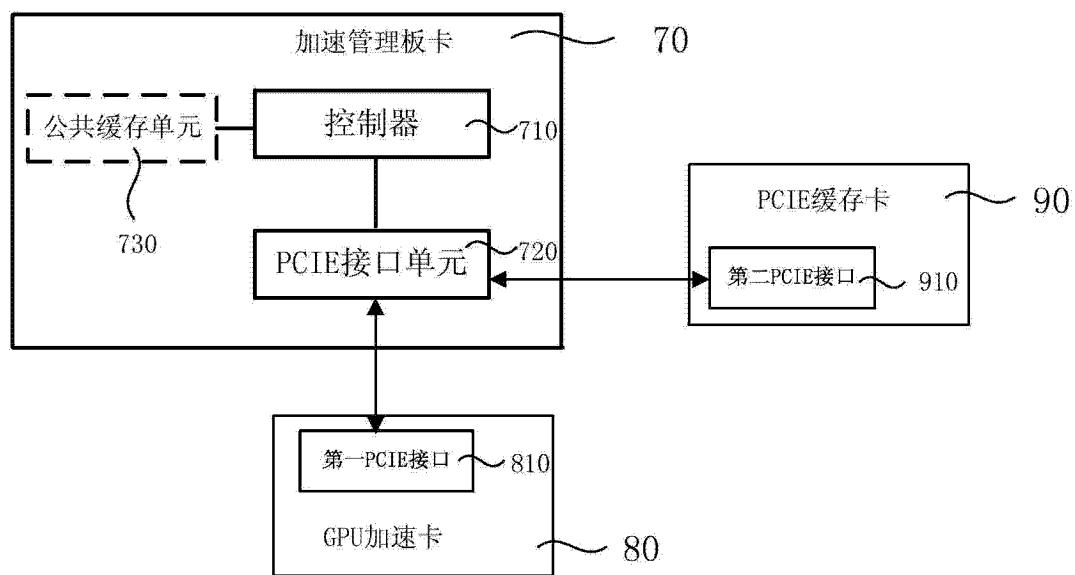


图 7