(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111095202 A
(43)申请公布日 2020.05.01

(21)申请号 201780094429.3
(22)申请日 2017.09.30
(85)PCT国际申请进入国家阶段日
2020.02.28
(86)PCT国际申请的申请数据
PCT/US2017/054663 2017.09.30
(87)PCT国际申请的公布数据
WO2019/066981 EN 2019.04.04
(71)申请人 英特尔公司
地址 美国加利福尼亚州
(72)发明人 K·瓦德雅纳坦 S·斯瑞哈兰 D·达斯
(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 黄嵩泉 何焜
(51)Int.Cl.
G06F 9/38(2006.01)

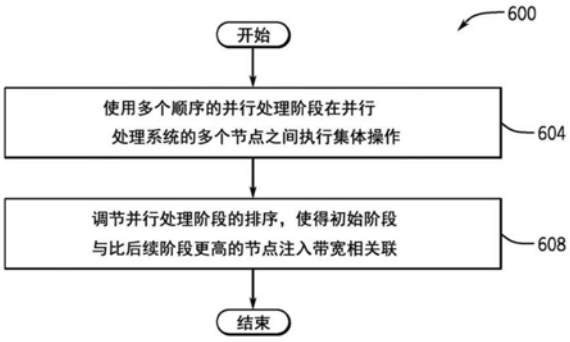
权利要求书2页 说明书6页 附图7页

(54)发明名称

基于注入节点带宽的并行处理

(57)摘要

技术包括使用多个并行处理阶段在并行处理计算机系统的多个节点之间执行集体操作。该技术包括调节并行处理阶段的排序,使得多个并行处理阶段的初始阶段与比多个并行处理阶段的后续阶段更高的节点注入带宽相关联。



CN 111095202 A

1.一种计算机实现的方法,包括:

使用多个并行处理阶段在并行处理系统的多个节点之间执行集体操作;以及

调节所述并行处理阶段的排序,其中所述多个并行处理阶段的初始阶段与比所述多个并行处理阶段的后续阶段更高的节点注入带宽相关联。

2.如权利要求1所述的方法,其中:

执行所述集体操作包括在所述多个节点之间传递消息;以及

调节所述排序包括调节所述排序使得与所述初始阶段相关联的消息尺寸大于与另一阶段相关联的消息尺寸。

3.如权利要求1所述的方法,其中执行所述集体操作包括执行归约分散操作。

4.如权利要求1所述的方法,其中执行所述集体操作包括:在所述多个节点之间并行处理数据向量的元素以对所述元素进行归约,并将经归约的元素跨所述多个节点分散。

5.如权利要求1所述的方法,进一步包括:

对于多个并行处理阶段的初始阶段,将多个消息从所述多个节点中的第一节点传递到所述多个节点中的另一节点以将数据从所述另一节点传递到所述第一节点,并在所述第一节点中处理所传递的数据以对所述所传递的数据进行归约操作。

6.如权利要求1所述的方法,其中所述多个节点包括节点的集群,所述方法进一步包括:

在所述初始阶段中在每个集群的节点之间传递消息;以及

在所述后续阶段中在所述集群之间传递消息。

7.如权利要求1所述的方法,其中所述多个节点包括布置在超级节点中的节点的子集,所述方法进一步包括:

在所述初始阶段中在每个超级节点的节点之间传递消息;以及

在所述后续阶段中在所述超级节点之间传递消息。

8.如权利要求1所述的方法,其中所述多个节点包括布置在超级节点中的节点的子集,以及布置在网格中的超级节点的子集,所述方法进一步包括:

在所述初始阶段中在每个超级节点的节点之间传递消息;

在所述多个并行处理阶段的第二阶段中,在每个网格的超级节点之间传递消息;以及

在所述多个并行处理阶段的第三阶段中,在所述网格之间传递消息。

9.如权利要求1所述的方法,其中所述多个节点包括布置在超级节点中的节点的子集,以及布置在网格中的超级节点的子集,所述方法进一步包括:

在所述初始阶段中在每个超级节点的节点之间传递消息;

在所述多个并行处理阶段的第二阶段中,在每个网格的超级节点之间传递消息;以及

在所述多个并行处理阶段的多个其他阶段中,在所述网格之间传递消息。

10.如权利要求9所述的方法,其中在所述多个并行处理阶段的多个其他阶段中在所述网格之间传递消息包括根据基于Rabenseifner的算法进行传递。

11.一种用于存储指令的非暂态计算机可读存储介质,所述指令在由并行处理机器执行时,使得所述机器:

在多个并行处理阶段的每个阶段中,在所述机器的多个处理节点之间传递消息,以交换和对数据进行归约,其中每个处理阶段与注入带宽相关联,并且所述注入带宽不同;以及

对阶段进行排序，使得所述多个并行处理阶段的初始阶段与相关联的注入带宽中的最高注入带宽相关联。

12.如权利要求11所述的计算机可读存储介质，其中所述计算机可读存储介质存储指令，所述指令在由所述并行处理机器执行时使得所述机器提供消息接口库，所述消息接口库提供允许对阶段进行排序的函数，并且其中所述初始阶段与所述最高注入带宽相关联。

13.如权利要求11所述的计算机可读存储介质，其中所述计算机可读存储介质存储指令，所述指令在由所述并行处理机器执行时使得所述机器根据相关联的注入带宽对阶段进行排序，以便在与相对较低的注入带宽相关联的阶段之前执行与相对较高的注入带宽相关联的阶段。

14.如权利要求11所述的计算机可读存储介质，其中：

所述多个处理节点包括布置在超级节点中节点的子集；

布置在网格中的超级节点的子集；以及

所述计算机可读存储介质存储指令，所述指令在由所述并行处理机器执行时使得每个超级节点的节点彼此通信以在所述初始阶段对数据进行归约，使每个网格的超级节点彼此通信以在所述多个并行处理阶段的第二阶段中对数据进行归约，并且使所述网格彼此通信以在所述多个并行处理阶段中的至少另一个第三阶段中对数据进行归约。

15.一种系统，包括：

多个处理网格，以对第一数据集执行归约分散并行处理操作，其中：

每个网格包括多个超级节点；以及

每个超级节点包括多个计算机处理节点；以及

用于将归约分散并行处理操作分成多个并行处理阶段的协调器，所述多个并行处理阶段包括第一阶段、第二阶段和至少一个附加阶段，

其中：

在所述初始阶段，每个超级节点的计算机处理节点彼此传递消息以对所述第一数据集进行归约，以提供第二数据集；

在所述第二阶段中，每个网格的超级节点彼此传递消息以对所述第二数据集进行归约，以产生第三数据集；以及

在所述至少一个附加的阶段，所述网格彼此传递消息，以进一步对所述第三数据集进行归约。

16.如权利要求15所述的系统，其中所述协调器包括消息传递接口(MPI)。

17.如权利要求15所述的系统，其中所述计算机处理节点包括多个处理核。

18.如权利要求15所述的系统，其中在所述初始阶段，给定超级节点的给定计算机处理节点与给定超级节点的另一个计算机处理节点传递多个消息。

19.如权利要求18所述的系统，其中，所述至少一个附加阶段包括第三阶段，并且在所述第三阶段，每个网格将单个消息与另一个网格进行通信。

20.如权利要求15所述的系统，其中所述计算机处理节点包括服务器刀片。

基于注入节点带宽的并行处理

背景技术

[0001] 并行计算系统可包括多个硬件处理节点，诸如中央处理单元(CPU)、图形处理单元(GPU)等。通常，给定节点独立于并行计算系统的其他节点执行其处理。

[0002] 为并行处理系统编写的给定应用可包括集体操作，其中节点彼此通信以交换数据。一种类型的集体操作是归约分散(reduce-scatter)操作，其中输入数据可以在一系列的并行处理时段(phase)或阶段(stage)中进行处理。以该方式，每个处理节点可以以代表输入数据向量的一部分的数据向量或数组开始操作；并且在每个阶段中，成对的处理节点可以交换一半的数据，并将数据合并(例如，将数据加在一起)以对数据进行归约。以该方式，集体处理将最初存储在每个节点上的数据数组归约为表示集体操作的结果的最终数据数组，并且该最终数据数组可以在处理节点上分布或分散。

附图说明

[0003] 图1是根据示例实施方式的并行处理计算机系统的示意图。

[0004] 图2是根据示例实施方式的节点执行环境的图示。

[0005] 图3是根据示例实施方式的并行处理计算机系统用于执行集体操作的处理阶段的图示。

[0006] 图4A是根据示例实施方式的存储在处理网格的处理节点上的初始数据数组的图示。

[0007] 图4B是根据示例实施方式的由处理节点阶段应用的归约的图示。

[0008] 图5是示出根据示例实施方式的服务器系统的示意图。

[0009] 图6是描绘根据示例实施方式的在并行处理系统中执行集体操作的技术的流程图。

具体实施方式

[0010] 并行计算机系统可包括并行处理节点，在集体并行处理操作中，并行处理节点可以使用消息收发来交换数据。例如，并行计算机系统可以执行称为“归约分散操作”的集体操作，其中处理节点出于交换和对所交换的数据进行归约操作的目的而使用消息收发进行通信。

[0011] 例如，处理节点最初可以存储输入数据的集合的一部分，该输入数据经受归约分散操作。例如，每个处理节点最初可以存储索引的输入数据数组，诸如例如包括从1到8索引的块的数据数组。在归约分散操作中，处理节点可以通过消息收发来交换其数据并应用归约操作。例如，归约分散可以是数学上的加法，并且由归约分散操作产生的输出数据数组可以是例如八元素数据数组，其中第一元素是输入数据数组的全部的第一元素的总和，输出数据数组的第二元素是输入数据数组的第二元素的总和等。此外，在归约分散操作结束时，输出数据数组的元素跨处理节点平均分散或分布。例如，在归约分散操作结束时，一个处理节点可以存储输出数据数组的第一元素，另一个处理节点可以存储输出数据数组的第二元

素，依此类推。

[0012] 并行处理系统执行集体操作的一种方式是将处理划分为一系列并行处理时段或阶段；在阶段中，成对的处理节点交换其一半数据(成对的一个节点从成对的另一个节点接收一半的数据，反之亦然)并对所交换的数据进行归约。以该方式，在给定的阶段中，给定的成对处理节点中的第一处理节点可以接收存储在该对的第二处理节点上的数据的一半，将所接收的数据与存储在第一处理节点上的数据的一半结合(例如，相加)，并将结果存储在第一处理节点上。该对的第二处理节点转而可以在同一给定阶段中，接收存储在第一节点上的数据的一半，将所接收的数据与存储在第二节点上的数据的一半结合，并且将所得的归约的数据存储在第二处理节点上。处理在一个或多个后续阶段中继续进行，在这些阶段中，处理节点交换其数据的一些(例如，其数据的一半)，对数据进行归约并存储所得的归约的数据，直到每个处理节点存储所得的输出数据数组的元素为止。

[0013] 根据本文所述的示例实施方式，选择用于集体操作的节点的配对，使得初始并行处理阶段具有相关联的节点注入带宽，该相关联的节点注入带宽高于任何后续并行处理阶段的节点注入带宽。在本上下文中，“节点注入带宽”是指给定处理节点可用于与其他节点进行数据通信的带宽。作为示例，给定的处理阶段可包括节点交换数据，这些节点通过多个网络链路连接，使得每个处理节点可以同时与多个其他处理节点交换数据。更具体地，如本文中进一步描述的，对于一些阶段，每个处理节点能够与多个处理节点同时交换数据，而对于其他阶段，每个处理节点可以与单个其他处理节点交换数据。

[0014] 根据示例实施方式，给定的处理阶段可包括超级节点的处理节点交换数据，这允许超级节点的每个节点同时与多个其他处理节点交换数据。在本文中，“超级节点”是指处理节点的组或集合，它们可以通过比其他处理节点使用的带宽链路更高的带宽链路交换数据。以该方式，根据示例实施方式，给定超级节点的处理节点可以在一个并行处理时段或阶段(例如，初始阶段)期间在超级节点内交换数据，并且随后，给定的超级节点可以与另一个并行处理阶段(例如，第二阶段)期间的另一个超级节点交换数据。

[0015] 因为，如本文所述，集体操作构造并行处理阶段，使得初始阶段与最高注入带宽相关联，所以可以显著减少执行集体操作的总时间。以该方式，集体操作更快，因为最大数量的数据是通过最高带宽链路进行通信。这种减少的处理时间对于深度学习尤其有利，因为它应用于人工智能和机器学习领域，如图像识别、自主驾驶和自然语言处理。

[0016] 作为更具体的示例，图1描绘了根据一些实施方式的并行处理计算机系统100。通常，计算机系统100包括多个并行计算机处理节点102(作为示例，在图1中描绘了P个处理节点102-1、102-2、102-3……102-P-1)，这些处理节点可以通过网络结构110彼此通信。以该方式，取决于特定的实施方式，网络结构110可包括互连、总线、交换机或其他网络结构组件。处理节点102可以经由网络结构110彼此通信，以执行点对点并行处理操作以及集体处理操作。特别地，例如在本文中所描述的实施方式中，处理节点102以基于节点注入带宽来组织处理的方式，出于并行处理诸如归约分散操作之类的集体操作的目的而彼此通信。

[0017] 更具体地，对于集体操作的给定并行处理阶段或阶段，给定处理节点102可以根据该阶段的关联节点注入带宽，与一个或多个其他处理节点102传递消息。以该方式，作为示例，给定的处理节点102在初始阶段可以具有相对高的节点注入带宽，这允许节点102在该阶段期间(经由消息收发)与其他三个处理节点102进行通信。然而，其他后续阶段可以与相

对较低的节点注入带宽相关联。

[0018] 更具体地，根据示例实施方式，由于各种因素，处理节点102可以具有不同程度的节点注入带宽。例如，如本文进一步所描述，某些处理节点102可以是超级节点的节点，该处理节点可以在特定阶段期间与超级节点的其他三个节点(作为示例)进行通信。作为另一示例，对于特定处理阶段，与其他节点102相反，一些处理节点102可以通过较多数量的链路耦合到网络结构110。

[0019] 根据示例实施方式，处理节点102可以使用消息传递接口(MPI)彼此通信，该消息传递接口是允许点对点和集体并行处理应用的函数调用库。以该方式，如图1所示，给定的处理节点102(在此，处理节点102-0)可包括一个或多个处理核140、网络结构接口142和存储器150。通常，存储器150可以是非暂态存储器，其可以存储数据152和机器可执行指令(或“软件”)154。存储器150可以由一个或多个不同的存储设备(诸如半导体存储设备；磁存储设备；相变存储设备；忆阻器；非易失性存储设备；易失性存储设备；由一种或多种前述存储技术形成的存储设备；等等)形成。通常，指令154在由一个或多个处理核140执行时，可以使处理核140执行与并行处理集体操作有关的操作，以用于并行处理计算机系统100。

[0020] 通常，MPI在处理节点102之间执行的进程之间提供虚拟拓扑、同步和通信功能。结合图1参考图2，根据示例实施方式，处理节点102可包括分层归约分散(HRS)协调器160，其可以至少部分地由节点102的MPI 210形成。如本文所述，出于执行集体操作(诸如归约分散操作)的目的，HRS协调器160对节点102的进程204和在其他节点102上执行的进程之间的消息和数据的传递进行协调。根据示例实施方式，处理节点102的HRS协调器160形成分布式HRS协调引擎，以按照允许具有最高相关联的注入带宽的阶段为在集体操作中的初始阶段的顺序来为给定的集体操作安排并行处理阶段。

[0021] 根据示例实施方式，HRS协调器160可以全部或部分地由执行机器可执行指令(存储在存储器150中的指令(图1))的处理节点102的一个或多个处理核140(图1)形成。根据进一步的示例实施方式，HRS协调器160可以全部或部分由不执行机器可执行指令的电路(诸如专用集成电路(ASIC)或现场可编程门阵列(FPGA))形成。

[0022] 图3是根据示例实施方式的集体并行处理操作(诸如归约分散操作)的示例阶段的图示300。通常，集体处理操作300遵循处理顺序301，并且具有一个或多个基于HRS的初始阶段(图3中的两个示例HRS阶段360和370)，随后是一个或多个基于Rabenseifner算法的后续处理阶段(图3中所描绘的三个基于Rabenseifner算法的阶段380、382和384)。通常，与基于Rabenseifner算法的阶段380、382和384相比，基于HRS的阶段360和370与更高的节点注入带宽相关联；并且初始的基于HRS的阶段360与最高的节点注入带宽相关联。

[0023] 在图3中，消息尺寸为“N”个字节；而由每个处理节点102交换的数据在每个阶段减少。以该方式，如图4所示，在HRS阶段360(初始阶段)中，每个处理节点102交换N/4字节的数据；在下一阶段，即HRS阶段370，每个处理节点102交换N/8字节的数据；在下一阶段380中，每个处理节点102交换N/4字节的数据；依次类推。

[0024] 由于在基于HRS的阶段360期间，同一超级节点310的处理节点102(图3中所描绘的两个超级节点310-1和310-2)交换数据，因此HRS阶段360具有最高的注入节点带宽。由于交换数据的处理节点102是同一超级节点310的节点，因此每个处理节点102在基于HRS的阶段360期间与其他三个处理节点102交换数据。以该方式，对于图3的特定示例，超级节点310-1

包括四个处理节点102-0、102-2、102-4和102-6;而超级节点310-2包括四个处理节点102-1、102-3、102-5和102-7。此外,如图3所示,可以将这些超级节点310-1和310-2编组在一起以形成相对应的网格308;而在下一阶段(HRS阶段370)期间,超级节点310-1、310-2中的一个的处理节点102与超级节点310-1、310-2中的另一个的处理节点102交换数据。

[0025] 更具体地,给定超级节点310的处理节点102能够与超级节点310的其他三个处理节点102同时传递消息。例如,超级节点310-1的处理节点102-0可以(通过相对应的链路320)与超级节点310-1的其他三个处理节点102-4、102-6和102-2传递消息。因此,在初始HRS阶段360期间,对于给定的超级节点310,具有N个字节的尺寸的消息被分为四个部分。每个处理节点102与超级节点310的其他三个处理节点102交换其相对应的N/4字节的数据,并执行相对应的归约操作。

[0026] 同样如图3所示,每个超级节点310的处理节点102在下一个HRS阶段370期间进行通信。在该方面,如图3所示,处理节点102-0和102-1通过相对应的链路320进行通信;处理节点102-6和102-7通过相对应的链路320进行通信;节点102-4和102-5通过相对应的处理链路320进行通信;而节点102-2和102-3通过相对应的链路320进行通信。对于HRS阶段370,每个处理节点102与另一个处理节点102交换其N/8部分并执行相对应的归约操作。

[0027] 对于基于Rabenseifner算法的后续阶段380、382和382,网格的处理节点102交换数据并执行相对应的归约操作,如图3所示。

[0028] 作为更具体的示例,图4A描绘了根据示例实施方式的由给定网格308的处理节点102处理的示例输入数据集400。输入数据集400包括八个数据数组,其中网格108的每个处理节点102最初存储八个输入数据数组中的一个。为了简化以下讨论,对于该示例,每个输入数据数组为<1,2,3,4,5,6,7,8>。

[0029] 图4B是针对图4A的输入数据集400的网格308内的输入数据集400的数据交换和归约的图示420。在第一HRS阶段360中,如附图标记422所示,网格308的每个处理节点102与网格108的其他三个处理节点102交换数据并执行相对应的归约。在此,作为示例,归约操作是数学上的加法操作。因此,如图4B所示,在HRS阶段360结束时,处理节点102-0和102-1各自为第一数组元素存储“4”。换句话说,处理节点102-0将其第一数据元素的值“1”与从其他三个处理节点102-2、102-4和102-6获得的值“1”相加。以类似的方式,在阶段360结束时,处理节点102-1存储通过将第一元素的“1”与从处理节点102-3、102-5和102-7接收的“1”值相加而得到的“4”。以类似的方式,在HRS处理阶段160结束时,处理节点102-2为元素三存储“12”,这代表输入数据集400的第三数据元素的一半的总和。

[0030] 对于第二阶段370,如参考数字444所指示,四对处理节点102(每个超级节点310中的一个)交换它们数据元素的一半并执行相对应的归约。例如,处理节点102-2与处理节点102-3交换数据,导致在处理节点102-2上存储的第三数据元素的值为“24”,而在处理节点102-3上存储的第四数据元素的值为“32”。

[0031] 参照图5,根据示例实施方式,网格可以是服务器500(例如,刀片服务器卡)的一部分。在该方面,服务器500可包括图形处理单元(GPU)510,其被布置成用于形成两个超级节点520-1和520-2。而且,每个GPU可包括HRS协调器512。如图5所示,服务器500可包括PCIe交换机560,该PCIe交换机560允许中央处理单元(CPU)570与每个超级节点520进行通信。

[0032] 因此,参考图6,根据示例实施方式,技术600包括使用多个顺序的并行处理阶段在

并行处理系统的多个节点之间执行集体操作(框604)。依据技术600的框608,调节并行处理阶段的排序,使得初始阶段与比后续阶段更高的节点注入带宽相关联。

[0033] 预期在所附权利要求的范围内的其他实施方式。例如,根据进一步的实施方式,本文所描述的系统和技术可以被应用于除归约分散操作之外的集体并行处理操作(诸如全部归约、全部对全部和全部聚集操作)。

[0034] 以下示例涉及进一步的实施方式。

[0035] 示例1包括一种计算机实现的方法,该方法包括使用多个并行处理阶段在并行处理系统的多个节点之间执行集体操作。方法包括调节并行处理阶段的排序,其中多个并行处理阶段的初始阶段与比多个并行处理阶段的后续阶段更高的节点注入带宽相关联。

[0036] 在示例2中,示例1的主题可任选地包括在多个节点之间传递消息,并调节排序使得与初始阶段相关联的消息尺寸大于与后续阶段相关联的消息尺寸。

[0037] 在示例3中,示例1和2的主题可任选地包括执行归约分散操作。

[0038] 在示例4中,示例1-3的主题可任选地包括在多个节点之间并行处理数据向量的元素,以对元素进行归约并将经归约的元素跨多个节点分散。

[0039] 在示例5中,示例1-4的主题可进一步包括:对于多个处理阶段的初始阶段,将多个消息从多个节点中的第一节点传递到多个节点中的另一节点以将数据从另一节点传递到第一节点,并在第一节点中处理所传递的数据以对所传递的数据进行归约操作。

[0040] 在示例6中,示例1-5的主题可任选地包括多个节点,该多个节点包括节点集群,在初始阶段中在每个集群的节点之间传递消息,并且在后续阶段中在集群之间传递消息。

[0041] 在示例7中,示例1-6的主题可任选地包括多个节点,该多个节点包括布置在超级节点中的节点的子集,在初始阶段中在每个超级节点的节点之间传递消息,并且在后续阶段中在超级节点之间传递消息。

[0042] 在示例8中,示例1-7的主题可任选地包括多个节点,该多个节点包括布置在超级节点中的节点的子集以及布置在网格中的超级节点的子集。方法可包括在初始阶段在每个超级节点的节点之间传递消息;在多个并行处理阶段的第二阶段中,在每个网格的超级节点之间传递消息;以及在多个并行处理阶段的第三阶段中,在网格之间传递消息。

[0043] 在示例9中,示例1-8的主题可任选地包括布置在超级节点中的节点的子集,以及布置在网格中的超级节点的子集。方法可进一步包括在初始阶段在每个超级节点的节点之间传递消息;在多个并行处理阶段的第二阶段中,在每个网格的超级节点之间传递消息;以及在多个并行处理阶段的多个其他阶段中,在消息之间传递消息。

[0044] 在示例10中,示例1-9的主题可任选地包括在多个并行处理阶段的多个其他阶段中传递网格之间的消息,包括根据基于Rabenseifner的算法进行通信。

[0045] 示例11包括用于存储指令的非暂态计算机可读存储介质,在指令在由并行处理机器执行时,使得机器在多个并行处理阶段的每个阶段中,在机器的多个处理节点之间传递消息以交换和对数据进行归约,其中每个处理阶段与注入带宽相关联,并且注入带宽不同。指令在由并行处理机器执行时,导致机器对阶段进行排序,使得多个并行处理阶段的初始阶段与相关联的注入带宽中的最高注入带宽相关联。

[0046] 在示例12中,示例11的主题可任选地包括计算机可读存储介质,该计算机可读存储介质存储指令,该指令在由并行处理机器执行时使得机器提供消息接口库,该消息接口

库提供允许对阶段进行排序的函数，并且初始阶段与最高注入带宽相关联。

[0047] 在示例13中，示例11和12的主题可任选地包括计算机可读存储介质，该计算机可读存储介质存储指令，该指令在由并行处理机器执行时使得机器根据相关联的注入带宽对阶段进行排序，以便在与相对较低的注入带宽相关的阶段之前执行与相对较高的注入带宽相关联的阶段。

[0048] 在示例14中，示例11-13的主题可任选地包括多个处理节点，该多个处理节点包括布置在超级节点中的节点的子集；以及布置在网格中的超级节点的子集。计算机可读存储介质可以存储指令，该指令在由并行处理机器执行时使得每个超级节点的节点彼此通信以在初始阶段对数据进行归约，使每个网格的超级节点彼此通信以在多个并行处理阶段的第二阶段中对数据进行归约，并且使网格彼此通信以在多个并行处理阶段中的至少另一个第三阶段中对数据进行归约。

[0049] 示例15包括一种系统，该系统包括多个处理网格，以对第一数据集执行归约分散并行处理操作。每个网格包括多个超级节点；以及每个超级节点包括多个计算机处理节点。系统包括用于将归约分散并行处理操作分成多个并行处理阶段的协调器，该多个并行处理阶段包括第一阶段、第二阶段和至少一个附加阶段。在初始阶段，每个超级节点的计算机处理节点彼此传递消息以对第一数据集进行归约，以提供第二数据集；在第二阶段中，每个网格的超级节点彼此传递消息以对第二数据集进行归约，以产生第三数据集；以及在至少一个附加的阶段，网格彼此传递消息，以进一步对第三数据集进行归约。

[0050] 在示例16中，示例15的主题可任选地包括协调器，该协调器包括消息传递接口(MPI)。

[0051] 在示例17中，示例15和16的主题可任选地包括计算机处理节点，该计算机处理节点包括多个处理核。

[0052] 在示例18中，示例15-17的主题可任选地包括在初始阶段，给定超级节点的给定计算机处理节点与给定超级节点的另一个计算机处理节点传递多个消息。

[0053] 在示例19中，示例15至18的主题可任选地包括在至少一个附加阶段的第三阶段中，每个网格将单个消息与另一个网格进行通信。

[0054] 在示例20中，示例15至19的主题可任选地包括计算机处理节点，该计算机处理节点包括服务器刀片。

[0055] 尽管已经相对于有限数量的实施方式描述了本公开，但是受益于本公开的本领域技术人员将理解其众多修改和变化。所附权利要求旨在覆盖所有此类修改和变化。

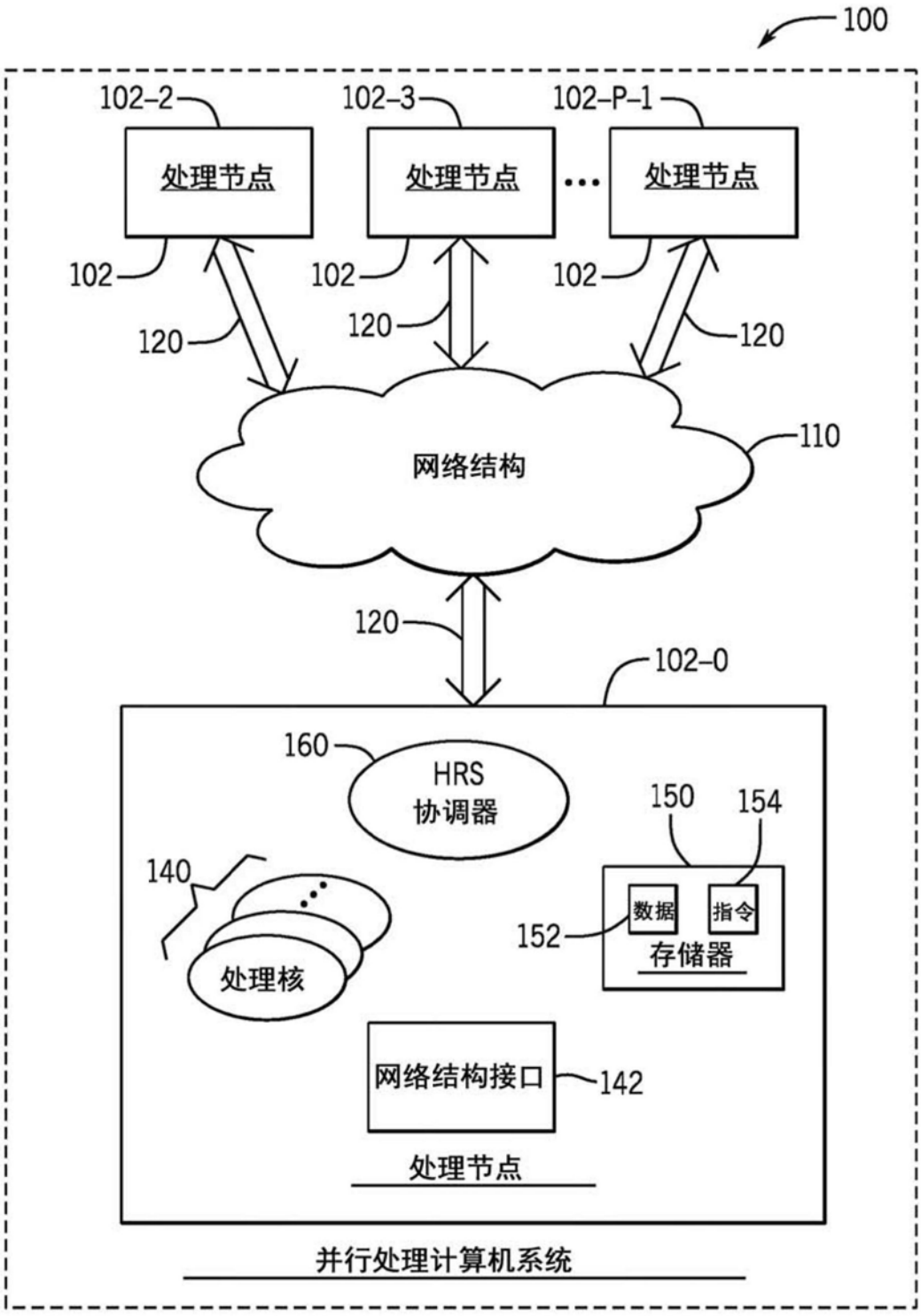

100
102-2
102-3
102-P-1
处理节点
处理节点
处理节点
102
102
102
120
120
120
网络结构
110
120
102-0
160
HRS
协调器
150
154
140
数据
指令
152
存储器
处理核
网络结构接口
142
处理节点
并行处理计算机系统

图1

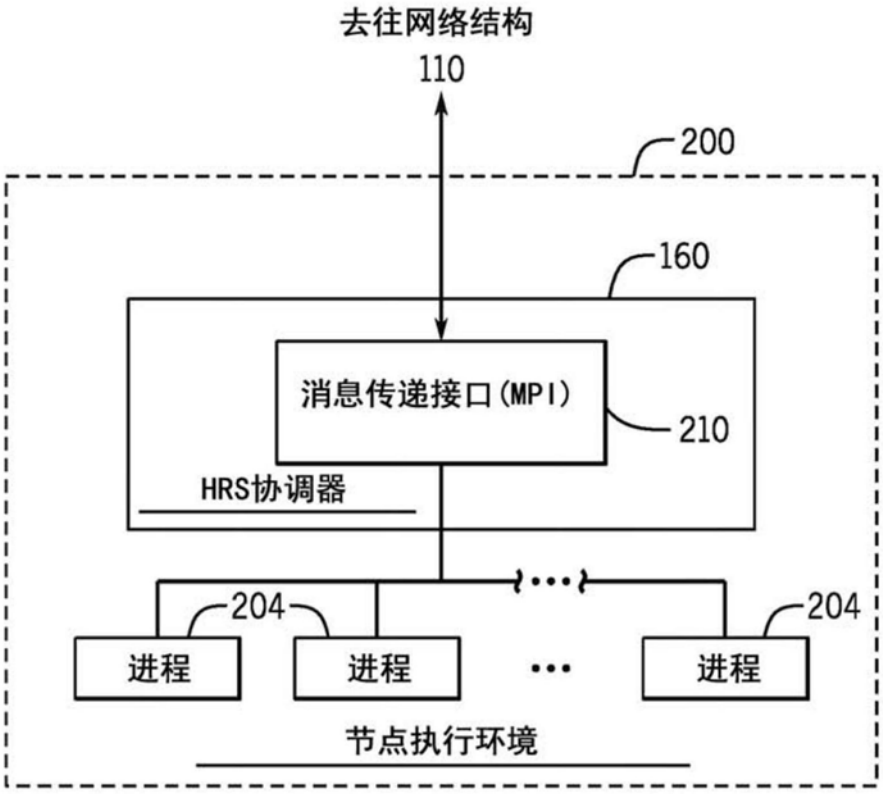

去往网络结构
110
200
160
消息传递接口(MPI)
210
HRS协调器
204
204
进程
进程
进程
节点执行环境

图2

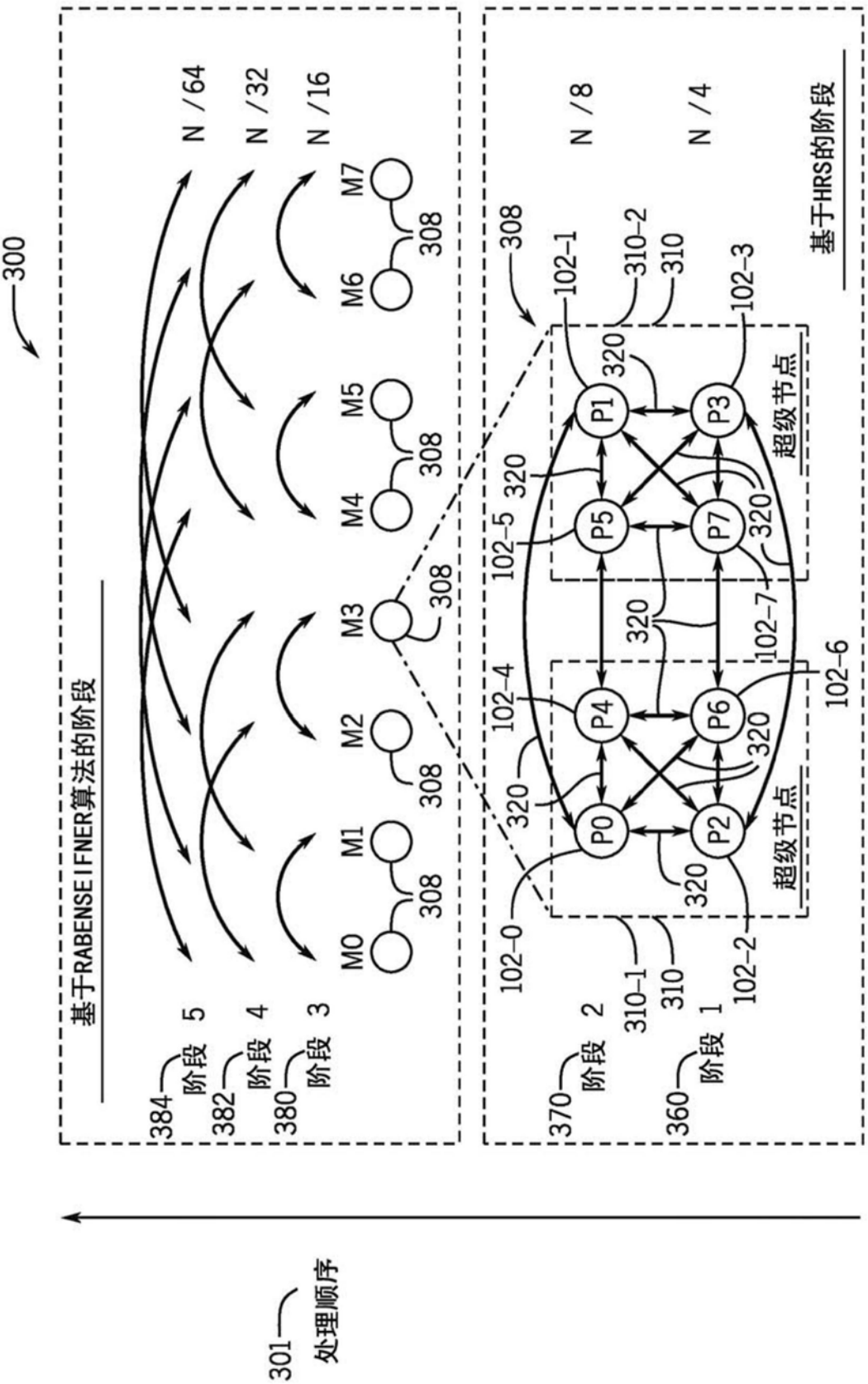

300
基于RABENSEIFNER算法的阶段
基于HRS的阶段
N /64
N /32
N /16
N /8
N /4
M0
M1
M2
M3
M4
M5
M6
M7
308
P0
P1
P2
P3
P4
P5
P6
P7
320
超级节点
超级节点
102-0
102-1
102-2
102-3
102-4
102-5
102-6
102-7
310
310-1
310-2
阶段 1
阶段 2
阶段 3
阶段 4
阶段 5
360
370
380
382
384
301
处理顺序

图3

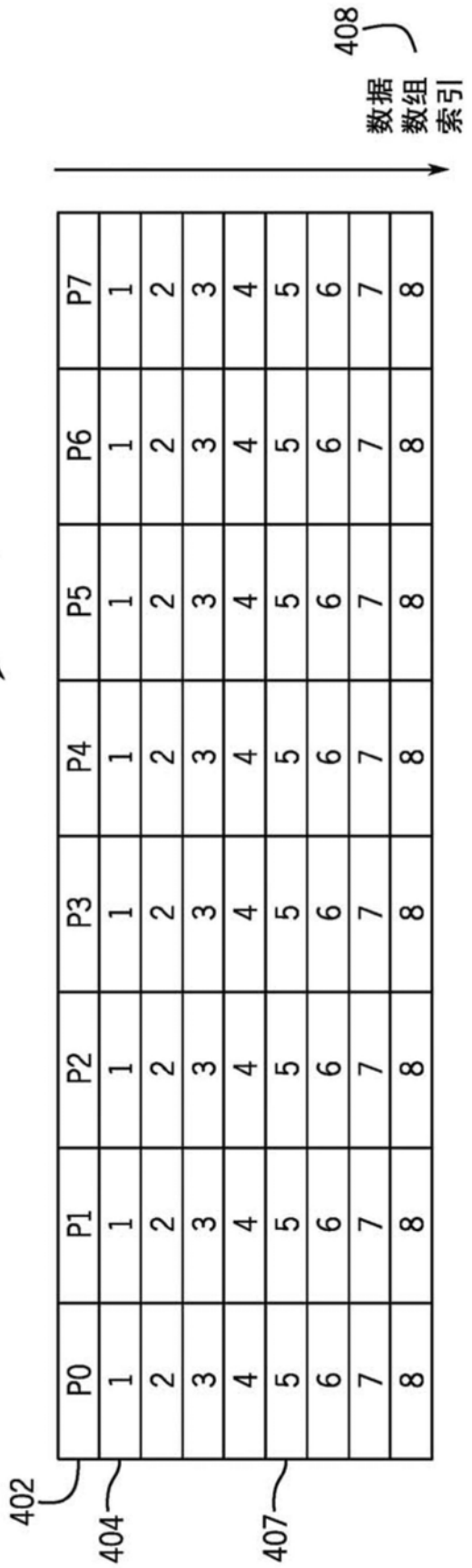

408
数据数组索引
400
P0 P1 P2 P3 P4 P5 P6 P7
1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8
402
404
407

图4A

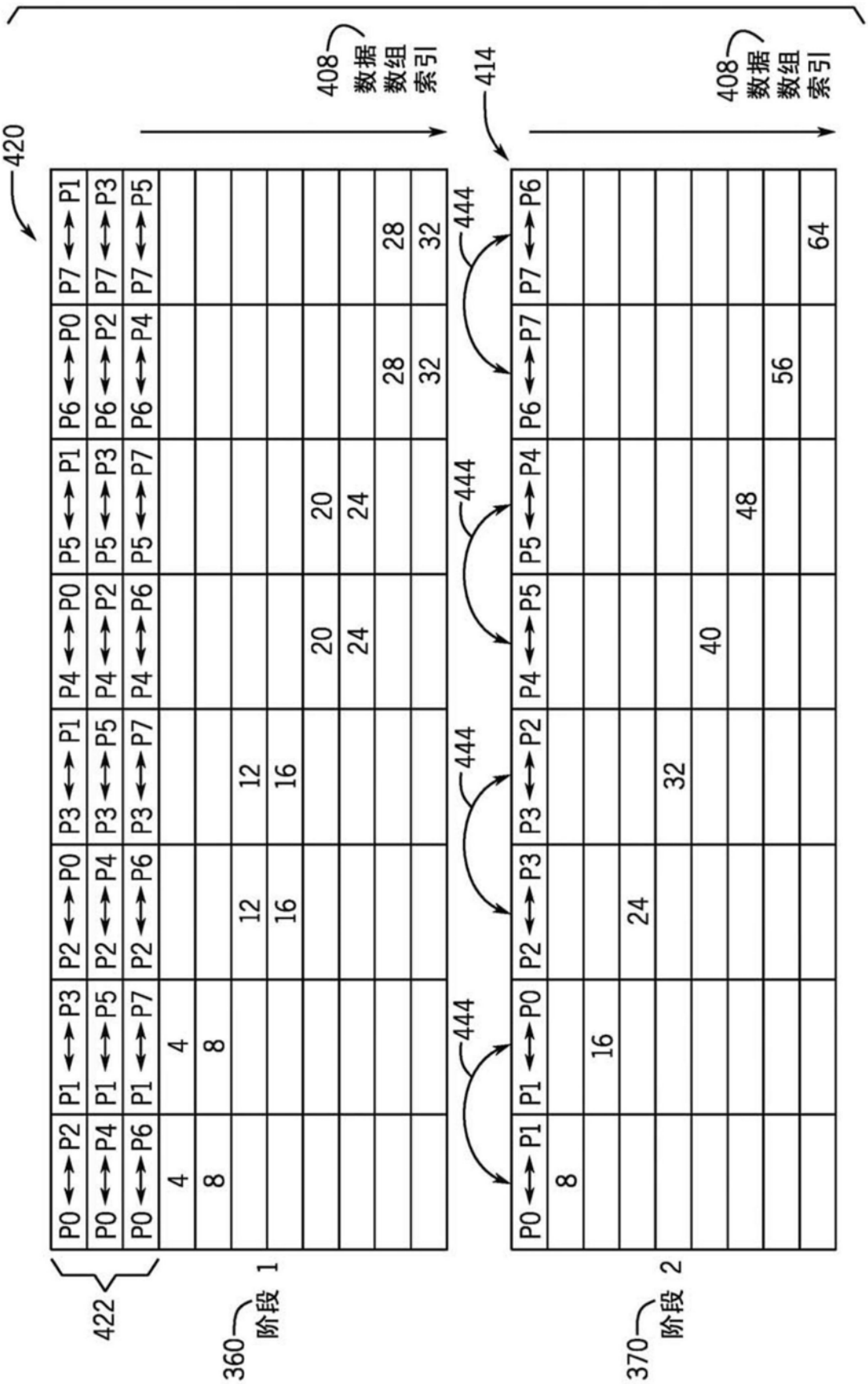

420
408 数据数组索引
414
408 数据数组索引
P0↔P2 P0↔P4 P0↔P6 P1↔P3 P1↔P5 P1↔P7 P2↔P0 P2↔P4 P2↔P6 P3↔P1 P3↔P5 P3↔P7 P4↔P0 P4↔P2 P4↔P6 P5↔P1 P5↔P3 P5↔P7 P6↔P0 P6↔P2 P6↔P4 P7↔P1 P7↔P3 P7↔P5
4 4 12 12 20 20 28 28
8 8 16 16 24 24 32 32
444
444
444
444
P0↔P1 P1↔P0 P2↔P3 P3↔P2 P4↔P5 P5↔P4 P6↔P7 P7↔P6
8 16 24 32 40 48 56 64
422
360 阶段 1
370 阶段 2

图4B

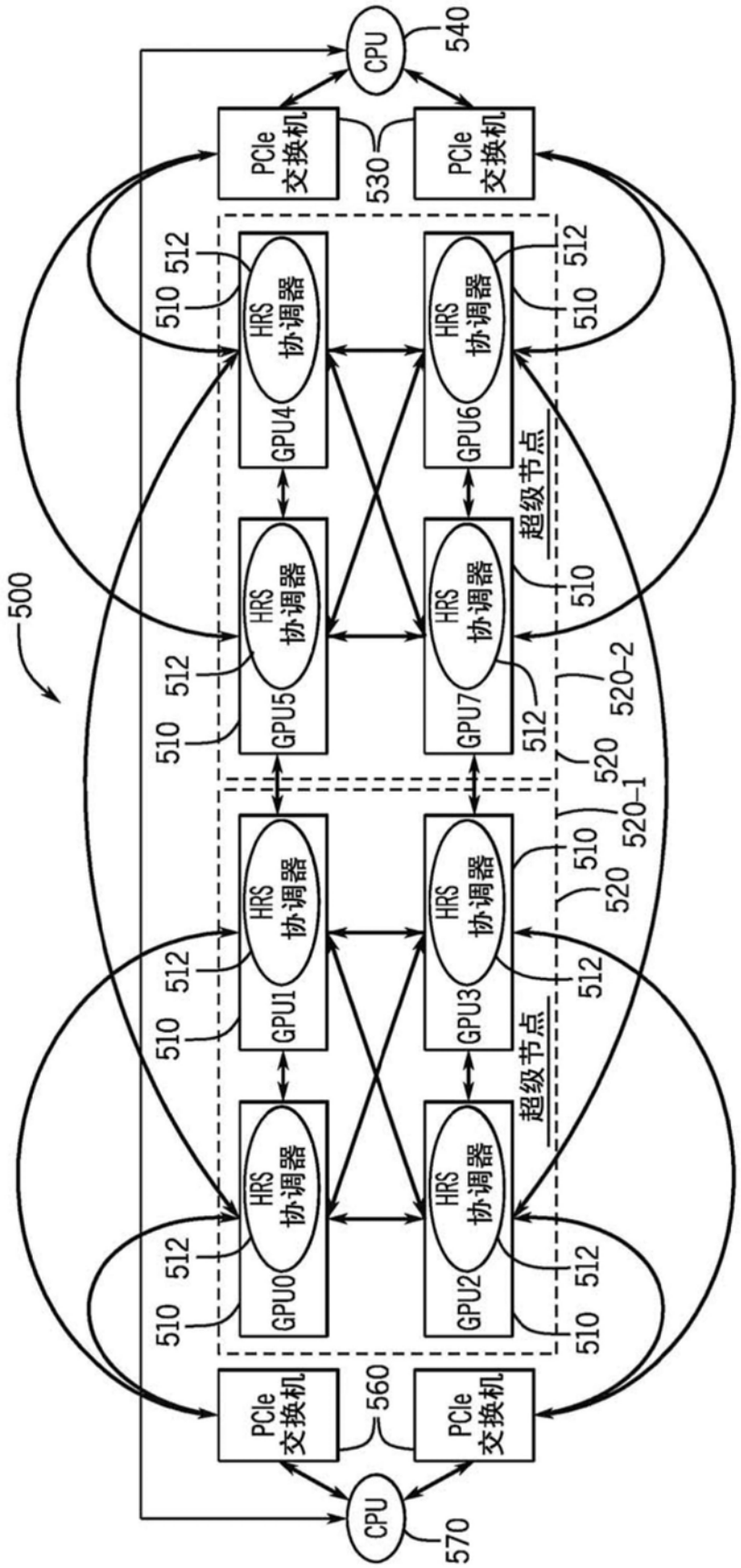

500
CPU
540
PCIe 交换机
530
PCIe 交换机
HRS 协调器
GPU4
GPU6
GPU5
GPU7
GPU1
GPU3
GPU0
GPU2
510
512
超级节点
520-2
520
520-1
PCIe 交换机
560
PCIe 交换机
CPU
570

图5

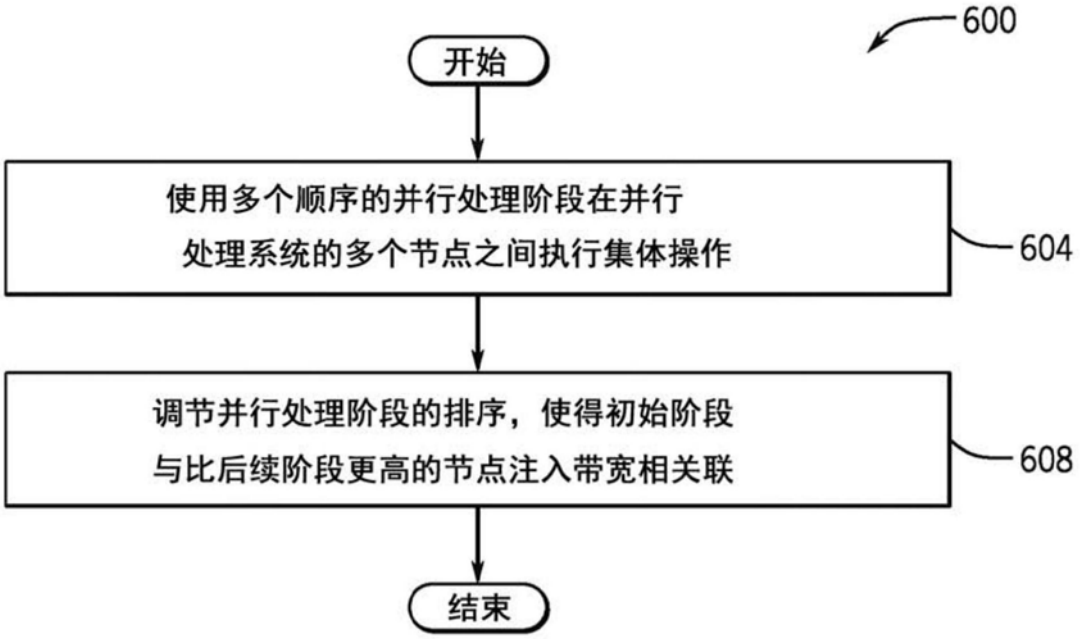

600
开始
使用多个顺序的并行处理阶段在并行
处理系统的多个节点之间执行集体操作
604
调节并行处理阶段的排序，使得初始阶段
与比后续阶段更高的节点注入带宽相关联
608
结束

图6