(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113254871 B
(45)授权公告日 2024.09.03

(21)申请号 202110566425.9
(22)申请日 2017.01.13
(65)同一申请的已公布的文献号
申请公布号 CN 113254871 A
(43)申请公布日 2021.08.13
(30)优先权数据
15/016,486 2016.02.05 US
(62)分案原申请数据
201710025656.2 2017.01.13
(73)专利权人 谷歌有限责任公司
地址 美国加利福尼亚州
(72)发明人 拉维·纳拉亚纳斯瓦米
拉胡尔·纳加拉扬 禹同爀
克里斯多佛·丹尼尔·利里
(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
专利代理师 周亚荣 邓聪惠
(51)Int.Cl.
G06F 17/16(2006.01)
G06F 17/14(2006.01)
H03M 7/30(2006.01)
(56)对比文件
CN 103262068 A,2013.08.21
审查员 王桂兰

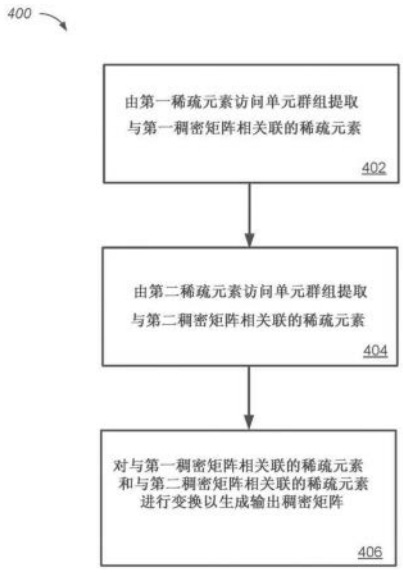
权利要求书4页 说明书13页 附图9页

(54)发明名称

矩阵处理装置

(57)摘要

本公开涉及矩阵处理装置。提供了方法、系统、和装置,其包括用于将稀疏元素变换为稠密矩阵的系统。所述系统包括数据提取单元,其包括多个处理器,所述数据提取单元被配置为基于对特定稀疏元素的子集的识别,确定用于提取所述特定稀疏元素的所述子集的处理器指定。所述系统包括串接单元,所述串接单元被配置为基于被应用于由所述数据提取单元提取的稀疏元素的变换来生成输出稠密矩阵。



CN 113254871 B

1.一种用于将稀疏元素变换为稠密矩阵的系统,所述系统包括:

沿着二维稀疏-稠密变换单元的相应的维度布置的多个稀疏元素访问单元,其中,每个稀疏元素访问单元包括:

相应的数据提取单元,所述相应的数据提取单元包括相应的多个处理器,所述数据提取单元被配置为:

从外部源接收一个或多个控制信号,所述一个或多个控制信号提供所述数据提取单元所位于的稀疏元素访问单元被指派来访问存储在一个或多个数据片中的多个特定稀疏元素的子集的指示;和

基于所述指示,提取所述多个特定稀疏元素的所述子集中的一个或多个稀疏元素;

相应的串接单元,所述相应的串接单元被配置为:

基于至少被应用于所述一个或多个稀疏元素的变换来生成输出稠密矩阵;以及

相应的压缩/解压缩单元,所述相应的压缩/解压缩单元被配置为:

压缩所述输出稠密矩阵以生成压缩输出稠密矩阵;和

向节点网络提供所述压缩输出稠密矩阵。

2.根据权利要求1所述的系统,每个稀疏元素访问单元还包括:

相应的请求识别单元,所述相应的请求识别单元被配置为:

通过所述节点网络来接收对于所述多个特定稀疏元素的请求;

确定所述请求识别单元所位于的稀疏元素访问单元被指派来处置所述多个特定稀疏元素的所述子集;以及

响应于确定所述请求识别单元所位于的稀疏元素访问单元被指派来处置所述多个特定稀疏元素的所述子集,生成对于所述数据提取单元的访问所述多个特定稀疏元素的所述子集的所述指示。

3.根据权利要求2所述的系统,其中,确定所述请求识别单元所位于的稀疏元素访问单元被指派来处置所述多个特定稀疏元素的所述子集包括:基于查找表来确定所述请求识别单元所位于的稀疏元素访问单元被指派来处置所述多个特定稀疏元素的子集。

4.根据权利要求1所述的系统,其中,每个稀疏元素访问单元进一步包括:

相应的稀疏降低单元,所述相应的稀疏降低单元被配置为:

接收包括来自第一处理器的第一稀疏元素的第一矩阵,所述第一矩阵具有第一维度;以及

生成包括所述第一稀疏元素的第二矩阵,所述第二矩阵具有小于所述第一维度的第二维度,

其中,所述串接单元进一步被配置为:

接收所述第二矩阵;

其中,生成所述输出稠密矩阵进一步包括基于所述第二矩阵来生成所述输出稠密矩阵。

5.根据权利要求1所述的系统,其中,所述串接单元被配置为:

在第一时间点接收第一稀疏元素;

在不同的第二时间点接收第二稀疏元素;以及

针对所述输出稠密矩阵来确定所述第一稀疏元素和所述第二稀疏元素的顺序,

其中,生成所述输出稠密矩阵进一步包括:基于所述第一稀疏元素和所述第二稀疏元素的所述顺序来生成所述输出稠密矩阵。

6.根据权利要求1所述的系统,

其中,所述串接单元进一步被配置为接收表示通过所述节点网络所发送的稠密矩阵的第一稠密矩阵,以及

其中,生成所述输出稠密矩阵进一步包括:基于所述第一稠密矩阵、第一稀疏元素、以及第二稀疏元素来生成所述输出稠密矩阵。

7.根据权利要求6所述的系统,其中,所述压缩/解压缩单元被配置为对压缩第一稠密矩阵解压缩以生成稠密矩阵,所述稠密矩阵与由所述串接单元接收的所述第一稠密矩阵相对应。

8.根据权利要求1所述的系统,其中,所述多个特定稀疏元素中的一个或多个稀疏元素为多维矩阵,并且其中,所述输出稠密矩阵为向量。

9.一种用于将稀疏元素变换为稠密矩阵的方法,包括:

由稀疏元素访问单元从外部源接收一个或多个控制信号以用于访问一个或多个稀疏元素,其中,多个稀疏元素访问单元沿着二维稀疏-稠密变换单元的相应的维度来布置,每个稀疏元素访问单元包括相应的数据提取单元、相应的串接单元以及相应的压缩/解压缩单元:

由具有相应的多个处理器的所述相应的数据提取单元,基于所述一个或多个控制信号,接收所述数据提取单元所位于的稀疏元素访问单元被指派来访问存储在一个或多个数据片中的多个特定稀疏元素的子集的指示;

基于所述指示,提取所述多个特定稀疏元素的所述子集中的一个或多个稀疏元素;

由所述相应的串接单元,基于至少被应用于所述一个或多个稀疏元素的变换来生成输出稠密矩阵;以及

由所述相应的压缩/解压缩单元压缩所述输出稠密矩阵以生成压缩输出稠密矩阵,并且向节点网络提供所述压缩输出稠密矩阵。

10.根据权利要求9所述的方法,进一步包括:

由稀疏降低单元接收包括来自第一处理器的第一稀疏元素的第一矩阵,所述第一矩阵具有第一维度;

由所述稀疏降低单元生成包括所述第一稀疏元素的第二矩阵,所述第二矩阵具有小于所述第一维度的第二维度;以及

由所述串接单元接收所述第二矩阵,

其中,生成所述输出稠密矩阵进一步包括基于所述第二矩阵来生成所述输出稠密矩阵。

11.根据权利要求9所述的方法,其中,生成所述输出稠密矩阵进一步包括:

在第一时间点接收第一稀疏元素;

在不同的第二时间点接收第二稀疏元素;

确定用于所述输出稠密矩阵的所述第一稀疏元素和所述第二稀疏元素的顺序;以及

基于所述第一稀疏元素和所述第二稀疏元素的所述顺序来生成所述输出稠密矩阵。

12.根据权利要求9所述的方法,其中,生成所述输出稠密矩阵进一步包括:

接收表示通过所述节点网络所发送的稠密矩阵的第一稠密矩阵;以及

基于所述第一稠密矩阵、第一稀疏元素、以及第二稀疏元素来生成所述输出稠密矩阵。

13.根据权利要求9所述的方法,进一步包括:

由请求识别单元通过所述节点网络来接收对于存储在一个或多个数据片中的多个特定稀疏元素的请求;

确定所述数据提取单元被指派来处置所述多个特定稀疏元素的子集;以及

响应于确定所述数据提取单元被指派来处置所述多个特定稀疏元素的子集,生成访问所述多个特定稀疏元素的所述子集的所述指示。

14.根据权利要求13所述的方法,其中,确定所述数据提取单元被指派来处置所述多个特定稀疏元素的所述子集包括:基于查找表来确定所述数据提取单元被指派来处置所述多个特定稀疏元素的子集。

15.一种用于将稀疏元素变换为稠密矩阵的系统,所述系统包括:

一个或多个处理器,所述处理器被配置为发送对于输出矩阵的请求,所述输出矩阵基于存储在一个或多个数据片中的多个特定稀疏元素;

通过节点网络连接并且沿着二维稀疏-稠密变换单元的相应的维度布置的多个稀疏元素访问单元,其中,每个稀疏元素访问单元包括:

相应的数据提取单元,所述相应的数据提取单元包括相应的多个处理器,所述数据提取单元被配置为:

从外部源接收一个或多个控制信号,所述一个或多个控制信号提供所述数据提取单元所位于的稀疏元素访问单元被指派来访问所述多个特定稀疏元素的子集的指示;和

基于所述指示,提取所述多个特定稀疏元素的所述子集中的一个或多个稀疏元素;

相应的串接单元,所述相应的串接单元被配置为:

基于至少被应用于所述一个或多个稀疏元素的变换来生成输出稠密矩阵;以及

相应的压缩/解压缩单元,所述相应的压缩/解压缩单元被配置为:

压缩所述输出稠密矩阵以生成压缩输出稠密矩阵;和

向所述节点网络提供所述压缩输出稠密矩阵。

16.根据权利要求15所述的系统,每个稀疏元素访问单元还包括:

相应的请求识别单元,所述相应的请求识别单元被配置为:

通过所述节点网络来接收对于所述多个特定稀疏元素的所述请求;

确定所述请求识别单元所位于的稀疏元素访问单元被指派来处置所述多个特定稀疏元素的所述子集;以及

响应于确定所述请求识别单元所位于的稀疏元素访问单元被指派来处置所述多个特定稀疏元素的子集,生成对于所述数据提取单元的访问所述多个特定稀疏元素的所述子集的所述指示。

17.根据权利要求15所述的系统,其中,每个稀疏元素访问单元进一步包括:

相应的稀疏降低单元,所述相应的稀疏降低单元被配置为:

接收包括来自第一处理器的第一稀疏元素的第一矩阵,所述第一矩阵具有第一维度;以及

生成包括所述第一稀疏元素的第二矩阵,所述第二矩阵具有小于所述第一维度的第二

维度，

其中，所述串接单元进一步被配置为：

接收所述第二矩阵；

其中，生成所述输出稠密矩阵进一步包括基于所述第二矩阵来生成所述输出稠密矩阵。

矩阵处理装置

[0001] 分案说明

[0002] 本申请属于申请日为2017年1月13日的中国发明专利申请201710025656.2的分案申请。

技术领域

[0003] 本公开涉及矩阵处理装置。

背景技术

[0004] 本说明书大体上涉及使用电路来处理矩阵。

发明内容

[0005] 根据本说明书中所描述的主题的一个创新方面,矩阵处理器能够被用来执行稀疏转稠密或者稠密转稀疏矩阵变换。大体上,高效能计算系统可以使用线性代数例程来处理矩阵。在一些实例中,矩阵的大小可能过大而无法存入一个数据存储中,并且该矩阵的不同部分可以被稀疏地存储在分布式数据存储系统的不同位置中。为了加载该矩阵,计算系统的中央处理单元可以指令多个矩阵处理器访问该矩阵的不同部分。每个矩阵处理器可以聚集稀疏数据,以执行对该稀疏数据的并行计算并且生成稠密矩阵,该稠密矩阵能够被串接在一起以供中央处理单元执行进一步的处理。

[0006] 大体上,本说明书中所描述的主题的一个创新方面能够被替选在用于将稀疏元素变换为稠密矩阵的系统中。所述系统包括请求识别单元,所述请求识别单元被配置为:通过节点网络来接收对于存储在一个或多个数据片(data shard)中的特定稀疏元素的请求;确定所述系统被指派来处置(handle)所述特定稀疏元素的子集;以及响应于确定所述系统被指派来处置所述特定稀疏元素的所述子集,生成访问所述特定稀疏元素的所述子集的指示。所述系统包括数据提取单元,其包括多个处理器,所述数据提取单元被配置为:从所述请求识别单元接收访问特定稀疏元素的所述子集的所述指示;基于对所述特定稀疏元素的所述子集的识别,确定用于提取所述特定稀疏元素的所述子集的处理器指定;基于所述指定由所述多个处理器中的第一处理器来提取所述特定稀疏元素的所述子集中的第一稀疏元素;以及基于所述指定由所述多个处理器中的第二处理器来提取所述特定稀疏元素的所述子集中的第二稀疏元素。所述系统包括串接单元,所述串接单元被配置为:基于至少被应用于所述第一稀疏元素和所述第二稀疏元素的变换来生成输出稠密矩阵。

[0007] 这些或其他实施方式均能够可选地包括以下特征中的一个或多个。例如,所述系统能够包括稀疏降低单元,所述稀疏降低单元被配置为接收包括来自所述第一处理器的所述第一稀疏元素的第一矩阵,所述第一矩阵具有第一维度;以及生成包括所述第一稀疏元素的第二矩阵,所述第二矩阵具有小于所述第一维度的第二维度。所述串接单元可以进一步被配置为接收所述第二矩阵。为了生成所述输出稠密矩阵,可以基于所述第二矩阵来生成所述输出稠密矩阵。

[0008] 所述串接单元可以被配置为在第一时间点接收所述第一稀疏元素;在不同的第二时间点接收所述第二稀疏元素;以及针对所述输出稠密矩阵来确定所述第一稀疏元素和所述第二稀疏元素的顺序。为了生成所述输出稠密矩阵,可以基于所述第一稀疏元素和所述第二稀疏元素的所述顺序来生成所述输出稠密矩阵。

[0009] 所述系统可以包括压缩/解压缩单元,所述压缩/解压缩单元被配置为压缩所述输出稠密矩阵以生成压缩输出稠密矩阵,以及向节点网络提供所述压缩输出稠密矩阵。所述串接单元可以进一步被配置为接收表示通过节点网络所发送的稠密矩阵的第一稠密矩阵。为了生成所述输出稠密矩阵,可以基于所述第一稠密矩阵、所述第一稀疏元素、以及所述第二稀疏元素来生成所述输出稠密矩阵。所述压缩/解压缩单元可以被配置为对压缩第一稠密矩阵解压缩以生成所述第一稠密矩阵。

[0010] 为了确定所述系统被指派来处置所述特定稀疏元素的所述子集,所述数据提取单元可以基于查找表来确定所述系统被指派来处置所述特定稀疏元素的子集。所述特定稀疏元素中的一个或多个稀疏元素可以是多维矩阵,并且所述输出稠密矩阵可以是向量。

[0011] 本说明书中所描述的主题能够以特定实施例来实现以便达到以下优点中的一个或多个。将稀疏转稠密数据加载任务从中央处理单元转移到专门的矩阵处理器提高了中央处理单元的计算带宽并且降低了所述系统的处理成本。该矩阵处理器能够被布置在存储数据的存储器附近,并且能够降低加载数据的时延。通过使用专门的矩阵处理器,能够避免将专用于稠密线性代数的处理器用于提取稀疏数据。通过每控制器通道具有多个单元,一次所服务的多个并发事务可以被并行化,并且事务可以被立即处理而无需等待先前的事务完成。

[0012] 以上和其他方面的其他实施方式包括对应的系统、装置、和计算机程序,其被配置来执行所述方法的动作,并且被编码在计算机存储设备上。一个或多个计算机的系统能够依靠安装在所述系统上的软件、固件、硬件、或它们的组合来被如此配置,从而在操作中使得所述系统执行所述动作。一个或多个计算机程序能够依靠具有指令来被如此配置,所述指令在由数据处理装置执行时,使得所述装置执行所述动作。

[0013] 在附图和下面的描述中阐述了本说明书中所描述的主题的一个或多个实施方式的细节。本主题的其他潜在特征、方面、和优点根据说明书、附图、和权利要求书将变得显而易见。

附图说明

[0014] 图1是示例计算系统的框图。

[0015] 图2A至图2D图示了示例稀疏-稠密变换单元。

[0016] 图3A至图3B图示了示例稀疏元素访问单元。

[0017] 图4是图示了用于生成稠密矩阵的过程的示例的流程图。

[0018] 图5是图示了用于将稀疏元素变换为稠密矩阵的过程的示例的流程图。

[0019] 各附图中相似的附图标记和名称指示相似的元素。

具体实施方式

[0020] 大体上,数据能够以矩阵的形式来表示并且计算系统可以使用线性代数算法来对

该数据进行操纵。矩阵可以是一维向量或者是多维矩阵。矩阵可以由数据结构来表示,诸如数据库表或变量。然而,当矩阵的大小过大时,将整个矩阵存储在一个数据存储中也许是不可能的。稠密矩阵可以被变换为多个稀疏元素(element),其中每个稀疏元素可以被存储在不同的数据存储中。稠密矩阵的稀疏元素可以是其中仅该矩阵的小的子矩阵(例如,单值元素、行、列、或子矩阵)具有非零值的矩阵。当计算系统需要访问该稠密矩阵时,中央处理单元(CPU)可以开始一个线程,该线程到达数据存储中的每一个以提取存储的稀疏元素并且应用稀疏转稠密变换以恢复该稠密矩阵。然而,提取全部的稠密元素所用的时间量可能很长,并且CPU的计算带宽可能因此未被充分利用。在一些情况下,计算系统可能需要访问若干稠密矩阵的稀疏元素以形成新的稠密矩阵,其中这些稠密矩阵可以不具有相等的维度。与到达数据存储中的每一个以提取不同稠密矩阵的稀疏元素的线程相关联的CPU空闲时间可能遇到不同的等待时间,并且可能进一步以不期望的方式影响计算设备的效能。在一些情况下,计算系统可能需要访问若干稠密矩阵的稀疏元素以形成新的稠密矩阵,其中这些稀疏元素可以不具有相等的维度。与到达数据存储中的每一个以提取不同稠密矩阵的稀疏元素的线程相关联的CPU空闲时间可能遇到不同的等待时间,并且可能进一步以不期望的方式影响计算设备的效能。与CPU分离的硬件稀疏-稠密变换单元可以通过独立于CPU操作来收集稀疏元素并且将该稀疏元素变换为稠密矩阵而提高处理器的计算带宽。

[0021] 图1示出了用于变换来自一个或多个稠密矩阵的稀疏元素以生成稠密矩阵的示例计算系统100的框图。计算系统100包括处理单元102、稀疏-稠密变换单元104、和数据片106a-106k,其中k是大于1的整数。大体上,处理单元102处理用于访问目标稠密矩阵的指令,并且将指令110发送至稀疏-稠密变换单元104以生成目标稠密矩阵。稀疏-稠密变换单元104访问来自数据片106a-106k中的一个或多个数据片的对应的稀疏元素108a-108n,其中n是大于1的整数。稀疏-稠密变换单元104使用对应的稀疏元素108a-108n来生成目标稠密矩阵112,并且将目标稠密矩阵112发送至处理单元102以供进一步处理。例如,稀疏元素108a-108n可以是具有不同大小的二维矩阵,并且稀疏-稠密变换单元104可以通过将稀疏元素108a-108n中的每一个变换为向量来生成稠密矩阵112,并且将n个向量串接为单个向量。

[0022] 在一些实施方式中,处理单元102可以处理用于更新目标稠密矩阵的指令并且将已更新的稠密矩阵发送至稀疏-稠密变换单元104。稀疏-稠密变换单元104可以将已更新的稠密矩阵变换为对应的稀疏元素并且相应地对存储在数据片106a-106k中的一个或多个稠密元素进行更新。

[0023] 处理单元102被配置为处理用于在计算系统100内执行的指令。处理单元102可以包括一个或多个处理器。在一些实施方式中,处理单元102被配置为处理由稀疏-稠密变换单元104生成的目标稠密矩阵112。在一些其他的实施方式中,处理单元102可以被配置为请求稀疏-稠密变换单元104生成目标稠密矩阵112,并且另一个处理单元可以被配置为处理该目标稠密矩阵112。数据片106a-106k存储包括稀疏元素108a-108n的数据。在一些实施方式中,数据片106a-106k可以是一个或多个易失性存储器单元。在一些其他实施方式中,数据片106a-106k可以是一个或多个非易失性存储器单元。数据片106a-106k也可以是另一形式的计算机可读介质,诸如存储区域网或其他配置中的设备。数据片106a-106k可以使用电连接、光连接、或无线连接来被耦合至稀疏-稠密变换单元104。在一些实施方式中,数据片

106a-106k可以是稀疏-稠密变换单元104的一部分。

[0024] 稀疏-稠密变换单元104被配置为基于稀疏元素来确定稠密矩阵。在一些实施方式中,稀疏-稠密变换单元104可以被配置为基于稠密矩阵来确定稀疏元素的位置。在一些实施方式中,稀疏-稠密变换单元104可以包括多个互连的稀疏元素访问单元,以下参照图2A来更详细描述。

[0025] 图2A示出了示例稀疏-稠密变换单元200。稀疏-稠密变换单元200可以对应于稀疏-稠密变换单元104。稀疏-稠密变换单元200包括物理或逻辑上被布置为M行和N列的M乘N稀疏元素访问单元$X_{1,1}$至$X_{M,N}$,其中M和N为等于或大于1的整数。在一些实施方式中,稀疏-稠密变换单元200可以包括被配置为处理数据的附加电路。大体上,稀疏-稠密变换单元200被配置为:接收对于稠密矩阵的请求,并且基于可由稠密元素访问单元$X_{1,1}$至$X_{M,N}$访问的对应的稀疏元素来确定稠密矩阵。大体上,每个稀疏元素访问单元被配置为访问指定的稀疏元素集合,并且在下面参照图3A-3B来更详细描述。在一些实施方式中,稀疏元素访问单元可以是单指令多数据(SIMD)处理设备。

[0026] 在一些实施方式中,稀疏元素访问单元$X_{1,1}$至$X_{M,N}$可以在物理上或逻辑上被布置为二维网状配置。例如,稀疏元素访问单元$X_{1,1}$直接耦合至稀疏元素访问单元$X_{1,2}$和$X_{2,1}$。作为另一个示例,稀疏元素访问单元$X_{2,2}$直接耦合至稀疏元素访问单元$X_{2,1}$、$X_{3,1}$、$X_{2,3}$、和$X_{1,2}$。两个稀疏元素访问单元之间的耦合可以是电连接、光连接、有线连接、或者任何其他合适的连接。

[0027] 在一些其他实施方式中,稀疏元素访问单元$X_{1,1}$至$X_{M,N}$可以在物理上或逻辑上被布置为二维环面(torus)配置。例如,稀疏元素访问单元$X_{1,1}$直接耦合至稀疏元素访问单元$X_{1,2}$、$X_{2,1}$、$X_{1,N}$和$X_{M,1}$。作为另一个示例,稀疏元素访问单元$X_{M,N}$直接耦合至稀疏元素访问单元$X_{M,N-1}$、$X_{M-1,N}$、$X_{M,1}$、和$X_{1,N}$。

[0028] 在一些实施方式中,稀疏-稠密变换单元200可以被配置为根据一组预定条件来对从稠密矩阵变换的稀疏元素进行划分(partition)。稀疏元素访问单元$X_{1,1}$至$X_{M,N}$中的每一行可以被划分以访问从特定稠密矩阵变换的稀疏元素。例如,稀疏-稠密变换单元200可以被配置为访问与计算机模型的1000个不同的数据库表相对应的稠密矩阵变换的稀疏元素。数据库表中的一个或多个可以具有不同的大小。稀疏元素访问单元的第一行202可以被配置为访问从数据库表No.1至数据库表No.100变换的稀疏元素,稀疏元素访问单元的第二行204可以被配置为访问从数据库表No.101至数据库表No.300变换的稀疏元素,并且稀疏元素访问单元的第M行206可以被配置为访问从数据库表No.751至数据库表No.1000变换的稀疏元素。在一些实施方式中,在处理器访问稀疏元素之前,可以使用稀疏-稠密变换单元200通过硬件指令来对所述划分进行配置。

[0029] 稀疏元素访问单元$X_{1,1}$至$X_{M,N}$中的每一列可以被划分以访问从特定稠密矩阵变换的稀疏元素的子集。例如,与数据库表No.1相对应的稠密矩阵可以被变换为1000个稀疏元素,其中该1000个稀疏元素如上所述可由第一行202访问。稀疏元素访问单元$X_{1,1}$可以被配置为访问数据库表No.1的稀疏元素No.1至No.200,并且稀疏元素访问单元$X_{1,2}$可以被配置为访问数据库表No.1的稀疏元素No.201至No.500。作为另一个示例,与数据库表No.2相对应的稠密矩阵可以被变换为500个稀疏元素,其中该500个稀疏元素如上所述可由第一行202访问。稀疏元素访问单元$X_{1,1}$可以被配置为访问数据库表No.2的稀疏元素No.1至No.50,

并且稀疏元素访问单元$X_{1,2}$可以被配置为访问数据库表No.2的稀疏元素No.51至No.200。作为另一个示例,与数据库表No.1000相对应的稠密矩阵可以被变换为10000个稀疏元素,其中该10000个稀疏元素如上所述可由第M行206访问。稀疏元素访问单元$X_{M,1}$可以被配置为访问数据库表No.1000的稀疏元素No.1至No.2000,并且稀疏元素访问单元$X_{M,N}$可以被配置为访问数据库表No.1000的稀疏元素No.9000至No.10000。

[0030] 图2B示出了稀疏-稠密变换单元200可以如何使用稀疏元素访问单元的二维网状网络来请求稀疏元素的示例。作为示例,处理单元可以执行指令以向稀疏-稠密变换单元200请求稠密一维向量,该稠密一维向量是使用数据库表No.1的稀疏元素No.1至No.50、数据库表No.2的稀疏元素No.100至No.200、以及数据库表No.1000的稀疏元素No.9050至No.9060来生成的。在稀疏-稠密变换单元200从该处理单元接收到该请求后,稀疏-稠密变换单元200可以指令稀疏元素访问单元$X_{1,1}$向网状网络中的其他稀疏元素访问单元广播对于所述稀疏元素的请求。稀疏元素访问单元$X_{1,1}$可以向稀疏元素访问单元$X_{1,2}$广播请求222并且向稀疏元素访问单元$X_{2,1}$广播请求224。在接收到请求222后,稀疏元素访问单元$X_{1,2}$可以向稀疏元素访问单元$X_{1,3}$广播请求226。在一些实施方式中,稀疏元素访问单元可以被配置为基于路由方案来向另一稀疏元素访问单元广播请求。例如,稀疏元素访问单元$X_{2,1}$可以不被配置为向稀疏元素访问单元$X_{2,2}$广播请求,因为稀疏元素访问单元$X_{2,2}$被配置为从稀疏元素访问单元$X_{2,1}$接收广播。路由方案可以是静态的或者被动态生成。例如,路由方案可以是查找表。在一些实施方式中,稀疏元素访问单元可以被配置为基于请求224来向另一稀疏元素访问单元广播请求224。例如,请求224可以包括对所请求的稀疏元素(例如数据库表No.1、稀疏元素No.1至No.50)的识别,并且稀疏元素访问单元$X_{1,2}$可以基于该识别来确定是否向稀疏元素访问单元$X_{2,2}$和/或稀疏元素访问单元$X_{1,3}$广播请求224。广播过程通过网状网络来传播,其中稀疏元素访问单元$X_{M,N}$从稀疏元素访问单元$X_{M,N-1}$接收请求230。

[0031] 图2C示出了稀疏-稠密变换单元200可以如何使用稀疏元素访问单元的二维网状网络来生成所请求的稠密矩阵的示例。在一些实施方式中,在稀疏元素访问单元接收到所广播的请求后,该稀疏元素访问单元被配置为确定是否其被配置来访问所请求的稀疏元素中的任何稀疏元素。例如,稀疏元素访问单元$X_{1,1}$可以确定其被配置来访问数据库表No.1的稀疏元素No.1至No.50,但其没有被配置来访问数据库表No.2的稀疏元素No.100至No.200或者数据库表No.1000的稀疏元素No.9050至No.9060。响应于确定其被配置来访问数据库表No.1的稀疏元素No.1至No.50,稀疏元素访问单元$X_{1,1}$可以从存储这些稀疏元素的数据片中提取数据库表No.1的稀疏元素No.1至No.50,并且基于这些稀疏元素来生成稠密矩阵242。

[0032] 作为另一个示例,稀疏元素访问单元$X_{2,1}$可以确定其没有被配置来访问数据库表No.1的稀疏元素No.1至No.50、数据库表No.2的稀疏元素No.100至No.200、以及数据库表No.1000的稀疏元素No.9050至No.9060中的任何稀疏元素。响应于确定其没有被配置来访问所请求的稀疏元素中的任何稀疏元素,稀疏元素访问单元$X_{2,1}$可以不执行进一步的动作。

[0033] 作为另一个示例,稀疏元素访问单元$X_{1,2}$可以确定其被配置来访问数据库表No.2的稀疏元素No.100至No.200,但是其没有被配置来访问数据库表No.1的稀疏元素No.1至No.50或者数据库表No.1000的稀疏元素No.9050至No.9060。响应于确定其被配置来访问数据库表No.2的稀疏元素No.100至No.200,稀疏元素访问单元$X_{1,2}$可以从存储这些稀疏元素

的数据片中提取这些稀疏元素,并且基于这些稀疏元素来生成稠密矩阵244。在一些实施方式中,在稀疏元素访问单元生成稠密矩阵后,该稀疏元素访问单元可以被配置为将该稠密矩阵转发至所广播的请求的发送者。此处,稀疏元素访问单元$X_{1,2}$将稠密矩阵244转发至稀疏元素访问单元$X_{1,1}$。

[0034] 作为另一个示例,稀疏元素访问单元$X_{M,N}$可以确定其被配置来访问数据库表No.1000的稀疏元素No.9050至No.9060,但是其没有被配置来访问数据库表No.1的稀疏元素No.1至No.50或者数据库表No.2的稀疏元素No.100至No.200。响应于确定其被配置来访问数据库表No.1000的稀疏元素No.9050至No.9060,稀疏元素访问单元$X_{M,N}$可以从存储这些稀疏元素的数据片中提取这些稀疏元素,并且基于这些稀疏元素来生成稠密矩阵246。在一些实施方式中,在稀疏元素访问单元生成稠密矩阵后,该稀疏元素访问单元可以被配置为将该稠密矩阵转发至所广播的请求的发送者。此处,稀疏元素访问单元$X_{M,N}$将稠密矩阵246转发至稀疏元素访问单元$X_{M,N-1}$。在下一循环中,稀疏元素访问单元$X_{M,N-1}$被配置为将稠密矩阵246转发至稀疏元素访问单元$X_{M,N-1}$。该过程继续,直到稀疏元素访问单元$X_{2,1}$已将稠密矩阵246转发至稀疏元素访问单元$X_{1,1}$为止。

[0035] 在一些实施方式中,稀疏-稠密变换单元200被配置为对由稀疏元素访问单元生成的稠密矩阵进行变换并且针对处理器单元生成稠密矩阵。此处,稀疏-稠密变换单元200针对处理器单元来将稠密矩阵242、244、和246变换为稠密矩阵。例如,稠密矩阵242可以具有100乘10的维度,稠密矩阵244可以具有20乘100的维度、并且稠密矩阵246可以具有3乘3的维度。稀疏-稠密变换单元200可以将稠密矩阵242、244、和246变换为具有1乘3009维度的向量。有利地,根据稠密矩阵(例如数据库表)的对行的划分允许稀疏-稠密变换单元200在所生成的稠密矩阵已经从列N传播至列1后获得全部所请求的稀疏元素。对列的划分减少了由仅使用稀疏元素访问单元中的一个来访问过多稀疏元素所导致的带宽瓶颈。

[0036] 图2D示出了稀疏-稠密变换单元200可以如何使用稀疏元素访问单元的二维网状网络基于稠密矩阵来更新稀疏元素的示例。作为示例,处理单元可以执行请求稀疏-稠密变换单元200稠密一维向量来更新所存储的稀疏元素,该稠密一维向量是使用数据库表No.1的稀疏元素No.1至No.50和数据库表No.1000的稀疏元素No.9050至No.9060来生成的。在稀疏-稠密变换单元200从处理单元接收到该请求后,稀疏-稠密变换单元200可以指令稀疏元素访问单元$X_{1,1}$向网状网络中的其他稀疏元素访问单元广播稀疏元素更新请求,其中该稀疏元素更新请求可以包括由该处理单元提供的稠密一维向量。在一些实施方式中,稀疏元素访问单元$X_{1,1}$可以确定是否其被指派来访问在该稠密一维向量中所包括的稀疏元素。响应于确定其被指派来访问该稠密一维向量中所包括的稀疏元素,稀疏元素访问单元$X_{1,1}$可以更新数据片中所存储的稀疏元素。此处,稀疏元素访问单元$X_{1,1}$确定其被指派来访问数据库表No.1的稀疏元素No.1至No.50,并且稀疏元素访问单元$X_{1,1}$执行指令以更新该数据片中的这些稀疏元素。

[0037] 稀疏元素访问单元$X_{1,1}$可以向稀疏元素访问单元$X_{1,2}$广播稀疏元素更新请求252并且向稀疏元素访问单元$X_{2,1}$广播稀疏元素更新请求254。在接收到稀疏元素更新请求252后,稀疏元素访问单元$X_{1,2}$可以确定其没有被指派来访问在稠密一维向量中所包括的稀疏元素。稀疏元素访问单元$X_{1,2}$向稀疏元素访问单元$X_{1,3}$广播请求256。广播过程通过网状网络来传播,其中稀疏元素访问单元$X_{M,N}$从稀疏元素访问单元$X_{M,N-1}$接收请求260。此处,稀疏元素访

问单元$X_{M,N}$确定其被指派来访问数据库表No.1000的稀疏元素No.9050至No.9060,并且稀疏元素访问单元$X_{M,N}$执行指令以更新数据片中的这些稀疏元素。

[0038] 图3A示出了示例稀疏元素访问单元300。稀疏元素访问单元300可以是稀疏元素访问单元$X_{1,1}$至$X_{M,N}$中的任何一个。大体上,稀疏元素访问单元300被配置为从节点网络320接收请求342以提取在一个或多个数据片中存储的稀疏元素并且将所提取的稀疏元素变换为稠密矩阵。在一些实施方式中,处理单元316向节点网络320中的稀疏元素访问单元发送对于使用稀疏元素生成的稠密矩阵的请求。稀疏元素访问单元可以向稀疏元素访问单元300广播请求342。对所广播的请求342的路由可以类似于图2B中的描述。稀疏元素访问单元300包括请求识别单元302、数据提取单元304、稀疏降低单元306、串接单元308、压缩/解压缩单元310、以及分裂单元312。节点网络320可以是二维网状网络。处理单元316可以类似于处理单元102。

[0039] 大体上,请求识别单元302被配置为接收请求342以提取在一个或多个数据片330中存储的稀疏元素,并且确定稀疏元素访问单元300是否被指派为访问由请求342指示的稀疏元素。在一些实施方式中,请求识别单元302可以通过使用查找表来确定稀疏元素访问单元300是否被指派为访问由请求342指示的稀疏元素。例如,如果对特定的所请求稀疏元素(例如数据库表No.1的No.1)的识别被包括在查找表中,则请求识别单元302可以向数据提取单元304发送信号344以提取该特定的所请求稀疏元素。如果对特定的所请求稀疏元素(例如数据库表No.1的No.1)的识别没有被包括在查找表中,则请求识别单元302可以丢弃所接收的请求。在一些实施方式中,请求识别单元302可以被配置为向节点网络320上的另一稀疏元素访问单元广播所接收的请求。

[0040] 数据提取单元304被配置为响应于接收到信号344而从数据片330中提取一个或多个所请求的稀疏元素。在一些实施方式中,数据提取单元304包括一个或多个处理器322a-322k,其中k为整数。处理器322a-322k可以是向量处理单元(VPU)、阵列处理单元、或者任何合适的处理单元。在一些实施方式中,处理器322a-322k被布置在数据片330附近以降低处理器322a-322k与数据片330之间的时延。基于稀疏元素访问单元300被指派来提取的所请求的稀疏元素的数目,数据提取单元304可以被配置为生成一个或多个请求以在处理器322a-322k当中分布。在一些实施方式中,处理器322a-322k中的每一个可以基于对稀疏元素的识别来被指派至特定稀疏元素,并且数据提取单元304可以被配置为基于对稀疏元素的识别来生成对于处理器322a-322k的一个或多个请求。在一些实施方式中,数据提取单元304可以通过使用查找表来确定处理器指派。在一些实施方式中,数据提取单元304可以针对处理器322a-322k而生成多个批次,其中每个批次是对于所请求的稀疏元素的子集的一个请求。处理器322a-322k被配置为从数据片330独立地提取所指派的稀疏元素,并且将所提取的稀疏元素346转发至稀疏降低单元306。

[0041] 稀疏降低单元306被配置来降低所提取的稀疏元素346的维度。例如,处理器322a-322k中的每一个可以生成具有100乘1的维度的稀疏元素。稀疏降低单元306可以接收具有100乘k的维度的所提取的稀疏元素346,并且通过利用逻辑操作、数学操作、或者这两者的组合来将所提取的稀疏元素346的维度降低为100乘1来生成稀疏降低的元素348。稀疏降低单元306被配置为向串接单元308输出稀疏降低的元素348。

[0042] 串接单元308被配置为重新布置并串接稀疏降低的元素348以生成所串接元素

350。例如,稀疏元素访问单元$X_{1,1}$可以被配置为访问数据库表No.1的稀疏元素No.1至No.200。处理器322a可以比被配置来返回所提取的稀疏元素No.5的处理器322b更快地向稀疏降低单元306返回所提取的稀疏元素No.10。串接单元308被配置为对较晚接收到的稀疏元素No.5重新布置以使其被排序在较早接收到的稀疏元素No.10之前,并且将稀疏元素No.1至No.200串接为串接元素350。

[0043] 压缩/解压缩单元310被配置为压缩串接元素350以针对节点网络320生成稠密矩阵352。例如,压缩/解压缩单元310可以被配置为对所串接元素350中的零值进行压缩以改善节点网络320的带宽。在一些实施方式中,压缩/解压缩单元310可以对所接收的稠密矩阵解压缩。例如,稀疏元素访问单元300可以经由节点网络320来从邻近稀疏元素访问单元接收稠密矩阵。稀疏元素访问单元300可以对所接收的稠密矩阵解压缩,并且可以将已解压缩的稠密矩阵与串接元素350串接以形成更新的串接元素,其能够被压缩并且之后被输出至节点网络320。

[0044] 图3B示出了稀疏元素访问单元300可以如何基于从节点网络320接收的稠密矩阵来更新稀疏元素的示例。作为示例,处理单元可以执行请求稀疏-稠密变换单元使用稠密一维向量来而更新所存储的稀疏元素的指令,该稠密一维向量是使用数据库表No.1的稀疏元素No.1至No.50和数据库表No.1000的稀疏元素No.9050至No.9060来生成的。在稀疏-稠密变换单元从处理单元接收到该请求后,该稀疏-稠密变换单元可以发送请求362以指令稀疏元素访问单元300确定是否其被指派来访问在稠密一维向量中所包括的稀疏元素。请求识别单元302被配置为确定是否稀疏元素访问单元300被指派来访问在稠密一维向量中所包括的稀疏元素。响应于确定稀疏元素访问单元300被指派来访问在稠密一维向量中所包括的稀疏元素,请求识别单元302可以向分裂单元312发送指示364以更新在数据片中存储的稀疏元素。

[0045] 分裂单元312被配置来将所接收的稠密矩阵变换为稀疏元素,该稀疏元素能够由数据提取单元304在数据片330中更新。例如,分裂单元312可以被配置为将稠密一维向量变换为多个稀疏元素,并且指令数据提取单元304更新稀疏元素访问单元300被指派来提取的数据片330中存储的稀疏元素。

[0046] 图4是图示了用于生成稠密矩阵的过程400的示例的流程图。过程400可以由诸如稀疏-稠密变换单元104或稀疏-稠密变换单元200的系统来执行。系统可以包括第一稀疏元素访问单元群组和第二稀疏元素访问单元群组。例如,参照图2A,稀疏-稠密变换单元200可以包括在物理上或逻辑上被布置为M行和N列的M乘N稀疏元素访问单元$X_{1,1}$至$X_{M,N}$。稀疏元素访问单元$X_{1,1}$至$X_{M,N}$中的每一行可以被划分以访问从特定稠密矩阵变换的稀疏元素。在一些实施方式中,第一稀疏元素访问单元群组可以包括第一稀疏元素访问单元和第二稀疏元素访问单元。例如,稀疏-稠密变换单元200的第一行可以包括稀疏元素访问单元$X_{1,1}$和$X_{1,2}$。在一些实施方式中,第一稀疏元素访问单元群组和第二稀疏元素访问单元群组可以以二维网状配置来布置。在一些实施方式中,第一稀疏元素访问单元群组和第二稀疏元素访问单元群组可以以二维环面配置来布置。

[0047] 系统接收对于基于稀疏元素的输出矩阵的请求,所述稀疏元素包括与第一稠密矩阵相关联的稀疏元素和与第二稠密矩阵相关联的稀疏元素。例如,参照图2B,处理单元可以执行向稀疏-稠密变换单元200请求稠密一维向量的指令,该稠密一维向量是使用数据库表

No.1的稀疏元素No.1至No.50、数据库表No.2的稀疏元素No.100至No.200、以及数据库表No.1000的稀疏元素No.9050至No.9060来生成的。

[0048] 在一些实施方式中,第一稀疏元素访问单元可以接收对于多个稀疏元素的请求,所述多个稀疏元素包括与第一稠密矩阵相关联的稀疏元素和与第二稠密矩阵相关联的稀疏元素。第一稀疏元素访问单元可以将该请求传送至第二稀疏元素访问单元。例如,参照图2B,在稀疏-稠密变换单元200从处理单元接收到该请求后,稀疏-稠密变换单元200可以指令稀疏元素访问单元$X_{1,1}$向网状网络中的其他稀疏元素访问单元广播对于所述稀疏元素的请求。稀疏元素访问单元$X_{1,1}$可以向稀疏元素访问单元$X_{1,2}$广播请求222。

[0049] 系统获得第一稀疏元素访问单元群组所提取的、与第一稠密矩阵相关联的稀疏元素(402)。在一些实施方式中,第一稀疏元素访问单元可以确定:所述多个稀疏元素中的特定稀疏元素的身份匹配与第一稠密矩阵相关联的稀疏元素的第一子集中的一个稀疏元素的身份。例如,参照图2C,稀疏元素访问单元$X_{1,1}$可以被配置为访问数据库表No.1的稀疏元素No.1至No.200。稀疏元素访问单元$X_{1,1}$可以确定其被配置为访问数据库表No.1的稀疏元素No.1至No.50,但是其没有被配置来访问数据库表No.2的稀疏元素No.100至No.200或者数据库表No.1000的稀疏元素No.9050至No.9060。响应于确定所述多个稀疏元素中的特定稀疏元素的身份匹配与第一稠密矩阵相关联的稀疏元素的第一子集中的一个稀疏元素的身份,第一稀疏元素访问单元可以提取与包括该特定稀疏元素的第一稠密矩阵相关联的稀疏元素的第一子集。例如,响应于确定其被配置来访问数据库表No.1的稀疏元素No.1至No.50,稀疏元素访问单元$X_{1,1}$可以从存储这些稀疏元素的数据片中提取数据库表No.1的稀疏元素No.1至No.50。

[0050] 第二稀疏元素访问单元可以提取与第一稠密矩阵相关联的稀疏元素的不同的第二子集。例如,参照图2C,稀疏元素访问单元$X_{1,2}$可以被配置来访问数据库表No.2的稀疏元素No.51至No.200。响应于确定其被配置来访问数据库表No.2的稀疏元素No.100至No.200,稀疏元素访问单元$X_{1,2}$可以从存储这些稀疏元素的数据片中提取这些稀疏元素。

[0051] 系统获得由第二稀疏元素访问单元群组提取的、与第二稠密矩阵相关联的稀疏元素(404)。例如,参照图2C,第二稀疏元素访问单元群组可以是M乘N稀疏元素访问单元的第M行,其中稀疏元素访问单元$X_{M,N}$可以被配置为访问数据库表No.1000的稀疏元素No.9000至No.10000。响应于确定其被配置来访问数据库表No.1000的稀疏元素No.9050至No.9060,稀疏元素访问单元$X_{M,N}$可以从存储这些稀疏元素的数据片中提取这些稀疏元素,并且基于这些稀疏元素来生成稠密矩阵246。

[0052] 在一些实施方式中,第一稀疏元素访问单元可以从第一数据片提取与第一稠密矩阵相关联的稀疏元素的第一子集,并且第二稀疏元素访问单元可以从不同的第二数据片提取与第一稠密矩阵相关联的稀疏元素的不同的第二子集。例如,参照图1,第一稀疏元素访问单元可以从数据片106a提取与第一稠密矩阵相关联的稀疏元素的第一子集,并且第二稀疏元素访问单元可以从数据片106b提取与第一稠密矩阵相关联的稀疏元素的不同的第二子集。

[0053] 系统对与第一稠密矩阵相关联的稀疏元素和与第二稠密矩阵相关联的稀疏元素进行变换以生成包括与第一稠密矩阵相关联的稀疏元素和与第二稠密矩阵相关联的稀疏元素的输出稠密矩阵(406)。例如,参照图2C,稀疏-稠密变换单元200针对处理器单元来将

稠密矩阵242、244、和246变换为一个稠密矩阵。

[0054] 在一些实施方式中,与第一稠密矩阵相关联的稀疏元素和与第二稠密矩阵相关联的稀疏元素可以是多维矩阵,并且输出稠密矩阵可以是向量。例如,稠密矩阵242可以具有100乘10的维度,稠密矩阵244可以具有20乘100的维度、并且稠密矩阵246可以具有3乘3的维度。稀疏-稠密变换单元200可以将稠密矩阵242、244、和246变换为具有1乘3009维度的向量。

[0055] 图5是图示了用于生成稠密矩阵的过程500的示例的流程图。过程500可以由诸如稀疏-稠密变换单元104或稀疏元素访问单元300的系统来执行。

[0056] 系统接收访问特定稀疏元素的子集的指示(502)。例如,参照图3A,数据提取单元304可以被配置为接收信号344以用于从数据片330提取一个或多个所请求的稀疏元素。在一些实施方式中,对于存储在一个或多个数据片中的特定稀疏元素的请求可以通过节点网络来接收。例如,参照图3A,请求识别单元302可以被配置为通过节点网络320来接收请求342以提取在数据片330中存储的稀疏元素。系统可以确定该数据提取单元被指派来处置(handle)该特定稀疏元素的子集。例如,请求识别单元302可以被配置为确定稀疏元素访问单元300是否被指派为访问由请求342指示的稀疏元素。响应于确定该数据提取单元被指派来处置该特定稀疏元素的子集,该指示可以被生成以用于访问该特定稀疏元素的子集。例如,如果对特定的所请求稀疏元素(例如数据库表No.1的No.1)的识别被包括在查找表中,则请求识别单元302可以向数据提取单元304发送信号344以提取该特定的所请求稀疏元素。

[0057] 系统基于对该特定稀疏元素的子集的识别来确定用于提取该特定稀疏元素的子集的处理器指定(504)。例如,参照图3A,数据提取单元304包括一个或多个处理器322a-322k。处理器322a-322k中的每一个可以基于对稀疏元素的识别来被指派至特定稀疏元素,并且数据提取单元304可以被配置为基于对稀疏元素的识别来生成对于处理器322a-322k的一个或多个请求。在一些实施方式中,系统可以确定该系统被指派来处置特定稀疏元素的子集包括:基于查找表来确定该系统被指派来处置该特定稀疏元素的子集。例如,数据提取单元304可以通过使用查找表来确定处理器指派。

[0058] 系统基于该指定由多个处理器中的第一处理器来提取该特定稀疏元素的子集中的第一稀疏元素(506)。例如,参照图3A,数据提取单元304可以指令处理器322a提取在信号344中包括的稀疏元素。

[0059] 系统基于该指定由多个处理器中的第二处理器来提取该特定稀疏元素的子集中的第二稀疏元素(508)。例如,参照图3A,数据提取单元304可以指令处理器322b提取在信号344中包括的不同的稀疏元素。

[0060] 在一些实施方式中,可以接收包括来自第一处理器的第一稀疏元素的第一矩阵,其中该第一矩阵可以具有第一维度。系统可以生成包括该第一稀疏元素的第二矩阵,该第二矩阵可以具有小于该第一维度的第二维度。例如,稀疏降低单元306可以被配置为降低所提取的稀疏元素346的维度。处理器322a-322k中的每一个可以生成具有100乘1的维度的稀疏元素。稀疏降低单元306可以接收具有100乘k的维度的所提取的稀疏元素346,并且通过利用逻辑操作、数学操作、或者这两者的组合来将所提取的稀疏元素346的维度降低为100乘1来生成稀疏降低的元素348。系统可以生成输出稠密矩阵,该输出稠密矩阵可以基于该

第二矩阵来被生成。例如,串接单元308可以被配置为重新布置并串接稀疏降低的元素348以生成串接元素350。

[0061] 在一些实施方式中,可以在第一时间点接收第一稀疏元素,并且可以在不同的第二时间点接收第二稀疏元素。系统可以针对该输出稠密矩阵来确定第一稀疏元素和第二稀疏元素的顺序。例如,参照图3A,处理器322a可以比被配置来返回所提取的稀疏元素No.5的处理器322b更快地向稀疏降低单元306返回所提取的稀疏元素No.10。串接单元308被配置为对较晚接收到的稀疏元素No.5重新布置以使其被排序在较早接收到的稀疏元素No.10之前,并且将稀疏元素No.1至No.200串接为串接元素350。

[0062] 系统基于至少被应用于第一稀疏元素和第二稀疏元素的变换来生成输出稠密矩阵(510)。在一些实施方式中,系统可以压缩该输出稠密矩阵以生成压缩输出稠密矩阵。系统可以向节点网络提供该压缩输出稠密矩阵。例如,压缩/解压缩单元310可以被配置来压缩串接元素350以生成针对节点网络320的稠密矩阵352。

[0063] 在一些实施方式中,系统可以接收表示通过节点网络所发送的稠密矩阵的第一稠密矩阵,并且基于该第一稠密矩阵、第一稀疏元素、和第二稀疏元素来生成输出稠密矩阵。例如,稀疏元素访问单元300可以经由节点网络320来从邻近的稀疏元素访问单元接收稠密矩阵。稀疏元素访问单元300可以对所接收的稠密矩阵解压缩,并且可以将已解压缩的稠密矩阵与串接元素350串接以形成更新的所串接元素,其能够被压缩并且之后被输出至节点网络320。

[0064] 在一些实施方式中,特定稀疏元素中的一个或多个稀疏元素是多维矩阵,并且输出稠密矩阵是向量。本说明书中描述的主题和功能性操作的实施例能够以数字电子电路、以有形体现的计算机软件或固件、以硬件——包括本说明书中公开的结构及其结构等同物、或者以它们中的一个或多个的组合来实现。本说明书中描述的主题能够被实现为一个或多个计算机程序,即计算机程序指令的一个或多个模块,其被编码在有形非暂时性程序载体上,以用于由数据处理装置执行或控制数据处理装置的操作。替选地或另外地,程序指令能够被编码在人工生成的传播信号上,该信号例如机器生成的电、光、或电磁信号,其被生成以对信息进行编码以用于传输到合适的接收器装置以供数据处理装置执行。计算机存储介质能够是机器可读存储设备、机器可读存储基底、随机或串行存取存储器设备、或它们中的一个或多个的组合。

[0065] 术语“数据处理装置”涵盖用于处理数据的各种装置、设备、和机器,包括例如可编程处理器、计算机、或者多个处理器或计算机。该装置能够包括专用逻辑电路,例如FPGA(现场可编程门阵列)或ASIC(专用集成电路)。除了硬件之外,该装置还能够包括创建用于所讨论的计算机程序的执行环境的代码,例如构成处理器固件、协议栈、数据库管理系统、操作系统、或它们中的一个或多个的组合的代码。

[0066] 计算机程序(也称为程序、软件、软件应用、模块、软件模块、脚本、或代码)能够以任何形式的编程语言编写,包括编译或解释语言、声明性或过程性语言,并且其能够以任何形式部署,包括作为独立程序或作为适于在计算环境中使用的模块、组件、子例程、或其它单元。计算机程序可以但不必对应于文件系统中的文件。程序能够被存储在保持其他程序或数据——例如存储在标记语言文档中的一个或多个脚本——的文件的一部分中、专用于所讨论的程序的单个文件中、或者多个协作文件——例如存储一个或多个模块、子程序、或

代码部分的文件中。计算机程序能够被部署为在一个计算机上或在位于一个地点或跨多个地点分布并通过通信网络互连的多个计算机上执行。

[0067] 本说明书中所描述的过程和逻辑流程能够由一个或多个可编程计算机来执行,所述可编程计算机执行一个或多个计算机程序以通过在输入数据上操作并且生成输出来执行功能。该过程和逻辑流程也能够由专用逻辑电路执行并且装置也能够被实现为专用逻辑电路,所述专用逻辑电路例如FPGA(现场可编程门阵列)、ASIC(专用集成电路)、或者GPGPU(通用图形处理单元)。

[0068] 适合于执行计算机程序的处理器包括例如通用微处理器或专用微处理器或者这两者,或者任何其他种类的中央处理单元。通常,中央处理单元将从只读存储器或随机存取存储器或这两者接收指令和数据。计算机的必需元件是用于实施或执行指令的中央处理单元和用于存储指令和数据的一个或多个存储器设备。通常,计算机还将包括用于存储数据的一个或多个大容量存储设备——例如磁盘、磁光盘、或光盘,或者可操作地耦合以从其接收数据或向其传输数据、或者这两者。然而,计算机不需要具有这样的设备。此外,计算机能够被嵌入另一设备中,例如移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏控制台、全球定位系统(GPS)接收器、或便携式存储设备——例如通用串行总线(USB)闪存驱动器,仅举数例。

[0069] 适合于存储计算机程序指令和数据的计算机可读介质包括所有形式的非易失性存储器、介质和存储器设备,包括例如半导体存储器设备,例如EPROM、EEPROM、和闪速存储器设备;磁盘,例如内部硬盘或可移动盘;磁光盘;以及CD ROM和DVD-ROM盘。处理器和存储器能够由专用逻辑电路补充或并入专用逻辑电路中。

[0070] 为了提供与用户的交互,本说明书中描述的主题的实施方式能够被实现在具有以下的计算机上:用于向用户显示信息的显示设备——例如CRT(阴极射线管)或LCD(液晶显示器)监视器;以及键盘和指示设备——例如鼠标或轨迹球,用户通过其能够向计算机提供输入。其他种类的设备也能够被用于提供与用户的交互;提供给用户的反馈能够是任何形式的感官反馈,例如视觉反馈、听觉反馈、或触觉反馈;并且能够接收来自用户的处于任何形式的输入,包括声音、语音、或触觉输入。此外,计算机能够通过向用户使用的设备发送文档以及从其接收文档来与该用户交互,例如通过响应于从用户的客户端设备上的web浏览器接收的请求而向该web浏览器发送网页。

[0071] 在本说明书中描述的主题的实施例能够被实现在计算系统中,该计算系统包括例如作为数据服务器的后端组件,或者包括例如应用服务器的中间件组件,或者包括例如具有图形用户界面或Web浏览器的客户端计算机的前端组件——用户通过该Web浏览器能够与本说明书中所描述的主题的实施方式交互,或者包括一个或多个这样的后端组件、中间件组件、或前端组件的任何组合。系统的组件能够通过例如通信网络的任何形式或介质的数字数据通信来互连。通信网络的示例包括局域网(“LAN”)和广域网(“WAN”),例如互联网。

[0072] 计算系统能够包括客户端和服务器。客户端和服务器通常彼此远离并且一般通过通信网络交互。客户端和服务器的关系依靠在相应计算机上运行并且彼此具有客户端-服务器关系的计算机程序而产生。

[0073] 虽然本说明书包含许多具体实施方式细节,但是这些不应被解释为对任何发明或可以要求保护的内容的范围的限制,而是应当被解释为对特定发明的特定实施例特有的特

征的描述。本说明书中在单独实施例的场境下所描述的某些特征也能够在单个实施例中组合实现。相反,在单个实施例场境下所描述的各个特征也能够在多个实施例中单独地或以任何合适的子组合来实现。此外,虽然上面可能将特征描述为以某些组合来起作用并且甚至最初如此要求保护,但是在一些情况下来自所要求保护的组合的一个或多个特征能够从组合中被删去,并且所要求保护的组合可以涉及子组合或子组合的变型。

[0074] 类似地,虽然在附图中以特定顺序描绘了操作,但是这不应被理解为需要以所示的特定顺序或以连续顺序来执行这样的操作、或者需要执行所有所图示的操作才能达到期望的结果。在某些情况下,多任务和并行处理可以是有利的。此外,上述实施例中的各种系统组件的分离不应被理解为在所有实施例中都需要这样的分离,并且应当理解,所描述的程序组件和系统通常能够一起集成在单个软件产品中或封装到多个软件产品内。

[0075] 已经描述了本主题的特定实施例。其他实施例落入所附权利要求书的范围内。例如,权利要求书中所记载的动作能够以不同的顺序执行并且仍然达到期望的结果。作为一个示例,附图中描绘的过程不一定需要所示的特定顺序或连续顺序来达到期望的结果。在某些实施方式中,多任务和并行处理可以是有利的。

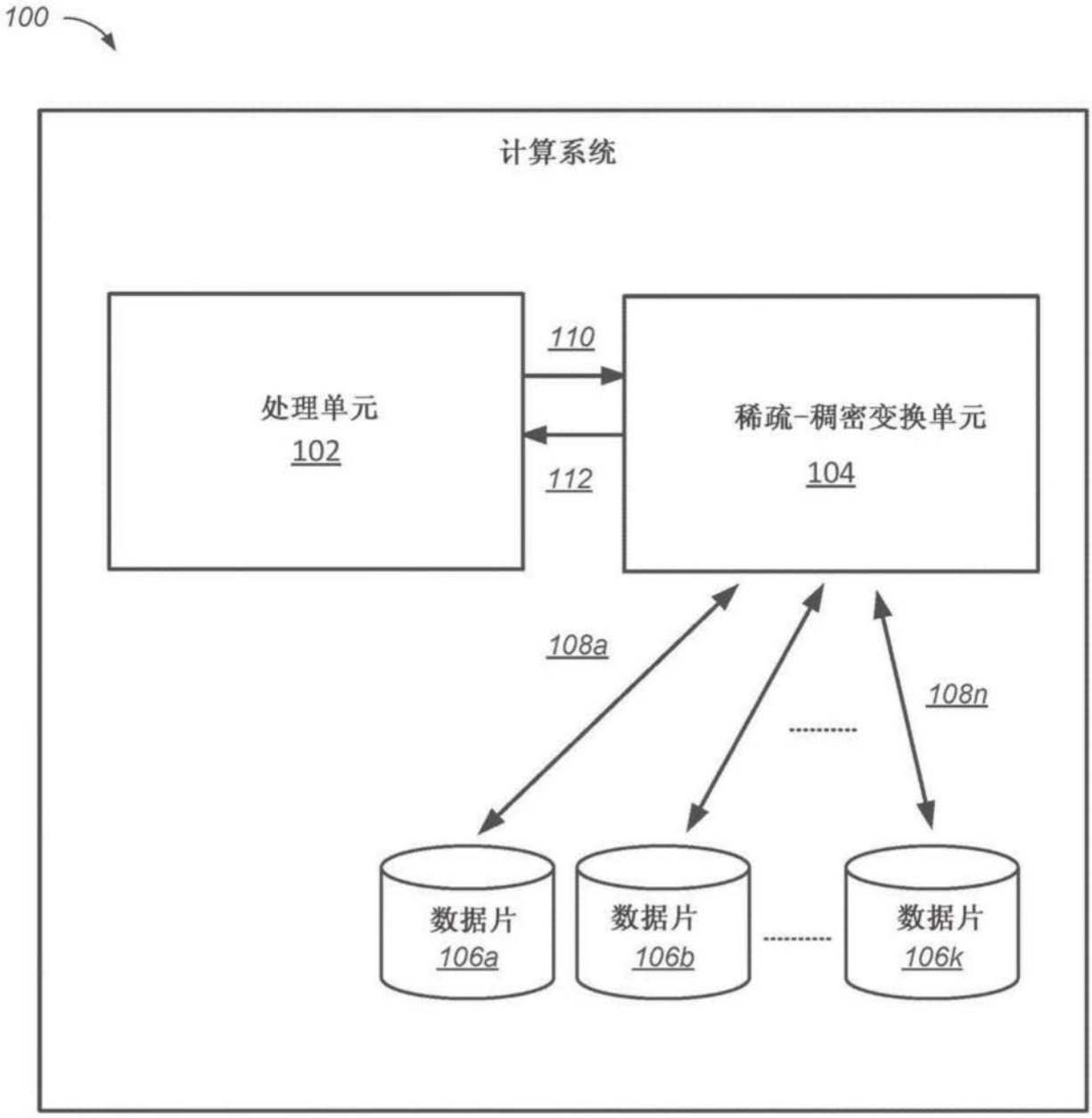

100
计算系统
处理单元
102
110
112
稀疏-稠密变换单元
104
108a
108n
数据片
106a
数据片
106b
数据片
106k

图1

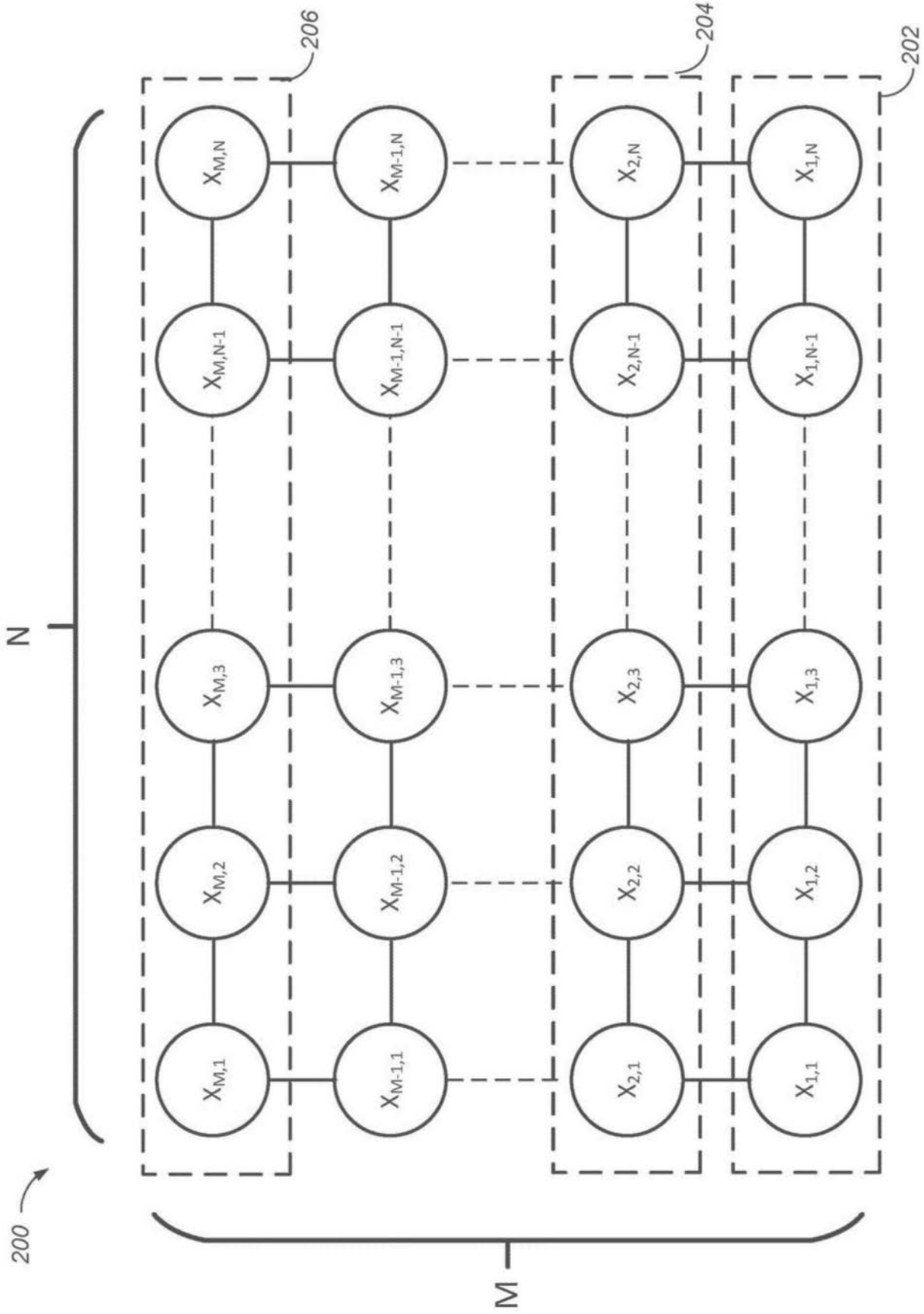

206
204
202
N
M
200
$X_{M,N}$
$X_{M-1,N}$
$X_{2,N}$
$X_{1,N}$
$X_{M,N-1}$
$X_{M-1,N-1}$
$X_{2,N-1}$
$X_{1,N-1}$
$X_{M,3}$
$X_{M-1,3}$
$X_{2,3}$
$X_{1,3}$
$X_{M,2}$
$X_{M-1,2}$
$X_{2,2}$
$X_{1,2}$
$X_{M,1}$
$X_{M-1,1}$
$X_{2,1}$
$X_{1,1}$

图2A

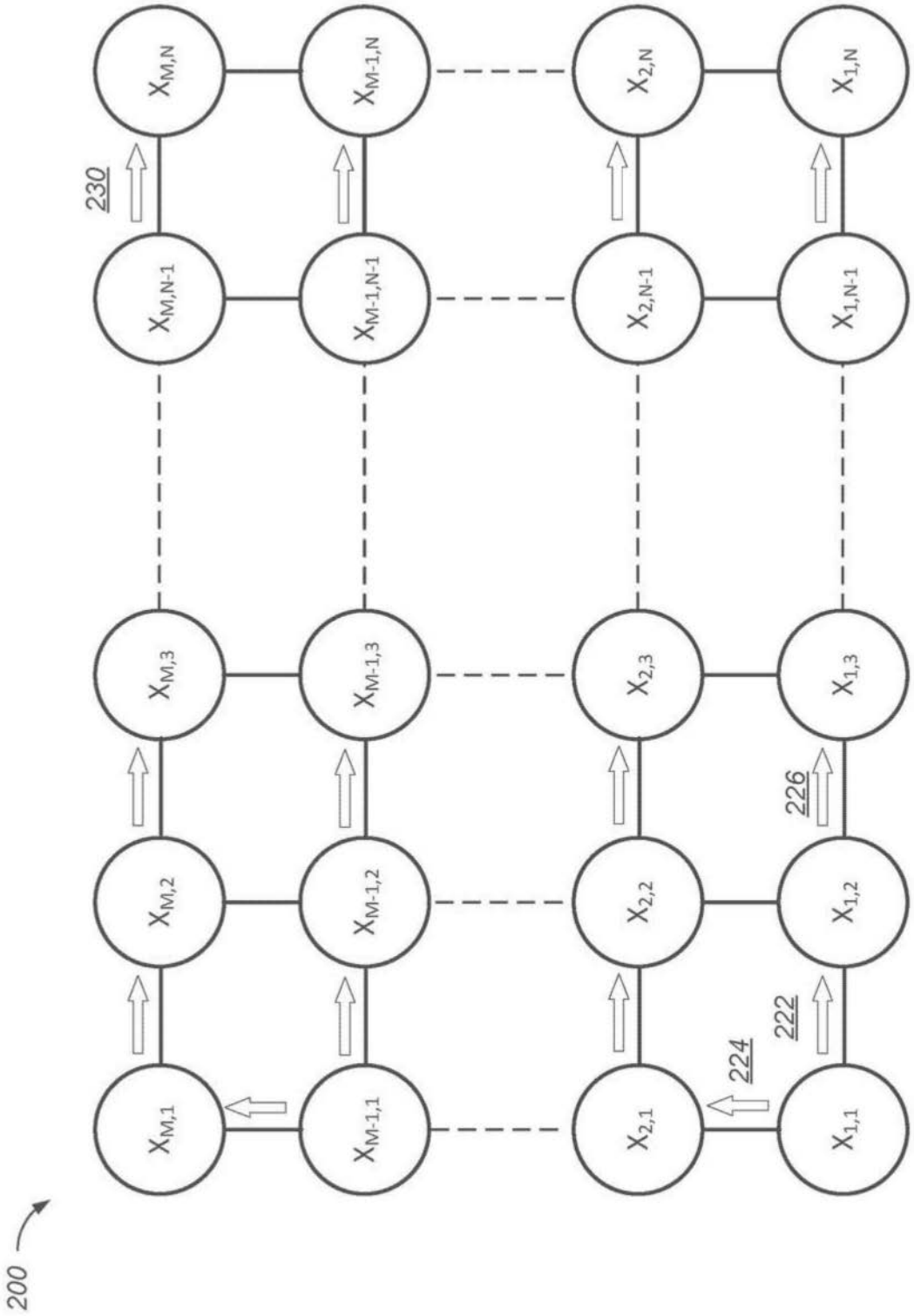

200
230
222
224
226
$X_{M,N}$
$X_{M-1,N}$
$X_{2,N}$
$X_{1,N}$
$X_{M,N-1}$
$X_{M-1,N-1}$
$X_{2,N-1}$
$X_{1,N-1}$
$X_{M,3}$
$X_{M-1,3}$
$X_{2,3}$
$X_{1,3}$
$X_{M,2}$
$X_{M-1,2}$
$X_{2,2}$
$X_{1,2}$
$X_{M,1}$
$X_{M-1,1}$
$X_{2,1}$
$X_{1,1}$

图2B

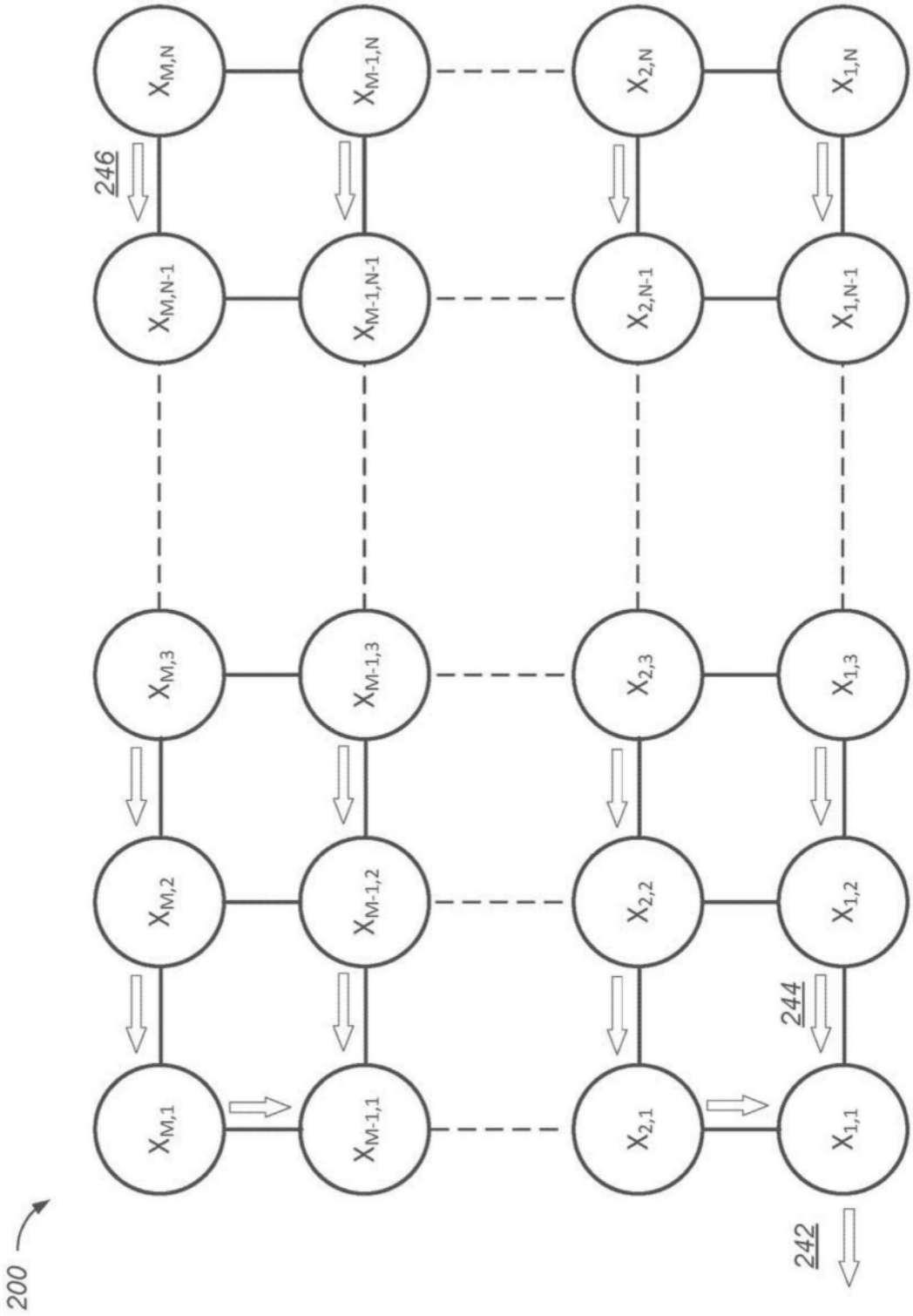

$X_{M,N}$
$X_{M-1,N}$
$X_{2,N}$
$X_{1,N}$
246
$X_{M,N-1}$
$X_{M-1,N-1}$
$X_{2,N-1}$
$X_{1,N-1}$
$X_{M,3}$
$X_{M-1,3}$
$X_{2,3}$
$X_{1,3}$
$X_{M,2}$
$X_{M-1,2}$
$X_{2,2}$
$X_{1,2}$
244
$X_{M,1}$
$X_{M-1,1}$
$X_{2,1}$
$X_{1,1}$
242
200

图2C

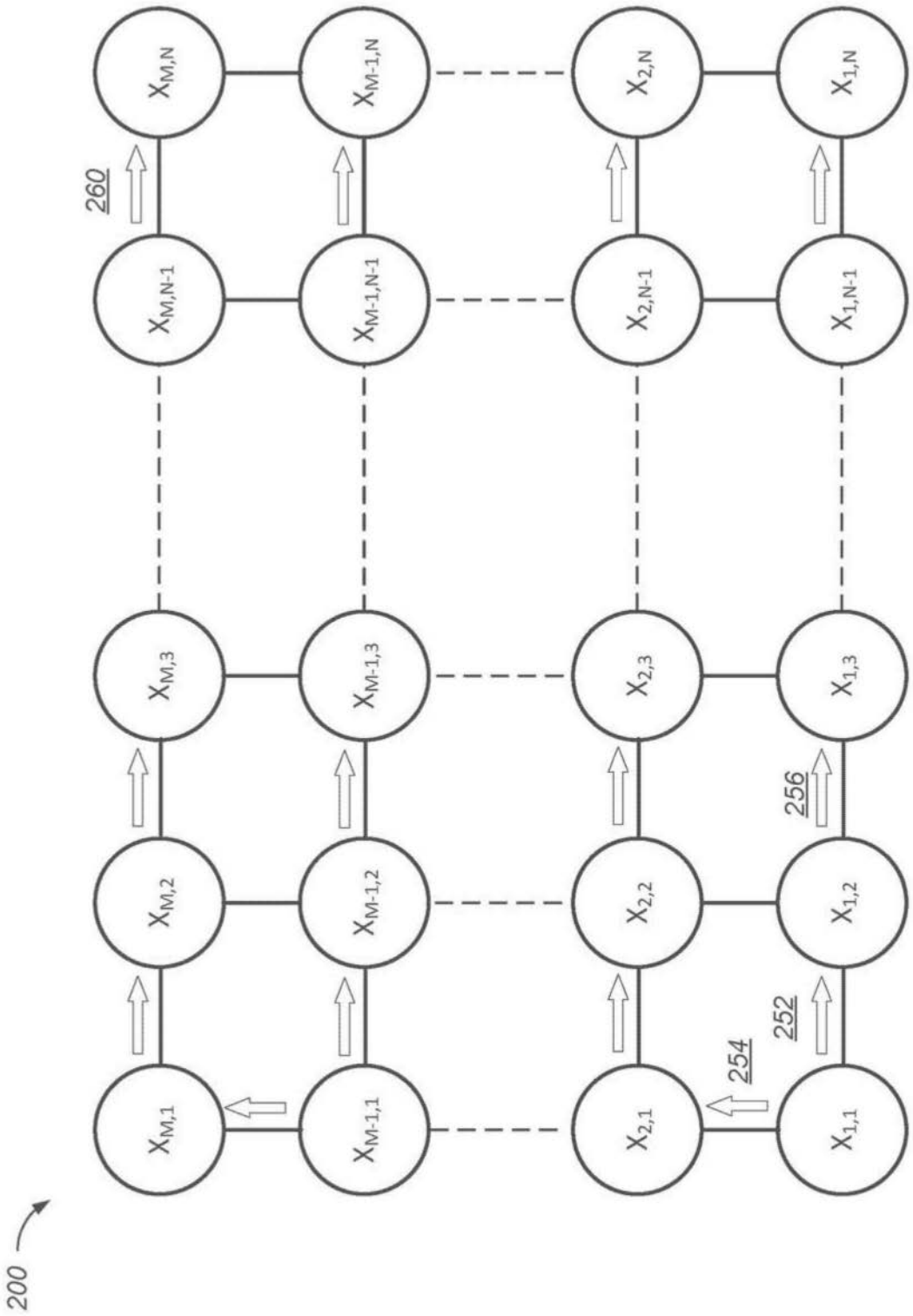

$X_{M,N}$
$X_{M-1,N}$
$X_{2,N}$
$X_{1,N}$
260
$X_{M,N-1}$
$X_{M-1,N-1}$
$X_{2,N-1}$
$X_{1,N-1}$
$X_{M,3}$
$X_{M-1,3}$
$X_{2,3}$
$X_{1,3}$
256
$X_{M,2}$
$X_{M-1,2}$
$X_{2,2}$
$X_{1,2}$
252
254
$X_{M,1}$
$X_{M-1,1}$
$X_{2,1}$
$X_{1,1}$
200

图2D

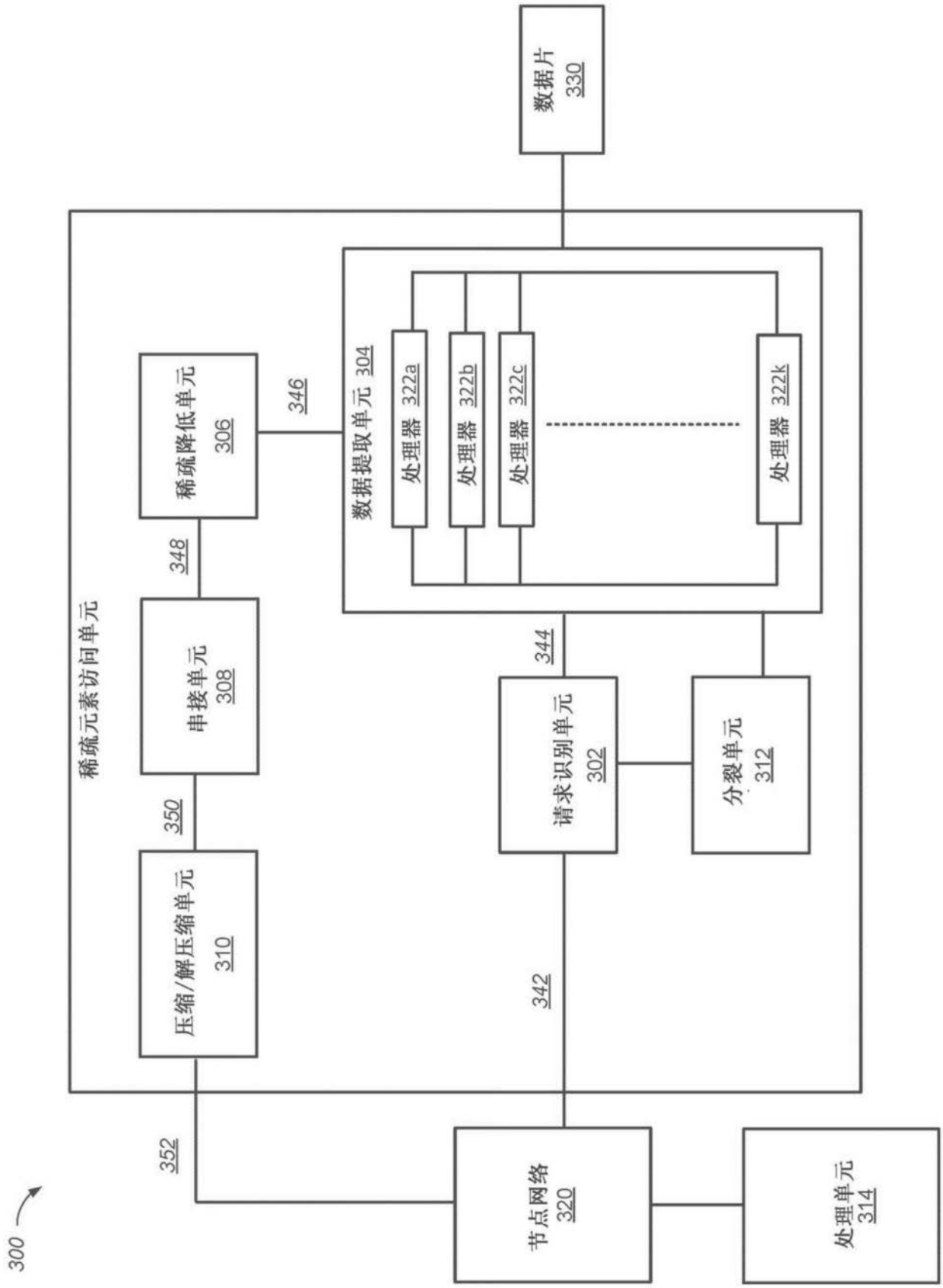

300
数据片 330
稀疏元素访问单元
数据提取单元 304
处理器 322a
处理器 322b
处理器 322c
处理器 322k
稀疏降低单元 306
346
348
串接单元 308
350
压缩/解压缩单元 310
344
请求识别单元 302
分裂单元 312
342
352
节点网络 320
处理单元 314

图3A

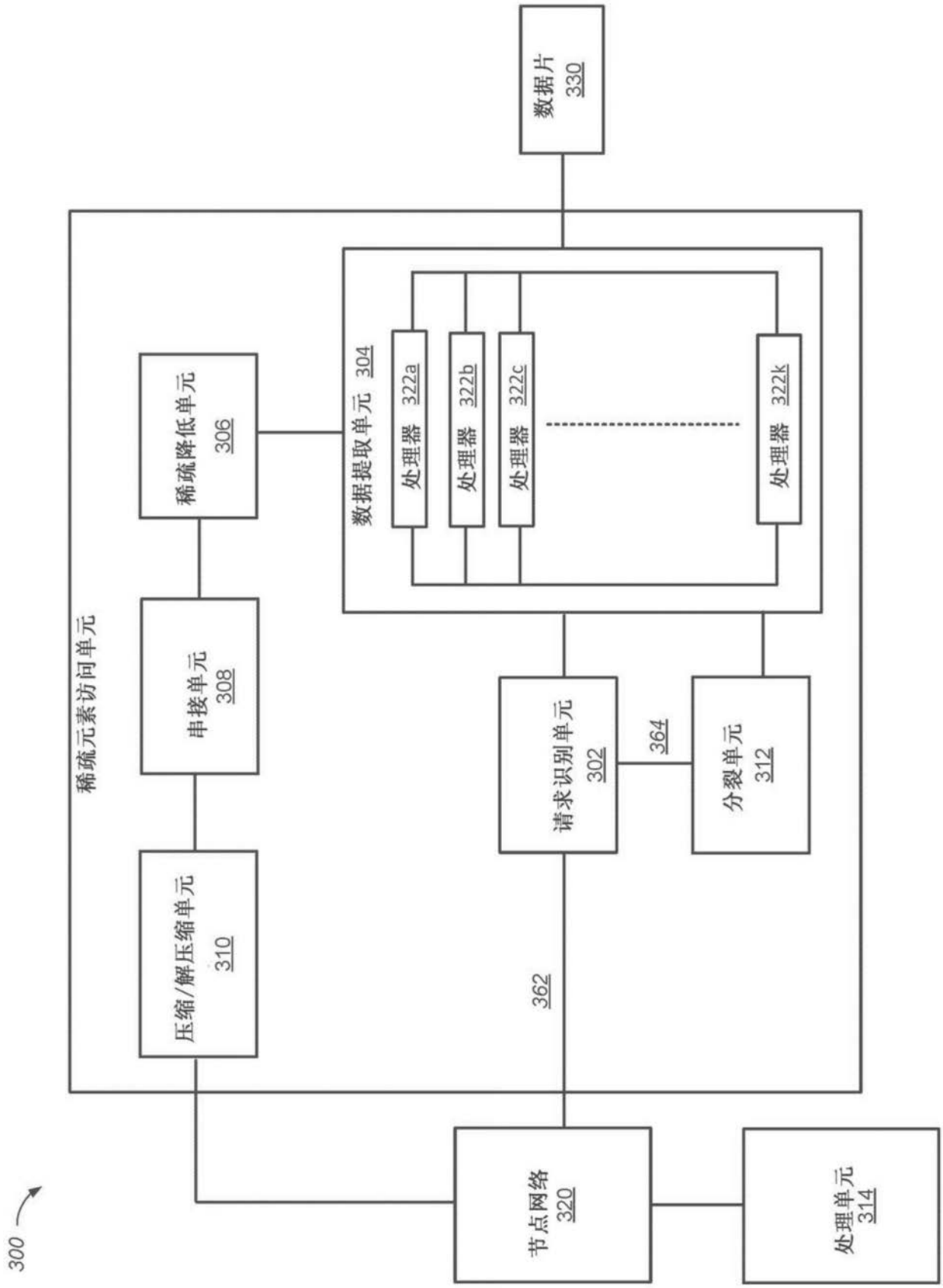

300
稀疏元素访问单元
数据片 330
数据提取单元 304
处理器 322a
处理器 322b
处理器 322c
处理器 322k
稀疏降低单元 306
串接单元 308
压缩/解压缩单元 310
请求识别单元 302
364
分裂单元 312
362
节点网络 320
处理单元 314

图3B

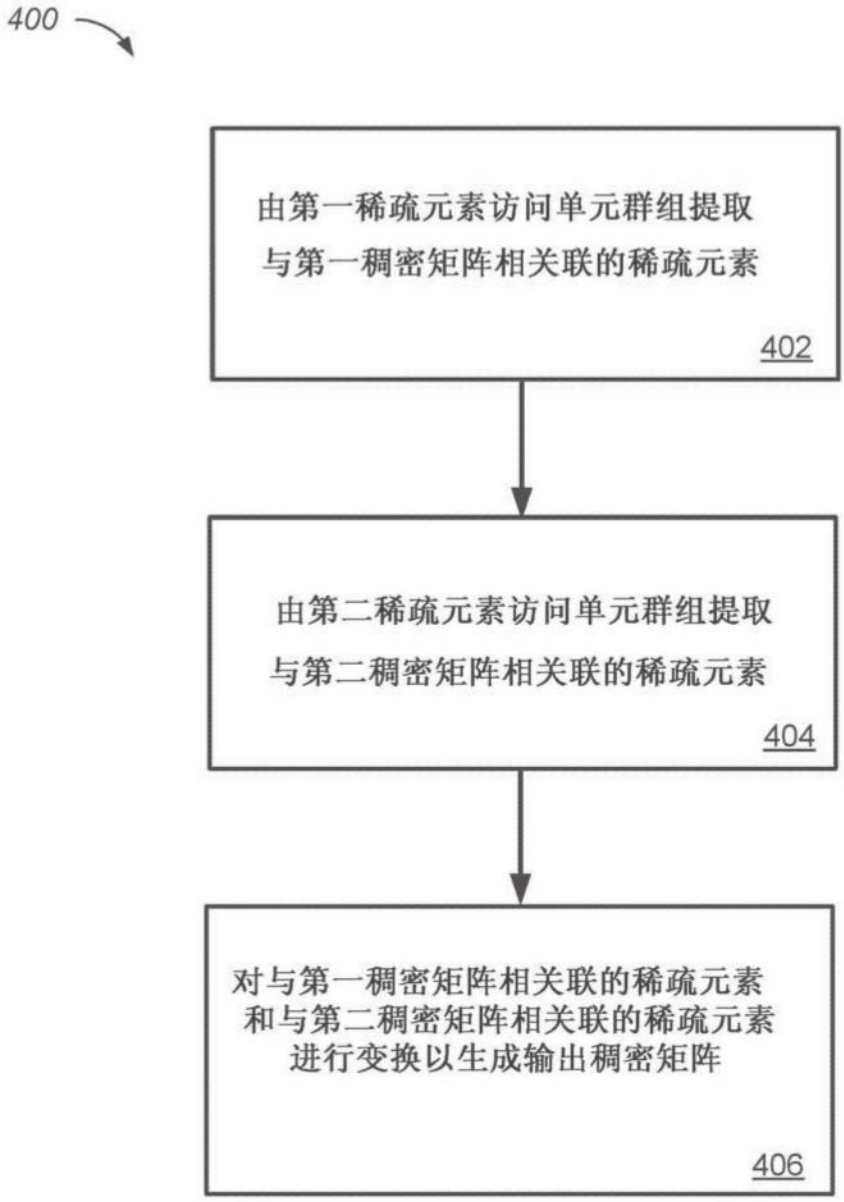

400
由第一稀疏元素访问单元群组提取
与第一稠密矩阵相关联的稀疏元素
402
由第二稀疏元素访问单元群组提取
与第二稠密矩阵相关联的稀疏元素
404
对与第一稠密矩阵相关联的稀疏元素
和与第二稠密矩阵相关联的稀疏元素
进行变换以生成输出稠密矩阵
406

图4

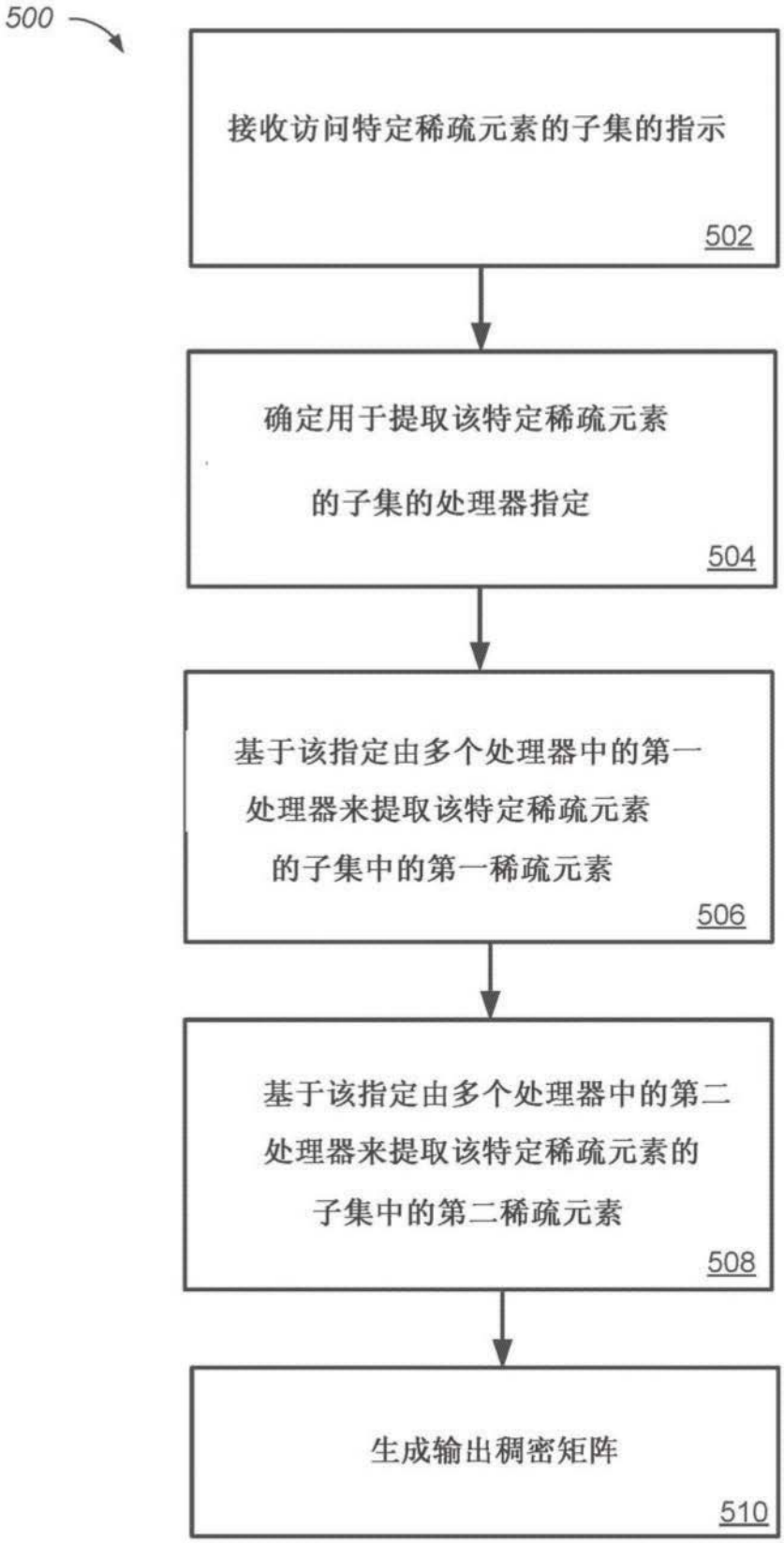

500
接收访问特定稀疏元素的子集的指示
502
确定用于提取该特定稀疏元素的子集的处理器指定
504
基于该指定由多个处理器中的第一处理器来提取该特定稀疏元素的子集中的第一稀疏元素
506
基于该指定由多个处理器中的第二处理器来提取该特定稀疏元素的子集中的第二稀疏元素
508
生成输出稠密矩阵
510

图5