(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 108008974 B
(45)授权公告日 2023.05.26

(21)申请号 201710823568.7

(22)申请日 2017.09.13

(65)同一申请的已公布的文献号
申请公布号 CN 108008974 A

(43)申请公布日 2018.05.08

(30)优先权数据
62/413,977 2016.10.27 US
62/418,155 2016.11.04 US
15/426,033 2017.02.06 US

(73)专利权人 三星电子株式会社
地址 韩国京畿道水原市

(72)发明人 李双辰 牛迪民
克里希纳·马拉丁 郑宏忠

(74)专利代理机构 北京铭硕知识产权代理有限公司 11286
专利代理师 张川绪 王兆赓

(51)Int.Cl.
*G06F 9/38*(2006.01)

(56)对比文件
US 2016232951 A1,2016.08.11
CN 105703765 A,2016.06.22
WO 2015187771 A2,2015.12.10

审查员 李娜

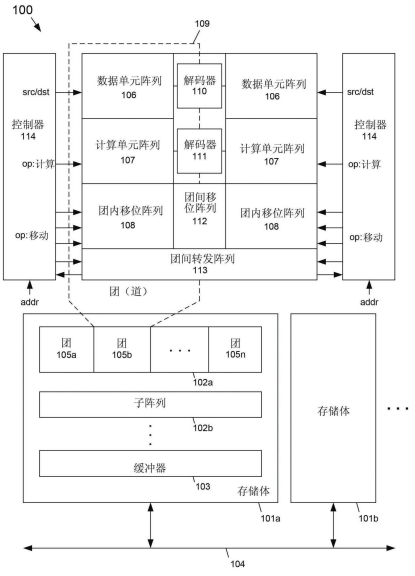
权利要求书3页 说明书10页 附图14页

(54)发明名称

基于动态随机存取存储器的处理单元架构

(57)摘要

提供一种基于动态随机存取存储器的处理单元架构。一种基于动态随机存取存储器的处理单元DPU可包括:具有多个基于动态随机存取存储器DRAM的计算单元的至少一个计算单元阵列,所述多个基于DRAM的计算单元以具有至少一列的阵列来布置,其中,所述至少一列可包括至少三行的基于DRAM的计算单元,所述至少三行的基于DRAM的计算单元被配置为提供对所述至少三行中的第一行和第二行进行运算的逻辑功能并且被配置为在所述至少三行中的第三行存储所述逻辑功能的结果;以及控制器,可连接到所述至少一个计算单元阵列,以配置所述至少一个计算单元执行DPU运算。




CN 108008974 B

1.一种动态随机存取存储器的处理单元DPU,包括:

至少一个计算单元阵列,所述至少一个计算单元阵列包括以具有至少第一预定数量的列和第二预定数量的行的阵列布置的多个基于动态随机存取存储器DRAM的计算单元,其中,第一预定数量大于或等于三并且第二预定数量大于或等于三,每列被配置为提供包括对所述列的第一行和第二行进行运算的逻辑功能的DPU运算,并且被配置为将逻辑功能的结果存储在所述列的第三行,所述至少一个计算单元阵列还包括第三预定数量的移位线,其中,第三预定数量是2的第一预定数量次的幂,每条移位线通过至少一个对应的第一晶体管连接到计算单元的列,移位线和对应的第一晶体管被配置为:在所述至少一个计算单元阵列中沿右方向或左方向将选择的列的两行计算单元的内容移位至少两列;以及

控制器,连接到所述至少一个计算单元阵列,以配置所述至少一个计算单元阵列执行DPU运算。

2.如权利要求1所述的DPU,其中,控制器通过地址总线接收用于DPU运算的指令。

3.如权利要求1所述的DPU,其中,至少一列的多个基于DRAM的计算单元中的每个包括三晶体管一电容器(3T1C)DRAM存储器单元。

4.如权利要求3所述的DPU,其中,至少一列的多个基于DRAM的计算单元提供NOR逻辑功能。

5.如权利要求1所述的DPU,其中,至少一列的多个基于DRAM的计算单元中的每个包括一晶体管一电容器(1T1C)DRAM存储器单元。

6.如权利要求5所述的DPU,其中,每个基于DRAM的计算单元还包括:连接到该基于DRAM的计算单元的位线的算术逻辑单元,所述算术逻辑单元提供逻辑功能。

7.如权利要求6所述的DPU,其中,所述算术逻辑单元提供NOR逻辑功能。

8.如权利要求1所述的DPU,还包括:

至少一个数据单元阵列,包括以第一预定数量的列布置的至少一个基于DRAM的存储器单元,所述至少一个数据单元阵列的基于DRAM的存储器单元的每列对应于对应的计算单元阵列的列;

感测放大器,连接到计算单元的每个列,每个感测放大器包括电连接到所述列的计算单元的读取位线的输入端以及电连接到所述列的计算单元的写入位线的输出端,

其中,控制器还连接到所述至少一个数据单元阵列,以配置所述至少一个数据单元阵列执行存储器操作,

其中,控制器还通过地址总线接收用于存储器操作的指令。

9.如权利要求1所述的DPU,还包括:

至少一个随机计算单元阵列,所述至少一个随机计算单元阵列包括以第一预定数量的列和第二预定数量的行布置的多个基于DRAM的随机计算单元,所述至少一个随机计算单元阵列的基于DRAM的随机计算单元的每列对应于对应的计算单元阵列的列,每列被配置为提供对由所述列的第一行接收的数据的第一流和由所述列的第二行接收的数据的第二流进行运算的随机逻辑功能,并且被配置为在所述列的第三行存储所述随机逻辑功能产生的数据的流,

其中,控制器还连接到所述至少一个随机计算单元阵列,以配置所述至少一个随机计算单元阵列执行与所述随机逻辑功能对应的随机逻辑运算,

其中,控制器还通过地址总线接收用于随机逻辑运算的指令。

10.一种基于动态随机存取存储器的处理单元DPU,包括:

至少一个计算单元阵列,所述至少一个计算单元阵列包括以具有至少第一预定数量的列和第二预定数量的行的阵列布置的多个基于动态随机存取存储器DRAM的计算单元,其中,第一预定数量大于或等于三并且第二预定数量大于或等于三,每列被配置为提供对所述列的第一行和第二行进行数位运算的逻辑功能,并且被配置为将逻辑功能的结果存储在所述列的第三行,所述至少一个计算单元阵列还包括第三预定数量的移位线,其中,第三预定数量是2的第一预定数量次的幂,每条移位线通过至少一个对应的第一晶体管连接到计算单元的列,移位线和对应的第一晶体管被配置为:在所述至少一个计算单元阵列中沿右方向或左方向将选择的列的两行计算单元的内容移位至少两列;

至少一个数据单元阵列,包括以第一预定数量的列布置的至少一个基于DRAM的存储器单元,所述至少一个数据单元阵列的基于DRAM的存储器单元的每列对应于对应的计算单元阵列的列;以及

控制器,连接到所述至少一个计算单元阵列,以配置所述至少一个计算单元阵列执行DPU运算,并且连接到所述至少一个数据单元阵列以执行存储器操作。

11.如权利要求10所述的DPU,其中,控制器通过地址总线接收用于DPU运算的指令。

12.如权利要求10所述的DPU,其中,至少一列的多个基于DRAM的计算单元中的每个包括三晶体管一电容器(3T1C)DRAM存储器单元,

其中,至少一列的多个基于DRAM的计算单元提供NOR逻辑功能。

13.如权利要求10所述的DPU,其中,至少一列的多个基于DRAM的计算单元中的每个包括一晶体管一电容器(1T1C)DRAM存储器单元,

其中,每个基于DRAM的计算单元还包括:连接到该基于DRAM的计算单元的位线的算术逻辑单元,所述算术逻辑单元提供逻辑功能。

14.如权利要求13所述的DPU,其中,所述算术逻辑单元提供NOR逻辑功能。

15.如权利要求10所述的DPU,还包括:

至少一个随机计算单元阵列,所述至少一个随机计算单元阵列包括以第一预定数量的列和第二预定数量的行布置的多个基于DRAM的随机计算单元,所述至少一个随机计算单元阵列的基于DRAM的随机计算单元的每列对应于对应的计算单元阵列的列,每列被配置为提供对由所述列的第一行接收的数据的第一流和由所述列的第二行接收的数据的第二流进行运算的随机逻辑功能,并且被配置为在所述列的第三行存储所述随机逻辑功能产生的数据的流,

其中,控制器还连接到所述至少一个随机计算单元阵列,以配置所述至少一个随机计算单元阵列执行随机逻辑运算,

其中,控制器还通过地址总线接收用于随机逻辑运算的指令。

16.如权利要求15所述的DPU,其中,所述至少一列的多个基于DRAM的随机计算单元中的每个包括:三晶体管一电容器(3T1C)DRAM存储器单元,或者一晶体管一电容器(1T1C)DRAM存储器单元。

17.一种动态随机存取存储器的处理单元DPU,包括:

至少一个计算单元阵列,所述至少一个计算单元阵列包括以具有至少第一预定数量的

列和第二预定数量的行的阵列布置的多个基于动态随机存取存储器DRAM的计算单元，其中，第一预定数量大于或等于三并且第二预定数量大于或等于三，每列被配置为提供对所述列的第一行和第二行进行运算的逻辑功能，并且被配置为将逻辑功能的结果存储在所述列的第三行，所述至少一个计算单元阵列还包括第三预定数量的移位线，其中，第三预定数量是2的第一预定数量次的幂，每条移位线通过至少一个对应的第一晶体管连接到计算单元的列，移位线和对应的第一晶体管被配置为：在所述至少一个计算单元阵列中沿右方向或左方向将选择的列的两行计算单元的内容移位至少两列；

至少一个随机计算单元阵列，所述至少一个随机计算单元阵列包括以第一预定数量的列和第二预定数量的行布置的多个基于DRAM的随机计算单元，所述至少一个随机计算单元阵列的基于DRAM的随机计算单元的每列对应于对应的计算单元阵列的列，每列被配置为提供对由所述列的第一行接收的数据的第一流和由所述列的第二行接收的数据的第二流进行运算的随机逻辑功能，并且被配置为在所述列的第三行存储所述随机逻辑功能产生的数据的流；

控制器，连接到所述至少一个计算单元阵列，以配置所述至少一个计算单元阵列执行DPU运算，并且连接到所述至少一个随机计算单元阵列以执行与所述随机逻辑功能对应的随机逻辑运算。

18.如权利要求17所述的DPU，其中，控制器通过地址总线接收用于DPU运算的指令。

19.如权利要求18所述的DPU，其中，至少一列的多个基于DRAM的计算单元中的每个包括三晶体管一电容器(3T1C)DRAM存储器单元，

其中，至少一列的多个基于DRAM的计算单元提供NOR逻辑功能。

20.如权利要求17所述的DPU，其中，至少一列的多个基于DRAM的随机计算单元中的每个包括三晶体管一电容器(3T1C)DRAM存储器单元。

# 基于动态随机存取存储器的处理单元架构

[0001] 本申请要求于2016年10月27日提交的第62/413,977号美国临时专利申请、于2016年11月4日提交的第62/418,155号美国临时专利申请和于2017年2月6日提交的第15/426,033号美国专利申请的优先权,所述申请的公开通过引用全部包含于此。

## 技术领域

[0002] 在此公开的主题涉及可使用动态随机存取存储器(DRAM)技术形成的并且可重构和可编程的基于DRAM的处理单元(DPU)。

## 背景技术

[0003] 图形处理单元(GPU)和张量处理单元(TPU)通常用于深度学习处理。深度学习处理包括不能被GPU或TPU有效执行的高度并行的处理。

## 发明内容

[0004] 一个示例实施例提供一种基于动态随机存取存储器(DRAM)的处理单元(DPU),所述DPU可包括:具有以具有至少一列的阵列布置的多个基于DRAM的计算单元的至少一个计算单元阵列,其中,所述至少一列可包括至少三行的基于DRAM的计算单元,所述至少三行的基于DRAM的计算单元被配置为提供对所述至少三行中的第一行和第二行进行运算的逻辑功能并且被配置为在所述至少三行中的第三行存储所述逻辑功能的结果;以及控制器,可连接到所述至少一个计算单元阵列,以配置所述至少一个计算单元执行DPU运算。

[0005] 一个示例实施例提供一种DPU,所述DPU可包括:至少一个计算单元阵列,所述至少一个计算单元阵列可包括以具有至少一列的阵列布置的多个基于DRAM的计算单元,其中,所述至少一列可包括至少三行的基于DRAM的计算单元,所述至少三行的基于DRAM的计算单元被配置为提供对所述至少三行中的第一行和第二行进行运算的逻辑功能并且被配置为在所述至少三行中的第三行存储所述逻辑功能的结果;至少一个数据单元阵列,可包括以至少一列布置的至少一个基于DRAM的存储器单元;以及控制器,连接到所述至少一个计算单元阵列,以配置所述至少一个计算单元阵列执行DPU运算,并且连接到所述至少一个数据单元阵列以执行存储器操作。在一个实施例中,至少一列的多个基于DRAM的计算单元中的每个可包括三晶体管一电容器(3T1C)DRAM存储器单元,并且所述至少一列的多个基于DRAM的计算单元可提供NOR逻辑功能。在另一实施例中,所述至少一列的多个基于DRAM计算单元中的每个可包括一晶体管一电容器(1T1C)DRAM存储器单元,每个基于DRAM的计算单元还可包括:连接到该基于DRAM的计算单元的位线的算数逻辑单元(ALU),其中,所述ALU可提供逻辑功能。

[0006] 一个示例实施例提供一种DPU,所述DPU可包括:至少一个计算单元阵列,所述至少一个计算单元阵列可包括以具有至少一列的阵列布置的多个基于DRAM的计算单元,其中,所述至少一列可包括第一至少三行的基于DRAM的计算单元,所述至少三行的基于DRAM的计算单元被配置为提供对所述至少三行中的第一行和第二行进行运算的逻辑功能并且被配

置为在所述至少三行中的第三行存储所述逻辑功能的结果;至少一个随机计算单元阵列,所述至少一个随机计算单元阵列可包括以具有至少一列的阵列布置的多个基于DRAM的随机计算单元,其中,所述至少一列可包括至少三行的基于DRAM的随机计算单元,所述至少三行的基于DRAM的随机计算单元被配置为提供对所述至少三行中的第一行和第二行进行运算的逻辑功能并且被配置为在所述至少三行中的第三行存储所述逻辑功能的结果;以及控制器,连接到所述至少一个计算单元阵列,以配置所述至少一个计算单元阵列执行DPU运算,并且连接到所述至少一个随机计算单元阵列以执行随机逻辑运算。

## 附图说明

[0007] 在下面的部分中,将参照附图中示出的示例性实施例来描述在此公开的主题的方面,其中:

[0008] 图1描述根据在此公开的主题的基于动态随机存取存储器(DRAM)的处理单元(DPU)的示例实施例的框图;

[0009] 图2A描述可用于计算单元阵列中的计算单元的三晶体管一电容器DRAM计算单元分布(topography)的示例实施例;

[0010] 图2B描述可用于计算单元阵列中的计算单元的一晶体管一电容器DRAM计算单元分布的可选的示例实施例;

[0011] 图3描述根据在此公开的主题的团内移位阵列(intra-mat shift array)的示例实施例;

[0012] 图4A描述根据在此公开的主题的团间移位阵列(inter-mat shift array)的实施例;

[0013] 图4B概念性地描述根据在此公开的主题的用于右团间移位的在相邻的计算单元列中两个相同放置的计算单元之间的团间移位内部连接构造;

[0014] 图4C概念性地描述根据在此公开的主题的用于左团间移位的在相邻的计算单元列中两个不同放置的计算单元之间的团间移位内部连接构造;

[0015] 图5描述根据在此公开的主题的团间转发阵列的实施例;

[0016] 图6A至图6G描述根据在此公开的主题的可由DPU提供的基于NOR逻辑的运算;

[0017] 图7描述根据在此公开的主题的包括随机数据区域的DPU的示例实施例的框图;

[0018] 图8A和图8B分别描述用于可被转换为多路复用运算的加法运算以及可被转换为AND逻辑运算的乘法运算的随机计算操作;

[0019] 图9描述根据在此公开的主题的包括DPU的系统架构。

## 具体实施方式

[0020] 在以下的详细描述中,为了提供本公开的全面理解,阐述许多特定的细节。然而,本领域技术人员将理解公开的方面可在没有这些特定细节的情况下被实践。在其他实例中,公知的方法、过程、组件和电路不被详细地描述,从而不模糊在此公开的主题。

[0021] 贯穿本说明书提及的“一个实施例”或“实施例”表示与该实施例相关地描述的具体的特征、结构或特性可被包括在这里公开的至少一个实施例。因此,贯穿本说明书在各个位置出现的短语“在一个实施例中”或“在实施例中”或“根据一个实施例”(或具有相似含义

的其他短语)可不必全部参照同一实施例。此外,具体的特征、结构或特性可在一个或多个实施例中以任意合适的方式来合并。就这一点而言,如在此使用的词语“示例性”表示“用作示例、实例或说明”。在此描述为“示例性”的任意实施例不被解释为必须优选于或优于其他实施例。此外,根据在此讨论的上下文,单数术语可包括相应的复数形式,复数术语可包括相应的单数形式。还应注意在此示出和讨论的各个附图(包括组件示图)仅为了说明性目的,并且不按比例绘制。相似地,仅为了说明性目的示出各种波形和时序示图。例如,为了清楚一些元件的尺寸可相对于其他元件被夸大。此外,适当情况下,参考标号可在附图中被重复,以指示相应的和/或类似的元件。

[0022] 在此使用的技术仅是为了描述具体的示例性实施例的目的,而不意在限制要求保护的主题。除非上下文另外清楚地指示,否则如在此使用的单数形式也意图包括复数形式。还将理解当在本说明书中使用术语“包括”和/或“包含”时,说明存在叙述的特征、整体、步骤、操作、元件和/或组件,但不排除存在或添加一个或多个其他特征、整体、步骤、操作、元件、组件和/或它们的组。如在此使用的术语“第一”、“第二”等被用作它们后面的名词的标号,而不意味着任意类型的顺序(例如,空间、时间、逻辑等),除非如此明确地定义。此外,相同的参考标号可贯穿两个或更多个附图被使用,以表示具有相同或相似功能的部分、组件、块、电路、单元或模块。然而,这样的使用仅是为了说明的简单和讨论的方便;不意味着这样的组件或单元的结构或架构细节贯穿所有的实施例是相同的,或者这样的通常引用的部分/模块仅是用于实现在此公开的具体实施例的教义的方式。

[0023] 除非另外定义,否则在此使用的全部术语(包括技术术语和科技术语)具有与本主题所属领域中的普通技术人员通常理解的含义相同的含义。还将理解,除非在此明确地定义,否则诸如在通用字典中定义的术语应被解释为具有在与相关领域的上下文中的含义一致的含义,将不被解释为理想化或过于正式的意义。

[0024] 在此公开的主题提供针对不同的运算(诸如,加、乘、移位、最大/最小和比较,但不限于此)的可编程和可重构的基于动态随机存取存储器(DRAM)的处理单元(DPU)。在一个实施例中,DPU基于三晶体管一电容器(3T1C)DRAM处理和结构。在另一实施例中,DPU基于具有微小改变的一晶体管一电容器(1T1C)DRAM处理和结构。因此,DPU不包含特定的计算逻辑电路(例如,加法器),但是提供使用利用高度并行运算的存储器单元的计算。在一个实施例中,DPU可包括随机计算阵列,在随机计算阵列中,加法可转换为多路复用运算,乘法可转换为与(AND)逻辑运算。

[0025] 在此公开的主题还提供包括具有框架扩展、库、驱动器、编译器和指令集架构(ISA)的环境(生态系统)的系统架构,以编程和重构DPU。

[0026] 此外,在此公开的主题提供适合于数据中心和/或移动应用的系统架构,并且该系统架构针对二进制和定点计算二者,提供用于机器学习应用的存储器内处理器(PIM)方案,这该PIM方案是针对GPU/ASIC(TPU)/FPGA机器学习应用的替换物。在一个实施例中,在此公开的主题提供高性能、节能且低成本的系统,该系统提供加速的深度学习,例如,二进制加权神经网络(Binary Weight Neural Network)。

[0027] 在此公开的主题涉及可使用动态随机存取存储器(DRAM)技术形成的并且可重构和可编程的基于DRAM的处理单元(DPU)。在一个实施例中,DPU可包括:基于DRAM的存储器单元阵列、以及可被配置为执行不同的运算(诸如,加、乘、排序等)的基于DRAM的计算单元阵

列。

[0028] DPU的内部架构可包括与具有子阵列的多个存储体连接的总线。在一个实施例中，总线可被配置为提供具有子阵列的H树连接的存储体。每个子阵列可包括局部控制器，每个单独的子阵列可被单独地或同时地激活。在一个实施例中，基于DRAM的计算单元可被划分为两个阵列，即，数据单元阵列和计算单元阵列。在一个实施例中，可通过基于DRAM的存储器单元来实现计算单元阵列。在另一实施例中，可通过具有逻辑电路的基于DRAM的存储器单元来实现计算单元阵列。DPU内部架构还可包括数据移位和数据移动电路。在一些实施例中，可存在可被配置为用于随机数据计算的基于DRAM的单元阵列。

[0029] 图1描述根据在此公开的主题的DPU 100的示例实施例的框图。DPU 100可包括一个或多个存储体101a至101m，在图1中仅描述其中的存储体101a和存储体101b。每个存储体101可包括一个或多个子阵列102a至102n，在图1中仅描述其中的子阵列102a和子阵列102b。每个存储体101还可包括缓冲器103。缓冲器103可连接到各个子阵列102并且连接到系统总线104。缓冲器103可读取存储体101中的整行，然后将该行写回到同一存储体或另一存储体。缓冲器103还可将行数据的复制体传播到子阵列102中的一个或多个团(mat)105a至105n。在一个实施例中，存储体101和系统总线104可被配置为提供H树连接的存储体。

[0030] 每个子阵列102可包括一个或多个团(或，道(lane))105，在图1中示出子阵列102a的团105a至团105n。每个团105是DPU 100的可包括数据单元阵列106、计算单元阵列107和团内移位阵列108的区域。图1中如虚线109所包围的指示示例性的团105。每个团105与相邻的团共享数据单元阵列解码器110、计算单元阵列解码器111、团间移位阵列112和团间转发阵列113。在一个实施例中，数据单元阵列解码器110、计算单元阵列解码器111和团间移位阵列112可在相邻的团105之间与子阵列控制器114交替地物理布置。在一个实施例中，解码器110和111可作为传统的DRAM型存储器解码器进行操作。

[0031] 在一个实施例中，每个团105可通信地连接到子阵列控制器114。每个子阵列控制114可被配置为独立于其他子阵列控制器114。子阵列控制器114可从DRAM地址总线接收指令作为地址(addr)。响应于地址(即，地址信号)，子阵列控制器114可将解码的地址作为输出提供给数据单元阵列106和计算单元阵列107中的任一个或二者。也就是说，子阵列控制器114可输出由数据单元阵列解码器110针对相关联的数据单元阵列106解码的源地址/目的地址(src/dst)，在计算单元阵列107的情况下，子阵列控制器114可输出由计算单元阵列解码器111解码的运算/计算(op/calc)地址。子阵列控制器114还可从使得两个或更多个子阵列控制器114以协调的方式进行操作的DRAM总线接收指令作为地址。子阵列控制器114还可控制数据移动电路，诸如，控制团内移位阵列108、团间移位阵列112以及团间转发阵列113。

[0032] 每个数据单元阵列106可包括以至少一列和至少一行布置的一个或多个动态随机存取存储器(DRAM)单元。在一个实施例中，数据单元阵列106可被配置为传统的DRAM单元阵列。在一个实施例中，数据单元阵列106可包括2000(即，2K)列和16行。在另一实施例中，数据单元阵列106可包括：比2K少或多的列，和/或比16少或多的行。

[0033] 每个计算单元阵列107可包括以至少一列和至少一行布置的一个或多个计算单元。计算单元阵列107中的列数与数据单元阵列106中的列数相同。在一个实施例中，计算单元阵列107可包括2K列和16行。在另一示例中，计算单元阵列107可包括：比2K少或多的列，

和/或比16少或多的行。

[0034] 图2A描述可用于计算单元阵列107中的计算单元的三晶体管一电容器(3T1C)DRAM计算单元分布(topography)201的示例实施例。如在图2A中所示,行X中的3T1C DRAM计算单元包括第一晶体管$T_1$,其中,第一晶体管$T_1$具有电连接到写入位线(Write BL)的源极端、电连接到电容器$C_1$的第一端和第二晶体管$T_2$的栅极端二者的漏极端以及电连接到写入使能(WEN)线(例如,$WEN_X$、$WEN_Y$和$WEN_R$)的栅极端。电容器$C_1$的第二端电连接到地线。第二晶体管$T_2$包括:电连接到地线的源极端、以及电连接到第三晶体管$T_3$的源极端的漏极端。第三晶体管$T_3$包括:电连接到字线(WL,例如,$WL_X$、$WL_Y$和$WL_R$)的栅极端、以及电连接到读取位线(Read BL)的漏极端。3T1C DRAM计算单元分布201包括感测放大器(sense amplifier)SA,其中,SA具有电连接到Read BL的输入端以及电连接到Write BL的输出端。

[0035] 行Y中的计算单元和行R中的计算单元二者还可包括与行X中的计算单元的布置相似的以3T1C DRAM构造布置的三个晶体管$T_1$至$T_3$和一个电容器C。图2A中描述的示例的三个计算单元和感测放大器SA被配置为提供或非(NOR)逻辑运算(即,行X NOR行Y逻辑运算),其中,结果被存储在行R中。虽然在图2A中仅明确地描述了3T1C DRAM计算单元的一列,但是应该理解在另一实施例中,3T1C DRAM计算单元可被配置为多个列(即,2K列)。还应该理解在另一实施例中,可提供多于3的行。此外,尽管图2A中描述的3T1C DRAM计算单元构造提供NOR逻辑运算,但是应该理解3T1C DRAM计算单元分布201的NOR逻辑运算可被用于提供多个函数运算,诸如,异或非(XNOR)、加(ADD)、选择(SET)、最大(MAX)、取符号(SIGN)、复用(MUX)、条件和加法逻辑(CSA)、乘、种群计数(popcount)和比较(COMPARE),但不限于此。移位阵列108和112还可提供移位功能。

[0036] 图2B描述可用于图1的计算单元阵列107中的计算单元的一晶体管一电容器(1T1C)DRAM计算单元分布202的可选的示例实施例。如在图2中所示,1T1C DRAM计算单元包括晶体管$T_4$,其中,$T_4$具有电连接到电容器$C_2$的第一端的源极端、电连接到位线(BL)的漏极端以及电连接到字线(WL)的栅极端。电容器$C_2$的第二端电连接到地线。位线BL电连接到感测放大器SA的输入端。感测放大器SA的输出端电连接到多路复用器(MUX)的第一输入端、第五晶体管$T_5$的漏极端以及算术逻辑单元(ALU)的输入端。MUX的输出端电连接到锁存器(LATCH)的输入端。第五晶体管T5的源极端电连接到LATCH的输出端。ALU的输出端电连接到MUX的第二输入端。图2B中的第五晶体管$T_5$、MUX和LATCH以及ALU分别从控制器114接收控制信号CNTL1至CNTL4。在一个实施例中,ALU可被配置为提供NOR功能。虽然在图2B中电连接到位线BL的逻辑电路提供NOR逻辑运算,但是应该理解,电连接到位线BL的逻辑电路(即,ALU)可提供其他的功能运算,诸如,异或非(XNOR)、加(ADD)、选择(SET)、最大(MAX)、SIGN、复用(MUX)、条件和加法逻辑(CSA)、乘、种群计数(popcount)和比较(COMPARE),但不限于此。移位阵列108和112还可提供移位功能。应该理解在图2B中仅描述了一个1T1C DRAM计算单元,但是可提供多个列和行的1T1C DRAM计算单元。

[0037] 在图2A和图2B中可看到,DPU的计算单元不包括特定的、复杂的计算逻辑,而是包括具有提供执行多个不同类型的计算的能力的可再编程属性的相对简单的分布(topography)。此外,DPU的分布可被布置为具有存储器结构中固有的大量并行性的优点,以更快且更有效地执行更多的计算。

[0038] 图3描述根据在此公开的主题的团内移位阵列108的示例实施例。为了简化团内移

位阵列108的描述，考虑如图3中所示的宽度为4列计算单元107的团105。团内移位阵列108包括：以阵列布置的多个第六晶体管$T_6$(图3中仅指示其中的一个晶体管$T_6$)、$2^n$条移位线SL(n是团105中的计算单元的列数)、n+2条左移位控制线SLcL、2条右移位控制线SRcL以及n条移位屏蔽线(shift mask line)SML。团内移位阵列108的多个第六晶体管$T_6$中的一些第六晶体管$T_6$电连接在Write BL与2n条移位线SL之间，团内移位阵列108的其他第六晶体管$T_6$电连接在Read BL与2n条移位线SL之间。这些第六晶体管$T_6$的栅极电连接到n+2条左移位控制线SLcL和2条右移位控制线SRcL。团内移位阵列108的其他第六晶体管$T_6$电连接在n条移位屏蔽线SML与2n条移位线SL之间。团内移位阵列108的控制线电连接到与团105相关联的子阵列控制器114。

[0039] 团内移位阵列108可通过控制线SLcL和SRcL上的合适的信号，将数据在团105内进行左移位或右移位。针对左移位，数据可被填充符号位，并且每次操作移位1位或(n-1)位，其中，n是每个团105的列数。针对右移位，数据可根据指令的控制被填充0或1，并且被移位$2^0$、$2^1$、…$2^{k-1}$、$2^k$直到每个团的列数，其中，$2^k$是列数。

[0040] 图4A描述根据在此公开的主题的团间移位阵列112的实施例。为了简化团间移位阵列112的描述，考虑如图4A至图4C中所示的团105为2列计算单元107的构造。也就是说，每个团105包括第一列计算单元107a和第二列计算单元107b。团间移位阵列112包括晶体管$T_{112a}$和$T_{112b}$、晶体管$T_{112c}$和$T_{112d}$、数据移位线112e和112f以及团间移位控制线ISLcL。在一个团内，晶体管$T_{112a}$包括：电连接到第一列计算单元107a的Read BL的源极端、电连接到数据移位线112e的漏极端。晶体管$T_{112b}$包括：电连接到第二列计算单元107b的Read BL的源极端、电连接到数据移位线112f的漏极端。数据移位线112e和112f电连接到缓冲器103(图4A中未示出)。在不同的团之间，晶体管$T_{112c}$包括分别电连接到相邻团中的数据移位线112e的源极端和漏极端。晶体管$T_{112d}$包括分别电连接到相邻团中的数据移位线112f的源极端和漏极端。在不同的团中的晶体管$T_{112c}$和晶体管$T_{112d}$的栅极分别电连接到各自不同的团间移位控制线ISLcL。团间移位阵列112可通过团间移位控制线ISLcL上的合适的信号，将数据在不同的团之间进行左移位或右移位。团间移位阵列112的控制线电连接到与团105相关联的子阵列控制器114。

[0041] 图4B概念性地描述根据在此公开的主题的用于右团间移位的在相邻的计算单元列105a和105b中两个相同放置的计算单元之间的团间移位内部连接构造。可通过加重的起作用的内部连接节点来概念性地描述图4B的内部连接构造。例如，晶体管$T_{112c}$和晶体管$T_{112d}$被激活使得导电路径存在于每个晶体管之间，从而在团105a(位于左侧)与团105b(位于右侧)之间连接数据移位线112e和数据移位线112f。晶体管$T_{112c}$和晶体管$T_{112d}$的栅极端电连接到有源的团间移位控制线ISLcL。团105b中的晶体管$T_{112a}$和晶体管$T_{112b}$被激活，使得团105b中的计算单元107a的Read BL电连接到团105b的左侧的团105a中的计算单元107a的Write BL，并且使得团105b中的计算单元107b的Read BL电连接到团105b的左侧的团105a中的计算单元107b的Write BL。

[0042] 图4C概念性地描述根据在此公开的主题的用于左团间移位的在相邻的计算单元列105a和105b中两个不同放置的计算单元之间的团间移位内部连接构造。可通过加重的起作用的内部连接节点来概念性地描述图4C的内部连接构造。例如，晶体管$T_{112c}$和晶体管$T_{112d}$被激活使得导电路径存在于每个晶体管之间，从而在团105a(位于左侧)与团105b(位于右

侧)之间连接数据移位线112e和数据移位线112f。晶体管$T_{112c}$和晶体管$T_{112d}$的栅极端电连接到有源的团间移位控制线ISLcL。团105a中的晶体管$T_{112a}$和晶体管$T_{112b}$被激活,使得团105a中的计算单元107a的Read BL电连接到团105a的右侧的团105b中的计算单元107a的Write BL,并且使得团105a中的计算单元107b的Read BL电连接到团105a的右侧的团105b中的计算单元107b的Write BL。

[0043] 图5描述根据在此公开的主题的团间转发阵列113的实施例。为了简化团间转发阵列113的描述,考虑如图5中所示的宽度为2列计算单元107的团105的构造。也就是说,每个团105包括第一列计算单元107a和第二列计算单元107b。在团105内,团间转发阵列113包括:晶体管$T_{113a}$和$T_{113b}$、晶体管$T_{113c}$和$T_{113d}$、晶体管$T_{113e}$和$T_{113f}$、$2^n$条转发数据线FDL(n是团中的计算单元列的数量)、2条转发控制线FCL、以及$2^m$条转发部分线FSL(m是部分的数量)。晶体管$T_{113a}$和晶体管$T_{113b}$的源极端分别电连接到第一列计算单元107a的Write BL和Read BL。晶体管$T_{113a}$和晶体管$T_{113b}$的漏极端电连接到第一转发数据线FDL 113g。晶体管$T_{113c}$和晶体管$T_{113d}$的源极端分别电连接到第二列计算单元107b的Write BL和Read BL。晶体管$T_{113c}$和晶体管$T_{113d}$的漏极端电连接到第二转发数据线FDL113h。晶体管$T_{113e}$和晶体管$T_{113f}$的源极端分别电连接到的晶体管$T_{113c}$和晶体管$T_{113d}$栅极端。晶体管$T_{113e}$和晶体管$T_{113f}$的漏极端均连接到同一转发部分线FSL。晶体管$T_{113e}$和晶体管$T_{113f}$的栅极端分别连接到不同的转发控制线FCL。团间转发阵列113可通过转发控制线FCL上的合适的信号,在团之间转发数据。团间转发阵列113的控制线电连接到与在其之间正在转发数据的团105相关联的子阵列控制器114。

[0044] 图6A至图6G描述根据在此公开的主题的可由DPU提供的基于NOR逻辑的运算。在图6A至图6G中,第一操作数(operand)可被存储在行X,第二操作数可被存储在行Y或行W。图6A至图6G中的箭头表示针对整行的计算单元的NOR逻辑运算的输入流和输出流。例如,图6A中的行X可表示存储在行X的计算单元中的整行的操作数。对存储在行X中的操作数和存储在行Y中的操作数进行NOR逻辑运算的结果被存储在结果行R中。例如,在一个实施例中,行X和行Y中的操作数可包括100列(即,$x_1$、$x_2$…$x_{100}$和$y_1$、$y_2$…$y_{100}$),所述结果可被存储在行R(即,$r_1$、$r_2$…$r_{100}$)中。也就是说,$x_i$ nor $y_i$=$r_i$,其中,i是列索引。在另一实施例中,行X可仅表示行中的选择的计算单元的组。

[0045] 图6B描述基于前缀Kogge-Stone加法器的针对N位数的示例的全加法器操作。在图6B中,第一N位操作数被存储在行X中,第二N位操作数被存储在行Y中。针对图6B中所示的示例的加法运算,计算中间项$G_0$、$P_0$、$G_1$、$P_1$、$G_2$、$P_2$…$G_{logN+1}$和$P_{logN+1}$。图6B的最上面的框表示使用来自行X和行Y的输入操作数确定$G_0$和$P_0$的五个分离的运算。在第一运算中,最上面的框确定行X的逆(即,～X),其被存储在行1中。第二运算确定行Y的逆(即,～Y),其被存储在行2中。第三运算确定运算:行X NOR行Y,运算结果被存储在行3中。第四运算确定运算$G_0$=行1NOR行2,运算结果被存储在行4中。第五运算确定$P_0$=行3NOR行4,其被存储在行5中。

[0046] 在图6B的中间框中,来自最上面的框的中间结果$G_0$和$P_0$用于确定中间结果$G_{i+1}$和$P_{i+1}$,其中,i是列索引。也就是说,在图6B的最上面的框确定的中间结果$G_0$和$P_0$用于确定中间结果$G_1$和$P_1$。中间结果$G_1$和$P_1$用于确定中间结果$G_2$和$P_2$,依此类推,以确定中间结果$G_{logN+1}$和$P_{logN+1}$。在图6B的最下面的框中,结果行R1和行R2分别存储针对全加法器操作的进位结果和求和结果。

[0047] 图6C描述可由3T1C DRAM计算单元分布201提供的示例的选择器操作。行1存储中间结果:行X的逆(即,～X)。行2存储中间结果:行Y的逆(即,～Y)。行3存储中间结果:行S的逆(即,～S)。行4存储中间结果:行1NOR行3。行5存储中间结果:行2NOR行S。行6存储中间结果:行4NOR行5。行R存储行6的逆的结果,即,S?X:Y。

[0048] 图6D描述可由3T1C DRAM计算单元分布201提供的可选示例的选择器操作。行1存储中间结果:行X的逆(即,～X)。行2存储中间结果:行S的逆(即,～S)。行3存储中间结果:行1NOR行S。行4存储行X的操作数(即,X)。行R存储结果:行3NOR行4,即,S?X:～X。

[0049] 图6E描述可由3T1C DRAM计算单元分布201提供的示例的最大/最小操作。行1存储中间结果:行Y的逆(即,～Y)。行2存储中间结果:X+(～Y+1)。行3存储中间结果:$C_{out}$>>n。行4存储中间结果:$C_{out}$?X:Y。行R存储结果:MAX(X:Y)。

[0050] 图6F描述可由3T1C DRAM计算单元分布201提供的示例的1位乘法操作。行1存储中间结果:行X NOR行W。行2存储中间结果:行X NOR行1。行3存储中间结果:行W NOR行1。结果行R存储结果:行2NOR行3,即,结果:行X XNOR行W。

[0051] 图6G描述可由3T1C DRAM计算单元分布201提供的示例的多位乘法操作。在图6G的上面的框中,行1存储中间结果:行W的逆(即,～W)。行2存储中间结果:行X的逆左移位$2^i$次(即,～X<<$2^i$),其中,i是索引。行3存储中间结果:行1NOR行2,即,$PP_i$＝～W NOR～X<<$2^i$。在图6G下面的框中,行1存储中间结果:行$PP_0$与行$PP_i$的和,即,$\Sigma PP_i$。行2存储中间结果:行2NOR行$W_{sign}$。行R存储结果:X×W。

[0052] 图7描述根据在此公开的主题的包括随机数据区域715的DPU 700的示例实施例的框图。具有与图1中所示的DPU 100的组件相同的参考指示符的DPU 700的各种组件是相似的,这样相似的组件的描述在此已经被省略。DPU 700的子阵列102包括:随机数据阵列715、转换器至随机阵列(converter-to-stochastic array)716、(实)数据单元阵列106、计算单元阵列107以及团内移位阵列108。

[0053] 每个随机数据阵列715可包括以至少一列和至少一行布置的一个或多个随机计算单元。随机数据阵列715中的列数与数据单元阵列106和计算单元阵列107中的列数相同。在一个实施例中,随机数据阵列715可包括2K列和16行。在另一实施例中,随机数据阵列715可包括:比2K少或多的列,和/或比16少或多的行。在随机数据阵列715中,使用“1”出现的概率,$2^n$位用于表示n位值。转换器至随机阵列716中的随机数产生器可用于将实数转换为随机数。种群计数运算可用于将随机数转换回实数。

[0054] 通过使用随机计算方法,加法可被转换为多路复用运算,乘法可被转换为AND逻辑运算。例如,图8A描述提供随机加法运算作为多路复用运算的电路,图8B描述提供随机乘法运算作为AND逻辑运算的电路。传统的用于随机计算的技术需要大量的存储器容量;然而,在此公开的主题因为基于DRAM的DPU能够执行大量并行的AND和MUX运算,所以可被用于提供高效的随机计算。使用在此公开的DPU的随机计算还使得加速深度学习为典型应用的复杂运算是可能的。

[0055] 图9描述根据在此公开的主题的包括DPU的系统架构900。系统架构900可包括:硬件层910、库和驱动层920、框架层930和应用层940。

[0056] 硬件层910可包括具有嵌入的DPU(诸如,在此描述的DPU)的硬件装置和/组件。装置和/或组件的一个实施例可以是可包括一个或多个嵌入的DPU的外围组件互联快速

(PCIe)装置911。装置和/或组件的另一实施例可以是可包括一个或多个嵌入的DPU的双列直插存储器模块(DIMM)912。应该理解,系统架构900的硬件层910不限于PCIe装置和/或DIMM,而是可包括可包含DPU的片上系统(SOC)装置或其他存储器类型的装置。可在硬件层910嵌入在装置和/或组件中的DPU可被配置为与图1中的DPU 100相似,和/或与图7中的DPU 700相似。在任意实施例中,DPU的具体的计算单元阵列可被配置为包括3T1C DRAM计算单元分布201(图2A)或1T1C DRAM计算单元分布202(图2B)。

[0057] 系统架构900的库和驱动层920可包括:DPU库921、DPU驱动器922和DPU编译器923。DPU库921可被配置为:针对可在应用层940操作的不同的应用,在硬件层910中针对DPU中的每个子阵列提供最优映射功能、资源分配功能和调度功能。

[0058] 在一个实施例中,DPU库921可针对框架层930提供高级应用编程接口(API),其可包括诸如移动、加、乘等的运算。例如,DPU库921还可包括用于标准型例程的实施,诸如,可应用于加速深度学习处理的向前和向后卷积层、池化层、归一化层、激活层,但不限于此。在一个实施例中,DPU库921可包括映射用于卷积神经网络(CNN)(例如,cudnn(NVIDIA CUDA® Deep Neural Network))的整个卷积层的计算的API类似功能。此外,DPU库921可包括用于优化卷积层计算到DPU的映射的API类似功能。

[0059] DPU库921还可包括以下的API类似功能:通过将任务(批量、输出通道、像素、输入通道、卷积核)内的单个或多个并行度映射到在芯片、存储体、子阵列和/或团等级上的相应的DPU并行度,来优化资源分配。此外,DPU库921可包括提供权衡性能(即,数据移动流)和功耗的在初始化和/或运行时优化的DPU构造的API类似功能。由DPU库921提供的其他API类似功能可包括:设计旋钮式(design-knob-type)功能,诸如,设置每个存储体的有源子阵列的数量、每个有源子阵列的输入特征映射的数量、特征映射的分割和/或卷积核的重复使用方案。其他API类似功能可包括:通过针对每个子阵列分配具体的任务(诸如,卷积计算、通道求和和/数据分发)的额外的资源分配优化。如果操作数将在整数与随机数之间转换,则DPU库921包括在满足精度约束的同时最小化开销的API类似功能。在精度低于期望的事件中,DPU库921可包括以下的API类似功能:针对随机表达式使用额外的位再次计算值,或者将任务卸载到其他硬件(诸如,CPU)。

[0060] DPU库921还可包括同时调度DPU中的激活的多个子阵列并且调度数据移动使得其被计算操作隐藏的API类似功能。

[0061] DPU库921的另一方面包括用于进一步DPU开发的扩展接口。在一个实施例中,DPU库921可提供用于使用NOR和移位逻辑直接地编程功能的接口,使得除了标准型运算(即,加、乘、最大/最小等)之外的操作可被提供。扩展接口还可提供这样的接口:使得不被DPU库921专门支持的运算可从库和驱动层920卸载到SoC控制器(未示出)、中央处理单元/图形处理单元(CPU/GPU)组件和/或CPU/张量处理单元(CPU/TPU)组件。DPU库921的另一方面提供在DPU存储器不被用于计算时将DPU的存储器用作存储器的扩展的API类似功能。

[0062] DPU驱动器922可被配置为:提供在硬件层910的DPU、DPU库921和在更高层的操作系统(OS)之间接口连接,以将DPU硬件层集成在系统中。也就说,DPU驱动器922将DPU暴露于系统OS和DPU库921。在一个实施例中,DPU驱动器922可在初始化提供DPU控制。在一个实施例中,DPU驱动器922可以以DRAM型地址或DRAM型地址的序列的形式,将指令发送到DPU,并且可控制数据移动到DPU内,或者移出DPU。DPU驱动器922可提供多个DPU通信(multi-DPU

communication)，并且处理DPU-CPU和/或DPU-GPU通信。

[0063] DPU编译器923可将来自DPU库921的DPU码编译为DPU指令，该DPU指令具有存储器地址的形式，可被DPU驱动器922用于控制DPU。由DPU编译器923产生的DPU指令可以是对DPU中的一行和/或两行进行操作的单一指令、向量指令和/或聚集向量(gathered vector)、读运算指令。

[0064] 框架层930可被配置为:将用户友好接口提供给库和驱动层920以及硬件层910。在一个实施例中，框架层930可提供用户友好接口，该用户友好接口可与在应用层940的宽范围的应用兼容，并且使得DPU硬件层910对于用户是透明的。在另一实施例中，框架层930可包括将定量功能(quantitation function)加到现有的传统方法的框架扩展，诸如，Torch7式应用和张量流(TensorFlow)式应用，但不限于此。在一个实施例中，框架层930可包括将定量功能加到训练算法。在另一实施例中，框架层930可将置换(override)提供到除法、乘法和平方根的现有批量标准化方法，以成为除法、乘法和平方根的移位近似方法。在另一实施例中，框架层930可提供允许用户设置用于计算的位数的扩展。在另一实施例中，框架层930提供将来自DPU库和驱动层920的多个DPU API封入(wrap)到框架层930的能力，使得用户可类似于使用多个GPU在硬件层使用多个DPU。框架层930的另一特征允许用户将功能分配给在硬件层910的DPU或GPU。

[0065] 应用940可包括宽范围的应用，诸如(但不限于)，图像加标签处理、自驱动/驾驶车辆(self-driving/piloting vehicle)、AlphaGo式的深度思考(deep-mind)应用和/或语音研究。

[0066] 如本领域技术人员将认识到的，在此公开的新的构思可在宽范围应用之上进行改变和修改。因此，要求保护的主题的范围不应受限于以上讨论的特定的示例性教导，而是由权利要求来定义。

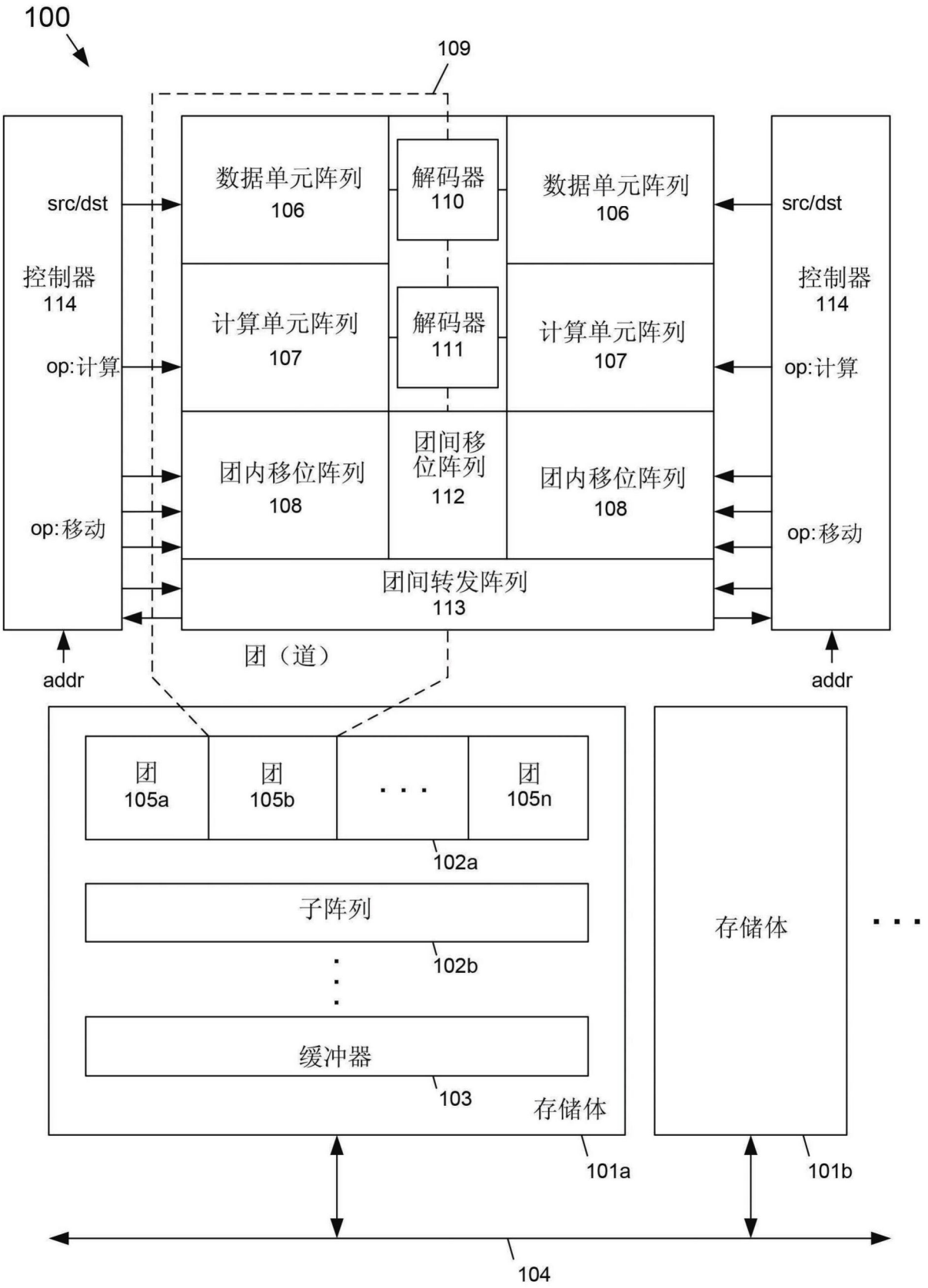

100
109
src/dst
控制器
114
op:计算
op:移动
数据单元阵列
106
解码器
110
数据单元阵列
106
计算单元阵列
107
解码器
111
计算单元阵列
107
团内移位阵列
108
团间移位阵列
112
团内移位阵列
108
团间转发阵列
113
src/dst
控制器
114
op:计算
op:移动
addr
团（道）
addr
团
105a
团
105b
团
105n
102a
子阵列
102b
缓冲器
103
存储体
101a
存储体
101b
104

图1

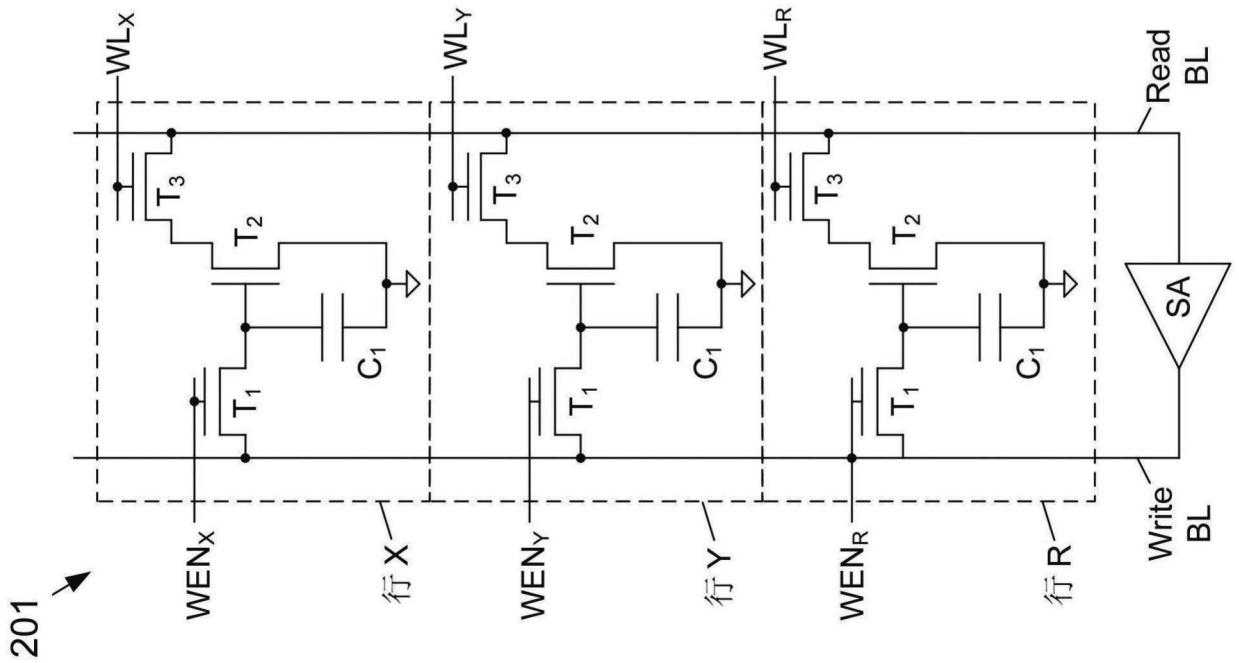

WLX
WLY
WLR
Read BL
T3
T2
T1
C1
SA
WENX
WENY
WENR
行 X
行 Y
行 R
Write BL
201

图2A

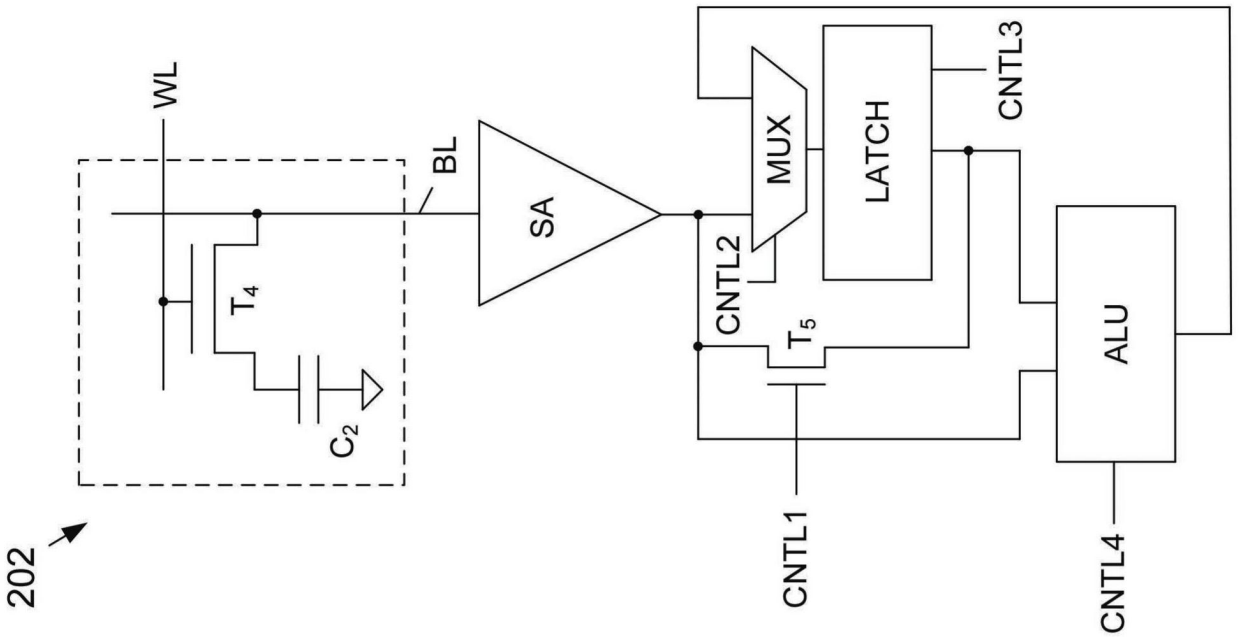

WL
BL
SA
MUX
LATCH
CNTL3
CNTL2
T4
T5
C2
ALU
CNTL1
CNTL4
202

图2B

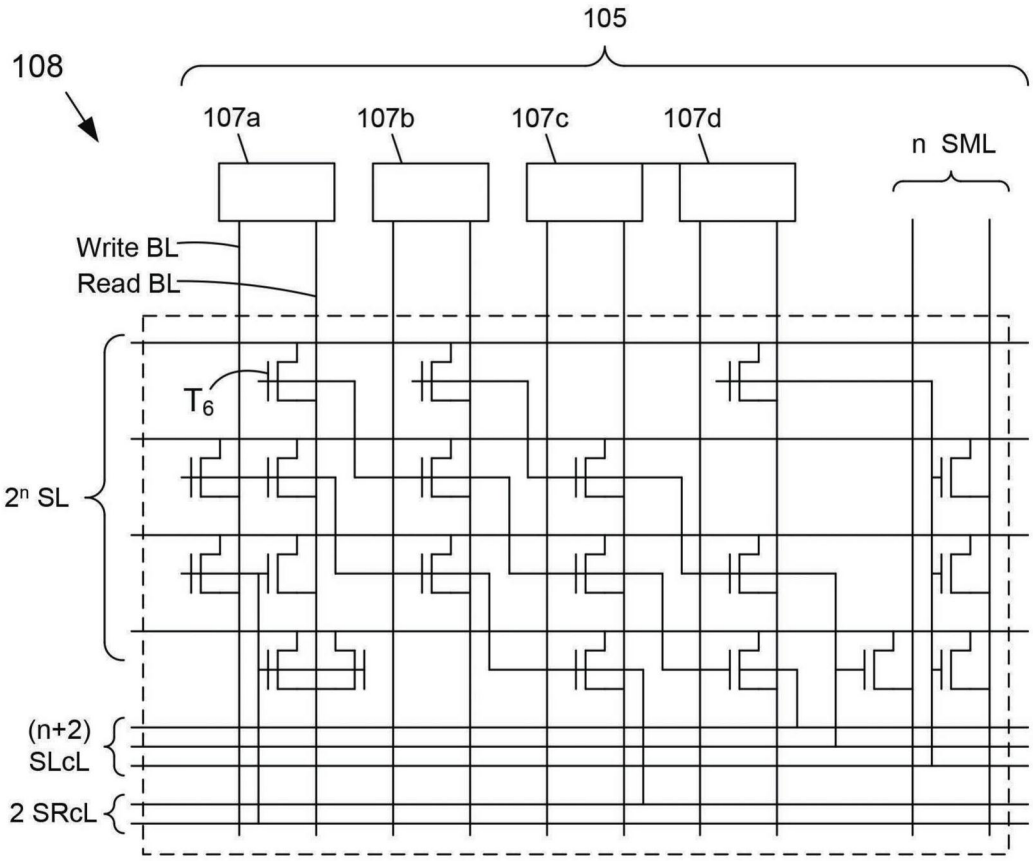

105
108
107a
107b
107c
107d
n SML
Write BL
Read BL
T6
2n SL
(n+2)
SLcL
2 SRcL

图3

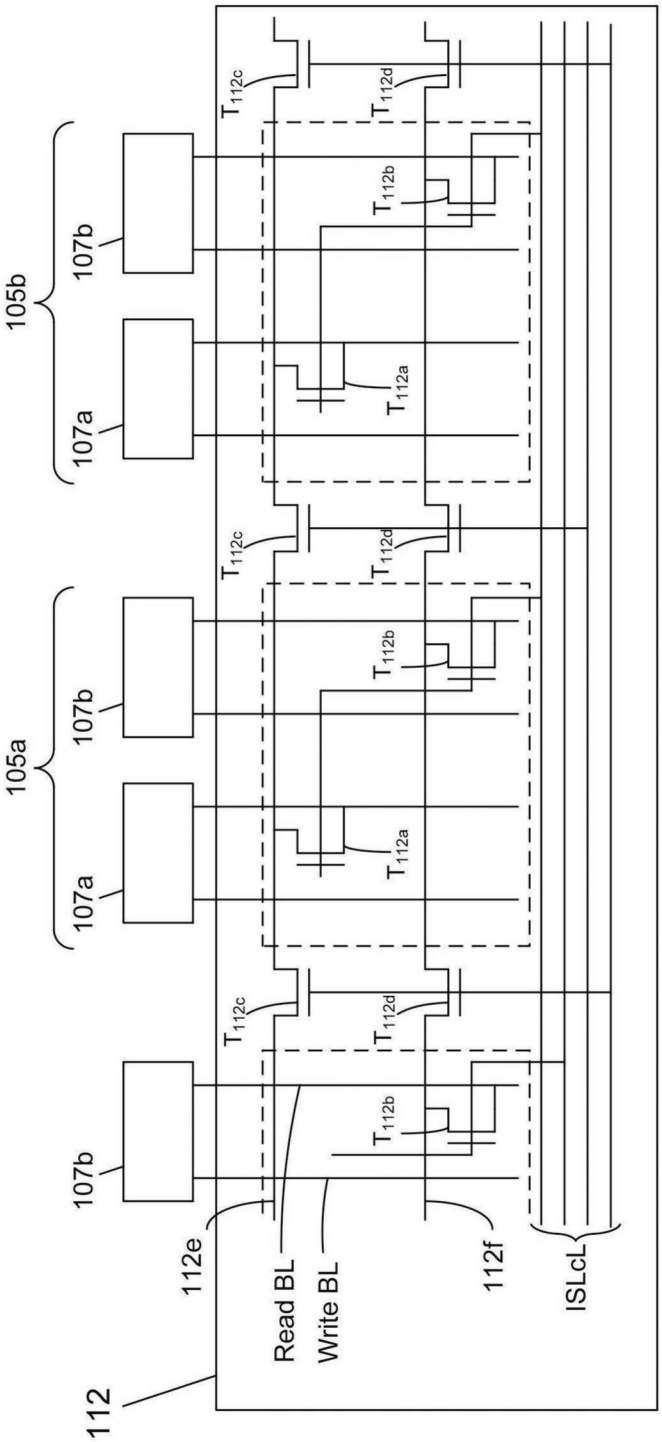

105b
107b
107a
105a
107b
107a
107b
$T_{112c}$
$T_{112d}$
$T_{112b}$
$T_{112a}$
$T_{112c}$
$T_{112d}$
$T_{112b}$
$T_{112a}$
$T_{112c}$
$T_{112d}$
$T_{112b}$
112e
Read BL
Write BL
112f
ISLcL
112

图4A

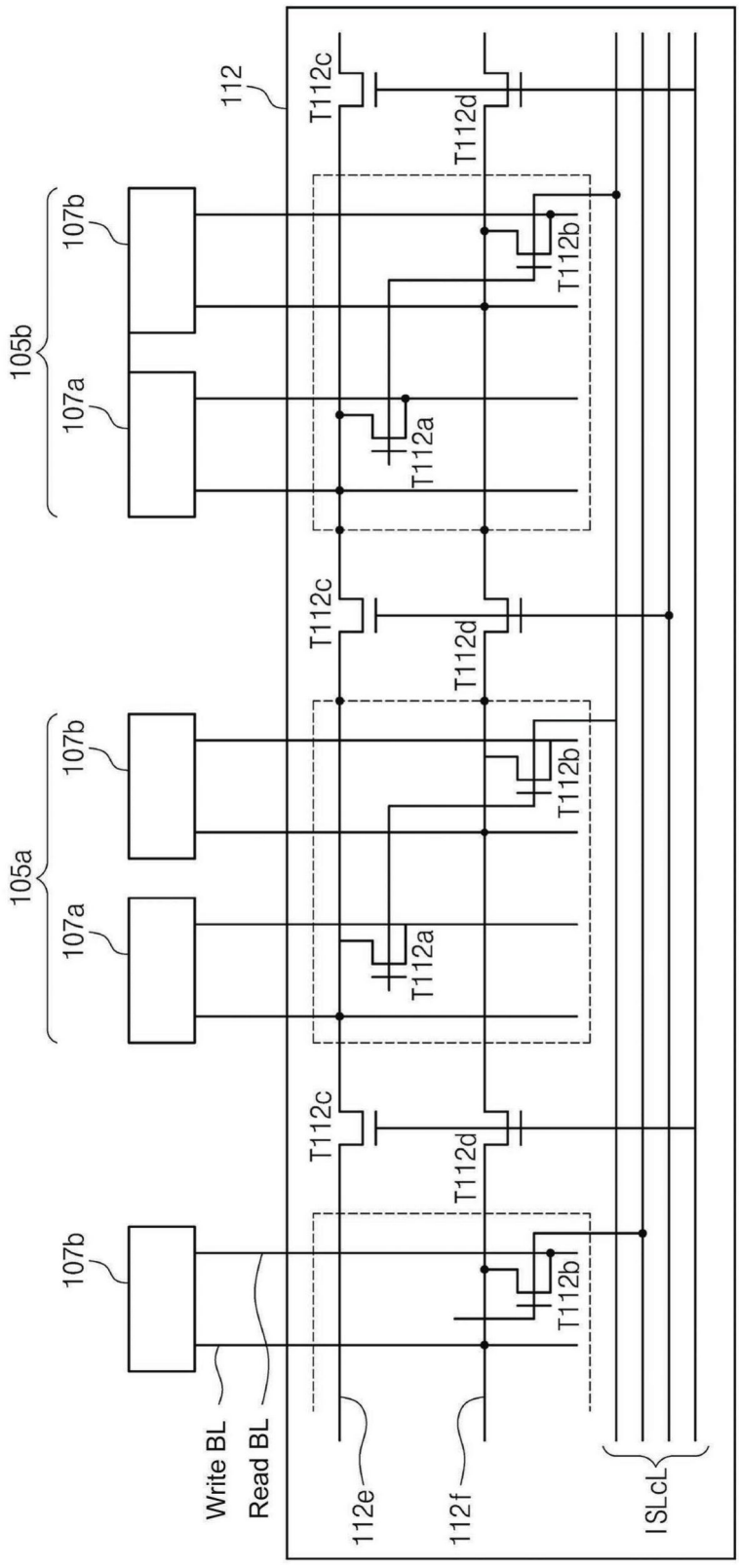

112
T112c
T112d
T112b
T112a
105b
107b
107a
105a
Write BL
Read BL
112e
112f
ISLcL

图4B

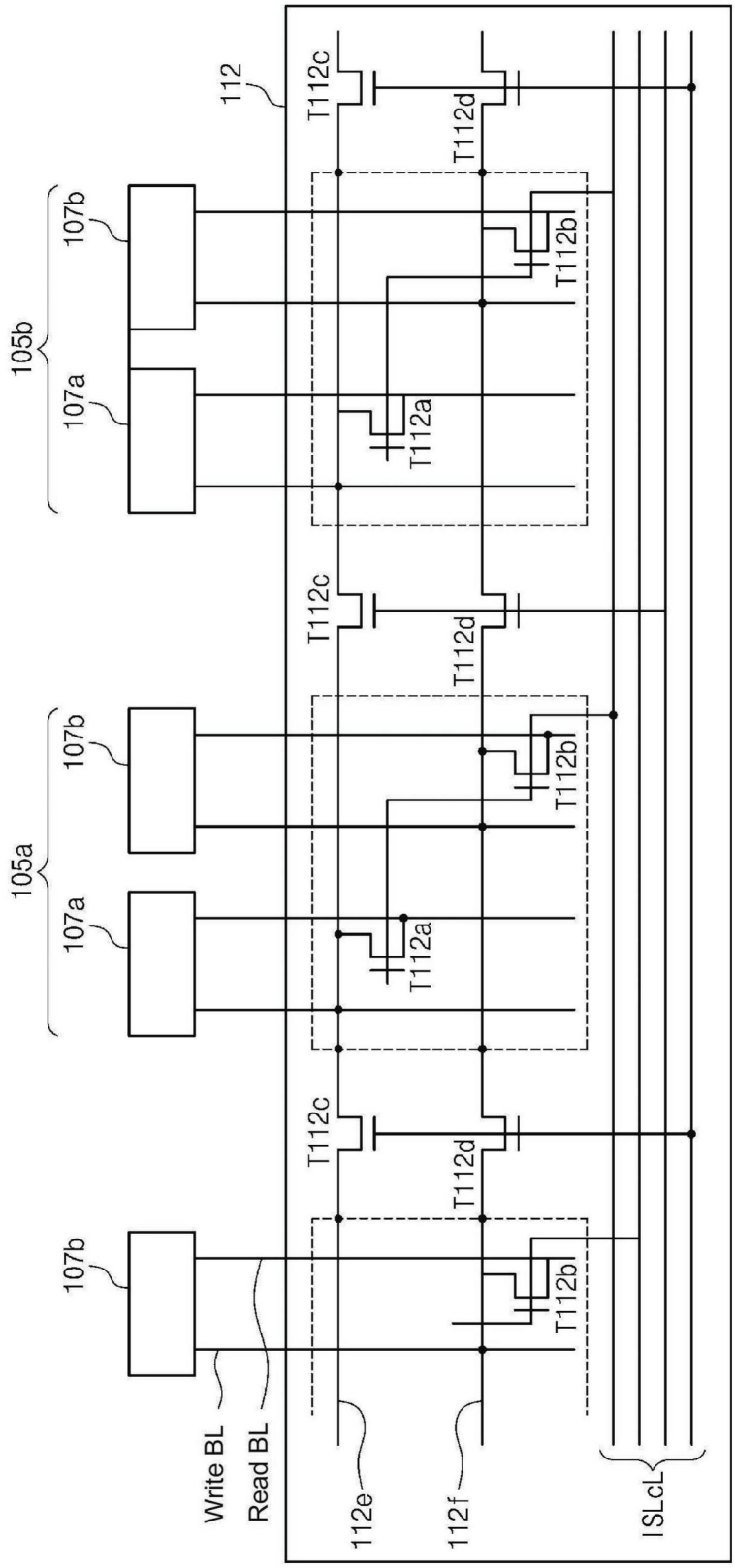

112
T112c
T112d
T112b
T112a
105b
107b
107a
105a
Write BL
Read BL
112e
112f
ISLcL

图4C

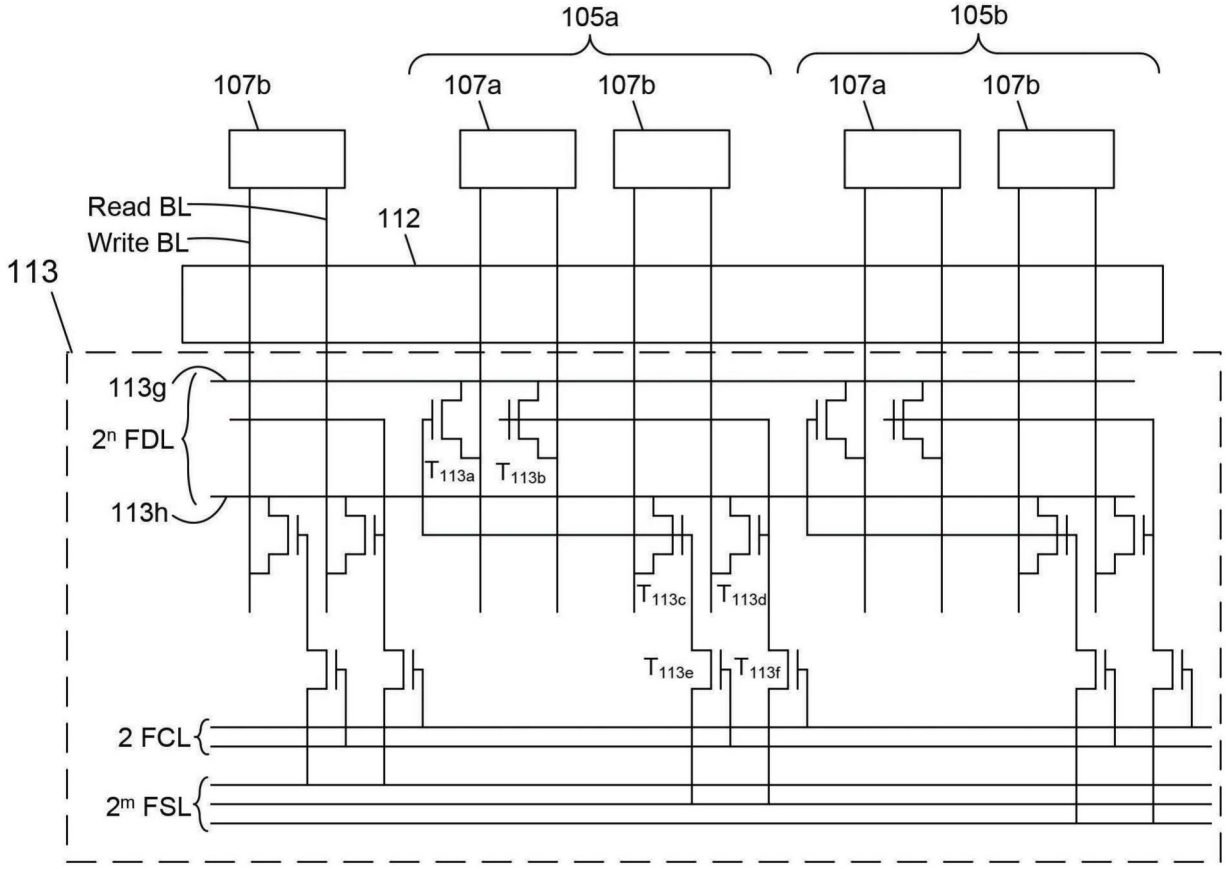

105a
105b
107b
107a
107b
107a
107b
Read BL
Write BL
112
113
113g
2n FDL
113h
T113a
T113b
T113c
T113d
T113e
T113f
2 FCL
2m FSL

图5

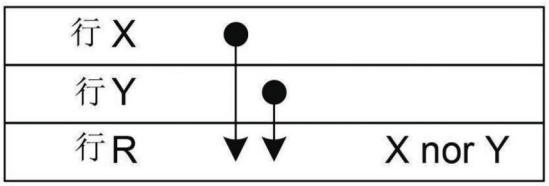

行 X
行 Y
行 R
X nor Y

图6A

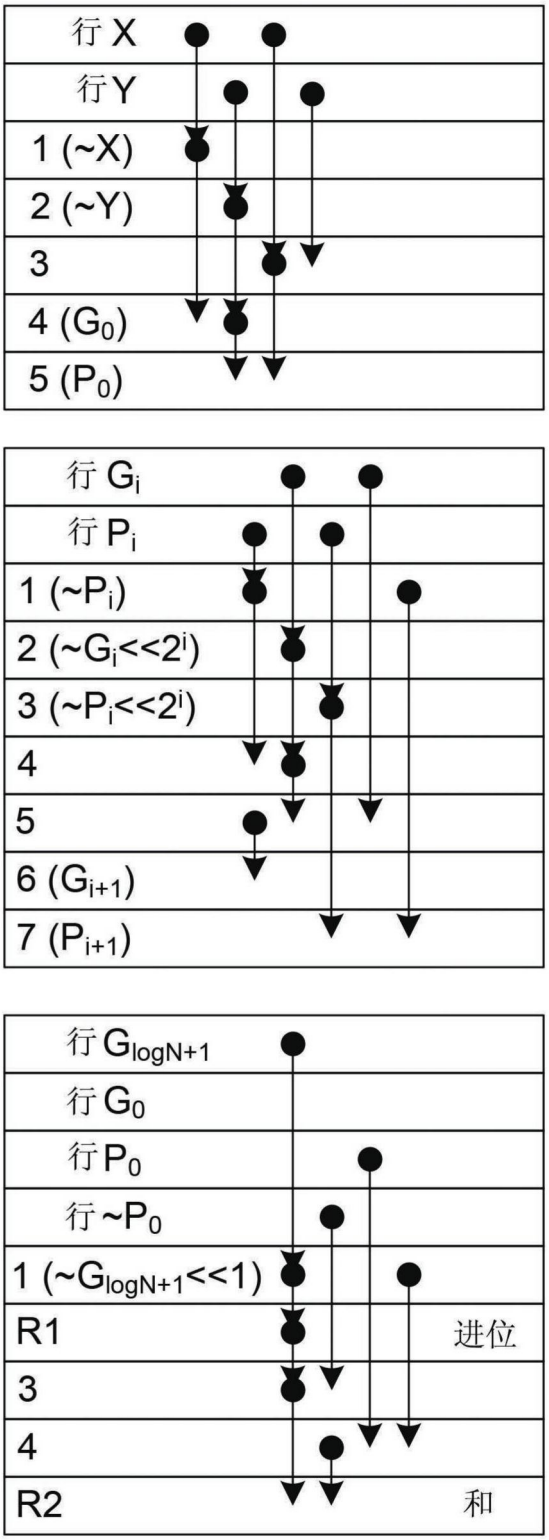

行 X
行 Y
1 (~X)
2 (~Y)
3
4 (G0)
5 (P0)
行 Gi
行 Pi
1 (~Pi)
2 (~Gi<<2i)
3 (~Pi<<2i)
4
5
6 (Gi+1)
7 (Pi+1)
行 GlogN+1
行 G0
行 P0
行 ~P0
1 (~GlogN+1<<1)
R1
进位
3
4
R2
和

图6B

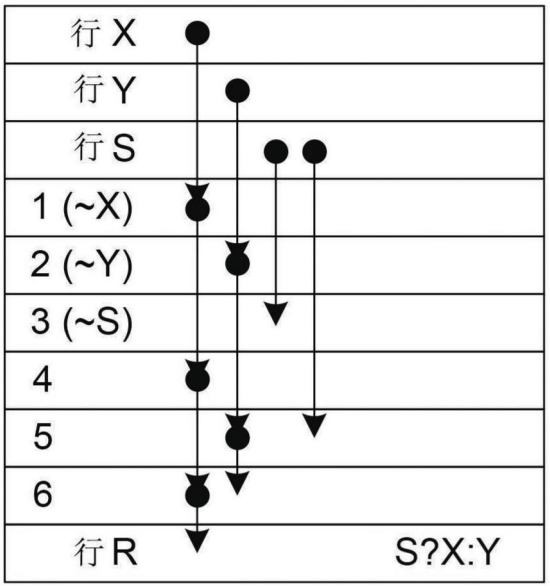



图6C

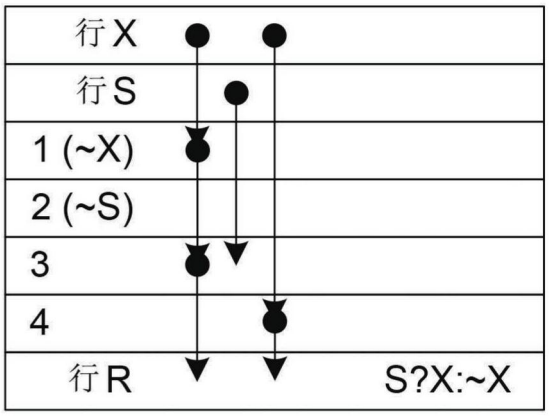



图6D

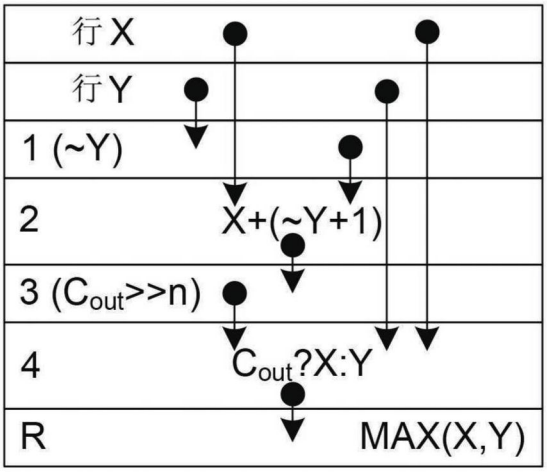



图6E

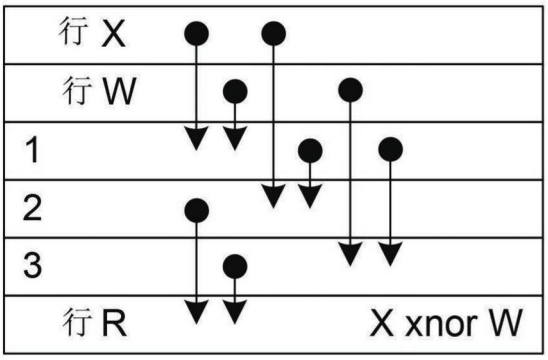



图6F

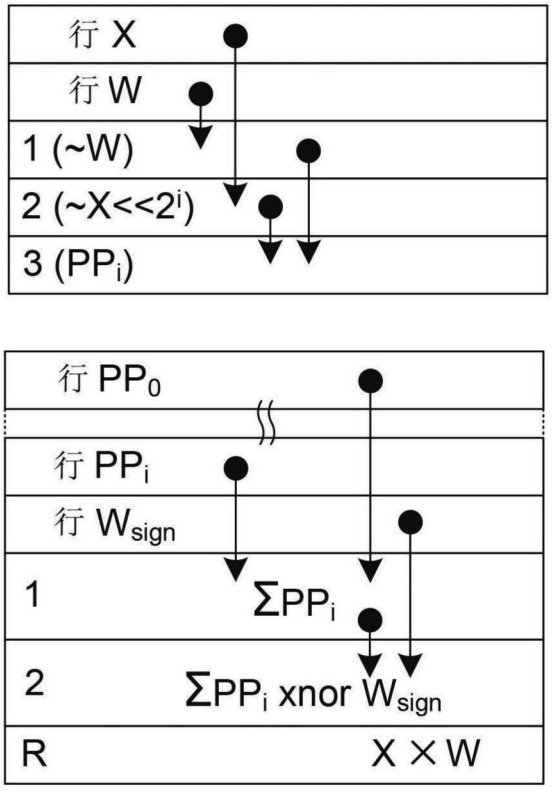

行 X
行 W
1 (~W)
2 (~X<<2^i)
3 (PP_i)
行 PP_0
行 PP_i
行 W_sign
1
ΣPP_i
2
ΣPP_i xnor W_sign
R
X × W

图6G

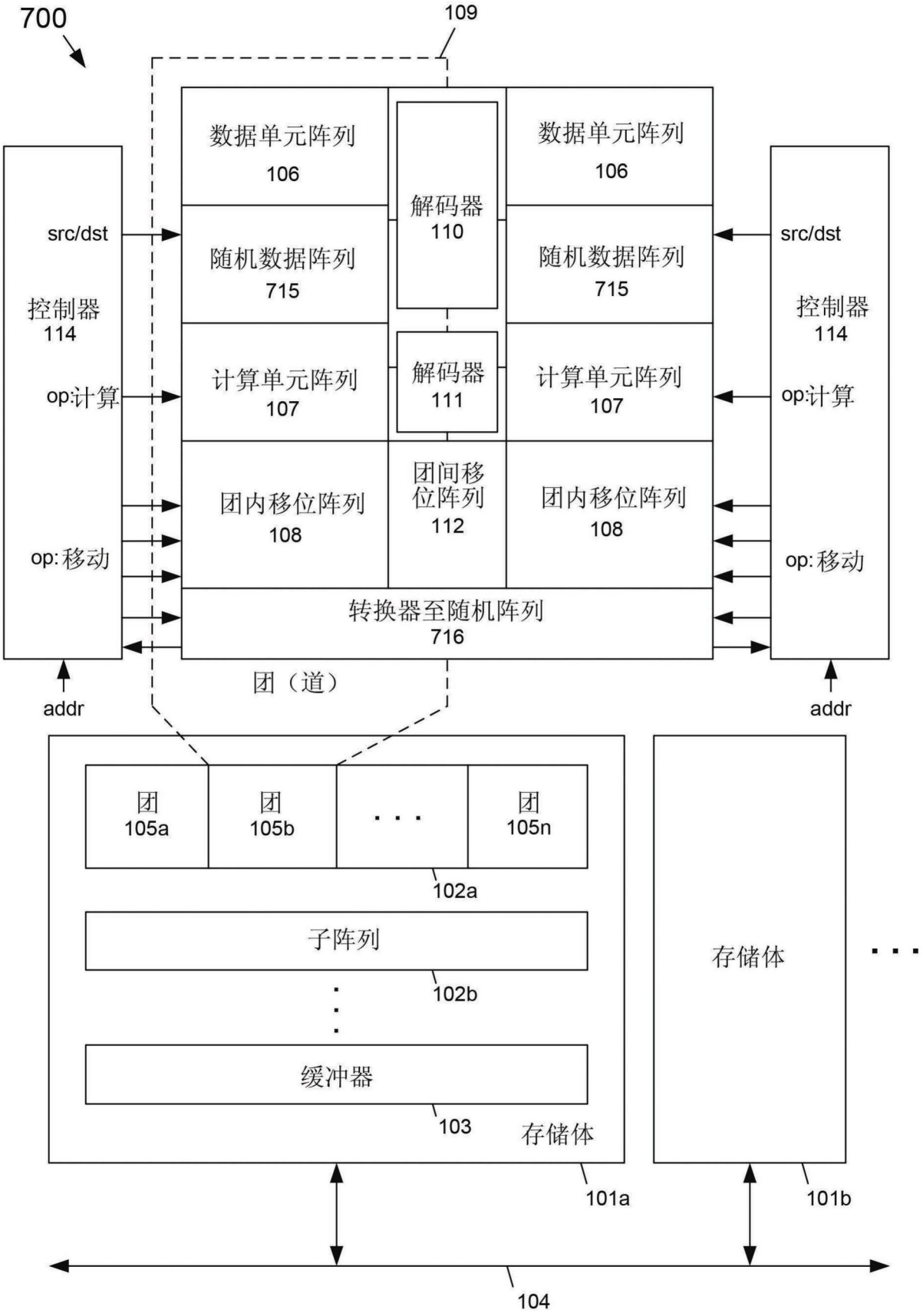

700
109
数据单元阵列
106
解码器
110
数据单元阵列
106
src/dst
随机数据阵列
715
随机数据阵列
715
src/dst
控制器
114
计算单元阵列
107
解码器
111
计算单元阵列
107
控制器
114
op:计算
op:计算
团内移位阵列
108
团间移位阵列
112
团内移位阵列
108
op:移动
op:移动
转换器至随机阵列
716
addr
团（道）
addr
团
105a
团
105b
团
105n
102a
子阵列
102b
缓冲器
103
存储体
存储体
101a
101b
104

图7

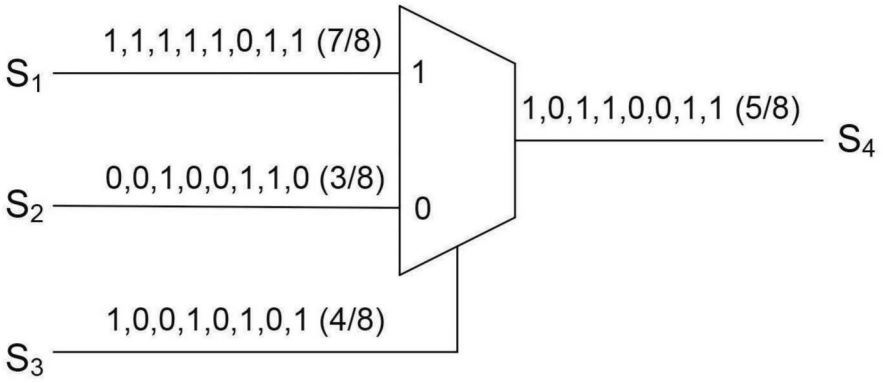

1,1,1,1,1,0,1,1 (7/8)
S1
1
1,0,1,1,0,0,1,1 (5/8)
S4
0,0,1,0,0,1,1,0 (3/8)
S2
0
1,0,0,1,0,1,0,1 (4/8)
S3

图8A

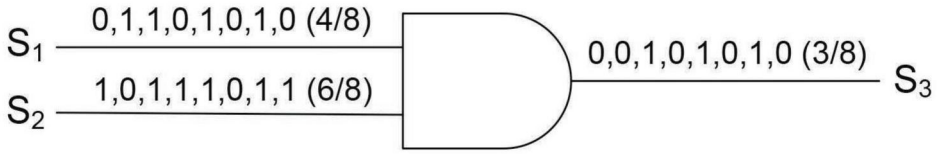

0,1,1,0,1,0,1,0 (4/8)
S1
0,0,1,0,1,0,1,0 (3/8)
S3
1,0,1,1,1,0,1,1 (6/8)
S2

图8B

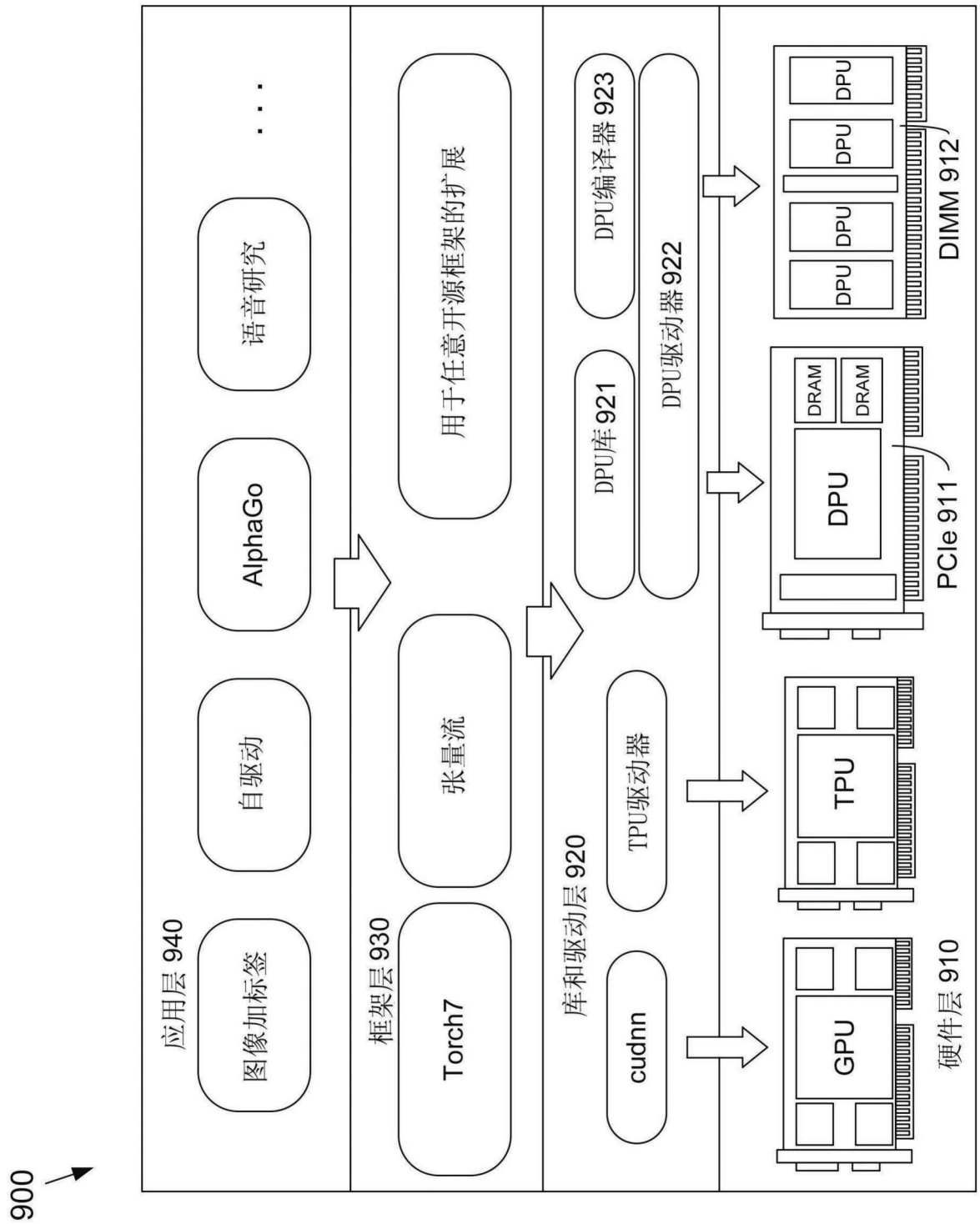

900
应用层 940
图像加标签
自驱动
AlphaGo
语音研究
. . .
框架层 930
Torch7
张量流
用于任意开源框架的扩展
库和驱动层 920
cudnn
TPU驱动器
DPU库 921
DPU驱动器 922
DPU编译器 923
硬件层 910
GPU
TPU
DPU
DRAM
DRAM
PCIe 911
DPU
DPU
DPU
DIMM 912

图9