(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 110168761 A
(43)申请公布日 2019.08.23

(21)申请号 201880006821.2

(22)申请日 2018.01.03

(30)优先权数据
15/405,555 2017.01.13 US

(85)PCT国际申请进入国家阶段日
2019.07.12

(86)PCT国际申请的申请数据
PCT/IB2018/050033 2018.01.03

(87)PCT国际申请的公布数据
WO2018/130914 EN 2018.07.19

(71)申请人 国际商业机器公司
地址 美国纽约阿芒克

(72)发明人 T·S·格申 K·W·布鲁 S·辛格 D·纽恩斯

(74)专利代理机构 北京市金杜律师事务所 11256
代理人 酆迅 辛鸣

(51)Int.Cl.
H01L 45/00(2006.01)

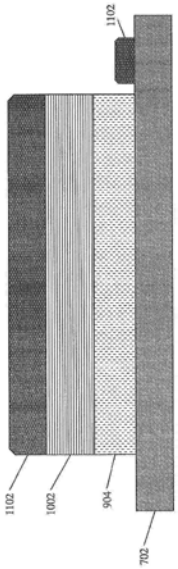
权利要求书2页 说明书12页 附图14页

(54)发明名称

基于忆阻器件过渡金属氧化物的碱性掺杂的忆阻器件

(57)摘要

忆阻器件包括第一导电材料层。氧化物材料层被布置在第一导电层上。第二导电材料层被布置在氧化物材料层上,其中第二导电层包括金属-碱合金。



CN 110168761 A

1.一种忆阻器件,包括:

第一导电材料层;

氧化物材料层,被布置在所述第一导电层上;以及

第二导电材料层,被布置在所述氧化物材料层上,其中所述第二导电材料层包括金属-碱合金。

2.根据权利要求1所述的忆阻器件,其中所述氧化物材料层被插入有碱金属。

3.根据权利要求1所述的忆阻器件,其中所述氧化物材料层包括过渡金属氧化物。

4.根据权利要求1所述的忆阻器件,其中所述第二导电材料层包括金属-碱合金。

5.根据权利要求1所述的忆阻器件,其中所述第二导电层和所述氧化物材料层被配置为响应于被施加到所述第二导电材料层的正电压脉冲而引起碱金属向所述氧化物材料层中的插入。

6.根据权利要求1所述的忆阻器件,还包括被布置在所述氧化物材料层上的扩散阻挡层。

7.根据权利要求3所述的忆阻器件,其中所述过渡金属氧化物包括氧化钛。

8.根据权利要求1所述的忆阻器件,其中所述第二导电层和所述氧化物材料层被配置为响应于被施加到所述第二导电层的正电压脉冲而引起所述碱金属向所述氧化物材料层中的插入。

9.一种形成忆阻器件的方法,所述方法包括:

在第一导电材料层的一部分上沉积氧化物材料层;以及

在所述氧化物材料层的一部分上沉积第二导电材料层,其中所述第二导电材料层包括金属-碱合金。

10.根据权利要求9所述的方法,其中所述氧化物材料层包括过渡金属氧化物。

11.根据权利要求9所述的方法,其中所述第二导电材料层包括金属碱合金。

12.根据权利要求9所述的方法,还包括将所述氧化物材料层暴露于碱金属一段持续时间。

13.根据权利要求9所述的方法,还包括配置所述第二导电层和所述氧化物材料层,从而使得响应于被施加到所述第二导电层的正电压脉冲,碱金属插入所述氧化物材料层中。

14.根据权利要求9所述的方法,还在所述氧化物材料层的一部分上沉积扩散阻挡层。

15.根据权利要求9所述的方法,其中所述第一导电材料层包括氟掺杂的氧化锡。

16.根据权利要求9所述的方法,其中所述金属-碱合金包括钼-锂合金。

17.一种忆阻器件,包括:

第一导电材料层;

氧化物材料层,被布置在所述第一导电材料层上,其中所述氧化物材料层被暴露于碱金属一段持续时间;以及

第二导电材料层,被布置在所述氧化物材料层上。

18.根据权利要求17所述的器件,其中所述氧化物材料层包括过渡金属氧化物。

19.根据权利要求18所述的器件,其中所述碱金属包括正丁基锂。

20.根据权利要求17所述的器件,其中所述第二导电材料层包括金属-碱合金。

21.根据权利要求19所述的器件,其中所述第二导电材料层和所述氧化物材料层被配

置为响应于被施加到所述第二导电层的负电压脉冲而引起所述碱金属向所述第二导电材料层中的。

基于忆阻器件过渡金属氧化物的碱性掺杂的忆阻器件

背景技术

[0001] 本发明涉及忆阻器件。更具体地，本发明涉及基于过渡金属氧化物的碱性掺杂的忆阻器件。

[0002] “机器学习”被用于广泛地描述从数据学习的电子系统的主要功能。在加速的机器学习和认知科学中，人工神经网络(ANN)是一系列统计学习模型，其灵感来自动物的生物神经网络，并且特别地是大脑。ANN可以被用于估计或近似取决于大量输入且通常未知的系统和功能。ANN架构、神经形态微芯片和超高密度非易失性存储器可以从被称为交叉(cross-bar)阵列的高密度低成本电路架构被形成。基本交叉阵列配置包括一组导电行线和被形成为与该一组导电行线交叉的一组导电列线。两组线之间的交叉点由所谓的交叉点器件分开，交叉点器件可以从薄膜材料被形成。交叉点器件可以被实现为所谓的忆阻器件。忆阻器件的特性包括非易失性、存储可变电阻值的能力以及使用电流或电压脉冲调高或调低电阻的能力。

发明内容

[0003] 根据本发明的实施例，忆阻器件包括第一导电材料层。氧化物材料层被布置在第一导电层上。并且第二导电材料层被布置在氧化物材料层上，其中第二导电材料层包括金属-碱合金。

[0004] 根据本发明的另一实施例，一种形成忆阻器件的方法包括在第一导电材料层的一部分上沉积氧化物材料层。第二导电材料层被沉积在氧化物材料的层的一部分上，其中第二导电材料层包括金属-碱合金。

[0005] 根据本发明的另一实施例，一种忆阻器件包括第一导电材料层。氧化物材料层被布置在第一导电材料层上。扩散阻挡层被布置在氧化物材料层上。第二导电材料层被布置在氧化物材料层上，其中第二导电材料层包括金属-碱合金。

[0006] 根据本发明的另一实施例，一种形成忆阻器件的方法包括在第一导电材料层的一部分上沉积氧化物材料层。扩散阻挡层被沉积在氧化物材料层的一部分上。第二导电材料层被沉积在氧化物材料层的一部分上，其中第二导电材料层包括金属-碱合金。

[0007] 根据本发明的另一实施例，一种忆阻器件包括第一导电材料层。氧化物材料层被布置在第一导电材料层上，其中氧化物材料层被暴露于碱金属持续一段持续时间。第二导电材料层被布置在氧化物材料层上。

[0008] 通过本文描述的技术实现了附加的特征和优点。本文详细描述了其他实施例和方面。为了更好地理解，参考说明和附图。

附图说明

[0009] 通过结合附图进行的以下详细描述，实施例的前述和其他特征和优点将变得显而易见，在附图中：

[0010] 图1描绘了生物神经元的输入和输出连接的简化示图；

[0011] 图2描绘了图1中所示的生物神经元的已知简化模型；

[0012] 图3描绘了并入图2中所示的生物神经元模型的ANN的已知简化模型；

[0013] 图4描绘了已知权重更新方法的简化框图；

[0014] 图5描绘了能够在一个或多个实施例中被使用的随机计算方法的简化框图；

[0015] 图6描绘了控制无源双端忆阻器的操作的已知方程；

[0016] 图7描绘了根据一个或多个实施例的具有外周神经元的突触的单个矩阵；

[0017] 图8描绘了根据一个或多个实施例的、在导电材料层的有源区上沉积氧化物层之后的忆阻器件的侧视图；

[0018] 图9描绘了根据一个或多个实施例的、随着氧化物层被插入碱金属的忆阻器件的侧视图；

[0019] 图10描绘了根据一个或多个实施例的、在氧化物层被插入碱金属之后的忆阻器件的侧视图；

[0020] 图11描绘了根据一个或多个实施例的、在合金层在氧化物层上的沉积之后的忆阻器件的侧视图；

[0021] 图12描绘了在金属触点在合金层和导电材料层上的沉积之后的忆阻器件的侧视图；

[0022] 图13描绘了在一个或多个正电压脉冲向金属触点的施加之后的忆阻器件的侧视图；

[0023] 图14描绘了在一个或多个负电压脉冲向金属触点的施加之后的忆阻器件的侧视图；

[0024] 图15描绘了忆阻器件的备选示例性实施例的侧视图；

[0025] 图16描绘了忆阻器件的备选示例性实施例的侧视图；以及

[0026] 图17描绘了忆阻器件的备选示例性实施例的侧视图。

具体实施方式

[0027] 这里参考相关附图描述了本发明的各种实施例。在不脱离本发明的范围的情况下，可以设计备选实施例。注意，在以下描述和在附图中，在元件之间阐述了各种连接和位置关系(例如，上方、下方、相邻等)。除非另有说明，否则这些连接和/或位置关系可以是直接的或间接的，并且本发明并不旨在在这方面中进行限制。因此，实体的耦合可以指直接或间接耦合，并且实体之间的位置关系可以是直接或间接的位置关系。作为间接位置关系的示例，本说明书中对在层“B”上形成层“A”的引用包括其中一个或多个中间层(例如，层“C”)在层“A”和层“B”之间的情况，只要层“A”和层“B”的相关特性和功能基本上不被(多个)中间层改变。

[0028] 以下定义和缩写将被用于解释权利要求和说明书。如这里所使用的，术语“包括(comprises)”，“包括(comprising)”，“包括(includes)”，“包括(including)”，“具有(has)”，“具有(having)”，“包含(contains)”或“包含(containing)”或它们的任何其他变型旨在涵盖非-排他包含。例如，包括元素的列表的组合物、混合物、过程方法、物品或装置不必仅限于那些元素，而是可以包括未被明确列出的其他元素或这样的组合物、混合物，过程、方法、物品或装置固有的元素。

[0029] 附加地，术语“示例性”在本文中被用于表示“用作示例、实例或说明”。本文中被描述为“示例性”的任何实施例或设计不必被解释为比其他实施例或者设计优选或有利。术语“至少一个”和“一个或多个”被理解为包括大于或等于一的任何整数，即，一个、两个、三个、四个等。术语“多个”被理解为包括大于或等于二的任何整数，即，两个、三个、四个、五个等。术语“连接”可以包括间接“连接”和直接“连接”。

[0030] 说明书中对“一个实施例”、“实施例”、“示例实施例”等的引用指示所描述的实施例可以包括特定特征、结构或特性，但是每个实施例可以或可以不包括特定特征、结构或特性。此外，这样短语不必是指同一实施例。此外，当结合实施例描述特定特征、结构或特性时，主张的是无论是否被明确描述，结合其他实施例来影响这样的特征、结构或特性在本领域技术人员的知识内。

[0031] 出于以下描述的目的，术语“上”、“下”、“右”、“左”、“竖直”、“水平”、“顶部”、“底部”以及它们的派生词应涉及所描述的结构和方法，如附图中所示。术语“覆盖”、“顶上”、“顶部”、“被定位在……上”或“被定位在……顶部”意味着第一元件(诸如第一结构)存在于第二元件(诸如第二结构)上，其中在第一元件和第二元件之间可以存在诸如接口结构的中间元件。术语“直接接触”是指第一元件(诸如第一结构)和第二元件(诸如第二结构)在两个元件的接口处没有任何中间导电、绝缘或半导体层的情况下被连接。应当注意，术语“选择性地”(诸如例如“对第二元素有选择性的第一元素”)意味着可以蚀刻第一元素，并且第二元素可以用作蚀刻停止。术语“大约”旨在包括与基于提交申请时可用设备的特定量的测量相关联的误差程度。例如，“大约”可以包括给定值的±8％或5％，或2％的范围。

[0032] 通常，被用于形成将被封装到IC中的微芯片的各种工艺落入四大类中，即，膜沉积、去除/蚀刻、半导体掺杂和图案化/光刻。沉积是生长、涂覆或以其他方式将材料转移到晶片上的任何过程。可用的技术包括物理气相沉积(PVD)、化学气相沉积(CVD)、等离子体增强化学气相沉积(PECVD)、电化学沉积(ECD)、分子束外延(MBE)以及最近和原子层沉积(ALD)等等。

[0033] 去除/蚀刻是从晶片去除材料的任何过程。示例包括蚀刻工艺(湿法或干法)和化学机械平坦化(CMP)等。湿法蚀刻工艺(诸如缓冲氢氟酸(BHF)蚀刻)是使用液体化学品或蚀刻剂以从表面去除材料的材料去除工艺。干蚀刻工艺(诸如反应离子蚀刻(RIE))通过将材料暴露于离子的轰击，使用化学反应性等离子体来去除材料(诸如半导体材料的掩模化图案)，该轰击从暴露表面逐出材料的部分。通过电磁场在低压(真空)下生成等离子体。

[0034] 半导体光刻是在半导体衬底上形成三维浮雕图像或图案，以用于随后将图案转移到衬底。在半导体光刻中，图案由被称为光致抗蚀剂的光敏聚合物形成。为了构建构成晶体管的复杂结构和连接电路的数百万个晶体管的许多导线，光刻和蚀刻图案转移步骤被重复多次。被印刷在晶片上的每个图案与先前形成的图案对齐，并且缓慢地构建导体、绝缘体和选择性掺杂区域以形成最终器件。

[0035] 现在转向与本发明相关的技术的更详细描述，如在此先前所述，变形神经网络(ANN)通常被体现为互连处理器元件的所谓的“神经形态”系统，其充当模拟“神经元”并以电子信号的形式在彼此之间交换“信息”。类似于在生物神经元之间携带消息的突触神经递质连接的所谓的“可塑性”，在模拟神经元之间携带电子消息的ANN中的连接被提供有对应于给定连接的强度或弱度(weakness)的数字权重。可以基于经验来调整和调节权重，使得

ANN适应于输入并且能够学习。例如，用于手写识别的ANN由一组输入神经元定义，该一组输入神经元可以由输入图像的像素激活。在通过由网络的设计者确定的函数加权和变换之后，这些输入神经元的激活然后被传递到其他下游神经元，这些下游神经元通常被称为“隐藏”神经元。这一过程被重复直到输出神经元被激活。激活的输出神经元确定哪个字符被读取。

[0036] 交叉阵列(也被称为交叉点阵列或交叉线阵列)是被用于形成各种电子电路和器件的高密度低成本电路架构，包括ANN架构、神经形态微芯片和超高密度非易失性存储器。基本交叉阵列配置包括一组导电行线和被形成为与该一组导电行线交叉的一组导电列线。两组线之间的交叉点由所谓的交叉点器件分开，交叉点器件可以从薄膜材料被形成。

[0037] 实际上，交叉点器件充当ANN在神经元之间的加权连接。纳米级双端器件(例如具有“理想”导通状态切换特性的忆阻器)通常被用作交叉点器件，以便以高能效模拟突触可塑性。通过控制在行和列线的各个线之间被施加的电压，可以改变理想忆阻器材料的导通状态(例如，电阻)。可以通过改变忆阻器材料在交叉点处的导通状态来存储数字数据，以实现高导通状态或低导通状态。忆阻器材料还可以被编程为通过选择性地设置材料的导通状态来维持两个或更多个不同的导通状态。可以通过跨材料施加电压并测量通过目标交叉点器件的电流来读取忆阻器材料的导通状态。

[0038] 为了限制功耗，ANN芯片架构的交叉点器件通常被设计为利用离线学习技术，其中一旦初始训练阶段已被解决，目标函数的近似就不会改变。离线学习允许简化交叉型ANN架构的交叉点器件，从而使得它们汲取非常少的功率。

[0039] 尽管存在针对较低功耗的可能性，但执行离线训练可能是困难且资源密集的，因为在训练期间通常必须修改ANN模型中的大量可调参数(例如，权重)以匹配针对训练数据的输入-输出。因此，简化ANN架构的交叉点器件以使省电的离线学习技术优先化通常意味着训练速度和训练效率不是最佳的。

[0040] 尽管本发明的实施例涉及电子系统，但为了便于参考和说明，例如使用诸如神经元、可塑性和突触的神经学术语来描述电子系统的各个方面。将理解，对于本文对电子系统的任何讨论或说明，使用神经学术语或神经学简写符号是为了便于参考，并且旨在涵盖所描述的神经功能或神经学部件的(多个)神经形态ANN等同物。

[0041] ANN并入来自各种学科的知识，这些学科包括神经生理学、认知科学/心理学、物理学(统计力学)、控制理论、计算机科学、人工智能、统计学/数学、模式识别、计算机视觉、并行处理和硬件(例如，数字/模拟/VLSI/光学)。代替利用操纵零和一的传统数字模型，ANN在处理元素之间建立连接，这些连接基本上是被估计或近似的核心系统功能的功能等价物。例如，IBM™的SyNapse™计算机芯片是电子神经形态机器的核心部件，它试图向哺乳动物的大脑提供类似的形式、功能和结构。虽然IBM SyNapse计算机芯片使用与传统计算机芯片相同的基本晶体管部件，但它的晶体管被配置为模仿神经元以及它们的突触连接的行为。IBM SyNapse计算机芯片使用仅超过一百万个模拟“神经元”的网络处理信息，这些神经元使用类似于生物神经元之间的突触通信的电尖峰彼此通信。IBM SyNapse架构包括读取存储器(即，模拟的“突触”)以及执行简单操作的处理器(即，模拟的“神经元”)的配置。这些处理器之间的通信通常位于不同的核心中，由片上网络路由器执行。

[0042] 现在将参考图1、图2和图3提供典型ANN如何操作的一般描述。如前所述，典型的

ANN对人脑建模，人脑包括被称为神经元的大约1000亿个互连细胞。图1描绘了具有路径104、106、108、110的生物神经元102的简化示图，路径104、106、108、110将其连接到上游输入112、114、下游输出s116和下游的“其他”神经元118，如图所示被配置和布置。每个生物神经元102通过路径104、106、108、110发送和接收电脉冲。这些电脉冲的性质以及它们如何在生物神经元102中被处理主要负责整体脑功能。生物神经元之间的通路连接可能强或弱。当给定神经元接收输入脉冲时，神经元根据神经元的功能处理输入并将该功能的结果发送给下游输出和/或下游的“其他”神经元。

[0043] 生物神经元102在图2中被建模为具有由图2中所示的方程描绘的数学函数f(x)的节点202。节点202从输入212、214取得电信号，将每个输入212、214乘以它们的相应连接路径204、206的强度，取得输入的总和，将该总和通过函数f(x)，并且生成结果216，其可以是最终输出或对另一节点的输出，或两者。在本说明书中，星号(*)被用于表示乘法。弱输入信号被乘以非常小的连接强度数，因此弱输入信号对功能的影响非常小。类似地，强输入信号被乘以更高的连接强度数，因此强输入信号对功能的影响更大。函数f(x)是设计选择，并且可以使用各种函数。针对f(x)的典型设计选择是双曲正切函数，它取前一个和的函数，并且输出负1和正1之间的数字。

[0044] 图3描绘了被组织为加权方向图的简化ANN模型300，其中人工神经元是节点(例如，302、308、316)，并且其中加权有向边(例如，m1到m20)连接节点。ANN模型300被组织以使得节点302、304、306是输入层节点，节点308、310、312、314是隐藏层节点，并且节点316、318是输出层节点。每个节点通过连接路径被连接到相邻层中的每个节点，连接路径在图3中被描绘为具有连接强度m1到m20的方向箭头。尽管仅示出了一个输入层、一个隐藏层和一个输出层，但实际上，可以提供多个输入层、隐藏层和输出层。

[0045] 类似于人脑的功能，ANN 300的每个输入层节点302、304、306直接从源(未示出)接收输入x1、x2、x3，没有连接强度调整并且没有节点求和。因此，y1＝f(x1)，y2＝f(x2)并且y3＝f(x3)，如图3底部列出的方程所示。每个隐藏层节点308、310、312、314根据与相关连接路径相关联的连接强度从所有输入层节点302、304、306接收它的输入。因此，在隐藏层节点308中，y4＝f(m1*y1+m5*y2+m9*y3)，其中*表示乘法。对隐藏层节点310、312、314和输出层节点316、318执行类似的连接强度乘法和节点求和，如图3底部所示的定义函数y5至y9的方程所示。

[0046] ANN模型300一次处理一个数据记录，并且通过将记录的最初任意分类与记录的已知实际分类进行比较来“学习”。使用被称为“反向传播”(即“错误的向后传播”)的训练方法，来自第一记录的初始分类的错误被反馈到网络中并且被用于第二次修改网络的加权连接，并且这种反馈过程持续多次迭代。在ANN的训练阶段中，已知针对每个记录的正确分类，并且因此输出节点可以被分配“正确”值。例如，针对对应于正确类的节点的的为“1”(或0.9)的节点值，以及针对其他节点的为“0”(或0.1)的节点值。因此，可以将网络的针对输出节点的计算出的值与这些“正确”值比较，并且计算针对每个节点的误差项(即“delta”规则)。然后使用这些误差项来调整隐藏层中的权重，以便在下一次迭代中输出值将更接近“正确”值。

[0047] 存在许多类型的神经网络，但是两个最广泛的类别是前馈和反馈/循环网络。ANN模型300是具有输入、输出和隐藏层的非循环前馈网络。信号只能在一个方向中行进。输入

数据被传递到执行计算的处理元件的层上。每个处理元件基于它的输入的加权和来进行它的计算。新的计算值然后将成为馈送下一层的新输入值。这以过程持续直到它经过所有层并确定输出。阈值传递函数有时被用于量化输出层中的神经元的输出。

[0048] 反馈/循环网络包括反馈路径，这意味着信号可以使用环路在两个方向中行进。允许节点之间的所有可能连接。因为在这种类型的网络中存在环路，所以在某些操作下，它可以变成非线性动态系统，其不断变化直到达它到均衡的状态。反馈网络通常被用在关联性存储器和优化问题中，其中网络寻找互连因子的最佳布置。

[0049] 前馈和循环ANN架构中的机器学习的速度和效率取决于ANN交叉阵列的交叉点器件如何有效地执行典型机器学习算法的核心操作。尽管难以制定机器学习的精确定义，但是ANN上下文中的学习过程可以被视为更新交叉点器件连接权重的问题，从而使得网络可以有效地执行具体任务。交叉点器件通常从可用的训练模式学习必要的连接权重。通过迭代地更新网络中的权重，性能随着时间被改进。代替遵循人类专家指定的一组规则，ANN从给定的代表性示例集合“学习”基本规则(如输入-输出关系)。因此，学习算法通常可以被定义为使用学习规则来更新和/或调整相关权重的过程。

[0050] 三种主要学习算法范例是有监督的、无监督的和混合的。在有监督学习或利用“教师”的学习中，针对每个输入模式向网络提供正确的答案(输出)。确定权重以允许网络产生尽可能接近已知正确答案的答案。强化学习是监督学习的一种变体，其中网络仅被提供有对网络输出的正确性的批评，而不是正确答案本身。相反，无监督学习或没有教师的学习不需要与训练数据集中的每个输入模式相关联的正确答案。它探索数据中的基础结构或数据中的模式之间的相关性，并且根据这些相关性将模式组织成类别。混合学习结合了有监督和无监督的学习。部分权重通常通过有监督学习而被确定，而其他通过无监督学习而被获得。人工神经网络:A Tutorial,Anil K.Jain,Jianchang Mao和K.M.Mohiuddin,IEEE,1996年3月中描述了ANN和学习规则的附加细节，其全部描述通过引用整体并入本文中。

[0051] 如前所述，为了限制功耗，ANN芯片架构的交叉点器件通常被设计为利用离线学习技术，其中一旦初始训练阶段已经被解决，目标函数的近似就不会改变。离线学习允许简化交叉型ANN架构的交叉点器件，从而使得它们汲取非常少的功率。

[0052] 尽管存在针对降低功耗的可能性，但执行离线训练可能是困难且资源密集的，因为在训练期间通常必须修改ANN模型中的大量可调参数(例如，权重)以匹配针对训练数据的输入-输出。图4描绘了典型的读取-处理-写入权重更新操作的简化图示，其中CPU/GPU核心(即，模拟的“神经元”)读取存储器(即，模拟的“突触”)并且执行权重更新处理操作，然后将更新的权重写回内存。因此，简化ANN架构的交叉点器件将省电的离线学习技术优先化通常意味着训练速度和训练效率不是最佳的。

[0053] 提供将功耗保持在可接受范围内以及加速训练ANN架构的速度和效率的简单交叉点器件将改进整体ANN性能并且允许更广泛的ANN应用。

[0054] 现在将提供与本发明的实施例相关的心脏收缩阵列、随机计算以及线性和非线性忆阻器器件的概述。心脏收缩阵列由并行处理元件(PE)组成，PE试图加速某些高度使用的算法的学习。心脏收缩阵列通常是用于具体操作(诸如“乘法和累加”)的硬连线的，以执行大规模并行积分、卷积、相关、矩阵乘法或数据排序任务。在C.Lehmann等人的名称为“A Generic Systolic Array Building Block for Neural Networks with On-Chip

Learning,"IEEE Transactions on Neural Networks,Vol,1993年5月4日第3期的出版物中,提出使用心脏收缩阵列作为针对在线学习神经网络的构建块,其中心脏收缩阵列中的每个PE具有本地存储装置以存储单个权重值并且能够执行针对矩阵乘法和权重更新所必须的计算。Lehmann的文章中描述的PE的超大规模集成(VLSI)实现需要每个PE大约1800个晶体管,这增加了功耗并降低了可扩展性。因此,希望提供每个PE需要尽可能少的晶体管的PE。

[0055] 随机计算是通过随机比特的流表示连续值的技术的集合,其中可以通过对流的简单逐位运算来计算复杂计算。具体地说,如果有两个随机和独立的比特流S_1,S_2所需的计算称为随机数(即伯努利过程),其中第一个流中"一"的概率是p,并且第二个流中"一"的概率是是q,两个流的逻辑AND可以如图6所示被取得。输出流中"一"的概率是pq。通过观察足够的输出位并且测量"一"的频率,可以将pq估计为任意精度。由于这些所谓的"乘法和累加"运算的设计简单性(其可以利用几个逻辑门/晶体管而被实现),随机计算通常在针对神经网络的硬件设计中被使用。V.K.Chippa等人的名称为"StoRM:A Stochastic Recognition and Mining Processor",2014年国际低功耗电子与设计研讨会论文集的出版物展示了随机计算对二维(2D)心脏收缩阵列的应用,二维(2D)心脏收缩阵列可以被用作针对神经网络训练算法的硬件加速器。

[0056] 然而,在Chippa等人的文章中,针对计算的必要权重从外部位置被提供给心脏收缩阵列,并且对权重的更新不由阵列执行。Chippa等人的文章仅解决在神经网络训练期间大量使用的矢量矩阵乘法或矩阵-矩阵乘法运算的加速度。然而,没有本地存储装置的心脏收缩阵列不能并行执行权重更新,因为权重被存储在外部存储器位置。加速重量更新(Chippa等人的文章没有描述)是为了加速整体学习算法所必须的。

[0057] 术语"忆阻器"被用于描述无源双端电子部件,其中器件的电阻值取决于先前已被施加给器件的电压的历史。忆阻器的操作由图6中所示的方程[1]和[2]控制,其中i是通过器件的电流,v是被施加给器件的电压,g是器件的电导值(其为电阻的倒数),s是控制电导值的器件的内部状态变量,并且f是显示内部状态变量s的时间演变的函数。在Chua,L.O.的名称为"Resistance Switching Memories are Memristors",Applied Physics A(2011),102(4):765-783的出版物中,提出了忆阻器功能,其用于电阻式存储器件的操作,电阻式存储器件诸如电阻随机存取存储器(RRAM)、相变存储器(PCM)和导电桥接随机存取存储器(CBRAM)。因为忆阻器器件记住它的历史(即,所谓的"非易失性特性"),所以Chua文章提出这样的器件作为针对非易失性存储器技术的可能备选。

[0058] D.Soudry等人的名称为"Memristor-Based Multilayer Neural Networks With Online Gradient Descent Training",IEEE Transactions on Neural Networks and Learning Systems(2015)的出版物提出了使用忆阻器以用于反向传播神经元网络培训硬件。然而,Soudry等人的文章假定了理想的忆阻器操作,其中电阻的变化相对于被施加到器件的电压是线性的。Soudry等人的设计假设图6的方程[2]中的函数f(s,v)是由关系f(s,v)＝v给出的简单函数。Soudry等人的文章提出了一种类似于如上所述的2D心脏收缩阵列的架构,其中每个交叉点利用理想的忆阻器和一对晶体管而被实现。在Soudry等人的文章中,忆阻器实际上被用于存储权重值,并且该对晶体管被用于计算针对权重更新所需要的局部乘法运算,其中权重更新的结果修改忆阻器的传导状态。Soudry等人的文章实际上描述了

一种由忆阻器和两个晶体管组成的四端子器件，它们被用于制作4端子器件的2D阵列以便实现神经网络硬件的反向传播训练。

[0059] 图7图示了根据一个或多个实施例的具有外围“神经元”并且具有输入和输出的“突触”的单个矩阵。单个矩阵包括一组突触75、一组输入85和一组输出95。所图示的实施例被称为“感知器”70。感知器70的目标是识别一个或多个实体，诸如作为来自一组输入85的猫，诸如图片中的像素。训练涉及让系统从一系列给定的输入识别猫，从而使得它可以从后续的未知输入识别猫。在这一示例中，一个或多个实体是一只猫；然而，可以使用任何类型或数量的实体。

[0060] 给定作为电导w_{ji}的突触权重，通过将电压I_i置于输入85上并且将输出95处的电流相加以得到$\sum_i w_{ji}I_i$以获得输出O_j，然后将这一结果通过具有特性op＝g(ip)的饱和放大器以得到：$O_j = g\left(\sum_i w_{ji}I_i\right)$。

[0061] 这里，g(x)；tanh(x)。必须通过训练获得突触权重w_{ij}。通过将成本函数$E(\mathbf{w}) = \frac{1}{2}\sum_i\left[\zeta_j - g\left(\sum_i w_{ji}I_i\right)\right]^2$最小化来完成训练，其中$\zeta_j$是关于$w_{ji}$的期望输出(例如，猫)。结果是Hebbian更新规则：$\Delta w_{ji} = \eta I_i \delta_j$，其中$\delta_j = \left(\zeta_j - O_j\right)g'\left(\sum_i w_{ji}I_i\right)$。

[0062] 这里η是“学习率”。当多个阵列被串联连接(深度神经网络)时，饱和放大器起作用。如果系统保持线性(g(x)＝x)，则额外的阵列将连接成一个数组。

[0063] 当前系统以软件实现这一方案。但是，培训非常缓慢(数据中心中48小时)。使用阵列的硬件实现(突触75和神经元)，阵列可以实现速度提高，因为训练过程可以在O(1)时间中被完成。如果同时施加输入电压，则所有竖直线78中的电流立即开始(以电容充电时间为模)，并且可以在输出神经元中同时处理结果。结果是由输出神经元计算的输出集δ_j。

[0064] 为了在每个突触75处实现Hebbian更新规则$\Delta w_{ji} = \eta I_i \delta_j$，计算乘积$I_i\delta_j$。这可以通过使用随机数的概念在一个实现中被完成。

[0065] 例如，假设存在被称为随机数的两个随机的独立比特流。可以假设比特流是时钟脉冲序列，其中脉冲幅度是1或0。令第一流中的一的概率是p，并且第二流中的概率是q。我们可以将两个流的乘积pq作为两个流的逻辑AND的概率$p \wedge q$。通过观察足够的输出位并且测量一的频率，可以将pq估计为任意精度。

[0066] 应用这一技术，Hebbian更新规则可以被写为$\Delta w_{ji} = \Delta w_{\min}\sum_{n=1}^{M} I_i^n \wedge \delta_j^n$，其中$I_i^n$和$\delta_j^n$是二进制脉冲，长度为M的0和1的随机比特流的成员具有1的概率分别为I_i和δ_j。

[0067] 可以看出，更新的数量级是发射M个脉冲的时间，与阵列大小无关。对于4000x4000的数组，加速可以是10^4阶。

[0068] 能够在O(1)时间中更新突触矩阵的具有伴随神经元的突触阵列被称为电阻处理单元(RPU)。RPU可以被串联放置，从而实现计算效果(深层神经网络或DNN)。在这种情况中，

输入和输出神经元是可见的，中间神经元是隐藏的。算法被修改但原理保持不变，信息从输入被传播到输出，在那里与期望的输出比较，并且然后错误被反向传播回输入，在此过程中更新突触权重。

[0069] 需要一种突触器件，其存储可在输入(水平)线和输出(竖直)线之间被测量的电导。电导需要通过接收输入线和输出线上的重合脉冲而可更新。存在针对突触器件的规范。

[0070] 4000x4000阵列示例包括以下规范：可写入的状态p的数量～1000，脉冲时间为1ns，电导为w，w-1≈24MOhm。MAX/MIN比率≈10，读取时要分离的状态数，并且如果状态p+q＝r，则如果p通过q个正脉冲被转换为r，则可以通过应用q个擦除脉冲恢复到状态p(即对称标准)。在某些情况中，每个突触需要不止一条水平/竖直线。存在本机器件，例如，忆阻器，其试图在没有附加电路和具有单输入和单输出线的情况下实现所有功能。

[0071] 现在转向本发明各方面的概述，一个或多个实施例提供了一种忆阻器件，其包括可通过引入或去除碱性杂质(例如Li，Na或K)而被可控地“掺杂”的半传导层。半传导层是氧化物，诸如氧化钛，其具有宽带隙并且是半绝缘的。向这一氧化物层引入诸如锂的碱改变氧化物层的电阻率，从而使其更具导电性，而去除碱导致氧化物层导电性降低。通过将正电压脉冲施加到充当忆阻器件中的电极的金属-碱合金或金属间化合物来执行对碱的引入。这一金属-碱合金电极被布置在忆阻器件中的半导体层上。电压脉冲使碱(例如，Li)渗透到氧化物层中。电导率的增加由于锂离子向氧化物的导带捐献电子。相反，施加负电压脉冲从氧化物层除去导致氧化物层导电性降低的碱金属离子。忆阻器件可以在人工神经网络中被利用。

[0072] 现在转到对本发明的更详细描述，下面通过参考附图8至图14中的附图详细描述用于形成基于透射金属氧化物(例如TiO_2)的碱性掺杂的忆阻器件的一个或多个实施例及其所得到的结构。

[0073] 图8图示了在根据本发明的一个或多个实施例的、形成忆阻器件的中间操作期间在导电材料层702上沉积氧化物层704之后的忆阻器件的侧视图。导电材料层702可以是自立式导电材料层(例如，金属箔)或被涂覆在衬底(例如，Si晶片，玻璃等)上的导电材料的薄膜的形式，其中氧化物层704被沉积在导电材料的有源区域上。在所图示的示例中，导电材料层702包括氟掺杂的氧化锡(FTO)涂覆的玻璃(SnO_2:F)。

[0074] 在一些实施例中，氧化物层704可以使用化学气相沉积(CVD)、等离子体增强化学气相沉积(PECVD)、原子层沉积(ALD)、物理气相沉积(PVD)、化学溶液沉积或其他类似的过程而被沉积。

[0075] 图9图示了根据本发明的一个或多个实施例的、被插入有碱金属802的忆阻器件的侧视图。在一些实施例中，可以通过使用正丁基锂将锂插入到氧化物层中。氧化物层704被暴露于正丁基锂802，然后利用烃溶剂(例如，己烷(未显示))被冲洗以去除多余的。正丁基锂802对氧化物层704的暴露时间可以基于器件的应用而变化，从而使得针对应用需要氧化物层704的具体薄层电阻。在作为过渡金属氧化物基质的TiO_2的实例中，随着曝光时间增加，氧化物层704的薄层电阻降低。对于忆阻器件的各种应用，正丁基锂802的这种可变时间段的暴露被用于调节氧化物层704的初始薄层电阻。在一个或多个实施例中，其他过渡金属氧化物材料可以被用于氧化物层704。

[0076] 图10图示了根据本发明的一个或多个实施例的、在氧化物层被插入碱金属(诸如

正丁基锂)之后的忆阻器件的侧视图。插入的氧化物层904具有的薄层电阻低于在插入碱金属之前的氧化物层704的薄层电阻。正丁基锂的暴露时间被设置以调节插入的氧化物层904的所需薄层电阻。

[0077] 在另一实施例中,可以通过在同一层中沉积包含电极材料和Li(或Na)的顶部触点(未示出)来将氧化物层704引入到Li(或Na)。在这一示例中,未掺杂的TiO_2层被沉积,并且然后顶部触点被溅射,溅射包括金属材料和一定百分比的Li或Na。效果将是利用引入电压脉冲来驱入/驱出(drive in/out)TiO_2。

[0078] 图11图示了根据本发明的一个或多个实施例的、在合金层的沉积之后的忆阻器件的侧视图。合金层1002可以是任何金属-碱合金或金属间化合物。在所图示的示例中,合金中的金属是锡(Sn)并且碱是锂(Li)。合金层1002包含碱金属的储库(reservoir)。被存储在合金层1002中的碱金属材料的最大量取决于合金层的厚度和碱在金属中的热力学溶解度。

[0079] 为了便于说明和讨论,仅示出了包括导电材料层702、插入氧化物层904和合金层1002的一个忆阻器件。应当理解,可以在导电材料层702的分离的部分上形成任何数目的这些器件。在一些实施例中,忆阻器件被布置在具有与忆阻器件类似组成的其他忆阻器件的阵列中。在一些实施例中,忆阻器件在针对机器学习应用的人工神经网络中被利用。

[0080] 图12图示了在合金层1002和导电材料层702上沉积金属触点1102之后的忆阻器件的侧视图。金属触点1102可以被连接到电压源。在一些实施例中,金属触点1102可以被沉积,例如,通过首先沉积层间电介质层并且通过光刻图案化和蚀刻工艺(诸如反应离子蚀刻或任何类似工艺)在层间电介质中形成空腔。然后,金属触点1102可以被沉积在合金层1002以及在层间电介质中被蚀刻的腔中的导电材料层702上。而在说明性示例中,金属触点1102被示出为相对于忆阻器件的其他层的尺寸;本领域技术人员可以领会到,可以在忆阻器件的层上利用和布置任何尺寸的金属触点1102。

[0081] 金属触点可包括任何合适的导电材料,包括例如多晶硅或非晶硅、锗、硅锗、金属(例如,钨、钛、钽、钌、锆、钴、铜、铝、铅、铂、锡、银、金)、导电金属化合物材料(例如,氮化钽、氮化钛、碳化钽、碳化钛、碳化钛铝、硅化钨、氮化钨、氧化钌、硅化钴、硅化镍)、碳纳米管、导电碳、石墨烯或这些材料的任何适当组合。

[0082] 图13图示了在向金属触点1102施加一个或多个正电压脉冲之后的忆阻器件的侧视图。一个或多个正电压脉冲使合金层1002中的碱金属插入到氧化物中从而形成具有离子化的碱性材料的氧化物层1204。随着离子化的碱金属插入氧化物层1204中,电阻跨忆阻器件降低。器件的电阻的下限由施加电压脉冲后TiO_2层中Li的含量确定,而器件的电阻的上限由当通过施加负电压脉冲去除所有Li时跨膜电阻确定。电压脉冲的幅度、长度和持续时间被选择以优化每个脉冲的电阻变化和可以被存储在器件中的独立电阻状态的数目。

[0083] 图14图示了在向金属触点1002施加一个或多个负正电压脉冲之后的忆阻器件的侧视图。一个或多个负电压脉冲使得碱金属离子从氧化物层去除,其中碱性材料1304被转移到接口(1304和2003之间)或合金层(2003)(如果存在的话)。

[0084] 图15图示了忆阻器件的备选示例性实施例的侧视图,其中扩散阻挡层1402被沉积在合金层1002和插入的氧化物层904之间。扩散阻挡层1402限制碱性材料离子在空闲期间(即,当没有施加电压脉冲时)从合金材料1002泄漏到氧化物层904中。扩散阻挡层1402可以包括Al_2O_3、氮化硅或允许离子穿过并被存储在层1402和1002之间的任何其他材料。针对这

一扩散阻挡层的材料选择受到要求它不阻止Li到达与层1002的接口(即,它不应导致Li在904和1402之间的界面处的沉积)的影响。这样的扩散阻挡层的存在将能够使电阻状态存储更长的时间段(非易失性)。

[0085] 图16图示了忆阻器件的备选实施例。忆阻器件包括被布置在导电材料层1602上的氧化物层1604。如上所述,导电材料层1602可以是自立式导电材料层(例如,金属箔)或被涂覆到衬底(例如,Si晶片,玻璃等)中的导电材料的薄膜的形式,其中氧化物层1604被沉积在导电材料层1602上的有源区域上。在所图示的示例中,导电材料层1602包括氟掺杂的氧化锡(FTO)涂覆的玻璃(SnO_2:F)。

[0086] 在利用氢氧化物溶液(例如,己烷(未显示))被冲洗之前,氧化物层1604被暴露于正丁基锂(未示出)以去除多余的。正丁基锂暴露于氧化物层1604的时间可以基于器件的应用而变化,从而使得针对应用需要氧化物层1604的具体薄层电阻。

[0087] 金属触点1608被沉积在氧化物层上和导电材料层1608上。被布置在氧化物层1604上的金属触点1608可以是任何合适的金属触点材料,例如,锡(Sn)。金属触点被配置为接收对金属触点1608的一个或多个负正电压脉冲。一个或多个负电压脉冲使得碱性离子从氧化物层被去除,其中碱性材料1604被转移到接口(在1604和1608之间)或金属触点1608。

[0088] 图17图示了忆阻器件的备选实施例的侧视图。忆阻器件包括被布置在导电材料层1702上的氧化物层1704。如上所述,导电材料层1702可以是独立的导电材料层(例如,金属箔)或被涂覆到衬底(例如,Si晶片,玻璃等)中的导电材料的薄膜的形式,其中氧化物层1704被沉积在导电材料层1702上的有源区域上。在所图示的示例中,导电材料层1702包括氟掺杂的氧化锡(FTO)涂覆的玻璃(SnO_2:F)。

[0089] 忆阻器件包括被沉积在氧化物层1704上的合金层1706。合金层1706可以是任何金属-碱合金或金属间化合物。在所图示的示例中,合金中的金属是锡(Sn),并且碱是锂(Li)。合金层1706包含碱金属的储库。被存储在合金层1706中的碱性材料的量取决于合金层的厚度和碱在金属中的热力学溶解度。在一个或多个实施例中,金属-碱金属合金或金属间化合物包括作为金属的钼(Mo)和作为碱的锂(Li)(Mo:Li)。

[0090] 忆阻器件包括金属触点1708,其可以被沉积在合金层1706和导电材料层1702上。金属触点1708可以被连接到电压源。在一些实施例中,可以例如通过首先沉积层间电介质层并且通过光刻图案化和蚀刻工艺(诸如反应离子蚀刻或任何类似工艺)在层间电介质中形成空腔来沉积金属触点1708。金属触点1708然后可以被沉积在合金层1706和在层间电介质中蚀刻的腔中的导电材料层1702上。而在说明性示例中,金属触点1708被示出为相对于忆阻器件的其他层的尺寸;本领域技术人员可以领会到,任何尺寸的金属触点1708可以被利用并且被布置在忆阻器件的层上。

[0091] 金属触点1708被配置为接收对金属触点1708的一个或多个正电压脉冲。一个或多个正电压脉冲使得合金层1708中的碱金属插入到氧化物层中从而形成具有离子化的碱性材料(未示出)的氧化物层。随着离子化的碱金属插入到氧化物层1704中,电阻跨忆阻器件降低。器件的电阻的下限由施加电压脉冲后TiO_2层中Li的量确定,而器件的电阻的上限由当通过施加负电压脉冲去除所有Li时跨膜的电阻确定。电压脉冲的幅度、长度和持续时间被选择以优化每个脉冲的电阻变化以及可以被存储在器件中的独立电阻状态的数目。

[0092] 已经出于说明的目的给出了对本发明的各种实施例的描述,但是这些描述并不旨

在是穷举的或被限制于所描述的实施例。在不脱离本发明的范围和精神的情况下，许多修改和变化对于本领域普通技术人员来说是显而易见的。选择这里使用的术语是为了最好地说明实施例的原理、实际应用或对市场中发现的技术的技术改进，或者使本领域普通技术人员能够理解本文所述的实施例。

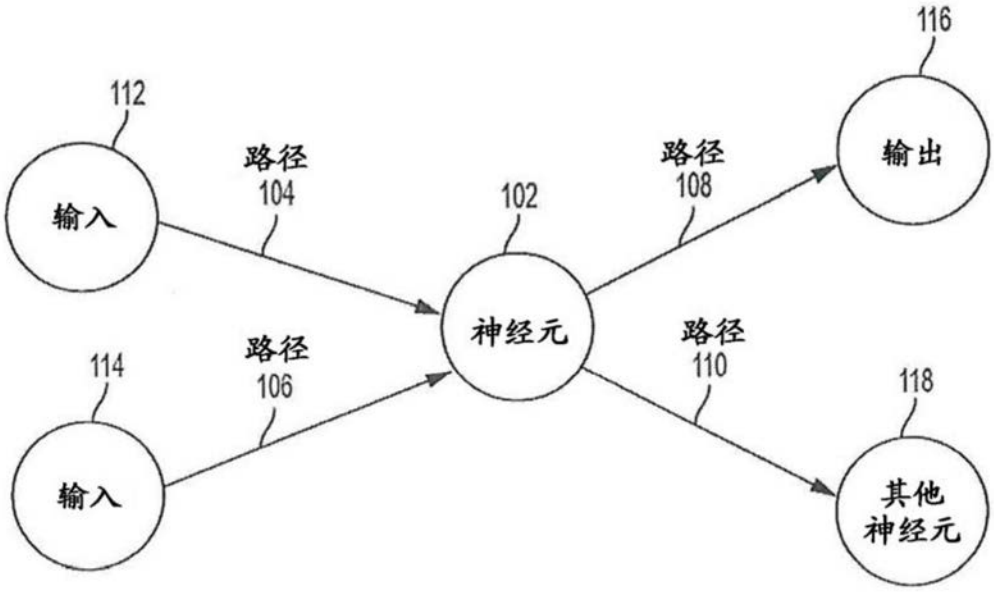
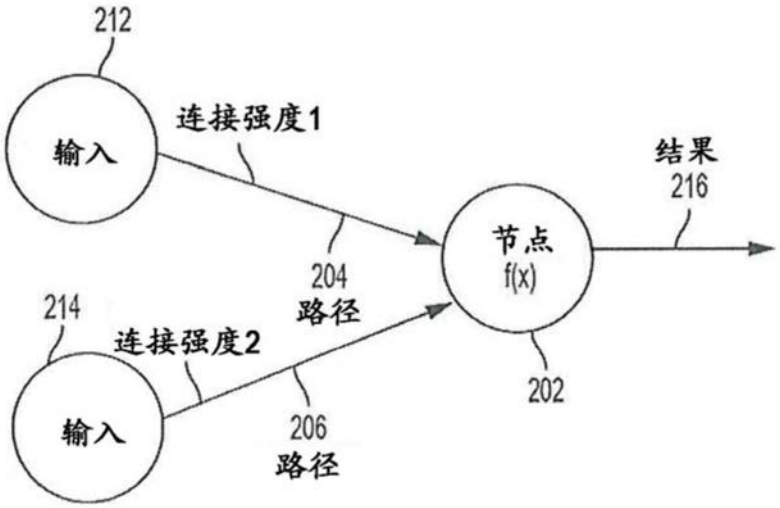


图1



$$f(x) = f(\text{输入}1 * \text{连接强度}1 + \text{输入}2 * \text{连接强度}2)$$

图2

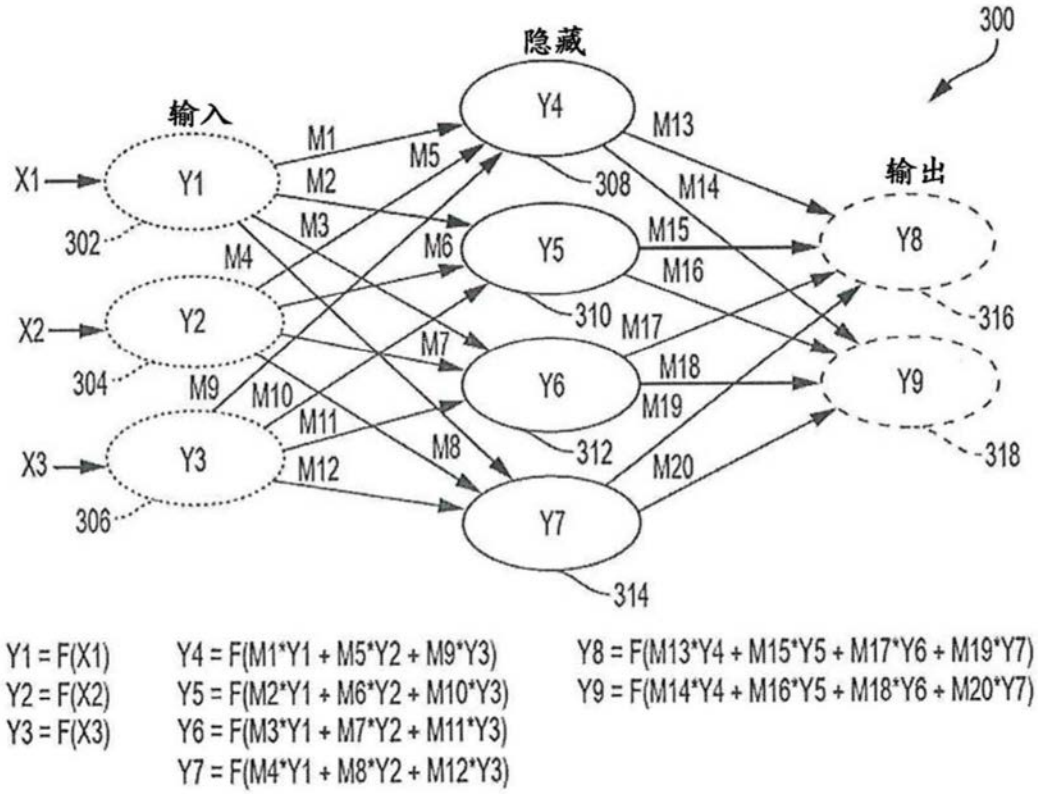

300
隐藏
输入
输出
X1
X2
X3
Y1
Y2
Y3
Y4
Y5
Y6
Y7
Y8
Y9
M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
M11
M12
M13
M14
M15
M16
M17
M18
M19
M20
302
304
306
308
310
312
314
316
318
Y1 = F(X1)
Y2 = F(X2)
Y3 = F(X3)
Y4 = F(M1*Y1 + M5*Y2 + M9*Y3)
Y5 = F(M2*Y1 + M6*Y2 + M10*Y3)
Y6 = F(M3*Y1 + M7*Y2 + M11*Y3)
Y7 = F(M4*Y1 + M8*Y2 + M12*Y3)
Y8 = F(M13*Y4 + M15*Y5 + M17*Y6 + M19*Y7)
Y9 = F(M14*Y4 + M16*Y5 + M18*Y6 + M20*Y7)

图3


存储器
读取-处理-写入
CPU/GPU
核心

图4

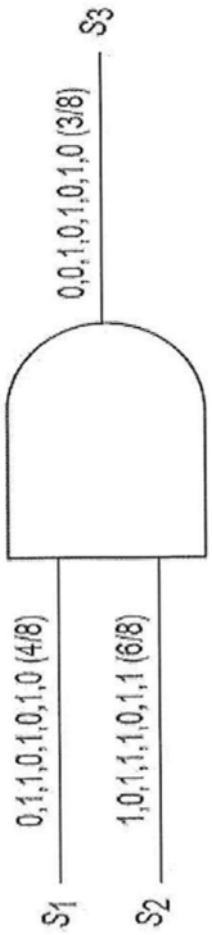


图5

方程[1]

$$i = g(s,v)v$$

方程[2]

$$\frac{\partial s(t)}{\partial t} = f(s,v)$$

图6

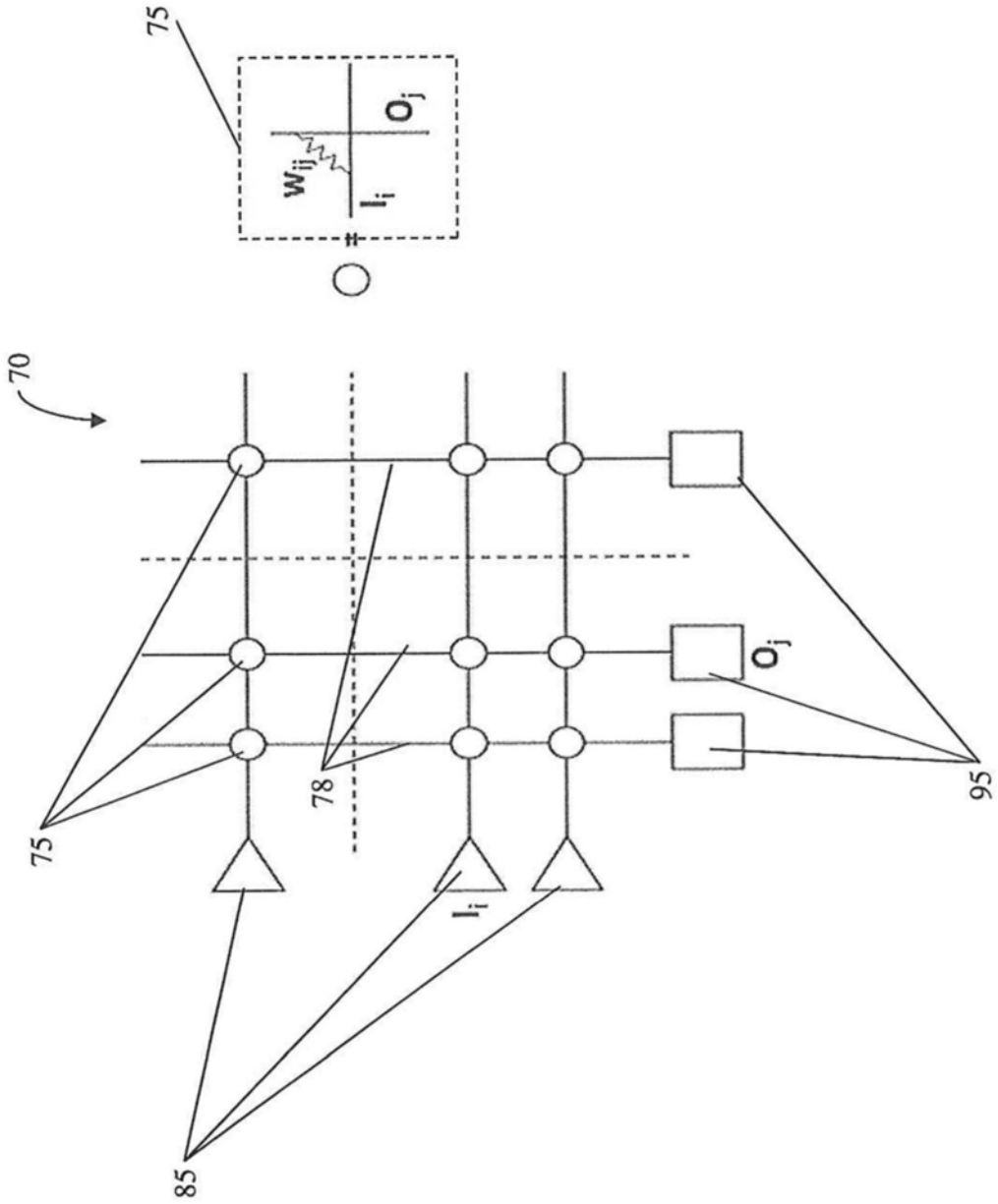


图7

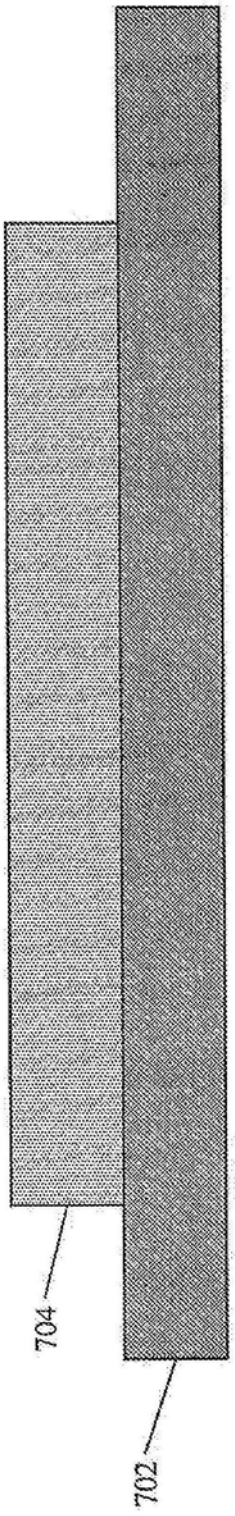


图8

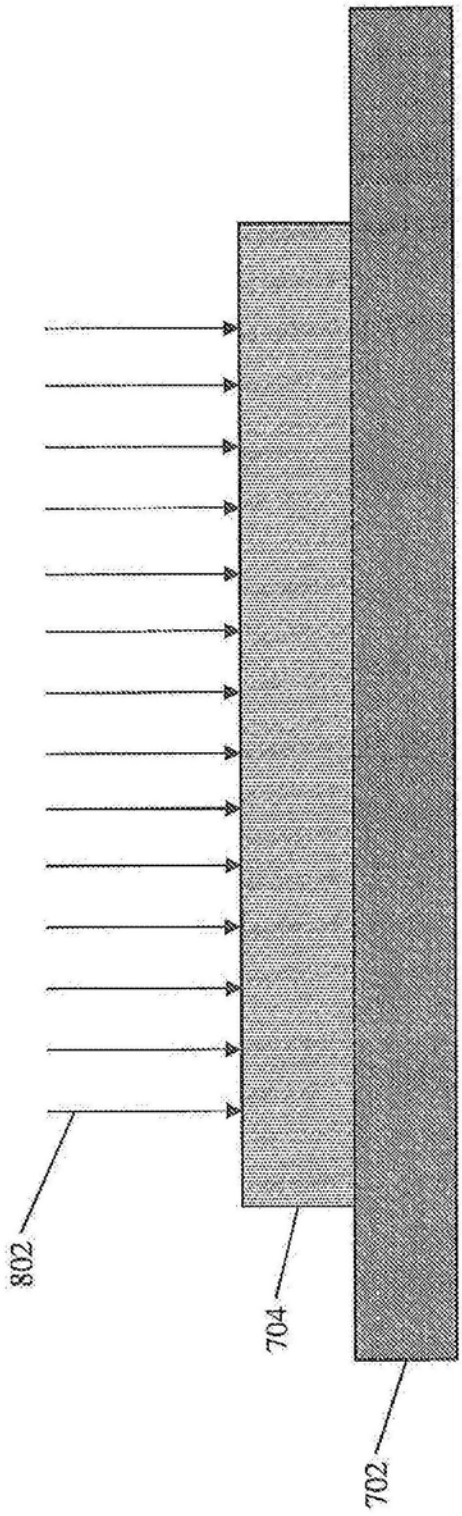


图9

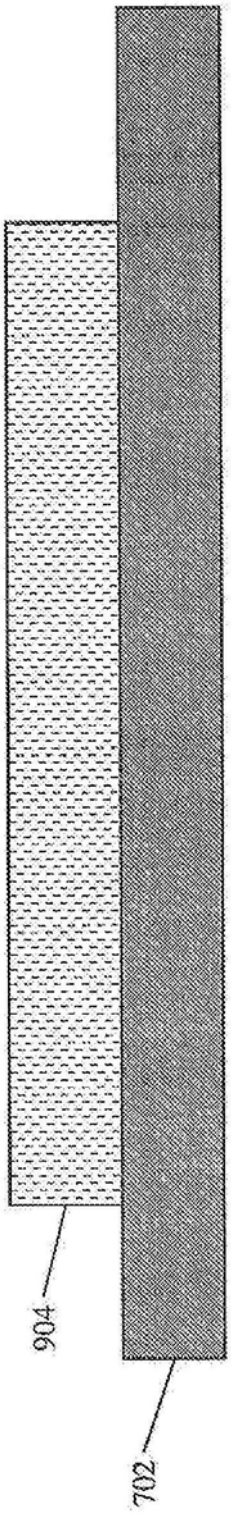


图10

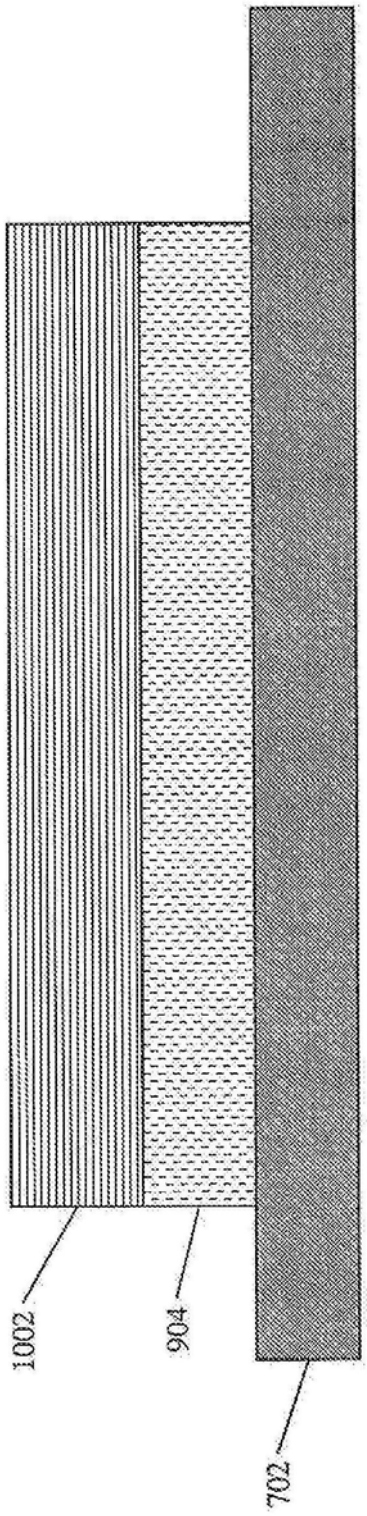


图11

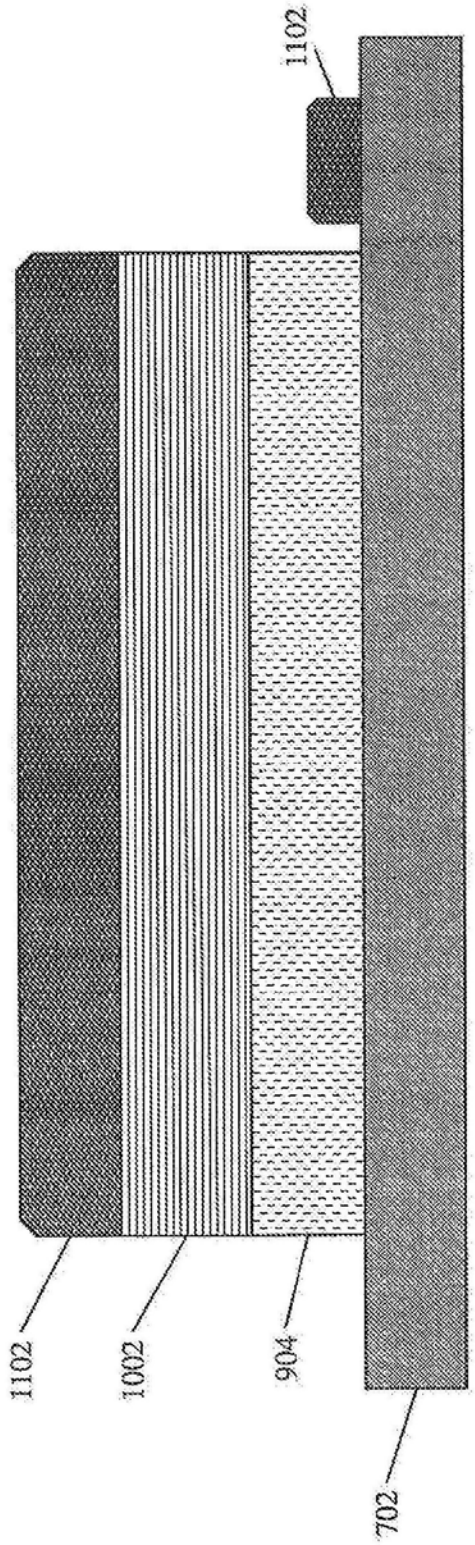


图12

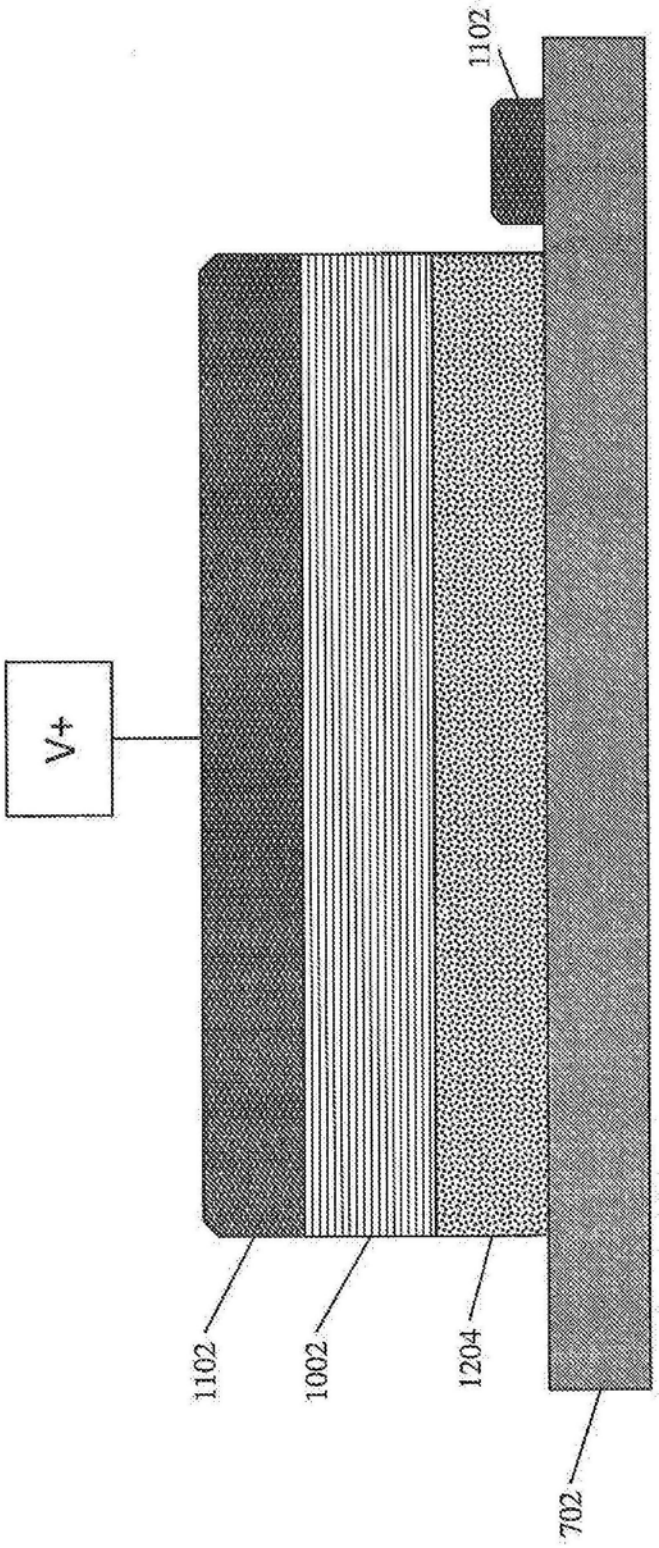


图13

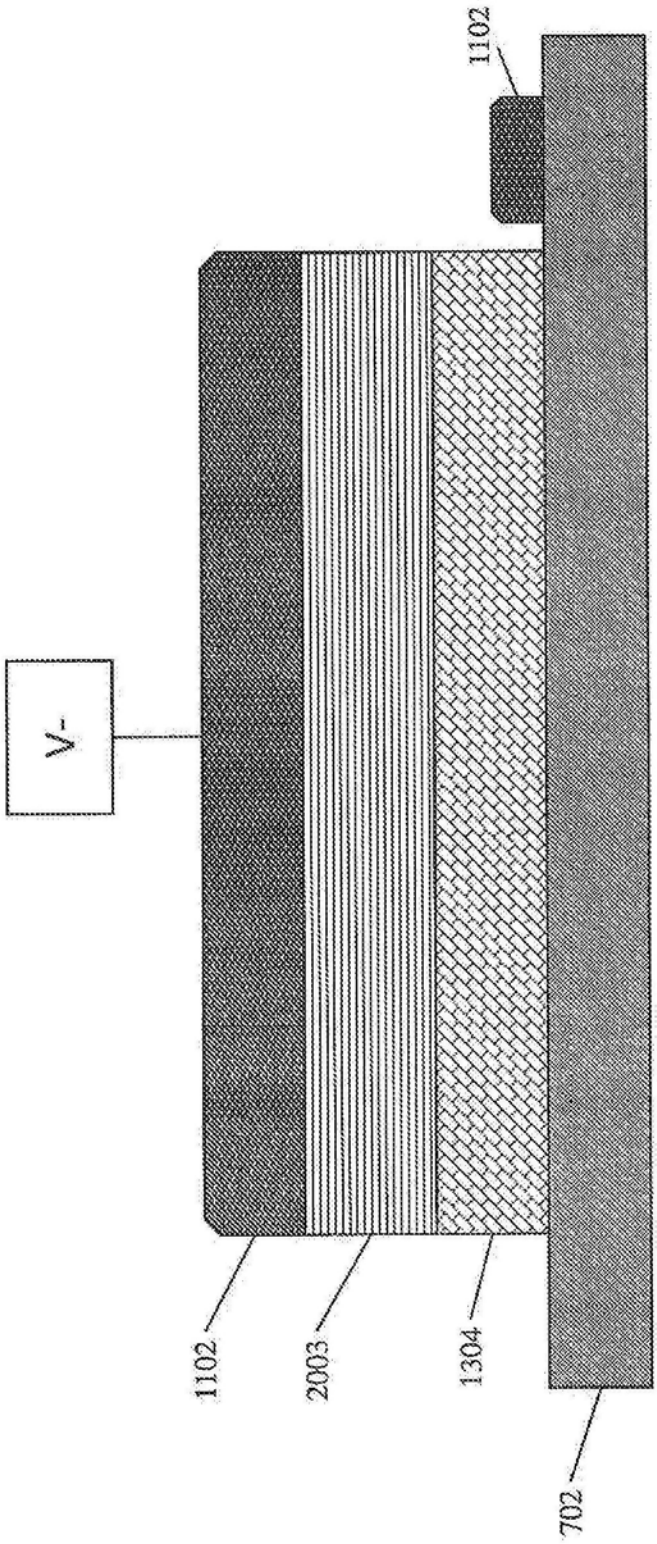


图14

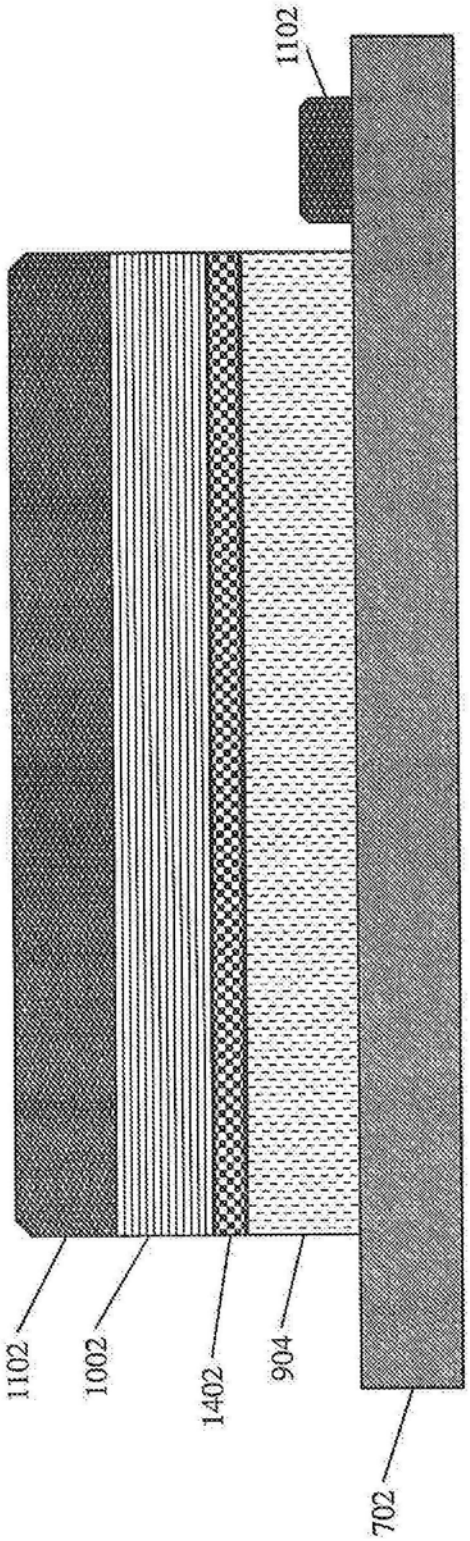


图15

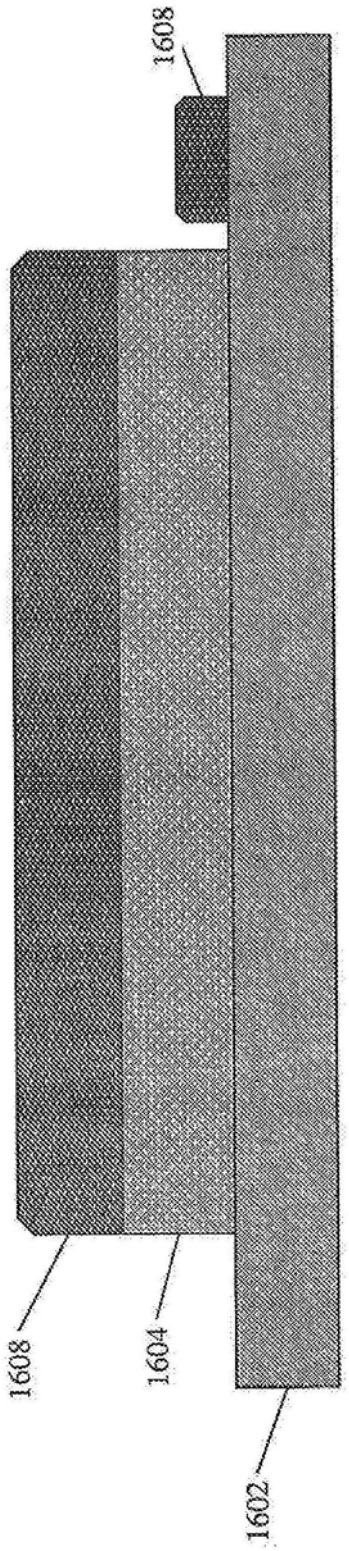


图16

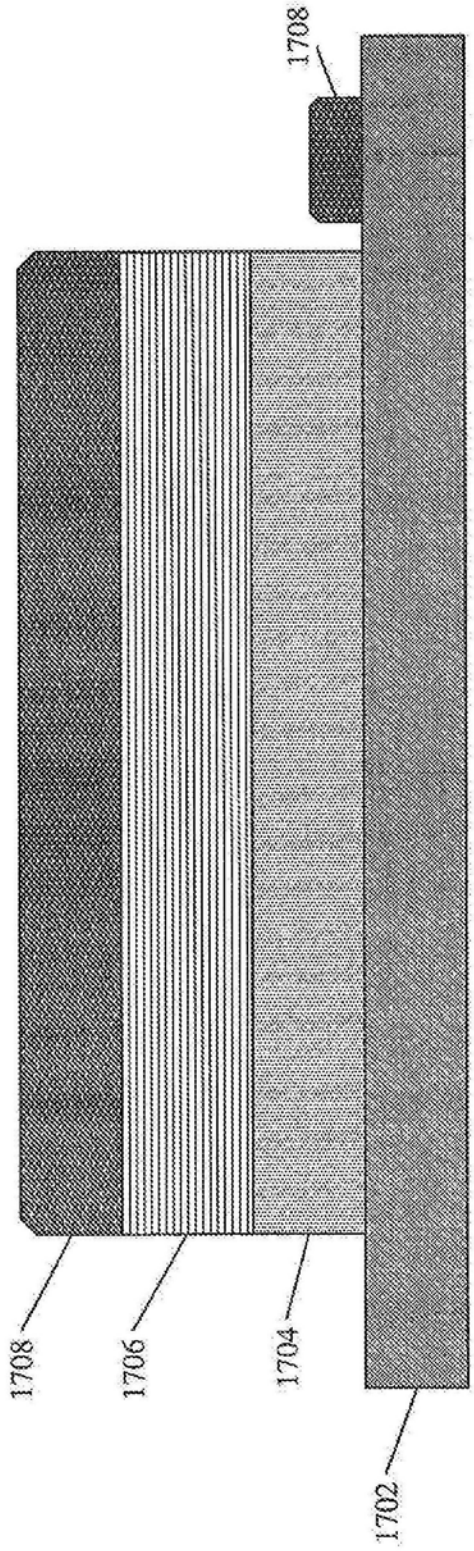


图17