(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114898792 A
(43)申请公布日 2022.08.12

(21)申请号 202210390722.7

(22)申请日 2022.04.14

(71)申请人 浙江大学
地址 310058 浙江省杭州市西湖区余杭塘路866号

(72)发明人 尹勋钊 刘哲恺 陈豪邦 卓成

(74)专利代理机构 杭州浙科专利事务所(普通合伙) 33213
专利代理师 吴昌楹

(51)Int.Cl.
G11C 16/04(2006.01)
G11C 16/06(2006.01)
G06F 7/575(2006.01)

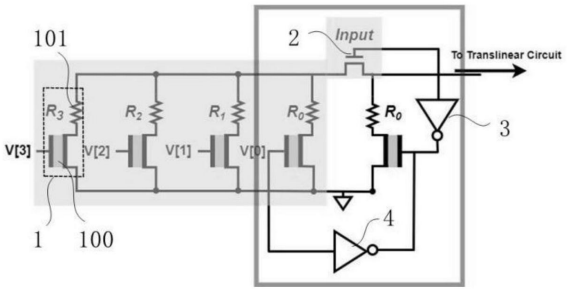
权利要求书1页 说明书5页 附图3页

(54)发明名称

多比特存内内积暨异或单元、异或向量及操作方法

(57)摘要

多比特存内内积暨异或单元、异或向量及操作方法，包括N个并联的1FeFET1R结构、输入晶体管、第一反相器和第二反相器，N为大于1的自然数，所述1FeFET1R结构包括电连接的FeFET和电阻，每个1FeFET1R结构的电阻均与输入晶体管电连接，所述输入晶体管的栅极通过第一反相器与其中一个1FeFET1R结构中FeFET的栅极电连接，该1FeFET1R结构中FeFET的栅极通过第二反相器与另一个1FeFET1R结构中FeFET的栅极电连接。本发明首次提出基于非易失存储器件且同时支持多比特存内内积暨异或的单元及其向量，在搜索能耗、搜索延时以及面积三大指标上均表现更优。



CN 114898792 A

1.一种多比特存内内积暨异或单元,其特征在于,包括N个并联的1FeFET1R结构、输入晶体管、第一反相器和第二反相器,N为大于1的自然数,所述1FeFET1R结构包括电连接的FeFET和电阻,每个1FeFET1R结构的电阻均与输入晶体管电连接,所述输入晶体管的栅极通过第一反相器与其中一个1FeFET1R结构中FeFET的栅极电连接,该1FeFET1R结构中FeFET的栅极通过第二反相器与另一个1FeFET1R结构中FeFET的栅极电连接。

2.根据权利要求1所述的一种多比特存内内积暨异或单元,其特征在于,每个1FeFET1R结构中电阻的阻值不同,形成一连串输出电流为一系列二进制2^{N-1},2^{N-2},…,2^{1},2^{0}存储单元。

3.根据权利要求1所述的一种多比特存内内积暨异或单元,其特征在于,所述1FeFET1R结构中电阻与FeFET的漏极或源极电连接。

4.根据权利要求1所述的一种多比特存内内积暨异或单元,其特征在于,所述输入晶体管工作于线性区,其将向量元素之权重映射为电压并输入于对应FeFET的栅极。

5.根据权利要求1所述的一种多比特存内内积暨异或单元,其特征在于,所述第一反相器用于输入向量元素的互补值。

6.根据权利要求1所述的一种多比特存内内积暨异或单元,其特征在于,所述第二反相器用于两对应FeFET存入互补值。

7.一种多比特存内内积暨异或向量,其特征在于,包括M个如权利要求1-6中任一所述多比特存内内积暨异或单元,该M个多比特存内内积暨异或单元并联。

8.一种如权利要求7所述多比特存内内积暨异或向量的操作方法,其特征在于,包括:

S1存储向量的每个向量元素先存入多比特存内内积暨异或单元,具体存入方法为:存入向量的每个向量元素为二进制,根据欲输入的向量元素二进制值,如果为‘1’,在对应的FeFET栅极输入高电压,使FeFET存入‘1’;如果为‘0’,则在对应的FeFET栅极输入低电压,使FeFET存入‘0’,同时,于另一异或1FeFET1R结构通过反相器存入$\bar{w}_0$;

S2存入向量的向量元素存入多比特存内内积暨异或单元后,当查询向量来临时,同时进行以下操作:

S2.1查询向量的向量元素以电压的形式施加于多比特存内内积暨异或单元中的输入晶体管的栅极;同时,查询向量的向量元素通过第一反相器对应1FeFET1R结构;

S2.2对于实现多比特内积功能,每个FeFET的栅极同时输入高电压,利用FeFET本身即可实现“与”的特点,当存储值为‘0’时,输出为‘0’;当存储值为‘1’时,输出为‘1’;

S2.3对于实现多比特功能,两个反相器处于关断状态;对于实现异或功能,两个反相器接上电源,处于工作状态,且前N-1个FeFET的栅极同时输入低电压,即存入‘0’。

多比特存内内积暨异或单元、异或向量及操作方法

技术领域

[0001] 本发明涉及存储、计算、电路领域，具体涉及一种多比特存内内积暨异或单元、异或向量及操作方法。

背景技术

[0002] 在人工智能大量数据密集计算的背景下，各种二值神经网络(Binary Neural Networks，BNN)及超高维度向量计算(Hyperdimensional Computing，HDC)已经被证明可以高效地应用于不同实际场景如：物体追踪，声音识别，图像聚类等等。由于传统冯-诺伊曼计算机架构计算单元与存储单元的分离会导致高延时和能耗，以存算一体架构替代传统冯-诺伊曼计算机架构成为研究热点；由各种新型非易失器件所组成的存算一体单元能实现不同的逻辑运算，如单一个铁电晶体管即可实现二值向量之间“与”的逻辑运算。

[0003] 然而，真实应用下，二值向量并不能满足数据密集的运算场景；多比特内积的运算单元可以更广泛应用于人工智能场景如卷积神经网络。基于传统SRAM的多比特存内多比特内积单元近年来被广泛提出，但其在延时、能耗、面积、可扩展性等仍存在诸多缺陷，并且基于新型非易失存储器件的多比特存内内积暨异或架构仍未被提出；同时，在实现多比特内积之于，实际场景如二值卷积神经网络仍然会需要实现异或功能，又如汉明码距离本身即为按位异或运算，因此，本发明提出同时适用于多比特向量内积以及异或功能的存算单元。

发明内容

[0004] 本发明的目的在于提出一种多比特存内内积暨异或单元及其异或向量的技术方案，首次提出基于FeFET的实现方式，且能耗、搜索延时、面积等指标与现在仅有的工作相比有所提升。

[0005] 为实现上述目的，本发明提供了如下方案：

[0006] 一种多比特存内内积暨异或单元，包括N个并联的1FeFET1R结构、输入晶体管、第一反相器和第二反相器，N为大于1的自然数，所述1FeFET1R结构包括电连接的FeFET和电阻，每个1FeFET1R结构的电阻均与输入晶体管电连接，所述输入晶体管的栅极通过第一反相器与其中一个1FeFET1R结构中FeFET的栅极电连接，该1FeFET1R结构中FeFET的栅极通过第二反相器与另一个1FeFET1R结构中FeFET的栅极电连接。

[0007] 进一步地，每个1FeFET1R结构中电阻的阻值不同，形成一连串输出电流为一系列二进制2^{N-1}，2^{N-2}，…，2^{1}，2^{0}存储单元。

[0008] 进一步地，所述1FeFET1R结构中电阻与FeFET的漏极或源极电连接。

[0009] 进一步地，所述输入晶体管工作于线性区，其将向量元素之权重映射为电压并输入于对应FeFET的栅极。

[0010] 进一步地，所述第一反相器用于输入向量元素的互补值。

[0011] 进一步地，所述第二反相器用于两对应FeFET存入互补值。

[0012] 本发明还提供一种多比特存内内积暨异或向量，包括M个如上所述多比特存内内

积暨异或单元，该M个多比特存内内积暨异或单元并联。

[0013] 本发明还提供一种如上所述多比特存内内积暨异或向量的操作方法，包括：

[0014] S1存储向量的每个向量元素先存入多比特存内内积暨异或单元，具体存入方法为：存入向量的每个向量元素为二进制，根据欲输入的向量元素二进制值，如果为‘1’，在对应的FeFET栅极输入高电压，使FeFET存入‘1’；如果为‘0’，则在对应的FeFET栅极输入低电压，使FeFET存入‘0’，同时，于另一异或1FeFET1R结构通过反相器存入$\bar{w}_0$；

[0015] S2存入向量的向量元素存入多比特存内内积暨异或单元后，当查询向量来临时，同时进行以下操作：

[0016] S2.1查询向量的向量元素以电压的形式施加于多比特存内内积暨异或单元中的输入晶体管的栅极；同时，查询向量的向量元素通过第一反相器对应1FeFET1R结构；

[0017] S2.2对于实现多比特内积功能，每个FeFET的栅极同时输入高电压，利用FeFET本身即可实现“与”的特点，当存储值为‘0’时，输出为‘0’；当存储值为‘1’时，输出为‘1’；

[0018] S2.3对于实现多比特功能，两个反相器处于关断状态；对于实现异或功能，两个反相器接上电源，处于工作状态，且前N-1个FeFET的栅极同时输入低电压，即存入‘0’。

[0019] 本发明的有益效果如下：

[0020] 本发明首次提出基于非易失存储器件且同时支持多比特存内内积暨异或的单元及其向量，在搜索能耗、搜索延时以及面积三大指标上均表现更优。

附图说明

[0021] 图1是N＝4比特的多比特存内内积暨异或单元应用于余弦搜索架构示意图；

[0022] 图2是本发明内容电路图，单个N＝4比特的多比特存内内积暨异或单元电路图；

[0023] 图3(a)是N＝4比特下，单个多比特存内内积暨异或单元存储值由0000至1111的结果示意图；

[0024] 图3(b)是N＝4比特下，单个多比特存内内积暨异或单元存储值由0000至1111经过100次蒙特卡洛的结果示意图；

[0025] 图4(a)和(b)分别为N＝4/N＝6扩展示意图，其中分析了多比特存内内积暨异或单元的可扩展性，图4(b)展示了即使到N＝6，最坏情况只会有一比特运算无法区分；

[0026] 图5是N＝4比特下，单个多比特存内内积暨异或单元内电阻值降低的结果示意图；

[0027] 图6是基于图1下多比特存内内积暨异或单元应用示意图。

具体实施方式

[0028] 下面结合附图和具体实施例对本发明作进一步详细说明。

[0029] 请参阅图1-6，一种多比特存内内积暨异或单元，包括N个并联的1FeFET1R结构1、输入晶体管2、第一反相器3和第二反相器4，N为大于1的自然数，所述1FeFET1R结构1包括FeFET100和电阻101，电阻101与FeFET100的漏极或源极电连接，每个1FeFET1R结构1的电阻101均与输入晶体管2电连接，所述输入晶体管2的栅极通过第一反相器3与其中一个1FeFET1R结构1中FeFET100的栅极电连接，该1FeFET1R结构1中FeFET100的栅极通过第二反相器4与另一个1FeFET1R结构1中FeFET100的栅极电连接。

[0030] 其中，对于工作于比特存内内积模式，要形成N+1比特的内积单元，只需要对N比特

结构新增一1FeFET1R结构1，1FeFET1R结构1的电阻101需有2^N倍或2^{-1}倍的饱和漏源电流。因此，本发明中每个1FeFET1R结构1中电阻101的阻值不同，形成一连串输出电流为一系列二进制2^{N-1}，2^{N-2}，…，2^1，2^0存储单元。

[0031] 其中，所述输入晶体管2工作于线性区，其将向量元素之权重映射为电压并输入于对应FeFET100的栅极。

[0032] 其中，所述第一反相器3用于输入互补值。

[0033] 其中，所述第二反相器4用于两对应FeFET100存入互补值。

[0034] 请参阅图2，本发明还提供一种多比特存内内积暨异或向量，包括M个如上所述多比特存内内积暨异或单元C，该M个多比特存内内积暨异或单元C并联，形成一拥有M个向量元素的的向量。

[0035] 一种如上所述多比特存内内积暨异或向量的操作方法，包括：

[0036] S1存储向量的每个向量元素先存入多比特存内内积暨异或单元，具体存入方法为：存入向量的每个向量元素为二进制，以N＝4比特的存内内积单元为例$W=w_3w_2w_1w_0$，高位为w_3，表示2^3；低位为w_0，表示2^0。根据欲输入的向量元素二进制值，如果为‘1’，在对应的FeFET栅极输入高电压，使FeFET存入‘1’；如果为‘0’，则在对应的FeFET栅极输入低电压，使FeFET存入‘0’。同时，于另一异或1FeFET1R结构通过反相器存入$\bar{w}_0$。

[0037] S2存入向量的向量元素存入多比特存内内积暨异或单元后，当查询向量来临时，同时进行以下操作：

[0038] S2.1查询向量的向量元素以电压的形式施加于多比特存内内积暨异或单元中的输入晶体管的栅极；同时，查询向量的向量元素通过第一反相器对应1FeFET1R结构。

[0039] S2.2对于实现多比特内积功能，每个FeFET的栅极同时输入高电压，利用FeFET本身即可实现“与”的特点，当存储值为‘0’时，输出为‘0’；当存储值为‘1’时，输出为‘1’。

[0040] S2.3对于实现多比特功能，两个反相器处于关断状态；对于实现异或功能，两个反相器接上电源，处于工作状态，且前N-1个(即图2的V[3]到V[1])FeFET的栅极同时输入低电压，即存入‘0’，使单元只有最右边两个1FeFET1R工作。

[0041] 单元应用及架构仿真操作流程说明

[0042] 多比特存内内积暨异或单元组成的向量计算出余弦计算电路输入；如图1所示，以N＝4比特为例，存储阵列中的每个存储单元的晶体管相连形成一含有M个向量元素的向量。此内积结果通过电流镜拷贝，作为余弦计算电路的输入。而图1右边的存储阵列用于计算出每个余弦值的L_2范数，即余弦表达式的分母；余弦计算电路的输出再经过Winner-Take-All电路，找出与查询向量间余弦距离最大值的存储向量。其中，余弦计算电路的表达式为：

$$I_z=\frac{I_x^2}{I_y}$$

[0043] 多比特存内内积暨异或单元具体运行过程如下

[0044] 1、在搜索开始前，对每个多比特存内内积暨异或单元输入存储向量；以N＝4比特为例，通过每个多比特存内内积暨异或单元的V[3]～V[0]，分别写入w_3、w_2、w_1、w_0；同时另一1FeFET1R通过反相器存入$\bar{w}_0$，‘0’用-4V电压脉冲写入，‘1’用+4V电压脉冲写入。将向量元素写入后，即可以开始搜索过程。

[0045] 2.1、搜索时，当单元工作在多比特存内内积模式下，每个实现多比特存内内积的1FeFET1R，即V[3]～V[0](图2)，用+4V电压脉冲写入，即写入'1'；同时，输入由输入晶体管的栅极输入(图2)。根据输入向量元素值的大小，在0～1.2V之间选取电压。

[0046] 2.2、搜索时，当单元工作在异或模式下，前N-1個实现多比特存内内积的1FeFET1R，即V[3]～V[1](图2)，用-4V电压脉冲写入，即写入'0'。

[0047] 本发明的功能和效果通过以下仿真实验进一步说明展示：

[0048] 1、仿真条件

[0049] 实验使用基于物理电路的兼容SPECTRE和SPICE模型对由1FeFET1R存储单元组成的存储阵列进行仿真，其中FeFET是基于Preisach模型。该模型实现了高效的设计与分析，已广泛应用于FeFET电路设计中。利用PTM45-HP作为其余晶体管的仿真模型。

[0050] 仿真架构以图1所示。图1实现人工智能场景的一项应用：基于余弦搜索的最近邻搜索。其原理为，寻找出与输入向量在余弦距离上最相近的存储向量。图1的存储单元(以C表示)即为本发明提出的多比特存内内积暨异或单元；其中图2是以N＝4比特为代表的多比特存内内积暨异或单元。

[0051] 2、仿真结果

[0052] (1)根据图2之多比特存内内积暨异或单元原理图，当电流在纳安培级别时，于SPECTRE仿真表明R_0:R_1:R_2:R_3为8:4:2:1。

[0053] (2)图3(a)的横坐标为输入于多比特存内内积暨异或单元内晶体管栅极的电压，为连续值。曲线由下至上为存储值由0000至1111；图3(b)为考虑了FeFET工艺误差(提取自非专利文献1T.Solimanetal.,"Ultra-LowPowerFlexiblePrecisionFeFETBasedAnalogIn-MemoryComputing",IEEEIEDM,2020.)，考虑了大电阻误差(提取自非专利文献2D.Saitoetal.,"AnalogIn-memoryComputinginFeFET-based 1T1RArrayforEdgeAIApplications",IEEE SymposiumonVLSICircuits,2021)和晶体管误差，即领域默认10％大小误差、10％阈值电压误差，所得到的结果。图3(a)的横坐标为输入于多比特存内内积暨异或单元内晶体管栅极的电压，曲线由下至上为存储值0001($1_{(10)}$)、0011($3_{(10)}$)、0101($5_{(10)}$)、0111($7_{(10)}$)、1001($9_{(10)}$)、1011($11_{(10)}$)、1101($13_{(10)}$)、1111($15_{(10)}$)。结果表明，本发明内积结果准确性高，内积结果相差大于2且内积结果范围足够大(输入大于0101($5_{(10)}$)，电压约为0.5V)，运算则不产生重叠。

[0054] (3)能耗和延時：

[0055] 将我们的结果与非专利文献3(M.Ali et al.,"IMAC:In-Memory Multi-Bit Multiplication and ACcumulation in 6T SRAM Array",TCAS-I,2020.)中提出的基于SRAM多比特存内内积暨异或单元进行对比，从图6对向量个数及向量维度分别扩展的结果，本发明得到了超过10^4倍每多比特存内内积暨异或单元能耗的下降，和4.67倍输出延时的下降。

[0056] (4)消耗面积：

[0057] 本发明的面积消耗相比于上述非专利文献3有显著减少，主要是因为利用了新型非易失存储器件FeFET且在设计上相比于传统SRAM更为简单。对于单个多比特存内内积暨异或单元，本发明相比于上述非专利文献3于面积上减少了488倍(SRAM 64.9μm^2/cell，本发明0.133μm^2/cell)。

[0058] (5)可扩展性：

[0059] 图4将N＝4比特的多比特存内内积暨异或单元扩展至N＝6比特，展示了最坏情况下，只会有1的运算不精确；具体来说，对于N＝6的单元，仿真结果表明000111($7_{(10)}$)以及001000($8_{(10)}$)会出现相同电流情况。

[0060] 图5展示了将1FeFET1R的电阻调小，失工作电流上升，增加了多比特存内内积暨异或单元中各个支路流出的电流差；表明了在不需要限制电流大小的应用中，如汉明计算，本发明可扩展性进一步上升。

[0061] 上述实施例用来解释说明本发明，而不是对本发明进行限制，在本发明的精神和权利要求的保护范围内，对本发明做出的任何修改和改变，都落入本发明的保护范围。

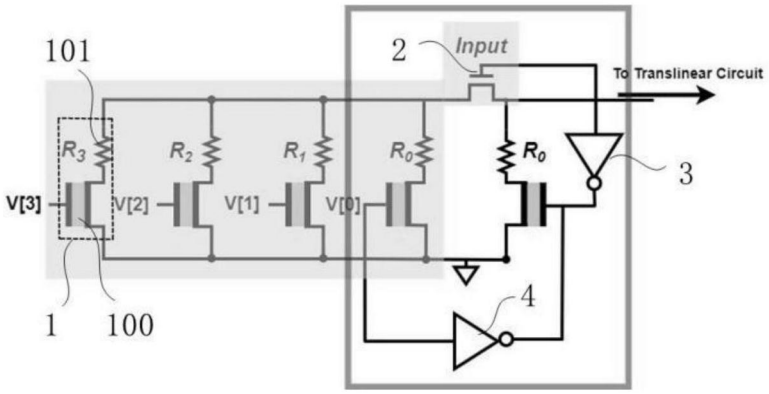

101
2
Input
To Translinear Circuit
R_3
R_2
R_1
R_0
R_0
V[3]
V[2]
V[1]
V[0]
3
4
1
100

图1

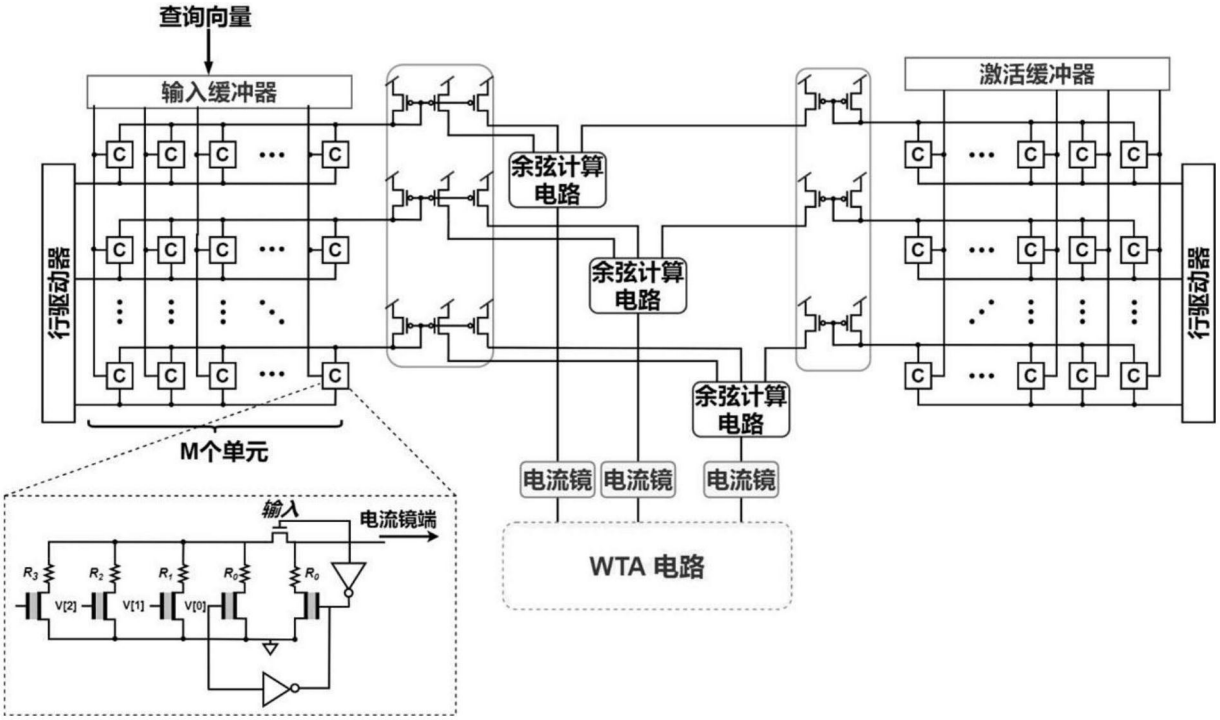

查询向量
输入缓冲器
激活缓冲器
行驱动器
行驱动器
余弦计算电路
余弦计算电路
余弦计算电路
M个单元
电流镜
电流镜
电流镜
WTA 电路
输入
电流镜端

图2

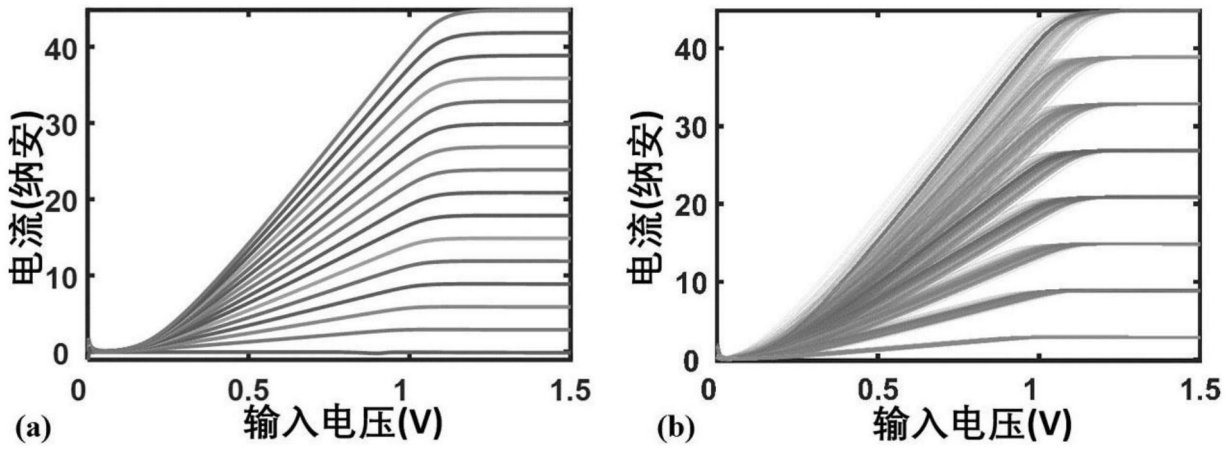


图3

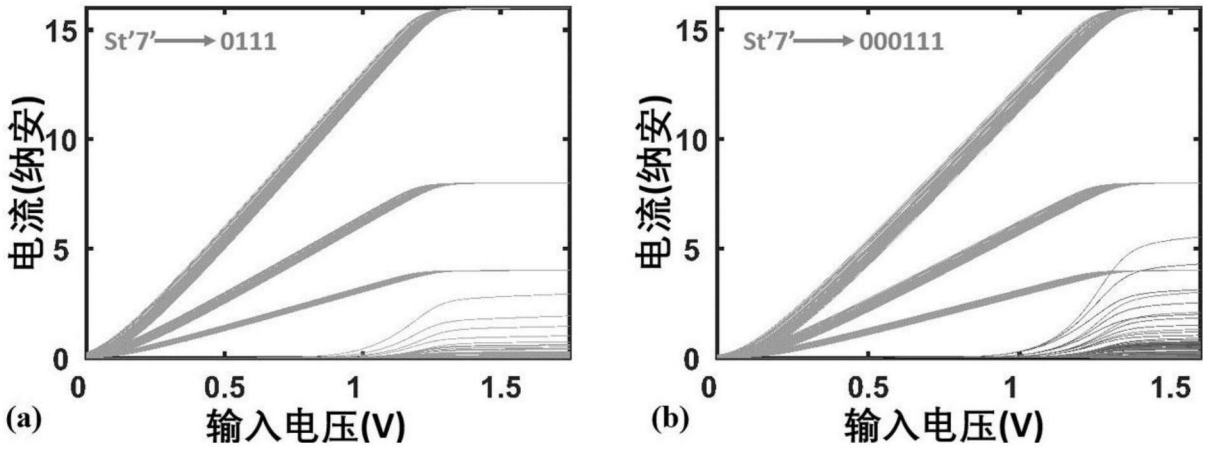


图4

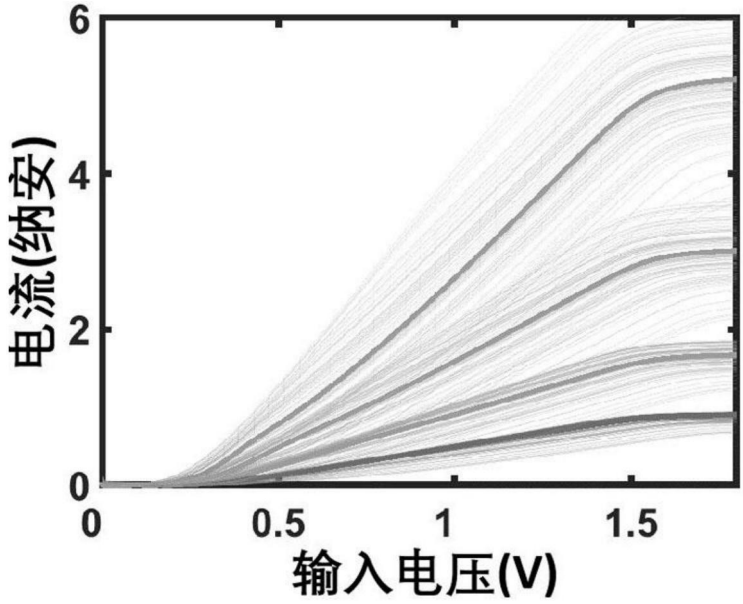

电流(纳安)
输入电压(V)
0
2
4
6
0
0.5
1
1.5

图5

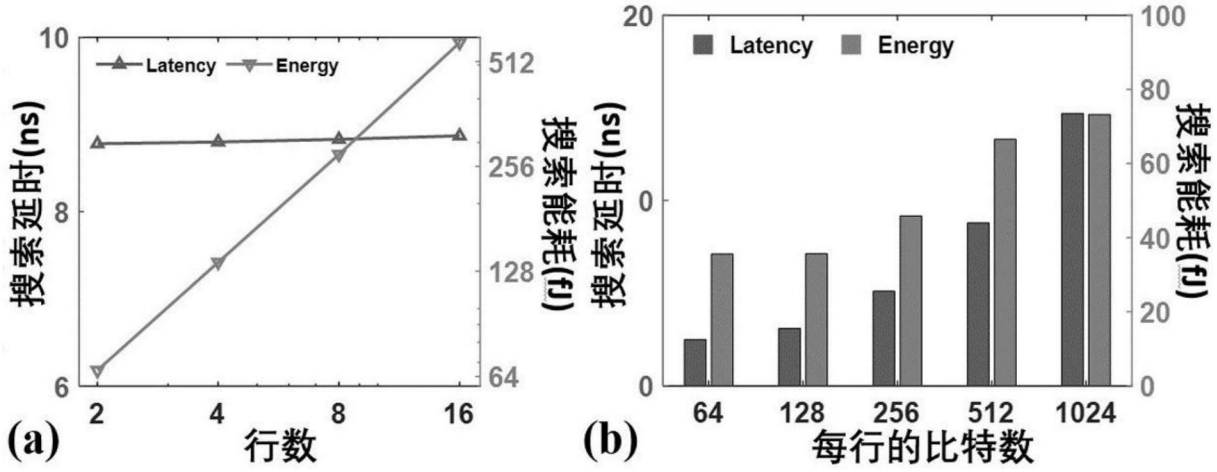

Latency
Energy
搜索延时(ns)
搜索能耗(fJ)
行数
(a)
每行的比特数
(b)

图6