(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 106372723 B
(45)授权公告日 2019.02.12

(21)申请号 201610849768.5

(22)申请日 2016.09.26

(65)同一申请的已公布的文献号
申请公布号 CN 106372723 A

(43)申请公布日 2017.02.01

(73)专利权人 上海新储集成电路有限公司
地址 201500 上海市金山区亭卫公路6505号2幢8号

(72)发明人 易敬军 陈邦明 王本艳

(74)专利代理机构 上海申新律师事务所 31272
代理人 俞涤炯

(51)Int.Cl.
G06N 3/063(2006.01)
H03M 1/46(2006.01)

(56)对比文件
CN 101807923 A,2010.08.18,
CN 7812757 B1,2010.10.12,
CN 102163973 A,2011.08.24,
US 2015/0280730 A1,2015.10.01,
CN 105281772 A,2016.01.27,
CN 103905049 A,2014.07.02,

审查员 翟紫伶

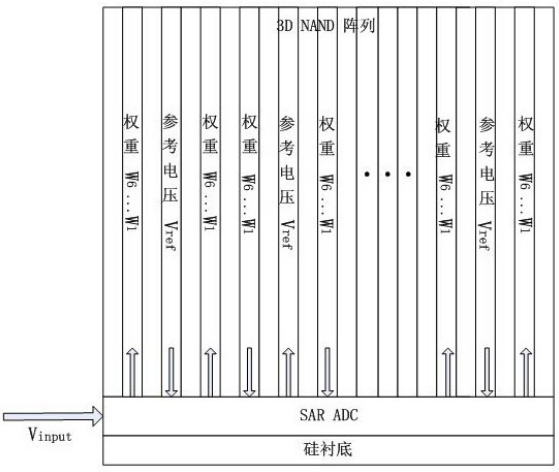
权利要求书1页 说明书5页 附图3页

(54)发明名称

基于神经网络芯片的存储结构及其存储方法

(57)摘要

本发明涉及存储器领域，尤其涉及一种基于神经网络芯片的存储结构及其存储方法。存储结构包括：衬底；N位模数转换电路，制备于衬底上；存储阵列，制备于N位模数转换电路上，包括至少一个存储单元，存储单元包括至少两个存储列；一存储列预存储有参考电压，另一存储列用于存储转换的M位进制信号的权重，N位模数转换电路利用参考电压得到M位进制信号的权重，通过读取M位进制信号的权重得到M位进制信号。存储方法包括：利用N位模数转换电路读取存储阵列中预存储的参考电压；N位模数转换电路通过比较参考电压和输入的模拟信号得到M位进制信号的权重；N位模数转换电路读取M位进制信号的权重得到M位进制信号。



CN 106372723 B

1.一种基于神经网络芯片的存储结构,其特征在于,应用于对输入的模拟信号进行M位进制的转换后存储,所述存储结构包括:

衬底;

N位模数转换电路,制备于所述衬底上;

存储阵列,制备于所述N位模数转换电路上,包括至少一个存储单元,所述存储单元包括至少两个存储列;其中,

一存储列预存储有参考电压,相邻的两个存储列用于存储转换的M位进制信号的权重,3≤M≤N,M和N均为整数;以及

所述N位模数转换电路对所述输入的模拟信号和所述参考电压进行逐次比较以得到所述M位进制信号的权重,通过读取所述M位进制信号的权重得到所述M位进制信号。

2.根据权利要求1所述的基于神经网络芯片的存储结构,其特征在于,所述N位模数转换电路为逐次逼近寄存器型模数转换器。

3.根据权利要求1所述的基于神经网络芯片的存储结构,其特征在于,所述存储阵列为3DNAND或相变存储器。

4.根据权利要求1所述的基于神经网络芯片的存储结构,其特征在于,所述N位模数转换电路与所述存储阵列集成于同一所述神经网络芯片上,和/或

所述N位模数转换电路与所述存储阵列为采用同一套半导体工艺制备。

5.根据权利要求1所述的基于神经网络芯片的存储结构,其特征在于,存储有所述M位进制信号的权重的存储列与存储有所述参考电压的存储列为相邻的存储列。

6.根据权利要求1所述的基于神经网络芯片的存储结构,其特征在于,所述N位模数转换电路包括:

比较器,两个输入端分别与所述模拟信号、所述参考电压连接,通过比较所述模拟信号和所述参考电压得到所述M位进制信号的权重。

7.一种基于神经网络芯片的存储方法,其特征在于,所述存储方法包括:

提供一衬底,在所述衬底上依次制备N位模数转换电路、存储阵列;

利用所述N位模数转换电路读取所述存储阵列中预存储的参考电压;

所述N位模数转换电路通过比较所述参考电压和输入的模拟信号得到M位进制信号的权重;

所述N位模数转换电路读取所述M位进制信号的权重得到所述M位进制信号;

其中,3≤M≤N,M和N均为整数。

8.根据权利要求7所述的基于神经网络芯片的存储方法,其特征在于,所述存储方法中,所述输入的模拟信号与所述M位进制信号的关系式为:

$V_{input}=W_M*V_{ref}+W_{M-1}*V_{ref}/2^1+W_{M-2}*V_{ref}/2^2+\ldots+W_{M-i}*V_{ref}/2^i+\ldots+W_1*V_{ref}/2^{M-1}$;

其中,Vinput为输入的模拟信号,WM为最高位,W1为最低位,Vref为参考电压。

基于神经网络芯片的存储结构及其存储方法

技术领域

[0001] 本发明涉及存储器领域，尤其涉及一种基于神经网络芯片的存储结构及其存储方法。

背景技术

[0002] 模数转换器(ADC)是一种计算机与人、与真实世界的沟通的重要工具，它可以将真实世界中广泛存在的模拟信号转换为计算机可以识别的数字信号。目前市面上有很多ADC的类型，其中逐次逼近寄存器型模数转换器(SAR ADC)是其中应用非常广泛的一种，它是一种中等速度、中等精度、低功耗、低成本的ADC。

[0003] 如图1所示，基本的SAR ADC由取样/保持电路、比较器、数模转换器(N位DAC)、N位寄存器和逻辑控制电路所组成。模拟输入电压(Vinput)由采样/保持电路采样并保持。为实现二进制搜索算法，N位寄存器首先设置在中间刻度(即：100....00，最高位MSB(most significant bit)设置为1)。这样，N位DAC输出(VDAC)被设为VREF/2，其中Vref是提供给比较器的基准电压。然后，比较判断Vinput是小于还是大于Vref。如果Vinput大于VDAC，则比较器输出逻辑高电平或1，N位寄存器的MSB保持为1。相反，如果Vinput小于VDAC，则比较器输出逻辑低电平，N位寄存器的MSB清0。以Dn表示最高位寄存器的数值。随后，SAR控制逻辑移至下一位，并将该位设置为高电平，进行下一次比较。这个过程一直持续到最低位LSB。以D1表示最低位寄存器的数值。上述操作结束后，也就完成了转换，N位转换结果储存在寄存器内，并且输出转换后的数字信号。输入模拟量就可以转换成N位数字量。

[0004] 但是，这种实现方法有两个缺点。第一，这种SAR ADC为平面结构，它占据了很大的面积，不能实现非常大的密度以及非常低的成本；第二，这种SAR ADC需要一个提供参考电压的电路，这也造成了很大的面积浪费。

[0005] 3D非易失性存储器主要有3D NAND和相变存储器PCM。如图2所示的在源线上制备有选择栅的3D NAND的结构，单个MLC(2bits/cell)闪存芯片上可以增加最高32GB的存储空间，而单个TLC(3bits/cell)闪存芯片可增加48GB。该技术可支持在更小的空间内容纳更高存储容量，进而带来很大的成本节约、能耗降低，以及大幅的性能提升以全面满足众多消费类移动设备和要求最严苛的企业部署的需求。但是，3D NAND也有缺点，它需要额外的感测放大器 (Sensed Amp)来读取3D NAND中存储的信息。这将带来许多额外的电路，加大整个存储器的面积。

发明内容

[0006] 针对现有技术存在的问题，本发明提供了一种基于神经网络芯片的存储结构及其存储方法，无需额外的参考电压产生电路和感测方法器，并且增加了存储密度。

[0007] 本发明采用如下技术方案：

[0008] 一种基于神经网络芯片的存储结构，应用于对输入的模拟信号进行M位进制的转换后存储，所述存储结构包括：

[0009] 衬底；

[0010] N位模数转换电路，制备于所述衬底上；

[0011] 存储阵列，制备于所述N位模数转换电路上，包括至少一个存储单元，所述存储单元包括至少两个存储列；其中，

[0012] 一存储列预存储有参考电压，相邻的两个存储列用于存储转换的M位进制信号的权重，3≤M≤N，M和N均为整数；以及

[0013] 所述N位模数转换电路利用所述参考电压得到所述M位进制信号的权重，通过读取所述M位进制信号的权重得到所述M位进制信号。

[0014] 优选的，所述N位模数转换电路为逐次逼近寄存器型模数转换器。

[0015] 优选的，所述存储阵列为3D NAND或相变存储器。

[0016] 优选的，所述N位模数转换电路与所述存储阵列集成于同一所述神经网络芯片上，和/或

[0017] 所述N位模数转换电路与所述存储阵列为采用同一套半导体工艺制备的。

[0018] 优选的，存储有所述M位进制信号的权重的存储列与存储有所述参考电压的存储列为相邻的存储列。

[0019] 优选的，所述N位模数转换电路包括：

[0020] 比较器，两个输入端分别与所述模拟信号、所述参考电压连接，通过比较所述模拟信号和所述参考电压得到所述M位进制信号的权重。

[0021] 一种基于神经网络芯片的存储方法，所述存储方法包括：

[0022] 提供一衬底，在所述衬底上依次制备N位模数转换电路、存储阵列；

[0023] 利用所述N位模数转换电路读取所述存储阵列中预存储的参考电压；

[0024] 所述N位模数转换电路通过比较所述参考电压和输入的模拟信号得到M位进制信号的权重；

[0025] 所述N位模数转换电路读取所述M位进制信号的权重得到所述M位进制信号；

[0026] 其中，3≤M≤N，M和N均为整数。

[0027] 优选的，所述存储方法中，所述输入的模拟信号与所述M位进制信号的关系式为：

[0028] $V_{input}=W_M*V_{ref}+W_{M-1}*\ V_{ref}/2^1+\ W_{M-2}*\ V_{ref}/2^2+\ \ldots+W_{M-i}*\ V_{ref}/2^i+\ \ldots+\ W_1*\ V_{ref}/2^{M-1}$；

[0029] 其中，V_{input}为输入的模拟信号，W_M为最高位，W_1为最低位，V_{ref}为参考电压。

[0030] 本发明的有益效果是：

[0031] 本发明通过将SAR ADC制作于3D存储器阵列与硅衬底之间，充分利用了硅片面积，并且通过SAR ADC来读取3D存储器阵列中的信息，就可以无需额外的参考电压产生电路与感测放大器；通过在每个存储单元中存储不仅仅是0和1两种情况的多位信号，大大增加了存储密度。

附图说明

[0032] 通过阅读参照以下附图对非限制性实施例所作的详细描述，本发明及其特征、外形和优点将会变得更加明显。在全部附图中相同的标记指示相同的部分。并未可以按照比例绘制附图，重点在于示出本发明的主旨。

[0033] 图1为现有技术SAR DAC的电路结构图；

[0034] 图2为现有技术3D NAND的结构示意图；

[0035] 图3为本发明3D存储器单元与SAR DAC的结构示意图；

[0036] 图4为本发明基于3D 存储器阵列的存储结构示意图。

具体实施方式

[0037] 下面结合附图和具体的实施例对本发明作进一步的说明，但是不作为本发明的限定。

[0038] 人工神经网络是20世纪80 年代以来人工智能领域兴起的研究热点。它从信息处理角度对人脑神经元网络进行抽象，建立某种简单模型，按不同的连接方式组成不同的网络。神经网络是一种运算模型，由大量的节点(或称神经元)之间相互联接构成。每个节点代表一种特定的输出函数，称为激励函数。每两个节点间的连接都代表一个对于通过该连接信号的加权值，称之为权重，这相当于人工神经网络的记忆。网络的输出则依网络的连接方式，权重值和激励函数的不同而不同。而网络自身通常都是对自然界某种算法或者函数的逼近，也可能是对一种逻辑策略的表达。

[0039] 在人工神经网络中，最基本模型的算法为 t=f(WA'+b)，通过上述存储列中存储的权重W，将其带入神经网络的基本模型算法中，求得神经网络的输出。

[0040] 其中，t为神经网络的输出；W为权重向量，分为W1~WN，为神经元各个突触的权值；A为输入向量，分为a1~an，为输入向量的各个分量，A'为A向量的转置；b为偏置；f为传递函数，通常为非线性函数。可见，一个神经元的功能是求得输入向量与权重向量的内积后，经一个非线性传递函数得到一个标量结果。

[0041] 下面结合具体实施例进行说明：

[0042] 实施例一：

[0043] 如图3所示，本实施例提供了一种基于神经网络芯片的存储结构，将SAR ADC制作于3D存储器阵列(存储阵列)与硅衬底之间，参考电压V_{ref}为3D存储器阵列中的存储单元的一存储列存储的SAR ADC的参考电压，外围输入的模拟信号经过SAR ADC转换为3D存储器阵列可识别的数字信号。这样，SAR ADC和非易失性3D存储器(存储阵列)就实现了模拟信号的采样、转换和保存，充分利用了硅片面积，充分利用了多位存储器高密度的优点，且无需额外的SAR ADC所需的参考电压产生电路。即以3个存储位可以实现2个M进制的M位数据存储，其数据存储效率就是2进制的2*M/3倍。M越大，存储效率越高。

[0044] 进一步的，3D存储器阵列和SAR ADC使用同一套半导体工艺制备，且集成于同一块芯片上。

[0045] 进一步的，3D存储器阵列采用3D NAND 或相变存储器(PCM)等非易失性存储器。

[0046] 进一步的，采用3D存储器阵列中相邻的存储单元做SAR ADC参考信号源，以抵消工艺过程的不匹配。

[0047] 实施例二

[0048] 本实施例提出了一种适用于神经网络芯片的多进制信号存储方法，即N(N≥3)位的SAR ADC实现利用对输入模拟信号V_{input}的M(3≤M≤N)位进制转换，并且将转换后的存储结果存储于3D存储器阵列的存储单元中。在本实施例中，每个3D存储单元可以存储多进制

信号,这样整个存储器的存储密度将远远大于原来的存储器。

[0049] 例如，3D存储器阵列的每个3D存储单元堆栈由三个存储列构成,两边两列存储SAR ADC的模拟输入信号,中间一列存储SAR ADC的参考电压。在这其中,每个存储单元将存储多位信号(而不是通常的0或1两种情况)。

[0050] 神经网络基本的模型算法为:以M=N位数据存储为例,SAR ADC实现了模拟信号的N位数字信号转换:

[0051] $V_{input}=W_N*V_{ref}+W_{N-1}*V_{ref}/2^1+W_{N-2}*V_{ref}/2^2+\ldots+W_{N-i}*V_{ref}/2^i+\ldots+W_1*V_{ref}/2^{N-1}$

[0052] 当M<N时：

[0053] $V_{input}=W_M*V_{ref}+W_{M-1}*V_{ref}/2^1+W_{M-2}*V_{ref}/2^2+\ldots+W_{M-i}*V_{ref}/2^i+\ldots+W_1*V_{ref}/2^{M-1}$

[0054] 其中,V_{input} 为外部输入的模拟信号,一存储列存储权重W,参考电压信号V_{ref}预先存储在紧密相邻存储单元的权重所在的存储列中参考电压信号V_{ref}为默认的1/2电源模拟量,如果用最高位电位为高电平其余都是低电平来表示这个权重信号,该信号为[M-1 M-2 ...0]=[1 0 ...0]。W 代表SAR ADC输出的数字标量，其中最高位是W_n,W_1代表最低位。这样就实现了对同一单元的M位数据的加权求和运算。

[0055] 下面举一个本实施例的具体应用来进一步说明。

[0056] 在人工神经芯片领域,信号被处理的速度不需要很快,处理的频率不需要很高,但是存储的容量必须很大。因此,高密度的存储器对于实现人工神经网络相当重要。在这其中,权重W很重要,它需要能够随时读写,并且需要很大的存储密度。而利用本实施例可以很容易地读取突触中的权值W,而不需额外附加太多电路,且有非常大的存储密度。具体实现方法为:如图4所示,有一个3D NAND阵列(存储阵列),其中每个3D NAND单元(存储单元)堆栈由两列构成,其中一列存储权重W6~W1,另一列存储SAR ADC的参考电压Vref=100,000。每个3D存储单元能存储6个可叠加信号,这样同样的单元数量存储的数据量就达到了2进制数据的64位的数据量。将SAR ADC制作于每个3D NAND单元堆栈与硅衬底之间,输入信号Vinput和参考电压Vref经过SAR ADC逐次比较,得到从最高位到最低位的6位权重W6~W1,SAR ADC每次读出一行字线中的权值,与此同时其他字线通高电平。这样SAR ADC就完成了对突触内权值W的读取。权重W6~W1的运算公式为：

[0057] Vinput=W6*Vref+W5* Vref/2+ W4* Vref/22+ W3* Vref/23+ W2* Vref/24+ W1* Vref/25。

[0058] 综上所述,本发明提出的这种适用于类神经网络的多位信号存储结构,将3D非易失性存储器与SAR ADC相结合,在每个存储单元中存储可叠加的多位信息,从而具有了高存储密度,且无需太多外部电路的优点。利用非易失性存储器的掉电数据不丢失,实现了数据记忆能力,又节省了数据保持的功耗。本发明的存储密度将远远超出普通3D存储器,而且进制的提高也意味着信号数据的加倍压缩,可以加快数据的传输效率,同时也大大节省了处理器的运算量。这是一种绝佳的脑神经网络芯片解决方案。

[0059] 以上对本发明的较佳实施例进行了描述。需要理解的是,本发明并不局限于上述特定实施方式,其中未尽详细描述的设备和结构应该理解为用本领域中的普通方式予以实施;任何熟悉本领域的技术人员,在不脱离本发明技术方案范围情况下,都可利用上述揭示

的方法和技术内容对本发明技术方案做出许多可能的变动和修饰，或修改为等同变化的等效实施例，这并不影响本发明的实质内容。因此，凡是未脱离本发明技术方案的内容，依据本发明的技术实质对以上实施例所做的任何简单修改、等同变化及修饰，均仍属于本发明技术方案保护的范围内。

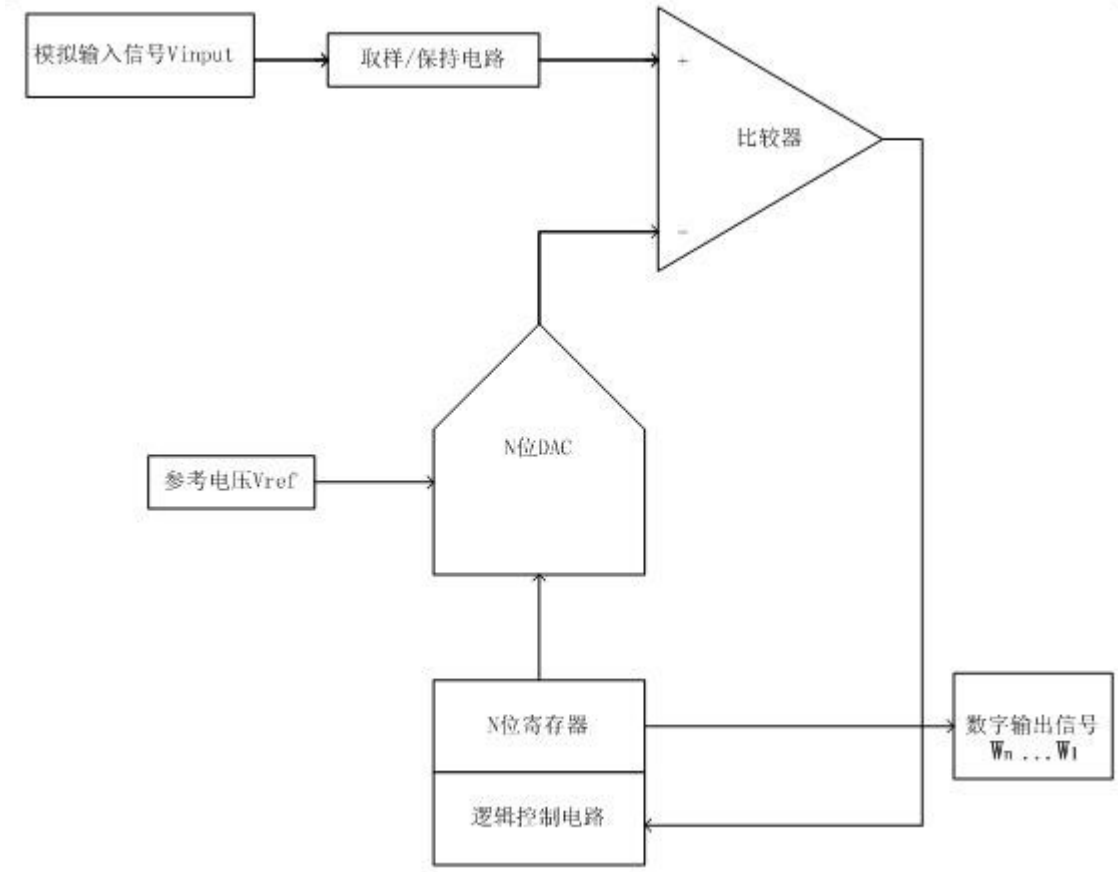

模拟输入信号Vinput
取样/保持电路
比较器
+
-
N位DAC
参考电压Vref
N位寄存器
逻辑控制电路
数字输出信号
Wn ...W1

图1

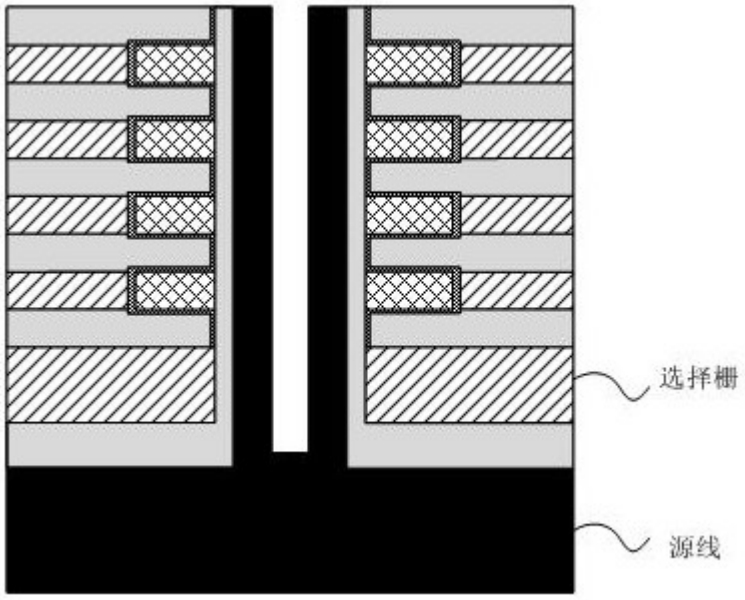

选择栅
源线

图2

图3

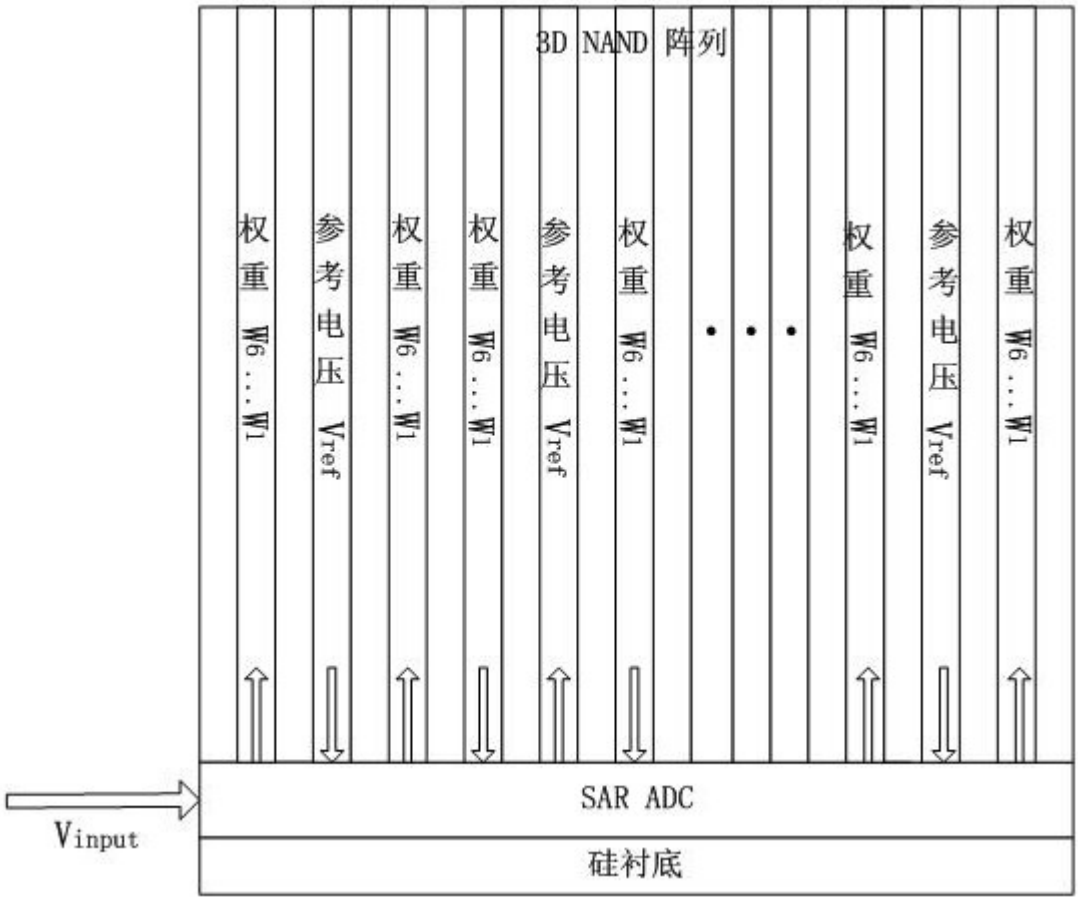

3D NAND 阵列
权重 W6 ...W1
参考电压 Vref
权重 W6 ...W1
权重 W6 ...W1
参考电压 Vref
权重 W6 ...W1
权重 W6 ...W1
参考电压 Vref
权重 W6 ...W1
SAR ADC
Vinput
硅衬底

图4