申请人：浙江大学

发明人：黄科杰 张赛 沈海斌

第一发明人身份证号：330222198007126952

联系人电话及邮箱：18202744825 784426657@qq.com

寄信地址及邮编：浙江大学玉泉校区老生仪楼3楼 邮编310027

**专利申请提交的技术交底材料**

**申请发明或者实用新型专利应提交交底材料应包括以下内容：**　　一、名称  
　　一种基于多位并行二进制突触阵列的神经形态计算电路  
　　二、技术领域  
　　本发明属于神经形态计算领域，涉及一种基于多位并行二进制神经网络突触阵列的神经形态计算电路。  
　　三、背景技术  
　　近年来深度神经网络在人工智能领域迅速发展，在图像识别、自然语言处理等方面取得优异的成果。目前很多先进的深度学习算法，通过增加网络的深度和参数的数量来提高网络的性能，对硬件的存储容量、计算能力以及能效提出了更高的要求。比如AlphaGo需要消耗一百万瓦的能量才能获得足够的算力，相比之下人脑只需要消耗20瓦的能量。

神经形态计算能够大幅提升人工神经网络计算的能效，通过模仿人脑的结构将存储单元和计算单元集成在一起，解决了传统冯诺依曼结构传输带宽和传输能耗的瓶颈问题。新兴的电阻式非易失性存储器（RRAM，Resistive Random-Access-Memory）是实现神经形态计算的最佳选择，利用RRAM阻值可以将输入信号的加权组合转变为输出电压，完成人工神经网络中的基本操作矩阵乘法和累加（MAC，Multiplication-and-Accumulation），从而实现超低功耗的存内并行计算。

当前提出的神经形态计算电路，大都需要高精度的数模转换器（DACs，Digital-to-Analog Converters）和模数转换器（ADCs，Analog-to-Digital Converters）作为接口器件，导致接口器件的能耗占整体能耗的80%以上，不利于在边缘计算设备里的应用。而且当前的神经形态计算解决方案，实现的权重量化精度和激活值量化精度低，只能面向Lenet等简单网络，对于Alexnet等规模较大的深度神经网络性能损失明显，很大程度上限制了其应用的范围。因此，本发明提出了一种基于多位并行二进制神经网络突触阵列的神经形态计算电路，能够在低能耗的情况下实现高精度高性能的深度神经网络。  
　　四、发明内容

针对现有技术存在的缺陷和对低功耗高精度的改进需求，本发明提出了一种新颖的神经网络突触阵列，能够执行大量乘和累加的并行计算。同时提出了一种高效能的神经形态计算架构，可配置成不同的深度神经网络，以满足不同的应用需求。

传统的神经形态计算电路如图1所示，DAC和ADC等接口部件会带来很大的功耗，而且以不同的输入电压作为RRAM的读取电压，RRAM阻值会产生较大的偏差，导致计算结果的精确度不高，限制了应用的范围。图2表示本发明提出的神经形态计算架构，包括神经轴突模块、多位并行的二进制RRAM突触阵列、时分复用器、多个积分器和一个共享的逐次逼近型模数转换器（SAR ADC，Successive Approximation Register Analog-to-Digital Converter）。来自神经网络上一层的输入信号，先进入神经轴突模块，神经轴突模块包括2个基本单元：时序调度器和加法器。时序调度器用于安排信号的时序，使输入信号采用树突优先的策略，依次输入到多位并行的二进制RRAM突触阵列；加法器可用于阵列规模的拓展，当配置的神经网络输入层大于1个RRAM阵列的输入时，可以利用轴突模块的加法器将多个阵列的计算结果相加，从而得到网络层的输出。积分器包括积分运放和开关电容电路，用来将输入信号和RRAM阵列权重的MAC计算结果转化为模拟积分电压，在下面积分电路的描述中将会给出详细的介绍。最后通过共享SAR ADC将模拟积分电压量化为N位数字形式的输出数据。其中的时分复用器用于将SAR ADC和积分器共享给网络层所有的输入，通过时序的调度最大化硬件资源的利用率。

本发明提出的多位并行的二进制RRAM突触阵列如图3所示。采用图4的1晶体管-1RRAM（1T1R，1 Transistor 1 RRAM）结构作为基本组成单元，晶体管（NMOS）用来控制开关行为，源级接地，漏级接二进制RRAM的一端，RRAM的另一端连入积分器电路，N个二进制RRAM以固定点数的形式来模拟神经突触的差异水平。栅极接输入信号线，神经网络层的输入也采用N位固定点数的形式，每位二进制的输入直接作为1T1R单元的控制电压Vc，从而消除了输入接口DAC的使用，大幅减少了能量的消耗和面积的占用。本发明采用氮掺杂氧化铝结构的实验数据对RRAM进行建模，每个RRAM具有2个电阻：低阻态（约10MΩ）和高阻态（约1G—10GΩ）。通过时序安排，使RRAM阵列只在积分阶段导通，大多数时间处于关断状态，从而大大降低了突触阵列的功耗。

相比传统的SRAM阵列，本发明提出的RRAM阵列具有高密度和一次性读取的特点，能大幅减少突触阵列的功耗和面积。相比传统的多位RRAM和采用不同输入电压作为读取电压的体制，本发明提出的多个二进制RRAM解决了单个多位RRAM非线性偏差大、量化精度低的问题，同时采用固定的运放参考电压作为RRAM的读取电压，能显著降低RRAM阻值在不同读取电压下的偏差，提升权重量化的精度。相比传统的二值神经网络（BNN，Binarized Neural Networks）突触阵列的方法，本发明提出的N个二进制RRAM能将权重量化精度提升至N位，而且RRAM阵列的每层激活值输入也是N位，因此可以将整个网络的精度提升到N位，克服了BNN阵列结构在Alexnet等深度神经网络中性能损失大的问题，能实现较高的准确率。

本发明提出的计算电路积分原理如图5所示，采用256路输入并行计算和树突优先的策略，将输入数据的每一位顺序输入积分电路，利用电荷重新分配的原理，完成如图6的积分体制。N位输入信号和N位RRAM阵列权重，可依次表示为数字形式x=An-1An-2…A0和w=an-1an-2…a0，利用欧姆定律和电流积分完成乘和累加过程，然后将模拟积分电压的结果通过共享SAR ADC量化为N位数字形式，便于信号的传输与存储。本发明提出的积分电路，利用电荷重新分配的原理，完成不同权重位和不同输入位的加权过程，结构简单且误差小易控制，能实现较高的积分精度和网络正确率。而当前提出的镜像电流源体制和动态阈值体制等，普遍存在结构复杂、电路误差大和功耗大的问题，导致只能应用在小规模的神经网络中。

本发明提出的8位共享SAR ADC结构如图7所示，可根据具体需要将SAR ADC配置为N位。在积分电路中用于数据暂存和电荷重新分配的电容也用于SAR ADC中的DAC电容阵列，通过资源共享的方式减少了面积的占用。本发明提出的8位共享SAR ADC采用高精度高功耗ADC和低精度低功耗ADC组合的结构，用低功耗ADC量化高4位结果，用高精度ADC量化低4位结果，在实现高精度的同时降低了能量的消耗。同时采用动态比较器结构和自激时钟的方法来降低比较器的功耗，采用分离DAC电容方法降低电容阵列的转换功耗，因此利于部署到低功耗的设计中。

本发明提出的多位并行的二进制RRAM神经网络突触阵列和神经形态计算电路，相比当前的体制具有高精度和低功耗的优势，可配置成大多数深度神经网络应用，特别适合部署于对能耗要求高的边缘计算设备中。

　　五、附图说明  
　　图1是传统神经形态计算结构图；

图2是本发明提出的高效能神经形态计算结构图；

图3是本发明提出的多位并行的二进制RRAM突触阵列；

图4是1T1R单元结构图；

图5是本发明提出的计算电路积分原理图；

图6是本发明提出的积分体制框图；

图7是本发明提出的8位共享SAR ADC结构图。

　　六、具体实施方式  
　　下面结合附图对本发明的具体实施方式做进一步详细地说明。

图3是本发明采用的神经突触阵列结构，用N个二进制的RRAM来模拟一个突触，因此可将一个N位固定点权重表示为w= an-1an-2…a0，进一步地可将树突输出结果表达为：

图5和图6是具体的计算电路积分原理和积分体制。每个积分器由积分运放、Cn电容、Cf-Cn电容以及S1、S2、S3、S4开关组成，具体连接关系如图所示。采用256路并行输入，每一个输入数据量化为N位固定点数，从低位到高位依次进入积分电路，换言之，A0,0A1,0…Ap-1,0依次被选为轴突线的输入，作为RRAM突触阵列中1T1R单元的控制电压。

当积分电路开启时，门控时钟打开，由开关S1、S2、S3、S4和S5控制积分过程和电荷重分配过程。

在积分阶段，开关S1、S2和SAR ADC中的采样开关S5闭合，同时将开关S3和S4关断以分隔开积分器的输出电压，得到的积分电压可表示为：

其中Vo是当前积分器的积分电压，Vo-是积分器前面的积分状态，T是固定的积分时间，Gi是二值化权重的电导值（RRAM在高阻态和低阻态对应的电阻值分别为1/RH和1/RL），Vref是基准读取电压，Cf是总的反馈电容。

当1位积分过程完成后，开关S2断开以维持积分电压恒定，同时关断运放以最小化功耗，然后开关S1关断使RRAM阵列的功耗接近于0。随后开关S3闭合利用电荷重分配方法得到MAC计算的等价模拟电压。同时开关S4关断完成积分电路的复位。一旦电荷重分配过程完成，开关S4关断、S2打开，为下一位输入数据的积分过程做好准备。

在电荷重分配阶段，同时完成了不同权重位和不同输入位的加权过程。首先用不同的电容来实现不同权重位的加权，电容从大到小依次为Cn-1, Cn-2…C0，可表示为如下关系：Cn-1=21 Cn-2=…=2n-1C0。不同权重位加权后，过程等效电压Vs可表示为：

等式（3）可以看作等式（1）在输入只有1位时的特殊情况。

SAR ADC中的共享DAC电容阵列Cf（Cf=2nC0）用于镜像Vs电压，以完成不同输入位的加权过程。输出积分电压Vout初始化为0，每位积分过程完成后，前面输入位的平分电压Vx-与当前位的积分等效电压Vs通过Cf和Cn-1Cn-2…C0进行电荷平分，由于输入数据从低位到高位依次进行输入，相当于分别被平分了2n-1,2n-2…20次，因此最后积分输出电压Vout可表示为：

等式（4）等价于等式（1）。通过上述积分过程和电荷重分配过程，完成了数字形式的N位固定点输入和N位固定点权重的乘和累加运算，得到模拟形式的输出电压。

最后，位于SAR ADC中的采样开关S5断开，在完成所有位积分过程得到Vout输出电压的同时，SAR ADC也通过共享DAC阵列完成了对Vout的采样，开始将模拟积分电压结果量化成N位数字形式，以便于存储和传输。在SAR ADC量化阶段，门控时钟关断，开关S1、S2、S3和S4断开以关断积分电路的能耗。

以上实施例仅用于说明本发明的技术方案而非对其限制，尽管参照上述实施例对本发明进行了详细的说明，所属领域的普通技术人员应当理解；其依然可以对本发明的具体实施方式进行修改或者等同替换，而未脱离本发明精神和范围的任何修改或者等同替换，其均应涵盖在本发明的权利要求范围当中。

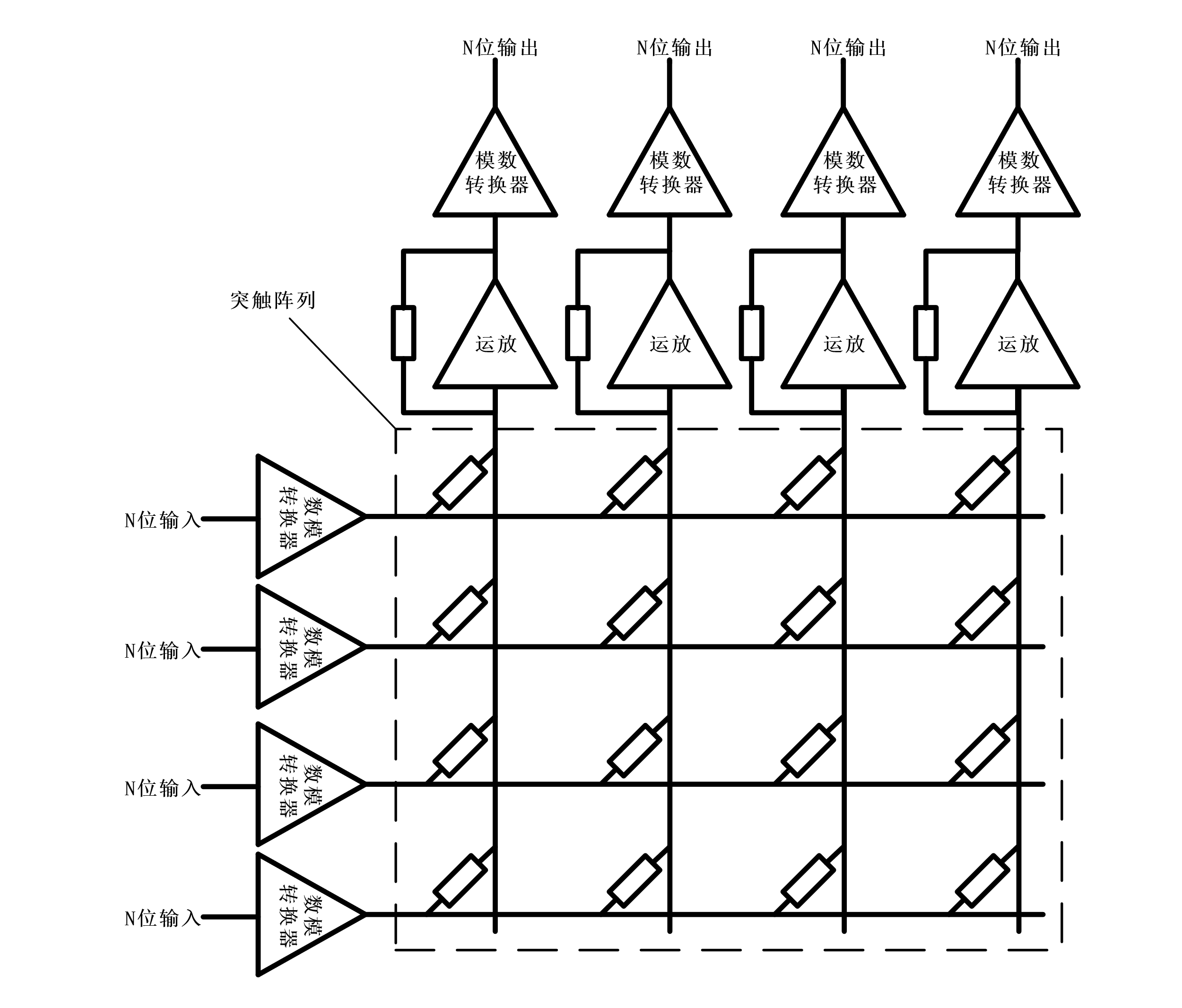


图1

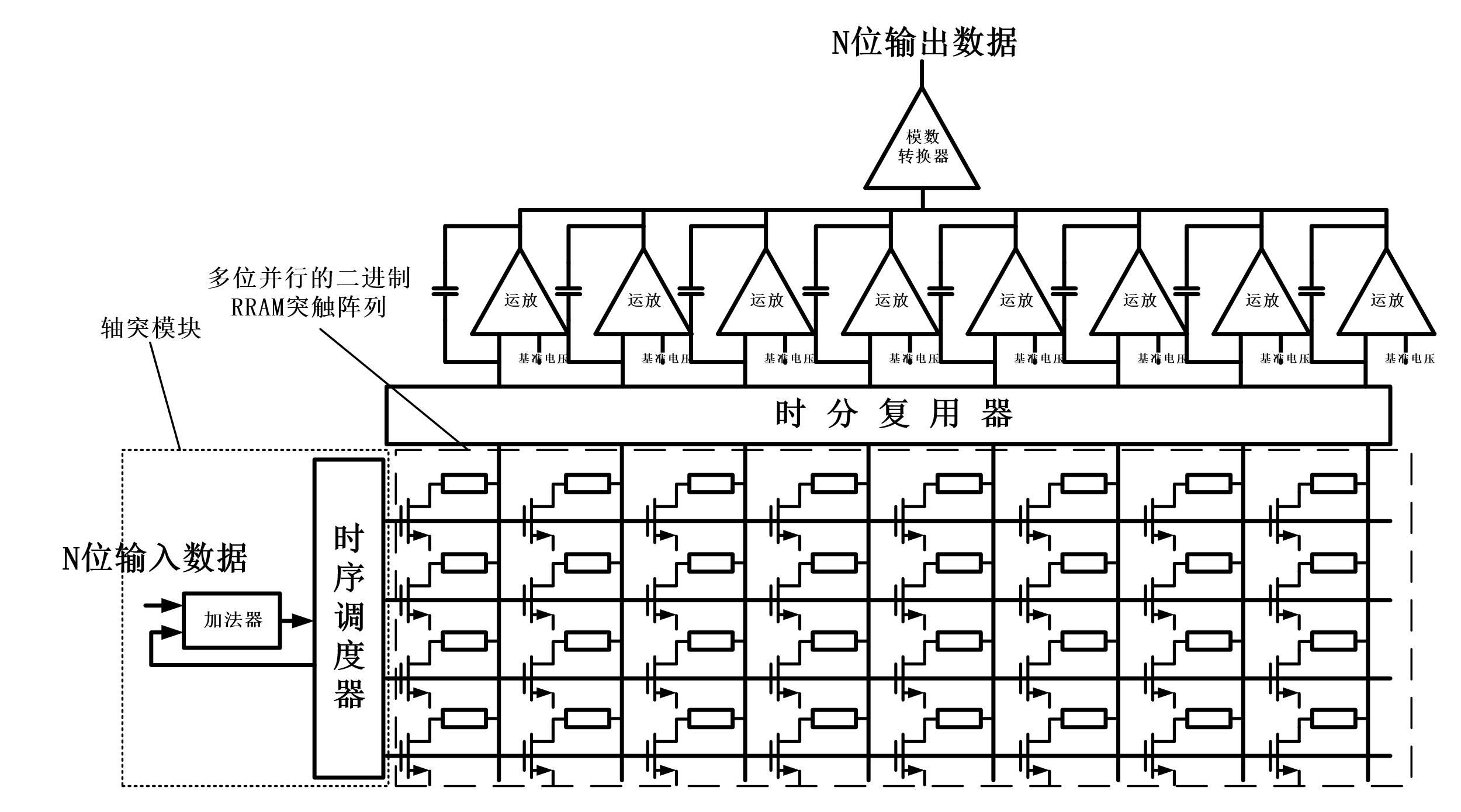


图2

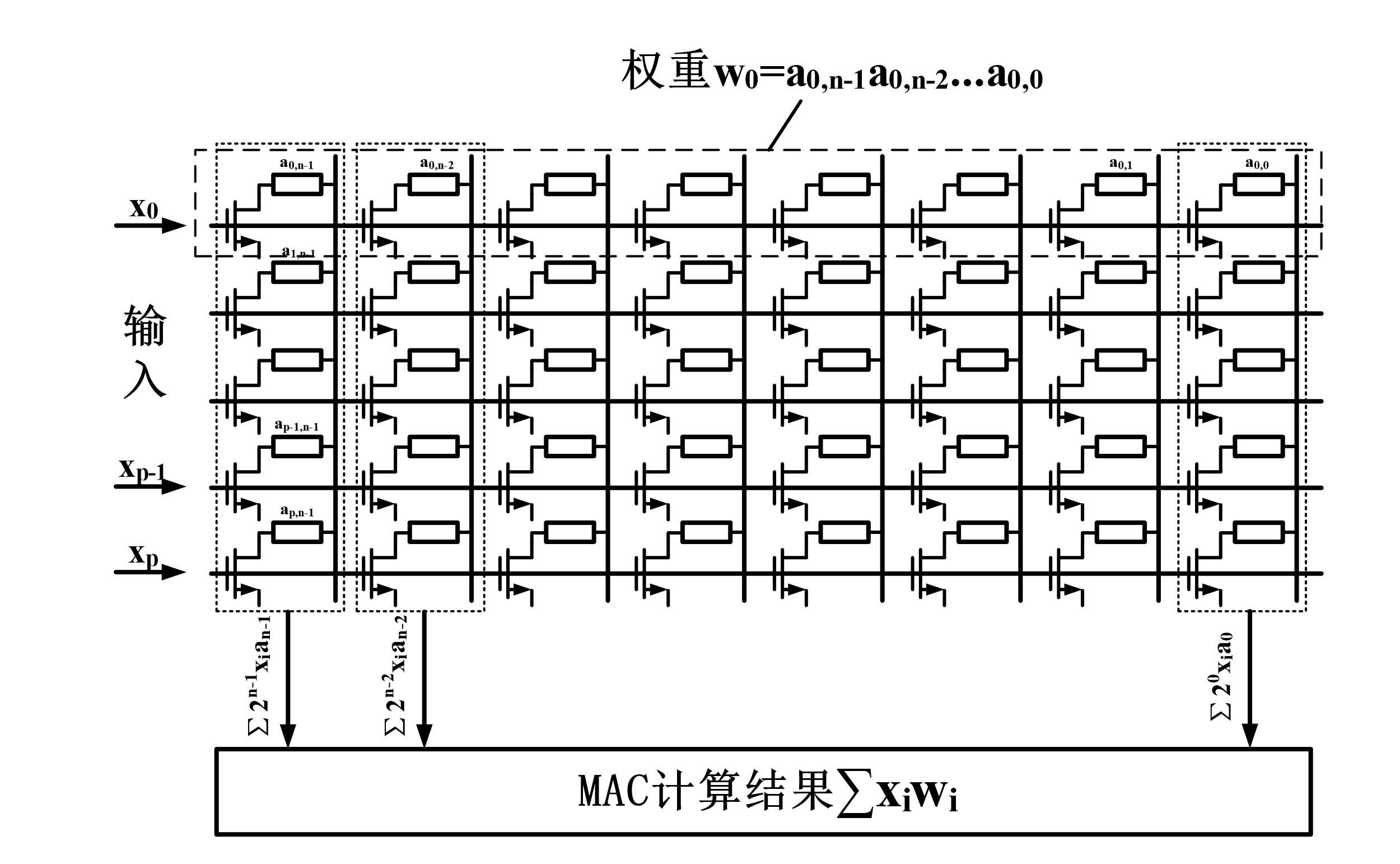


图3

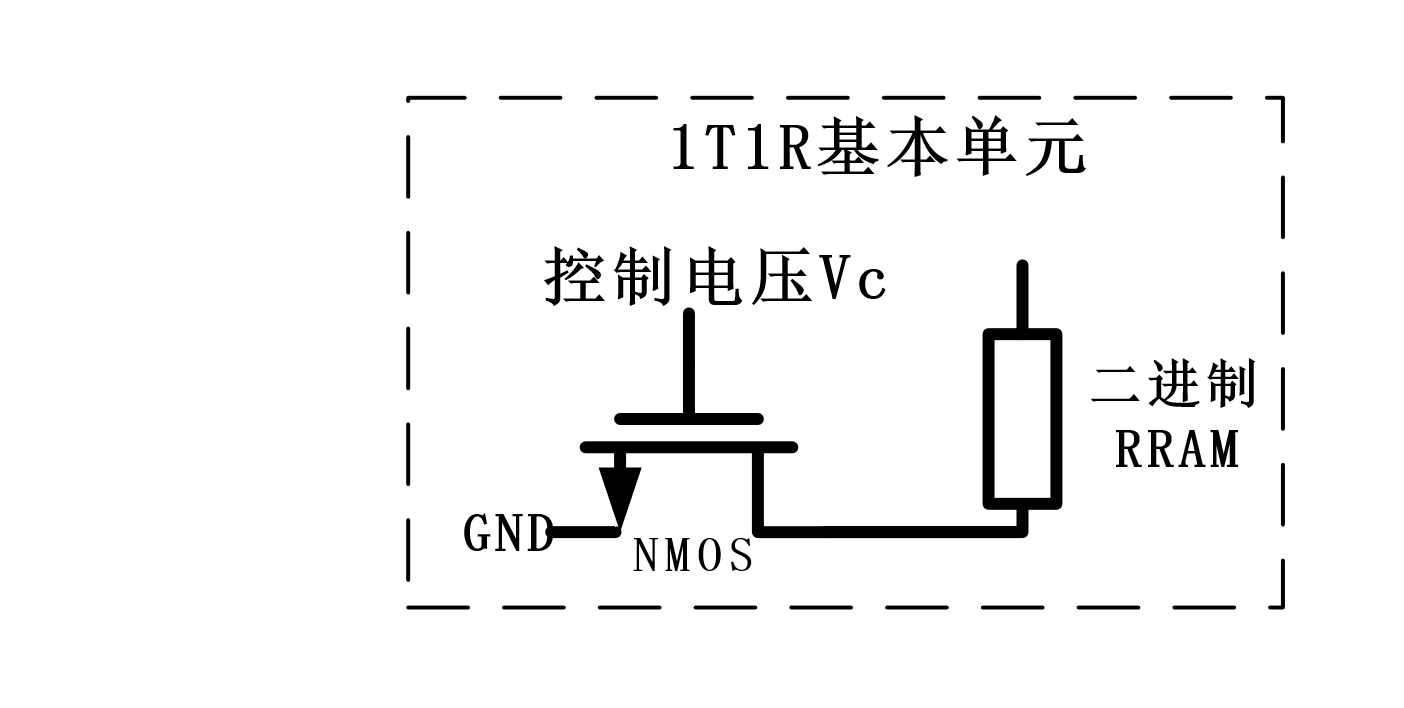


图4

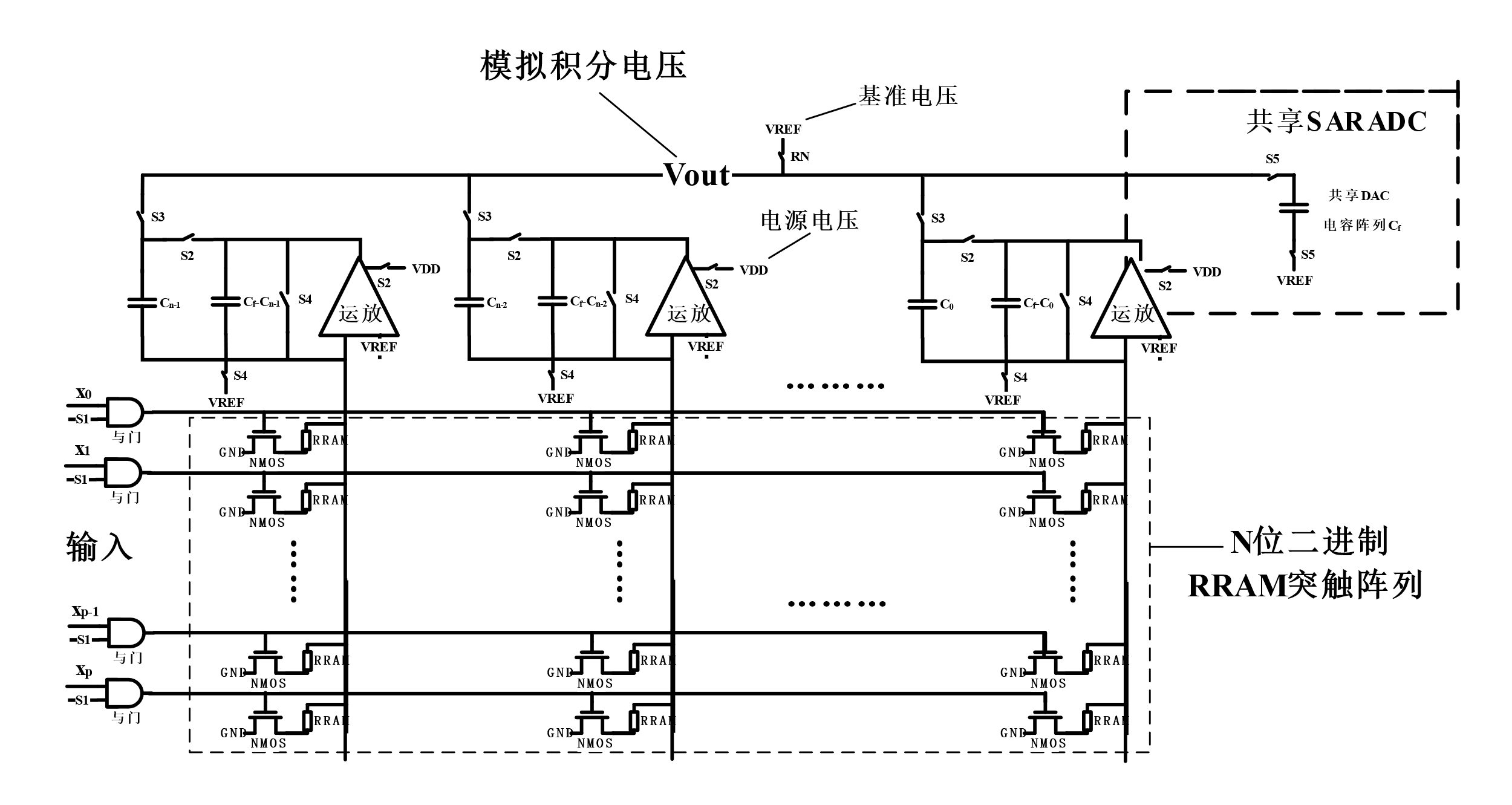


图5

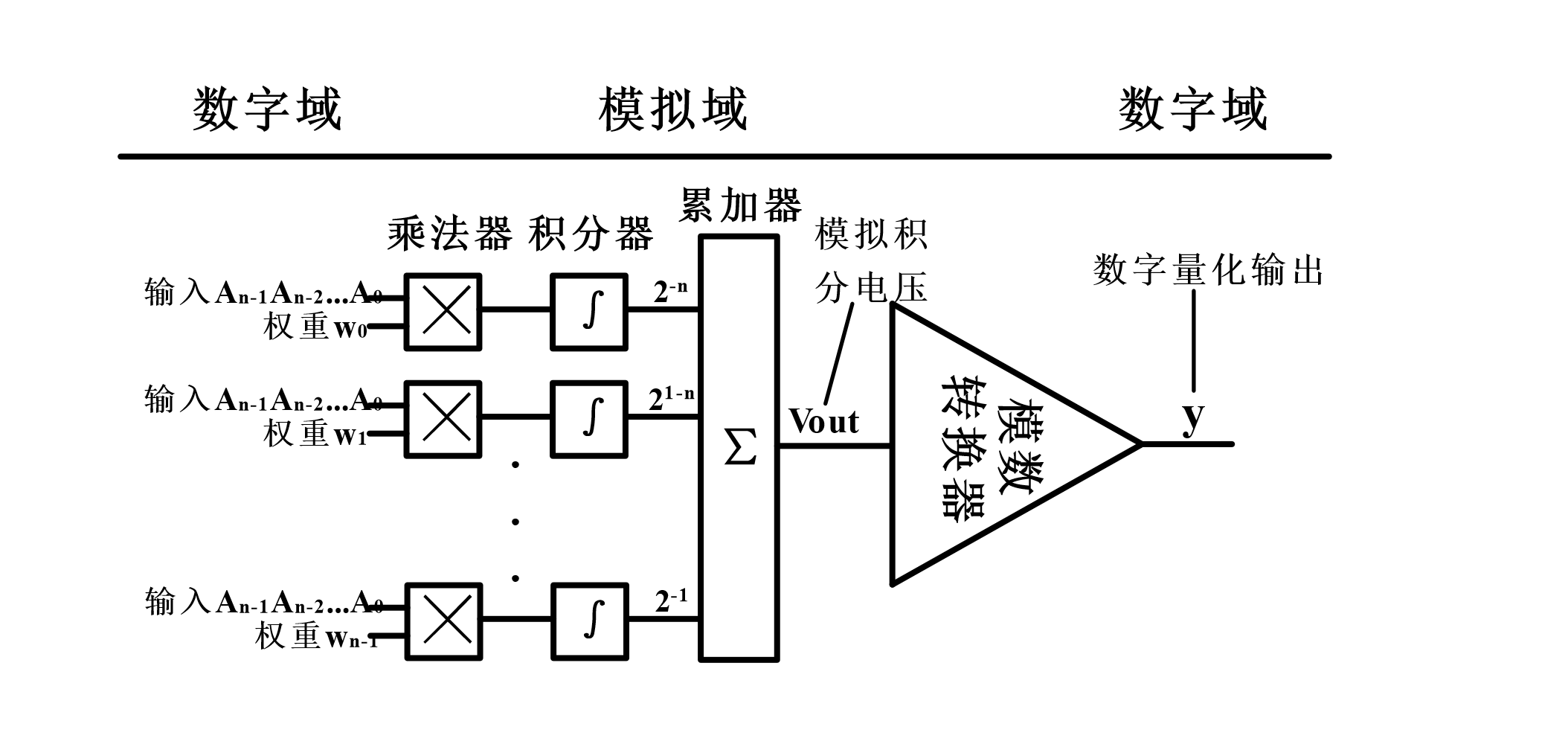


图6

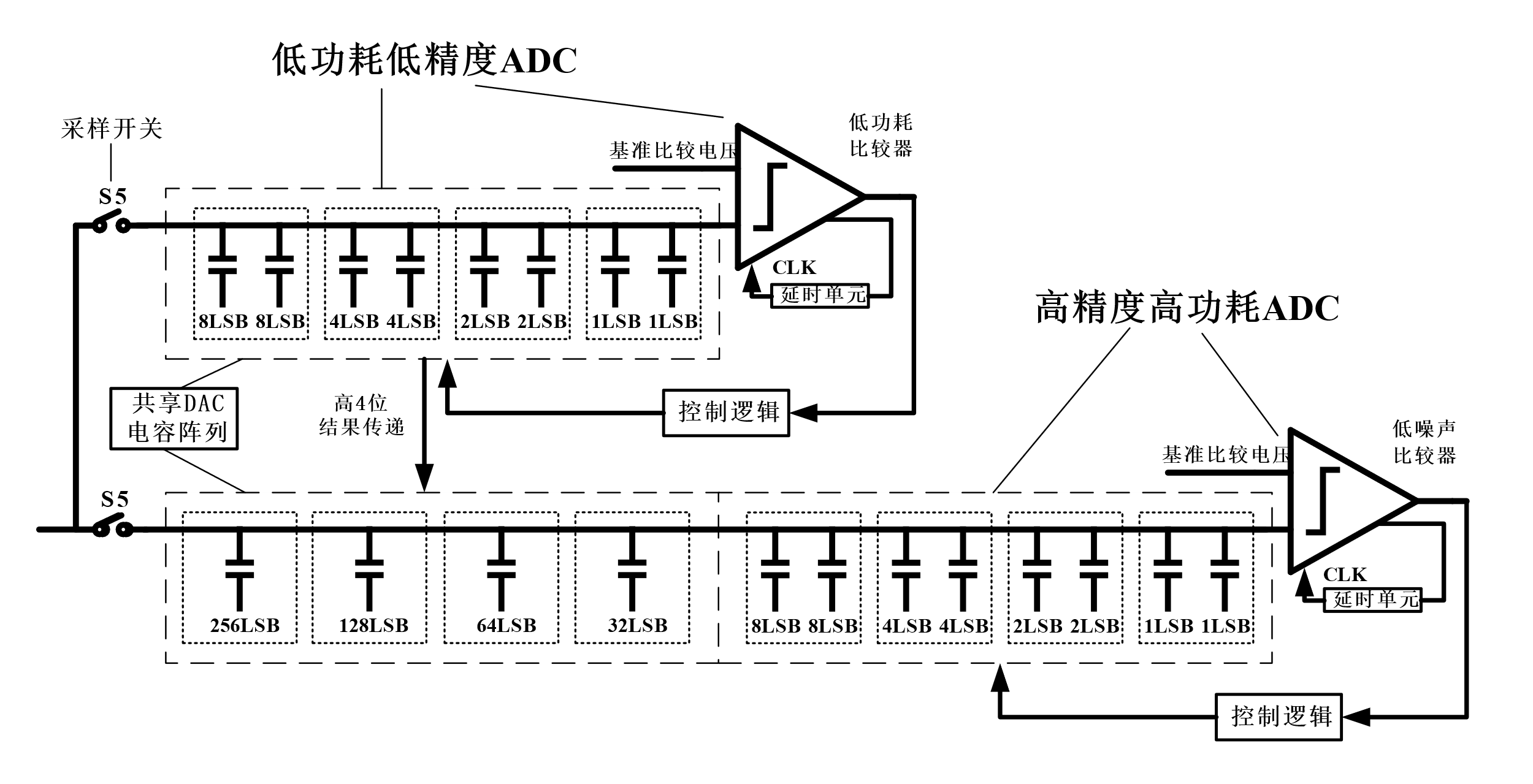


图7