(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115080501 A
(43)申请公布日 2022.09.20

(21)申请号 202210455820.4

(22)申请日 2022.04.27

(71)申请人 北京大学
地址 100871 北京市海淀区颐和园路5号

(72)发明人 王源 肖康林 崔小欣

(74)专利代理机构 北京路浩知识产权代理有限公司 11002
专利代理师 李文清

(51)Int.Cl.
G06F 15/78(2006.01)
G06F 7/544(2006.01)
G06N 3/063(2006.01)

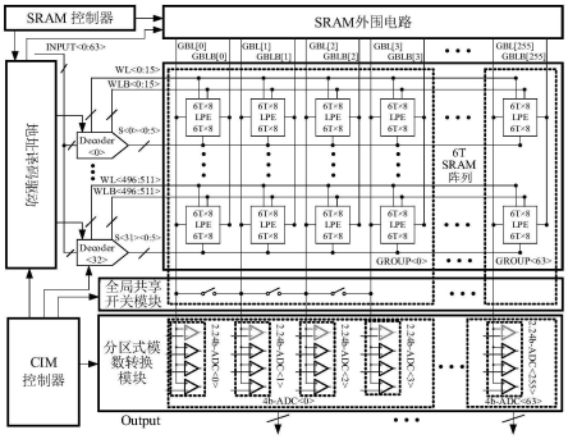
权利要求书2页 说明书16页 附图10页

(54)发明名称

基于局部电容电荷共享的SRAM存算一体芯片

(57)摘要

本发明提供一种基于局部电容电荷共享的SRAM存算一体芯片，通过译码模块确定输入数据，通过按位计算模块基于电荷共享原理，在与运算模式下实现输入数据与存储数据的乘法运算，在异或运算模式下实现输入数据与存储数据的异或运算，通过全局共享开关模块将乘法运算结果进行累加。最后通过分区式模数转换模块在与运算模式下将模拟累加结果进行量化输出，在异或运算模式下将异或运算结果进行量化输出。该芯片支持与运算模式以及异或运算模式，拓宽了应用范围。其中不存在用于接收输入数据的DAC结构，可以避免在芯片中出现多比特输入数据导致的计算的非线性和涨落现象。采用分区式模数转换模块，以分区方式减少工作比较器的数量，降低量化功耗。



CN 115080501 A

1.一种基于局部电容电荷共享的SRAM存算一体芯片,其特征在于,包括:顺次连接的译码模块、按位计算模块、全局共享开关模块以及分区式模数转换模块;

所述译码模块用于确定输入数据;

所述按位计算模块包括多个位计算单元,每个所述位计算单元均用于在与运算模式下基于局部电容的电荷共享原理,将所述输入数据与按位存储的存储数据的一位数据进行乘法运算,得到所述存储数据的一位数据对应的乘法运算结果;或者,用于在异或运算模式下基于所述局部电容的电荷共享原理,将所述输入数据与按位存储的存储数据的一位数据进行异或运算,得到所述存储数据的一位数据对应的异或运算结果;

所述全局共享开关模块用于在所述与运算模式下将所述存储数据的各位对应的乘法运算结果进行累加,得到模拟累加结果;

所述分区式模数转换模块包括多个位比较单元,所述位比较单元与所述位计算单元一一对应,所有所述位比较单元用于在所述与运算模式下将所述模拟累加结果进行量化输出;每个所述位比较单元用于在所述异或运算模式下将所述存储数据的一位数据对应的异或运算结果进行量化输出。

2.根据权利要求1所述的基于局部电容电荷共享的SRAM存算一体芯片,其特征在于,所述存储数据包括神经网络中的多个4位的权重数据;

所述位计算单元的数量包括4的倍数个,且每4个所述位计算单元组成位计算单元组,每个所述位计算单元组用于确定每个所述权重数据的各位对应的乘法运算结果或异或运算结果;

所述位比较单元包括对应位的全局位线上连接的4个比较器,所述位计算单元组对应的位比较单元组中,除最高位之外的其他位对应的所述位比较单元以及所述最高位的全局位线上的3个比较器,用于在所述与运算模式下将所述模拟累加结果进行量化输出。

3.根据权利要求2所述的基于局部电容电荷共享的SRAM存算一体芯片,其特征在于,在所述异或运算模式下,每条所述全局位线上连接的4个比较器,按序号由小到大的顺序,采用的参考电压由低至高,且分两阶段进行部分使能;

在所述与运算模式下,所述位比较单元组的各比较器按序号由小到大的顺序,采用的参考电压由低至高,且分两阶段进行部分使能。

4.根据权利要求2所述的基于局部电容电荷共享的SRAM存算一体芯片,其特征在于,所述全局共享开关模块包括一行全局共享开关阵列,所述全局共享开关阵列包括多个全局共享开关组,所述全局共享开关组与所述位计算单元组一一对应;

所述全局共享开关组中每个全局共享开关连接于对应的所述位计算单元组中相邻两个所述位计算单元之间。

5.根据权利要求1-4中任一项所述的基于局部电容电荷共享的SRAM存算一体芯片,其特征在于,所述位计算单元包括一列局部处理单元LPE阵列,所述LPE阵列包括多个LPE;

所述SRAM存算一体芯片还包括SRAM外围电路,所述多个LPE的全局位线均连接于所述SRAM外围电路。

6.根据权利要求5所述的基于局部电容电荷共享的SRAM存算一体芯片,其特征在于,每个LPE包括6T SRAM单元以及传输门开关,所述6T SRAM单元通过所述传输门开关与对应的全局位线连接;

每个LPE的局部位线之间存在寄生电容。

7.根据权利要求6所述的基于局部电容电荷共享的SRAM存算一体芯片,其特征在于,所述译码模块包括译码器阵列,所述译码器阵列中译码器与所述LPE阵列中所述6T SRAM单元一一对应连接;

所述SRAM存算一体芯片还包括SRAM控制器以及地址解码驱动,所述SRAM控制器分别与所述SRAM外围电路以及所述地址解码驱动连接。

8.根据权利要求1-3中任一项所述的基于局部电容电荷共享的SRAM存算一体芯片,其特征在于,所述SRAM存算一体芯片还包括存内计算控制器,所述存内计算控制器分别与所述译码模块、所述全局共享开关模块以及所述分区式模数转换模块连接。

9.一种通过权利要求1-8中任一项所述的基于局部电容电荷共享的SRAM存算一体芯片实现的存内计算方法,其特征在于,包括:

在与运算模式下,基于SRAM存算一体芯片中按位计算模块包括的每个位计算单元,采用局部电容的电荷共享原理,将输入数据与按位存储的存储数据的一位数据进行乘法运算,得到所述存储数据的一位数据对应的乘法运算结果,并基于所述SRAM存算一体芯片中全局共享开关模块,将所述存储数据的各位对应的乘法运算结果进行累加,得到模拟累加结果,基于所述SRAM存算一体芯片中分区式模数转换模块包括的所有位比较单元,将所述模拟累加结果进行量化输出;

在异或运算模式下,基于每个所述位计算单元,采用所述局部电容的电荷共享原理,将所述输入数据与所述存储数据的一位数据进行异或运算,得到所述存储数据的一位数据对应的异或运算结果,并基于所述分区式模数转换模块包括的每个所述位比较单元,将所述存储数据的一位数据对应的异或运算结果进行量化输出。

10.根据权利要求9所述的存内计算方法,其特征在于,还包括:

在所述异或运算模式下,从所述位比较单元中选取序号大于所述位比较单元中的第一序号中位数的第一类比较器进行第一阶段使能,并基于所述第一类比较器的输出结果,从所述位比较单元的各比较器中所述第一类比较器的一侧选取第二类比较器进行第二阶段使能;

在所述与运算模式下,从所述位比较单元组的各比较器中选取序号不等距的多个第三类比较器进行第一阶段使能,并基于所述多个第三类比较器的输出结果,从所述位比较单元组的各比较器中所述多个第三类比较器形成的区间之一选取第四类比较器进行第二阶段使能;

其中,序号大于所述位比较单元组中的第二序号中位数的第三类比较器的数量多于序号小于所述第二序号中位数的第三类比较器的数量。

基于局部电容电荷共享的SRAM存算一体芯片

技术领域

[0001] 本发明涉及集成电路设计技术领域,尤其涉及一种基于局部电容电荷共享的SRAM存算一体芯片。

背景技术

[0002] 存内计算(Compute-In-Memory,CIM)技术,是指把传统以计算为中心的架构转变为以数据为中心的架构,其直接利用存储器进行数据处理,从而把数据存储与计算融合在同一个芯片当中,即构成存算一体芯片,可以彻底消除冯诺依曼计算架构瓶颈,降低数据传输造成的额外功耗和性能损失。静态随机存取存储器(Static Random Access Memory,SRAM)因其高速、低功耗和高鲁棒性的特点,可被广泛用于构造存算一体芯片。

[0003] 目前,存算一体芯片可以作为神经网络模型的乘累加运算的硬件化实现,但是现有的存算一体芯片,在模拟域中,一般采用大的额外加权电容器阵列以及模数转换器(Analog-to-Digital Converter,ADC)进行多位权重的乘法累加(Multiply Accumulate,MAC)运算,这将造成较大的能量消耗,仅仅ADC就可以覆盖几乎40%的芯片功耗。

发明内容

[0004] 本发明提供一种基于局部电容电荷共享的SRAM存算一体芯片,用以解决现有技术中存在的缺陷。

[0005] 本发明提供一种基于局部电容电荷共享的SRAM存算一体芯片,包括:顺次连接的译码模块、按位计算模块、全局共享开关模块以及分区式模数转换模块;

[0006] 所述译码模块用于确定输入数据;

[0007] 所述按位计算模块包括多个位计算单元,每个所述位计算单元均用于在与运算模式下基于局部电容的电荷共享原理,将所述输入数据与按位存储的存储数据的一位数据进行乘法运算,得到所述存储数据的一位数据对应的乘法运算结果;或者,用于在异或运算模式下基于所述局部电容的电荷共享原理,将所述输入数据与按位存储的存储数据的一位数据进行异或运算,得到所述存储数据的一位数据对应的异或运算结果;

[0008] 所述全局共享开关模块用于在所述与运算模式下将所述存储数据的各位对应的乘法运算结果进行累加,得到模拟累加结果;

[0009] 所述分区式模数转换模块包括多个位比较单元,所述位比较单元与所述位计算单元一一对应,所有所述位比较单元用于在所述与运算模式下将所述模拟累加结果进行量化输出;每个所述位比较单元用于在所述异或运算模式下将所述存储数据的一位数据对应的异或运算结果进行量化输出。

[0010] 根据本发明提供的一种基于局部电容电荷共享的SRAM存算一体芯片,所述存储数据包括神经网络中的多个4位的权重数据;

[0011] 所述位计算单元的数量包括4的倍数个,且每4个所述位计算单元组成位计算单元组,每个所述位计算单元组用于确定每个所述权重数据的各位对应的乘法运算结果或异或

运算结果；

[0012] 所述位比较单元包括对应位的全局位线上连接的4个比较器，所述位计算单元组对应的位比较单元组中，除最高位之外的其他位对应的所述位比较单元以及所述最高位的全局位线上的3个比较器，用于在所述与运算模式下将所述模拟累加结果进行量化输出。

[0013] 根据本发明提供的一种基于局部电容电荷共享的SRAM存算一体芯片，在所述异或运算模式下，每条所述全局位线上连接的4个比较器，按序号由小到大的顺序，采用的参考电压由低至高，且分两阶段进行部分使能；

[0014] 在所述与运算模式下，所述位比较单元组的各比较器按序号由小到大的顺序，采用的参考电压由低至高，且分两阶段进行部分使能。

[0015] 根据本发明提供的一种基于局部电容电荷共享的SRAM存算一体芯片，所述全局共享开关模块包括一行全局共享开关阵列，所述全局共享开关阵列包括多个全局共享开关组，所述全局共享开关组与所述位计算单元组一一对应；

[0016] 所述全局共享开关组中每个全局共享开关连接于对应的所述位计算单元组中相邻两个所述位计算单元之间。

[0017] 根据本发明提供的一种基于局部电容电荷共享的SRAM存算一体芯片，所述位计算单元包括一列局部处理单元LPE阵列，所述LPE阵列包括多个LPE；

[0018] 所述SRAM存算一体芯片还包括SRAM外围电路，所述多个LPE的全局位线均连接于所述SRAM外围电路。

[0019] 根据本发明提供的一种基于局部电容电荷共享的SRAM存算一体芯片，每个LPE包括6T SRAM单元以及传输门开关，所述6T SRAM单元通过所述传输门开关与对应的全局位线连接；

[0020] 每个LPE的局部位线之间存在寄生电容。

[0021] 根据本发明提供的一种基于局部电容电荷共享的SRAM存算一体芯片，所述译码模块包括译码器阵列，所述译码器阵列中译码器与所述LPE阵列中所述6T SRAM单元一一对应连接；

[0022] 所述SRAM存算一体芯片还包括SRAM控制器以及地址解码驱动，所述SRAM控制器分别与所述SRAM外围电路以及所述地址解码驱动连接。

[0023] 根据本发明提供的一种基于局部电容电荷共享的SRAM存算一体芯片，所述SRAM存算一体芯片还包括存内计算控制器，所述存内计算控制器分别与所述译码模块、所述全局共享开关模块以及所述分区式模数转换模块连接。

[0024] 本发明还提供一种通过上述的基于局部电容电荷共享的SRAM存算一体芯片实现的存内计算方法，包括：

[0025] 在与运算模式下，基于SRAM存算一体芯片中按位计算模块包括的每个位计算单元，采用局部电容的电荷共享原理，将输入数据与按位存储的存储数据的一位数据进行乘法运算，得到所述存储数据的一位数据对应的乘法运算结果，并基于所述SRAM存算一体芯片中全局共享开关模块，将所述存储数据的各位对应的乘法运算结果进行累加，得到模拟累加结果，基于所述SRAM存算一体芯片中分区式模数转换模块包括的所有位比较单元，将所述模拟累加结果进行量化输出；

[0026] 在异或运算模式下，基于每个所述位计算单元，采用所述局部电容的电荷共享原

理，将所述输入数据与所述存储数据的一位数据进行异或运算，得到所述存储数据的一位数据对应的异或运算结果，并基于所述分区式模数转换模块包括的每个所述位比较单元，将所述存储数据的一位数据对应的异或运算结果进行量化输出。

[0027] 根据本发明提供的一种存内计算方法，还包括：

[0028] 在所述异或运算模式下，从所述位比较单元中选取序号大于所述位比较单元中的第一序号中位数的第一类比较器进行第一阶段使能，并基于所述第一类比较器的输出结果，从所述位比较单元的各比较器中所述第一类比较器的一侧选取第二类比较器进行第二阶段使能；

[0029] 在所述与运算模式下，从所述位比较单元组的各比较器中选取序号不等距的多个第三类比较器进行第一阶段使能，并基于所述多个第三类比较器的输出结果，从所述位比较单元组的各比较器中所述多个第三类比较器形成的区间之一选取第四类比较器进行第二阶段使能；

[0030] 其中，序号大于所述位比较单元组中的第二序号中位数的第三类比较器的数量多于序号小于所述第二序号中位数的第三类比较器的数量。

[0031] 本发明提供的基于局部电容电荷共享的SRAM存算一体芯片，包括：顺次连接的译码模块、按位计算模块、全局共享开关模块以及分区式模数转换模块，通过译码模块确定输入数据，通过按位计算模块基于局部电容的电荷共享原理，在与运算模式下实现输入数据与存储数据的乘法运算，得到乘法运算结果，在异或运算模式下实现输入数据与存储数据的异或运算，得到异或运算结果，并通过全局共享开关模块将乘法运算结果进行累加，得到模拟累加结果。最后通过分区式模数转换模块在与运算模式下将模拟累加结果进行量化输出，在异或运算模式下将异或运算结果进行量化输出。该芯片可以支持与运算模式以及异或运算模式，拓宽了应用范围。该芯片中不存在用于接收输入数据的DAC结构，可以避免在芯片中出现多比特输入导致的计算的非线性和涨落现象。该芯片采用了分区式模数转换模块，以分区方式减少其中工作的比较器的数量，降低量化功耗。

附图说明

[0032] 为了更清楚地说明本发明或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0033] 图1是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片的结构示意图之一；

[0034] 图2是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片中分区式模数转换模块中与位计算单元组对应的位比较单元组的结构示意图；

[0035] 图3是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片中LPE的结构示意图；

[0036] 图4是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片中XNOR模式以及AND11模式下LPE的配置结构图；

[0037] 图5是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片中AND21模式下LPE的配置结构图；

[0038] 图6是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片中LPE运算时存储节点的波动特性示意图；

[0039] 图7是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片的结构示意图之二；

[0040] 图8是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片在AND-1-4-4-64模式下的5个阶段的时序波形图；

[0041] 图9是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片中全局共享开关模块在实现乘法累加方案时乘法阶段结构图；

[0042] 图10是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片中全局共享开关模块在实现乘法累加方案时单位累加阶段结构图；

[0043] 图11是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片中全局共享开关模块在实现乘法累加方案时多位累加阶段结构图；

[0044] 图12是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片在XNOR模式下从理想MAC运算结果到模拟电压的传递函数示意图；

[0045] 图13是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片在AND模式下从理想MAC运算结果到模拟电压的传递函数示意图；

[0046] 图14是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片在MAC运算中模拟电压在1.8mV左右的平均涨落示意图；

[0047] 图15是本发明提供的基于局部电容电荷共享的SRAM存算一体芯片的ADC的传递函数和INL示意图；

[0048] 图16是本发明提供的存内计算方法的流程示意图。

具体实施方式

[0049] 为使本发明的目的、技术方案和优点更加清楚，下面将结合本发明中的附图，对本发明中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0050] 深度卷积神经网络(Deep Convolution Neural Network，DCNN)在广泛的应用中实现了前所未有的增长，包括云计算和边缘计算中的人工智能(Artificial Intelligence，AI)。乘法累加(Multiply Accumulate，MAC)运算具有高规律性和并行性，并在功率和延迟方面占主导地位，是DCNN中最常见的操作。对于大多数应用，AI边缘芯片需要具有多位和多模式MAC运算的功能。由于冯·诺依曼瓶颈，传统的基于冯·诺依曼架构的数字AI边缘处理器在运行MAC运算时容易出现过多的能耗和延迟。

[0051] 存内计算(Compute-In-Memory，CIM)是一种有吸引力的方法，可以通过启用并行计算、减少内存访问和中间数据来提高MAC运算的能量效率，从而消除冯·诺依曼瓶颈。

[0052] 早期基于静态随机存取存储器(Static Random Access Memory，SRAM)的CIM结构是基于6T单元在电流域中计算的，这会导致由严重的电流涨落引起的潜在写入干扰和错误。然后出现了具有更大动态传感范围的时域计算和更稳健的电荷域计算的CIM结构。多位输入的CIM结构通常通过数模转换器(Digital-to-Analog Converter，DAC)实现，这会导致

输入出现非线性和涨落。并且在模拟域中，一般采用大的额外加权电容器阵列进行多位权重累加，造成较大的能量和面积消耗。用于MAC运算的模数转换器(Analog-to-Digital Converter，ADC)覆盖了几乎40％的芯片功耗。注意到DCNN中低MAC运算结果的比例很高，自适应的低MAC运算结果感知ADC方案和优先混合ADC来提高能效的方案应运而生，二者均基于逐次逼近型ADC和单斜率ADC，需要进行大量比较。

[0053] 总而言之，现有的CIM结构仍然存在几个关键问题：(1)执行MAC运算时的写入干扰；(2)多位输入的DAC的非线性和涨落；(3)多位权重累积中过多的面积开销；(4)过多的ADC功耗和面积成本。基于此，本发明实施例中提供了一种基于局部电容电荷共享的SRAM存算一体芯片。

[0054] 图1为本发明实施例中提供的一种基于局部电容电荷共享的SRAM存算一体芯片的结构示意图，如图1所示，该芯片包括：顺次连接的译码模块1、按位计算模块2、全局共享开关模块3以及分区式模数转换模块4；

[0055] 所述译码模块1用于确定输入数据；

[0056] 所述按位计算模块2包括多个位计算单元21，每个所述位计算单元21均用于在与运算模式下基于局部电容的电荷共享原理，将所述输入数据与按位存储的存储数据的一位数据进行乘法运算，得到所述存储数据的一位数据对应的乘法运算结果；或者，用于在异或运算模式下基于所述局部电容的电荷共享原理，将所述输入数据与按位存储的存储数据的一位数据进行异或运算，得到所述存储数据的一位数据对应的异或运算结果；

[0057] 所述全局共享开关模块3用于在所述与运算模式下将所述存储数据的各位对应的乘法运算结果进行累加，得到模拟累加结果；

[0058] 所述分区式模数转换模块4包括多个位比较单元41，所述位比较单元41与所述位计算单元21一一对应，所有所述位比较单元41用于在所述与运算模式下将所述模拟累加结果进行量化输出；每个所述位比较单元41用于在所述异或运算模式下将所述存储数据的一位数据对应的异或运算结果进行量化输出。

[0059] 具体地，本发明实施例中提供的基于局部电容电荷共享的SRAM存算一体芯片，采用译码模块1实现输入数据的接收和确定。该译码模块1可以包括译码器(Decoder)阵列，译码器阵列中可以包括多个译码器，译码器阵列中译码器的数量可以根据需要进行设定，例如可以是32，则各译码器可以分布表示为Decoder<0>、Decoder<1>、…、Decoder<31>。

[0060] 输入数据的数量可以根据需要进行设定，例如可以是64，输入数据可以表示为INPUT<0:63>，每个输入数据的位数可以是2位。每个输入数据均分别输入至各译码器中，通过各译码器可以得到用于输入至每个位计算单元的译码结果。

[0061] 按位计算模块2可以包括多个位计算单元21，按位计算模块2中位计算单元21的数量可以根据需要进行设置，例如可以设置为256个，每个位计算单元21可以包括与译码器阵列中译码器数量相同的核心计算单元，译码器与核心计算单元一一对应连接。在每个核心计算单元内均可以存储有单个存储数据的对应位数据。每个核心计算单元可以包括两个核心计算子单元，每个核心计算子单元中也存储有单个存储数据的对应位数据。每个核心计算子单元可以对应于单个输入数据的一位数据计算。

[0062] 每个核心计算单元还可以包括用于控制两个核心计算子单元工作状态的传输门开关，各传输门开关可以为CMOS。传输门开关的数量可以根据需要进行设置，例如可以为7

个,此处不作具体限定。各传输门开关的工作状态可以相互独立,也可以存在工作状态相同的传输门开关,各传输门开关可以连接于两个计算子单元之间。

[0063] 按位计算模块2中4个相邻的位计算单元可以共同实现所有输入数据与按位存储的一个4b的存储数据的计算。

[0064] 按位计算模块2的工作模式可以包括与运算(AND)模式以及异或运算(XNOR)模式,与运算模式可以包括第一与运算(AND11)模式以及第二与运算(AND21)模式,第一与运算模式下,每个位计算单元21可以实现各输入数据的1b数据与单个存储数据的1b数据的乘法运算,此时每个位计算单元21中各核心计算单元中只有一个核心计算子单元工作;第二与运算模式下,每个位计算单元21可以实现各2b的输入数据与单个存储数据的1b数据的乘法运算,此时每个位计算单元21中各核心计算单元中两个核心计算子单元共同工作;异或运算模式下,每个位计算单元21可以实现各输入数据的1b数据与单个存储数据的1b数据的异或运算,此时每个位计算单元21中各计算单元中只有一个计算子单元工作。

[0065] 每个核心计算单元中存在局部电容,该局部电容是每个核心计算单元的左列局部位线与右列局部位线之间的等效电容,该局部电容可以是其中两个计算子单元的共享电容。通过局部电容的电荷共享原理,每个核心计算单元可以实现连接的译码器输出的译码结果与按位存储的单个存储数据的1b数据进行相应运算。

[0066] 也就是说,每个位计算单元21在与运算模式下基于局部电容的电荷共享原理,将输入数据与按位存储的存储数据的一位数据进行乘法运算,得到的存储数据的一位数据对应的乘法运算结果可以包括两个1b数据的乘法运算结果以及一个1b数据与一个2b数据的乘法运算结果。每个位计算单元21在异或运算模式下基于局部电容的电荷共享原理,将输入数据与按位存储的存储数据的一位数据进行异或运算,得到的存储数据的一位数据对应的异或运算结果可以包括两个1b数据的异或运算结果。

[0067] 芯片中全局共享开关模块3可以包括一行全局共享开关阵列,该全局共享开关阵列可以包括多个全局共享开关,每个全局共享开关可以连接于按位计算模块2中相邻两个位计算单元21之间。

[0068] 每个全局共享开关可以结合连接的位计算单元中的传输门开关,以实现与运算模式下将单个存储数据的一位数据对应的乘法运算结果进行累加,4个相邻的全局共享开关可以实现与运算模式下将单个存储数据的各位数据对应的乘法运算结果进行累加,256个相邻的全局共享开关共可以实现与运算模式下将64个存储数据的各位数据对应的乘法运算结果进行累加。最后,全局共享开关模块3可以得到并输出累加结果,该累加结果为模拟累加结果。

[0069] 芯片中分区式模数转换模块4可以包括多个位比较单元41,每个位比较单元41可以连接于按位计算模块2中每个位计算单元21的全局位线上,即位比较单元41与位计算单元21一一对应。该全局位线可以是左列全局位线,也可以是右列全局位线,此处不作具体限定。

[0070] 每4个位比较单元41可以在与运算模式下将每个存储数据的各位数据对应的模拟累加结果进行量化输出;每个位比较单元41可以在异或运算模式下将每个存储数据的一位数据对应的异或运算结果进行量化输出。

[0071] 每个位比较单元41可以包括多个比较器,分区式模数转换模块4的工作原理是:在

与运算模式下,从每个位比较单元41的多个比较器中选取部分比较器进行两个阶段使能,且两个阶段使能的比较器也只是每个位比较单元41中包含的所有比较器的一部分,而非所有比较器;在异或运算模式下,从每4个位比较单元41的多个比较器中选取部分比较器进行两个阶段使能,且两个阶段使能的比较器也只是每4个位比较单元41中包含的所有比较器的一部分,而非所有比较器。每个位比较单元41中包括的比较器均可以是强臂比较器。

[0072] 本发明实施例中提供的基于局部电容电荷共享的SRAM存算一体芯片,包括:顺次连接的译码模块、按位计算模块、全局共享开关模块以及分区式模数转换模块,通过译码模块确定输入数据,通过按位计算模块基于局部电容的电荷共享原理,在与运算模式下实现输入数据与存储数据的乘法运算,得到乘法运算结果,在异或运算模式下实现输入数据与存储数据的异或运算,得到异或运算结果,并通过全局共享开关模块将乘法运算结果进行累加,得到模拟累加结果。最后通过分区式模数转换模块在与运算模式下将模拟累加结果进行量化输出,在异或运算模式下将异或运算结果进行量化输出。该芯片可以支持与运算模式以及异或运算模式,拓宽了应用范围。该芯片中不存在用于接收输入数据的DAC结构,可以避免在芯片中出现计算的非线性和涨落现象。该芯片采用了分区式模数转换模块,以分区方式减少其中工作的比较器的数量,降低量化时耗和功耗。

[0073] 在上述实施例的基础上,本发明实施例中提供的基于局部电容电荷共享的SRAM存算一体芯片,所述存储数据包括神经网络中的多个4位的权重数据;

[0074] 所述位计算单元的数量包括4的倍数个,且每4个所述位计算单元组成位计算单元组,每个所述位计算单元组用于确定每个所述权重数据的各位对应的乘法运算结果或异或运算结果;

[0075] 所述位比较单元包括对应位的全局位线上连接的4个比较器,所述位计算单元组对应的位比较单元组中,除最高位之外的其他位对应的所述位比较单元以及所述最高位的全局位线上的3个比较器,用于在所述与运算模式下将所述模拟累加结果进行量化输出。

[0076] 具体地,本发明实施例中,神经网络中通常包括大量的4位的权重数据,均可以将其作为存储数据按位存储至芯片中的位计算单元,并将其与神经网络的输入数据进行乘法累加运算或异或运算。

[0077] 该芯片中,位计算单元的数量包括4的倍数个,例如包括256个,每4个相邻的位计算单元组成位计算单元组,每个位计算单元组用于确定每个权重数据的各位数据对应的乘法运算结果或异或运算结果。

[0078] 图2为芯片中分区式模数转换模块4中与位计算单元组对应的位比较单元组的结构示意图,该位比较单元组包括4个位比较单元41,每个位比较单元41均包括对应位的全局位线上连接的4个比较器。每个位比较单元41包括的4个比较器可以连接在左列全局位线GBL上,也可以连接在右列全局位线GBLB上,此处不做具体限定。图2中以4个比较器连接在GBL上为例,则每个位比较单元组中4个位比较单元41对应位的左列全局位线GBL分别表示为GBL<0>、GBL<1>、GBL<2>、GBL<3>,GBL<0>对应最低位,GBL<3>对应最高位。

[0079] 图2中S_GSHARE分别为全局共享开关模块中连接于位计算单元组中相邻两个位计算单元之间的全局共享开关,ADC_EN为使能信号。图2中各比较器的反相端上侧的VREF<i>表示为异或运算模式下每个位比较单元中第i个比较器的参考电压,0≤i≤3;各比较器的反相端下侧的VREF<j>表示为与运算模式下4个位比较单元中第j个比较器的参考电压,0≤

j≤14。

[0080] 从图2可以看出，在异或运算模式下，每个位比较单元中各比较器按与全局共享开关模块的距离由近到远的顺序进行标号，得到各比较器的序号分别为Comp<i>；在与运算模式下，如图2中42所示，涉及的位比较单元组中各比较器按逐行递增的方式进行标号，得到各比较器的序号分别为Comp<j>。由于第4个位比较单元中第4个比较器并未应用于与运算模式下，因此并非对其标号。

[0081] 在异或运算模式下，每个位比较单元中各比较器可以将输入的模拟累加结果量化为2.24位，此时该位比较单元可以构成2.24b-ADC<i>；在与运算模式下，4个位比较单元中各比较器可以将输入的模拟累加结果量化为4位，此时4个位比较单元可以构成4b-ADC<k>，0≤k≤K-1，K为位计算单元组的数量，可以为64。

[0082] 本发明实施例中，分区式模数转换模块可以在两种运算模式下实现不同尺度的量化，保证了不同运算模式下运算结果的量化输出，使芯片可以适用于不同应用需求。

[0083] 在上述实施例的基础上，本发明实施例中提供的基于局部电容电荷共享的SRAM存算一体芯片，在所述异或运算模式下，每条所述全局位线上连接的4个比较器，按序号由小到大的顺序，采用的参考电压由低至高，且分两阶段进行部分使能；

[0084] 在所述与运算模式下，所述位比较单元组的各比较器按序号由小到大的顺序，采用的参考电压由低至高，且分两阶段进行部分使能。

[0085] 具体地，本发明实施例中，如图2所示，在异或运算模式下，每条左列全局位线GBL<i>上连接的4个比较器，序号可以通过各比较器与全局共享开关模块的距离的远近进行标号得到，距离远序号大，距离近序号小。在与运算模式下，位比较单元组中各比较器的序号可以通过按逐行递增的方式进行标号得到，每一行中序号从最低位向最高位依次增大，每一列中与全局共享开关模块的距离由近到远序号依次增大。

[0086] 进而，每个位比较单元的4个比较器中，按序号由小到大的顺序，采用的参考电压由低至高，即在异或运算模式下，VREF<0>、VREF<1>、VREF<2>、VREF<3>依次升高；在与运算模式下，VREF<0>-VREF<14>依次升高。

[0087] 此外，在每种运算模式下，涉及的各比较器均分两阶段进行部分使能。ADC_EN可以包括ADC_EN<0>和ADC_EN<1>，分别表示第一阶段使能信号以及第二阶段使能信号。

[0088] 第一阶段使能可以粗略确定工作的第一类比较器，第二阶段使能时在第一阶段使能的基础上，基于第一类比较器的输出结果精确确定工作的比较器。

[0089] 在异或运算模式下，第一阶段使能策略可以是从位比较单元中选取序号大于位比较单元中的第一序号中位数的第一类比较器进行第一阶段使能，由于位比较单元中共包括4个比较器，序号分别为0-3，则第一序号中位数为1.5，此时可以选取第三个比较器Comp<2>作为第一类比较器进行第一阶段使能，使之工作。第一类比较器的输出结果可以是0或1，即out<2>＝0或1。

[0090] 在异或运算模式下，第二阶段使能策略可以是根据out<2>，从位比较单元的各比较器中第一类比较器的一侧选取第二类比较器进行第二阶段使能。当out<2>＝0时，可以选取序号大于第一类比较器的比较器作为第二类比较器进行第二阶段使能，即选取Comp<3>作为第二类比较器进行第二阶段使能；当out<2>＝1时，可以选取序号小于第一类比较器的比较器作为第二类比较器进行第二阶段使能，即选取Comp<0>和Comp<1>作为第二类比较器

进行第二阶段使能。此后，第二类比较器的输出结果即作为异或运算模式下位比较单元的输出结果。

[0091] 在与运算模式下，第一阶段使能策略可以是从位比较单元组的各比较器中选取序号不等距的多个第三类比较器进行第一阶段使能。序号不等距是指任意相邻两个第三类比较器的序号间距不同，也可以是序号越大的相邻两个第三类比较器的序号间距越小。由于位比较单元组中共包括15个比较器，序号分别为0-14，则可以选取第5个比较器Comp<5>、第10个比较器Comp<9>以及第13个比较器Comp<12>作为第三类比较器进行第一阶段使能，使之工作。各第三类比较器的输出结果按序号由大到小的顺序组合可以得到000、100、110或111，即out＝000、100、110或111。

[0092] 在与运算模式下，第二阶段使能策略可以是根据out，从位比较单元组的各比较器中多个第三类比较器形成的区间之一选取第四类比较器进行第二阶段使能。

[0093] 各第三类比较器形成的区间包括4个，分别为序号大于12的第一区间、序号在9和12之间的第二区间、序号在4和9之间的第三区间以及序号在小于4的第四区间。进而，当out＝000时，可以选取第一区间中的Comp<13>、Comp<14>作为第四类比较器进行第二阶段使能；当out＝100时，可以选取第二区间中的Comp<10>、Comp<11>作为第四类比较器进行第二阶段使能；当out＝110时，可以选取第三区间中的Comp<5>-Comp<8>作为第四类比较器进行第二阶段使能；当out＝111时，可以选取第四区间中的Comp<0>-Comp<3>作为第四类比较器进行第二阶段使能。此后，第四类比较器的输出结果即作为与运算模式下位比较单元的输出结果。

[0094] 在与运算模式下，序号大于位比较单元组中的第二序号中位数的第三类比较器的数量多于序号小于第二序号中位数的第三类比较器的数量。例如，第二序号中位数为7，序号小于7的第三类比较器有1个，而序号大于7的第三类比较器有2个。由于模拟累加结果通过比较器的正相端的电压进行表征，正相端的电压越大，对应的模拟累加结果越小，因此较密集的选取序号大的比较器作为第三类比较器，可以更快速的准确确定出工作的第四类比较器进行第二阶段使能，实现小值译码优先，此时该分区式模数转换模块为基于子分区小值译码优先的2.24b/4b ADC模块。该分区式模数转换模块可以适用于计算结果均比较小的神经网络，可以提高量化能量效率。

[0095] 在上述实施例的基础上，本发明实施例中提供的基于局部电容电荷共享的SRAM存算一体芯片，所述全局共享开关模块包括一行全局共享开关阵列，所述全局共享开关阵列包括多个全局共享开关组，所述全局共享开关组与所述位计算单元组一一对应；

[0096] 所述全局共享开关组中每个全局共享开关连接于对应的所述位计算单元组中相邻两个所述位计算单元之间。

[0097] 具体地，本发明实施例中，位计算单元组可以包括4个相邻的位计算单元，故每个全局共享开关组中可以包括3个全局共享开关。由于每个位计算单元均与对应位的左列全局位线GBL以及右列全局位线GBLB连接，因此每个全局共享开关可以连接于2个相邻的位计算单元的左列全局位线GBL之间。如此可以顺利实现每个位计算单元的乘法运算结果的累加。

[0098] 在上述实施例的基础上，本发明实施例中提供的基于局部电容电荷共享的SRAM存算一体芯片，所述位计算单元包括一列寄生参数提取LPE阵列，所述LPE阵列包括多个LPE；

[0099] 所述SRAM存算一体芯片还包括SRAM外围电路,所述多个LPE的全局位线均连接于所述SRAM外围电路。

[0100] 具体地,本发明实施例中,每个位计算单元均可以包括一列局部处理单元(Local Process Element,LPE)阵列,每个LPE阵列包括多个LPE。每个LPE均为一个核心计算单元,在每个核心计算单元内均可以存储有单个存储数据的对应位数据。

[0101] 该SRAM存算一体芯片还可以包括SRAM外围电路(SRAM Peripheral circuits),该SRAM外围电路可以通过左列全局位线GBL以及右列全局位线GBLB与每个LPE连接,以实现对每个LPE的驱动及控制。

[0102] 本发明实施例中,采用LPE阵列作为位计算单元,实现存储数据的一位数据与输入数据的乘法运算或异或运算,为芯片的两种工作模式提供基础,提高了芯片性能。

[0103] 在上述实施例的基础上,本发明实施例中提供的基于局部电容电荷共享的SRAM存算一体芯片,每个LPE包括6T SRAM单元以及传输门开关,所述6T SRAM单元通过所述传输门开关与对应的全局位线连接;

[0104] 每个LPE的局部位线之间存在寄生电容。

[0105] 具体地,本发明实施例中,如图3所示,每个LPE均包括6T SRAM单元以及传输门开关,6T SRAM单元以及传输门开关的数量可以根据需要进行设置,例如6T SRAM单元的数量有16个,传输门开关的数量有7个,8个6T SRAM单元构成一个核心计算子单元,两个核心计算子单元之间通过7个传输门开关连接。图3中传输门开关分别为2个S_LEFT、1个S_GLEFT、1个S_XNOR_U、1个S_XNOR_D以及2个S_GRIGHT。1个S_GLEFT与左列全局位线GBL连接,2个S_GRIGHT均与右列全局位线GBLB连接。7个传输门开关上侧的8个6T SRAM单元6T<0>-6T<7>构成第一核心计算子单元,第一核心计算子单元左右分别连接有左列局部位线LBL_U和右列局部位线LBLB_U,左列局部位线LBL_U和右列局部位线LBLB_U均存在固有位线电容C,因此在左列局部位线LBL_U和右列局部位线LBLB_U之间形成2C的局部电容。左行字线WL<0:7>以及右行字线WLB<0:7>均分别与第一核心计算子单元中每个6T SRAM单元连接。

[0106] 同理,7个传输门开关下侧的8个6T SRAM单元6T<8>-6T<15>构成第二核心计算子单元,第二核心计算子单元左右分别连接有左列局部位线LBL_D和右列局部位线LBLB_D,左列局部位线LBL_D和右列局部位线LBLB_D均存在固有位线电容C,因此在左列局部位线LBL_D和右列局部位线LBLB_D之间形成2C的局部电容。左行字线WL<8:15>以及右行字线WLB<8:15>均分别与第二核心计算子单元中每个6T SRAM单元连接。

[0107] 图3中,LPE支持三种工作模式:XNOR、AND11和AND21。AND11模式旨在实现1-b输入数据与1-b存储数据的乘法运算,而AND21模式旨在实现2-b输入数据与1-b存储数据的乘法运算。通常,在每种模式下,LBL_U/D和LBLB_U/D均先预充电到VDD。S_XNOR_U、S_XNOR_D和S_LEFT由不同的模式和输入数据控制,形成不同值和可能的放电路径的局部上限。然后,根据输入数据和地址激活8/16个WL/WLB之一。最后,根据储存数据,形成的局部电容将被排放到地面或保留其电荷,这代表了乘法运算结果。

[0108] XNOR模式以及AND11模式下LPE的配置结构图如图4所示。在XNOR模式下,S_XNOR_U或S_XNOR_D开启,形成一个局部电容LBL+LBLB和两个可能的放电路径(从电容到Q或QB)用于异或运算。在AND11模式下,仅采用LBL+LBLB和WL。当且仅当结果为1时,LBL+LBLB正在对地放电。

[0109] XNOR模式下,输入数据(input,IN)为±1,存储数据为±1,S_XNOR_U/D＝1时,LBL与LBLB连接,LBL+LBLB被预充电到VDD,该模式下LPE的真值表如表1所示。

[0110] AND11模式下,输入(input,IN)为0/1,存储数据为0/1,S_XNOR_U/D＝1时,LBL与LBLB连接,LBL+LBLB被预充电到VDD,该模式下LPE的真值表如表2所示。

[0111] 表1 XNOR模式下LPE的真值表

[0112]

IN	WL	WLB	W	Q	QB	Result	LBL+LBLB
-1	GND	VDD	-1	VDD	GND	+1	GND
+1	VDD	GND	-1	VDD	GND	-1	VDD
-1	GND	VDD	+1	GND	VDD	-1	VDD
+1	VDD	GND	+1	GND	VDD	+1	GND

[0113] 表2 AND11模式下LPE的真值表

[0114]

IN	WL	WLB	W	Q	QB	Result	LBL+LBLB
0	GND	idle	0	VDD	GND	0	VDD
1	VDD	idle	0	VDD	GND	0	VDD
0	GND	idle	1	GND	VDD	0	VDD
1	VDD	idle	1	GND	VDD	1	GND

[0115] 在AND21模式下,LPE的配置结构图如图5所示,除了输入数据＝0的情况下,WL被激活。当输入数据为1、2和3时,局部放电电容将设置为LBL_U/LBL_D、LBL_U+LBL_D和LBL_U+LBLB_U+LBL_D,分别显示用于多位输入实现的模式不敏感的2b-DAC功能。同时,考虑到最坏的放电条件,100k点蒙特卡罗(MC)模拟表明,对低存储节点的破坏很小,避免了写入干扰,如图6所示。图6中圈出的部分为100k点蒙特卡罗(MC)出现的小波动,最大的局部电容可以是6fF。

[0116] AND21模式下,2b输入数据(input,IN＜1:0＞)为0/1/2/3,1b存储数据为0/1,LBL_U+LBLB_U+LBL_D被预充电到VDD,input来自于WL_U/WL_D,结束于局部电容共享,此时,S_XNOR_U＝1,S_LEFT＝1。该模式下LPE的真值表如表3所示。

[0117] 表3 AND21模式下LPE的真值表

[0118]

IN <1:0>	WL	S_XNOR_U	S_LEFT	W	Q	QB	Result	LBL_U+LBLB_U+LBL_D
00	GND	1	1	0	VDD	GND	0	VDD
01	VDD	0	0	0	VDD	GND	0	VDD
10	VDD	1	0	0	VDD	GND	0	VDD
11	VDD	1	1	0	VDD	GND	0	VDD
00	GND	1	1	1	GND	VDD	0	VDD
01	VDD	0	0	1	GND	VDD	1	2/3 VDD
10	VDD	1	0	1	GND	VDD	2	1/3 VDD
11	VDD	1	1	1	GND	VDD	3	GND

[0119] 在上述实施例的基础上,本发明实施例中提供的基于局部电容电荷共享的SRAM存

算一体芯片,所述译码模块包括译码器阵列,所述译码器阵列中译码器与所述LPE阵列中所述6T SRAM单元一一对应连接;

[0120] 所述SRAM存算一体芯片还包括SRAM控制器以及地址解码驱动,所述SRAM控制器分别与所述SRAM外围电路以及所述地址解码驱动连接。

[0121] 具体地,本发明实施例中,译码模块包括译码器阵列,译码器阵列可以包括多个译码器,每个译码器与LPE阵列中每个6T SRAM单元一一对应连接。

[0122] SRAM存算一体芯片还包括SRAM控制器(SRAM Controller)以及地址解码驱动(Address Decoder&Driver),SRAM控制器分别与SRAM外围电路以及地址解码驱动连接,共同保证存储数据可以按位存储至每个LPE中。

[0123] 本发明实施例中,通过SRAM控制器可以实现芯片存储功能的自动化实现。

[0124] 在上述实施例的基础上,本发明实施例中提供的基于局部电容电荷共享的SRAM存算一体芯片,所述SRAM存算一体芯片还包括存内计算控制器(CIM Controller),存内计算控制器分别与译码模块、全局共享开关模块以及分区式模数转换模块连接。通过存内计算控制器,可以实现对芯片的计算功能的全局控制。

[0125] 图7为本发明实施例中提供的基于局部电容电荷共享的SRAM存算一体芯片的完整结构示意图,该芯片可以实现神经网络中与运算模式以及异或运算模式共存的需求。该芯片包括512×256的LPE阵列、SRAM控制器、SRAM外围电路、地址译码器驱动、CIM控制器、32个译码器、包含有1×192个全局共享开关的全局共享开关模块、包含有1×64个4b-ADC的分区式模数转换模块。LPE阵列的大小为512×256。LPE阵列的256列一共分为64个组,每个组包含4列,用于存放一个4b的权重数据。

[0126] 每列由32个LPE组成,每个LPE包括16个6T SRAM单元、7个传输门开关和一个用于SRAM和CIM运算的固有位线电容。在与运算模式下,64个输入数据通过译码器逐行馈送到每一列LPE,然后在电荷域中与LPE中的权重数据相乘。单位权重数据累加由32个LPE之间的局部电容共享来执行。使用乘法累加方案,最终的4-b权重数据乘法累积是通过每个组内的局部电容共享来完成的。最后,模拟电压由分区式模数转换模块量化。在XNOR模式下,乘法累加方案被跳过,最终结果被量化为2.24位。

[0127] 该芯片共可以支持64个4b的输入数据,可以表示为INPUT<0:63>。输入数据经对应的译码器后得到的译码结果经左行字线WL<0:15>输入至各LPE,右行字线WLB<0:15>闲置。同时,译码器还可以接收CIM控制器对各LPE内传输门开关的控制信号,并将控制信号S<0><0:5>输入至各LPE。

[0128] 该芯片中,每4个相邻的LPE阵列以及每两个LPE之间连接的全局共享开关可以构成GROUP,该芯片共可以包含有64个GROUP。最终经分区式模数转换模块得到输出结果(Output)。

[0129] 在上述实施例的基础上,本发明实施例中提供的基于局部电容电荷共享的SRAM存算一体芯片,采用的译码模块结构不涉及DAC输入模块,如图8所示,以AND-1-4-4-64模式(1-b输入数据,4-b权重数据,4-b输出数据,64并行度)的5个阶段的计算波形为例。其中,CLK为时钟信号,PE为预充电信号,WL_EN为WL使能信号,L_SHARE为局部累加信号,G_SHARE为全局累加信号。

[0130] LBL+LBLB预充电至VDD。当WL被激活时,LBL+LBLB将被释放到地或保持它们的电

荷。然后所有按列的LBL+LBLB将共享电荷。列共享后，乘法累加方案将被激活以进行全局共享，产生一个由分区式模数转换模块读出的平衡电压。7个输出门开关和WL/WLB的控制信号由译码器产生。在每个全局时序阶段，输出阵列信号和输入(模式、数据和地址)之间存在清晰的逻辑关系。因此，与传统的DAC输入模块不同，无DAC输入模块是一种设计友好的组合逻辑模块，没有非线性和PVT涨落问题。

[0131] 图8中，81为预充电阶段，82为乘法阶段，83为逐列累加阶段，84为多列累加阶段，85为分区式模数转换模块量化输出阶段。

[0132] 在上述实施例的基础上，本发明实施例中提供的基于局部电容电荷共享的SRAM存算一体芯片，所述全局共享开关模块在实现乘法累加方案时，需结合6T SRAM单元中的传输门开关，同一行中的四个6T SRAM单元组合为一个4-b权重数据。

[0133] 在乘法阶段，如图9所示，所有传输门开关以及所有全局共享开关都断开以进行乘法运算。

[0134] 在单位累加阶段，如图10所示，所有传输门开关均闭合，所有全局共享开关都断开以进行单位累加运算。

[0135] 在多位累加阶段，如图11所示，部分传输门开关断开，所有全局共享开关都闭合以进行4位累加运算。

[0136] 图9-图11中，S0<i>、S1<i>、S2<i>、S3<i>均表示6T SRAM单元中的S_GLEFT。

[0137] 在乘法和单位权重累加后，四个GBL通过全局共享开关连接。通过控制全局共享开关阵列，从LSB到MSB，相应的GBL以1:2:4:8的电容比连接到存储阵列内的电容。电荷共享发生在选定的电容之间，平均电压，这代表了4-b权重数据累积的最终结果。该芯片采用了内存中的上限和LPE的控制驱动程序。因此，与在存储器外增加一个加权电容阵列的传统方法相比，乘法累加方案消除了额外和过度的补偿电容，从而节省了面积和额外的充放电能量。

[0138] 基于以上内容，本发明实施例中在TSMC 28纳米标准CMOS工艺下前仿真和部分后仿真实现了该芯片。6T SRAM单元是基于逻辑规则设计。为了评估本发明实施例中提供的芯片，在不同模式和电源下仿真计算线性度和涨落、能耗组成和能量效率的性能。此外，基于从模拟中提取的电路级非理想性，对VGG-like CNN模型和CIFAR-10数据集进行了行为模拟，以评估由于非理想性导致的分类精度损失。

[0139] 图12是XNOR模式下从理想MAC运算结果到模拟电压的传递函数示意图，图13是AND模式下从理想MAC运算结果到模拟电压的传递函数示意图。扫描MAC运算结果并评估模拟输出，传递函数均具有大于0.999的拟合优度。由于GBL和ADC输入电容上的金属线电容不可忽略，因此动态范围损失了150mV左右。图12中传递函数方程为$y=-5.4272x+558.59$，$R^2=0.9998$；图13中传递函数方程为$y=-0.5037x+890.69$，$R^2=0.9997$。图12和图13中传递函数的纵坐标范围是分区式模数转换模块的量化范围。图12和图13中横坐标为MAC运算结果(MAC value)，纵坐标为平均模拟电压(average output)，单位为mV。

[0140] 图14显示了MAC运算中模拟电压在1.8mV左右的平均涨落。传递函数在动态范围内表现出良好的线性度，混合信号计算的可变性很小。ADC的传递函数和INL如图15所示。传递函数和INL的平均涨落标准差为8.1mV。ADC设计在目标动态范围上表现出良好的线性度。

[0141] 为了评估非理想情况下的分类精度损失，建立了行为级仿真模型。非理想是模拟的平均计算涨落(1.8mV)和平均PS-ADC INL涨落(8.1mV)。将计算涨落等同于分区式模数转

换模块的INL(9.9mV)，并将4-b量化精度的VGG-like CNN映射到该芯片。该芯片设置为AND-2-4-4-32模式。使用的CNN模型和拓扑、数据集以及不同条件下的推理精度如表4所示。精度损失为1.37％，可以接受。

[0142] 表4行为级仿真模型和结果比较

[0143]

数据集	CIFAR-10
CIM 模式	AND-2-4-4-32
神经网络	VGG-like CNN
网络技术	64C3-64C3-MP2-128C3-128C3-MP2-256C3-256C3-MP2-512FC-10FC
浮动精度	86.45%
理想的 4 位量化精度	86.92%
行为模拟精度	85.55%

[0144]

精度损失	1.37%

[0145] 如表5所示为本发明实施例中提供的芯片结构与现有芯片结构的性能对比情况表，该芯片支持具有最高通道数的XNOR模式和AND模式的MAC运算。计算获取时间为10ns，具有竞争力。该芯片在XNOR模式和AND-2-4-4-32模式下均实现了最高吞吐量，分别比现有芯片结构提高了5.42倍和3.28倍。XNOR模式下的能效为2.28倍，AND模式下为1.097倍。

[0146] 表5本发明实施例中提供的芯片结构与现有芯片结构的性能对比情况表

[0147]

芯片结构	ASSCC'21	ISSCC'21	ISSCC'20	TCAS-I'21	JSSC'20	本发明
技术	28 nm	28 nm	28 nm	40 nm	65nm	28 nm
尺寸	16 kb	384 kb	64 kb	8 kb	16 kb	128 kb
工作电压(V)	0.8	0.7-0.9	0.7-0.9	0.9-1	1	0.65-0.9
计算模式	Charge	Charge	Current	Time	Current	Charge
计算单元结构	8 6T+ CCU	32 6T+ SILMC	16 6T+ LCC	12T	12T	16 6T+ LPE
多模式	YES	NO	NO	NO	NO	YES
输入数据/存储数据/输出数据的位数(bit)	**1/1/3.46[5] 1/1/4[6]**	**4/4/12[6]**	**4/4/12[6]**	**1/1/2[5]**	**1/1.5/ 3.46[5]**	**1/1/2.24[5] 2/4/4[6]**
MAC 运算并行度	128/16	16/16	16/16	96/129	256/64	**64/256[5] 32/64[6]**
计算获取时间	20 ns	4 ns	4.1 ns	500 ns	54.21ns	10ns
吞吐量(GOPS)	20.48	N/A	124.88	49.15	604.46	**3276.8[6] 409.6[5]**
能效(TOPS/W)	**166[6] 32.14[1,5]**	**159.59[1,4]**	**116.3[1,8]**	**537[2,3]**	**604.5[2]**	**1379.4[6] 175.2[1,5]**
精度	80.26	N/A	91.9%	**98%[7]**	88.78%	85.55%

[0148] 其中,表3中尾注的含义如下:[1]表示归一化为2b/4b输入/权重操作(Normalized to 2b/4b input/weight operation);[2]表示归一化为1b/1b输入/权重操作(Normalized to 1b/1b input/weight operation);[3]表示有偿(With compensation),[4]表示根据论文中120.56-198.61TOPS/W的范围估计,平均为159.59 TOPS/W(159.59 TOPS/W in average by estimated from the range 120.56-198.61 TOPS/W in the paper);[5]表示XNOR模式(XNOR mode);[6]表示AND模式(AND mode);[7]表示MNIST数据集(MNIST dataset);[8]表示根据论文中的95.8-136.8TOPS/W范围估计,平均为116.3 TOPS/W(116.3 TOPS/W in average by estimated from the range 95.8-136.8 TOPS/W in the paper)。

[0149] 如图16所示,在上述实施例的基础上,本发明实施例中提供了一种通过上述各实施例提供的基于局部电容电荷共享的SRAM存算一体芯片实现的存内计算方法,包括:

[0150] S61,在与运算模式下,基于SRAM存算一体芯片中按位计算模块包括的每个位计算单元,采用局部电容的电荷共享原理,将输入数据与按位存储的存储数据的一位数据进行乘法运算,得到所述存储数据的一位数据对应的乘法运算结果,并基于所述SRAM存算一体芯片中全局共享开关模块,将所述存储数据的各位对应的乘法运算结果进行累加,得到模拟累加结果,基于所述SRAM存算一体芯片中分区式模数转换模块包括的所有位比较单元,将所述模拟累加结果进行量化输出;

[0151] S62,在异或运算模式下,基于每个所述位计算单元,采用所述局部电容的电荷共

享原理，将所述输入数据与所述存储数据的一位数据进行异或运算，得到所述存储数据的一位数据对应的异或运算结果，并基于所述分区式模数转换模块包括的每个所述位比较单元，将所述存储数据的一位数据对应的异或运算结果进行量化输出。

[0152] 在上述实施例的基础上，本发明实施例中提供的存内计算方法，还包括：

[0153] 在所述异或运算模式下，从所述位比较单元中选取序号大于所述位比较单元中的第一序号中位数的第一类比较器进行第一阶段使能，并基于所述第一类比较器的输出结果，从所述位比较单元的各比较器中所述第一类比较器的一侧选取第二类比较器进行第二阶段使能；

[0154] 在所述与运算模式下，从所述位比较单元组的各比较器中选取序号不等距的多个第三类比较器进行第一阶段使能，并基于所述多个第三类比较器的输出结果，从所述位比较单元组的各比较器中所述多个第三类比较器形成的区间之一选取第四类比较器进行第二阶段使能；

[0155] 其中，序号大于所述位比较单元组中的第二序号中位数的第三类比较器的数量多于序号小于所述第二序号中位数的第三类比较器的数量。

[0156] 具体地，该存内计算方法，其执行主体为上述各实施例提供的基于局部电容电荷共享的SRAM存算一体芯片，具体实现方法参见上述实施例，此处不再赘述。

[0157] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

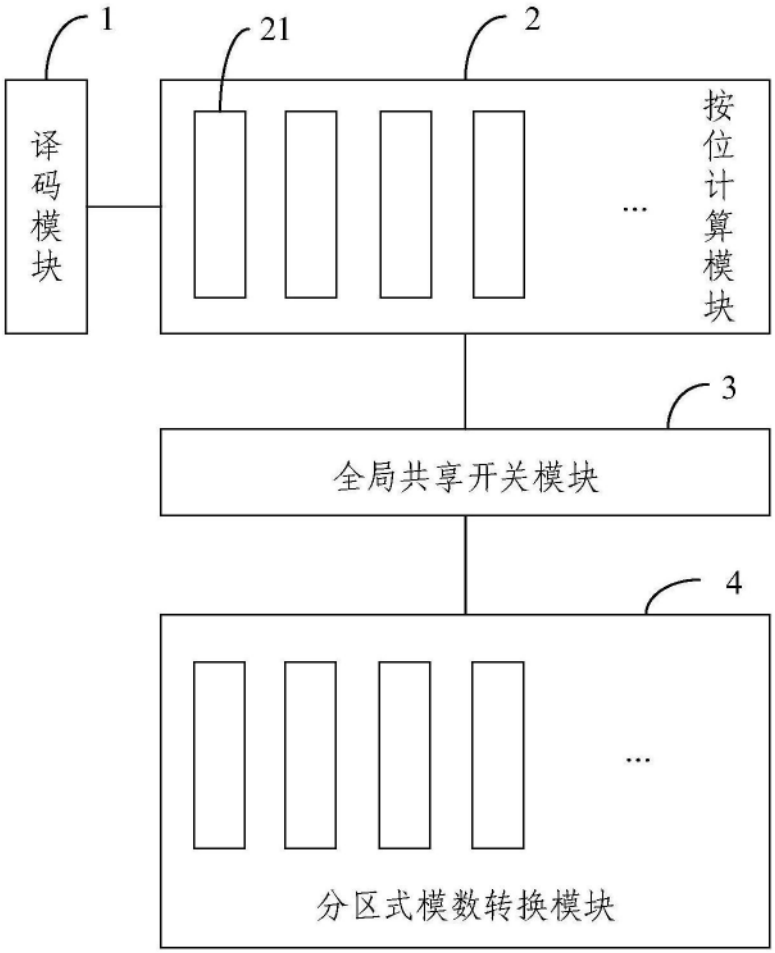

1
21
2
译码模块
按位计算模块
...
3
全局共享开关模块
4
...
分区式模数转换模块

图1

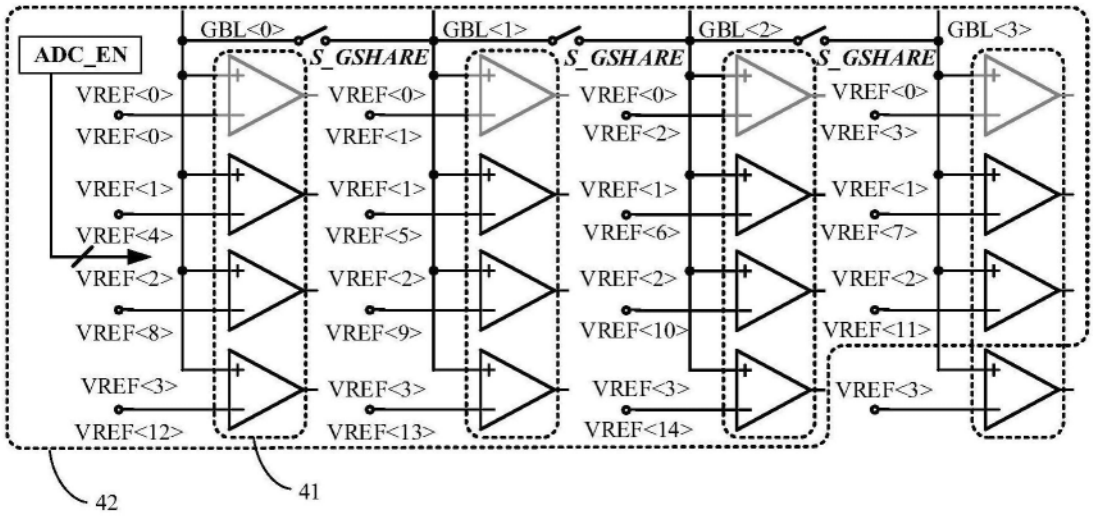

ADC_EN
GBL<0>
GBL<1>
GBL<2>
GBL<3>
S_GSHARE
S_GSHARE
S_GSHARE
VREF<0>
VREF<0>
VREF<1>
VREF<4>
VREF<2>
VREF<8>
VREF<3>
VREF<12>
VREF<0>
VREF<1>
VREF<5>
VREF<2>
VREF<9>
VREF<3>
VREF<13>
VREF<0>
VREF<2>
VREF<1>
VREF<6>
VREF<2>
VREF<10>
VREF<3>
VREF<14>
VREF<0>
VREF<3>
VREF<1>
VREF<7>
VREF<2>
VREF<11>
VREF<3>
42
41

图2

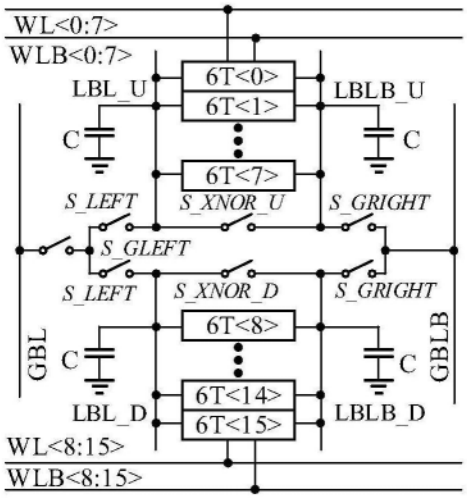

WL<0:7>
WLB<0:7>
LBL_U
LBLB_U
6T<0>
6T<1>
6T<7>
C
C
S_LEFT
S_XNOR_U
S_GRIGHT
S_GLEFT
S_LEFT
S_XNOR_D
S_GRIGHT
6T<8>
GBL
C
C
GBLB
6T<14>
6T<15>
LBL_D
LBLB_D
WL<8:15>
WLB<8:15>

图3

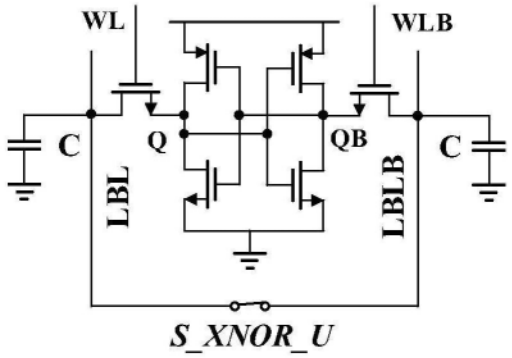

WL
WLB
C
Q
QB
C
LBL
LBLB
S_XNOR_U

图4

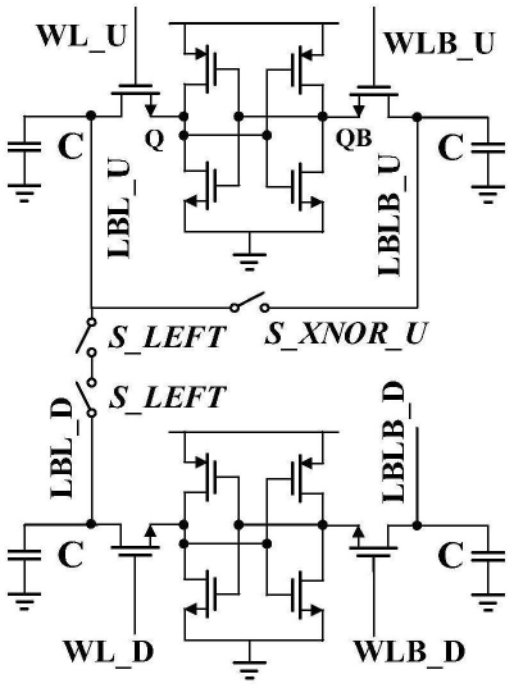

WL_U
WLB_U
C
Q
QB
C
LBL_U
LBLB_U
S_LEFT
S_XNOR_U
S_LEFT
LBL_D
LBLB_D
C
C
WL_D
WLB_D

图5

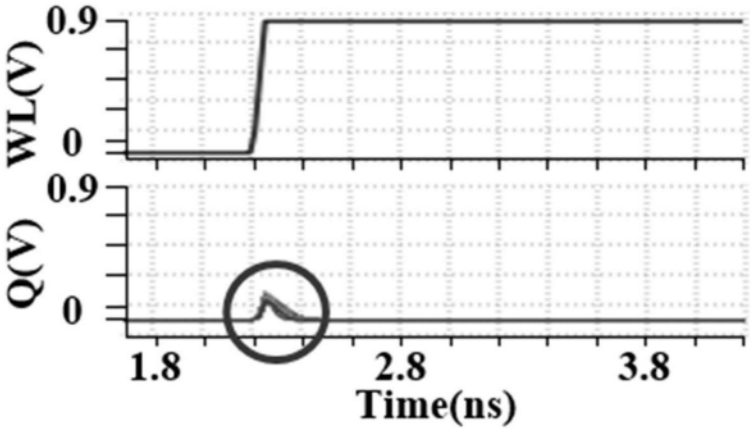

0.9
WL(V)
0
0.9
Q(V)
0
1.8
2.8
3.8
Time(ns)

图6

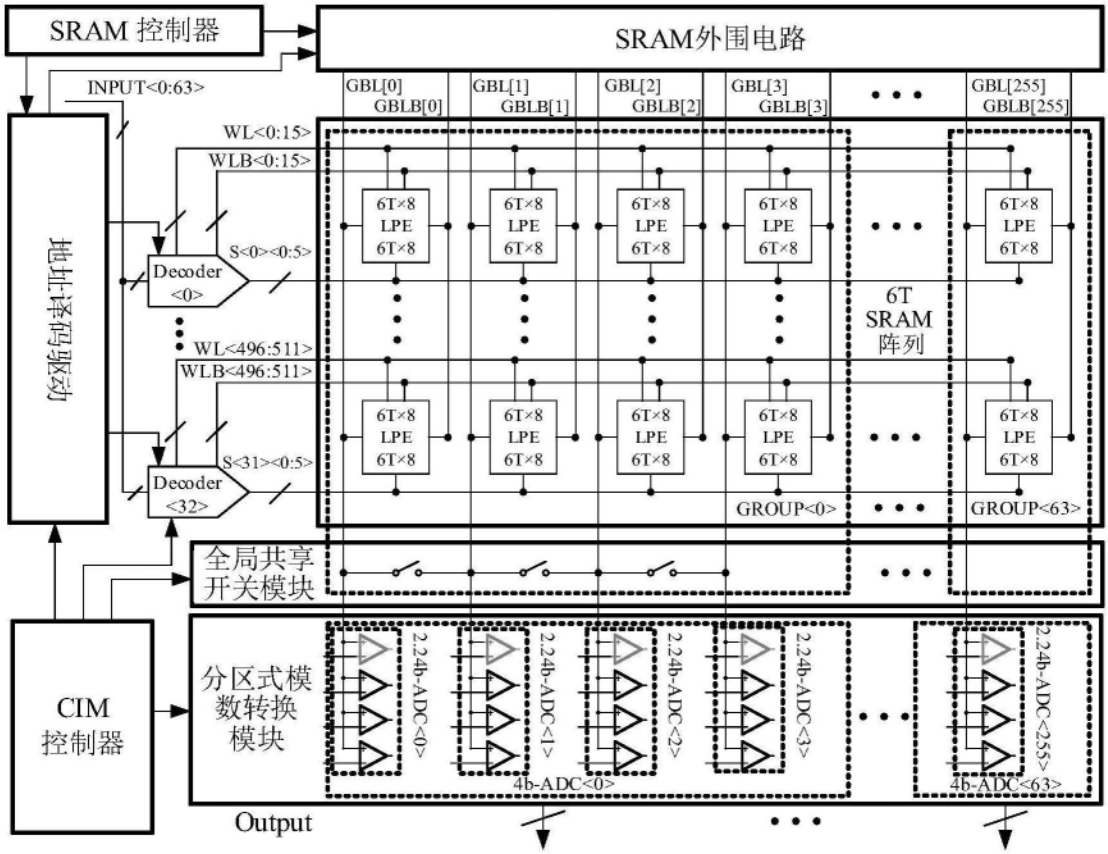

SRAM 控制器
SRAM外围电路
INPUT<0:63>
GBL[0]
GBLB[0]
GBL[1]
GBLB[1]
GBL[2]
GBLB[2]
GBL[3]
GBLB[3]
GBL[255]
GBLB[255]
WL<0:15>
WLB<0:15>
地址译码驱动
Decoder <0>
S<0><0:5>
WL<496:511>
WLB<496:511>
Decoder <32>
S<31><0:5>
6T×8 LPE 6T×8
6T SRAM 阵列
GROUP<0>
GROUP<63>
全局共享开关模块
CIM 控制器
分区式模数转换模块
2.24b-ADC<0>
2.24b-ADC<1>
2.24b-ADC<2>
2.24b-ADC<3>
2.24b-ADC<255>
4b-ADC<0>
4b-ADC<63>
Output

图7

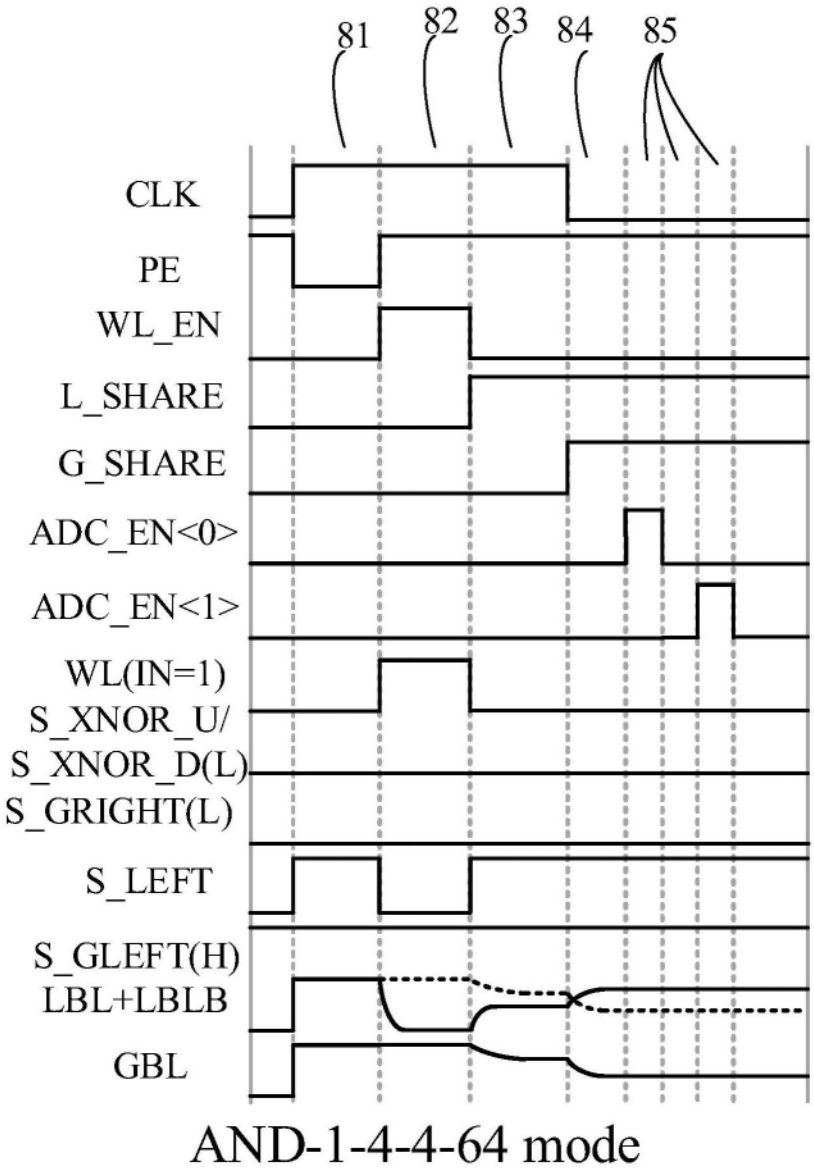

81
82
83
84
85
CLK
PE
WL_EN
L_SHARE
G_SHARE
ADC_EN<0>
ADC_EN<1>
WL(IN=1)
S_XNOR_U/
S_XNOR_D(L)
S_GRIGHT(L)
S_LEFT
S_GLEFT(H)
LBL+LBLB
GBL
AND-1-4-4-64 mode

图8

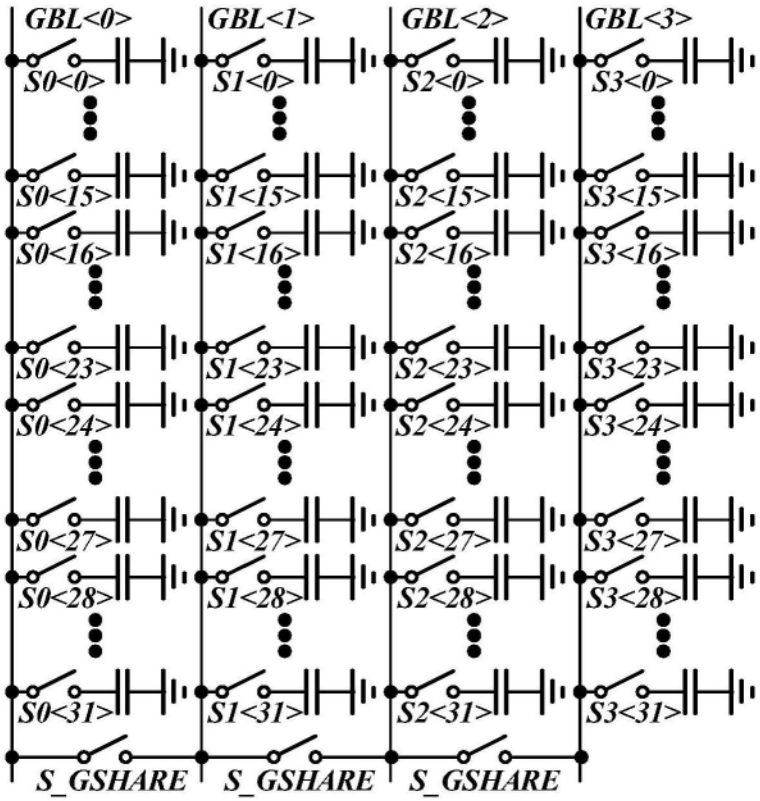

GBL<0>
GBL<1>
GBL<2>
GBL<3>
S0<0>
S1<0>
S2<0>
S3<0>
S0<15>
S1<15>
S2<15>
S3<15>
S0<16>
S1<16>
S2<16>
S3<16>
S0<23>
S1<23>
S2<23>
S3<23>
S0<24>
S1<24>
S2<24>
S3<24>
S0<27>
S1<27>
S2<27>
S3<27>
S0<28>
S1<28>
S2<28>
S3<28>
S0<31>
S1<31>
S2<31>
S3<31>
S_GSHARE
S_GSHARE
S_GSHARE

图9

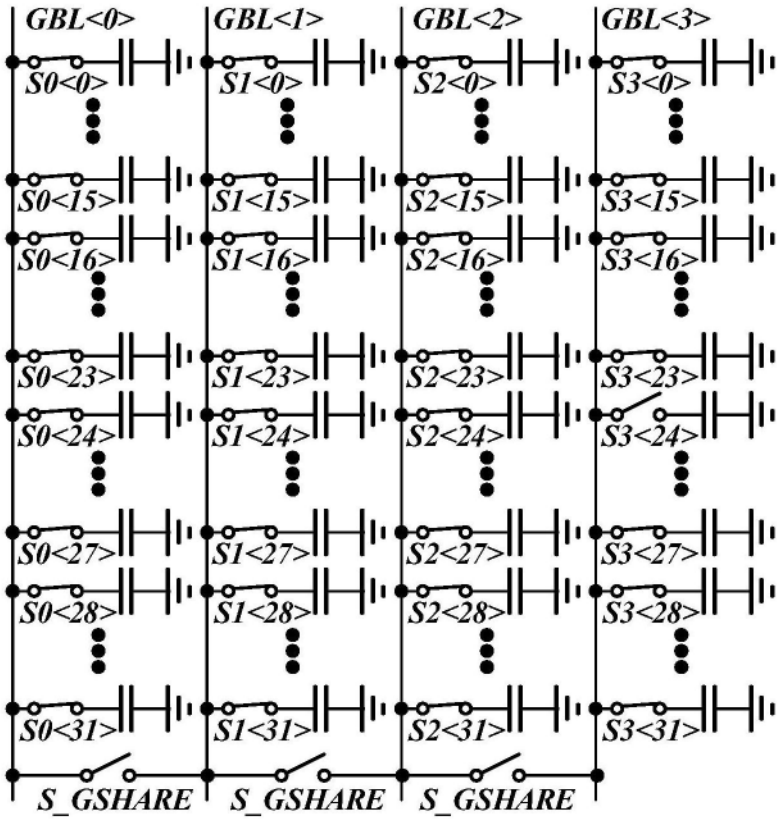

GBL<0>
GBL<1>
GBL<2>
GBL<3>
S0<0>
S1<0>
S2<0>
S3<0>
S0<15>
S1<15>
S2<15>
S3<15>
S0<16>
S1<16>
S2<16>
S3<16>
S0<23>
S1<23>
S2<23>
S3<23>
S0<24>
S1<24>
S2<24>
S3<24>
S0<27>
S1<27>
S2<27>
S3<27>
S0<28>
S1<28>
S2<28>
S3<28>
S0<31>
S1<31>
S2<31>
S3<31>
S_GSHARE
S_GSHARE
S_GSHARE

图10

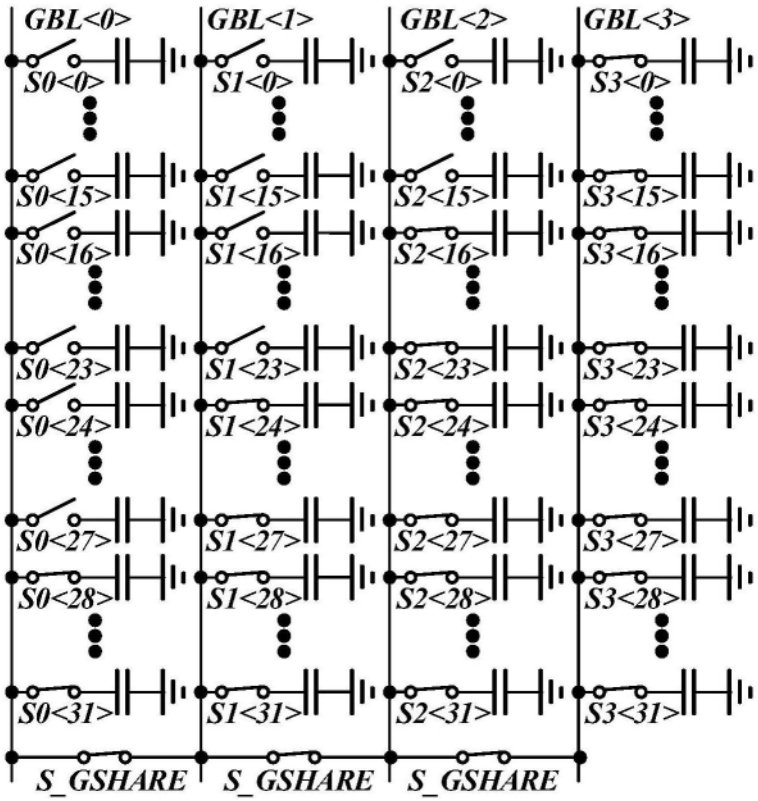

GBL<0>
GBL<1>
GBL<2>
GBL<3>
S0<0>
S1<0>
S2<0>
S3<0>
S0<15>
S1<15>
S2<15>
S3<15>
S0<16>
S1<16>
S2<16>
S3<16>
S0<23>
S1<23>
S2<23>
S3<23>
S0<24>
S1<24>
S2<24>
S3<24>
S0<27>
S1<27>
S2<27>
S3<27>
S0<28>
S1<28>
S2<28>
S3<28>
S0<31>
S1<31>
S2<31>
S3<31>
S_GSHARE
S_GSHARE
S_GSHARE

图11

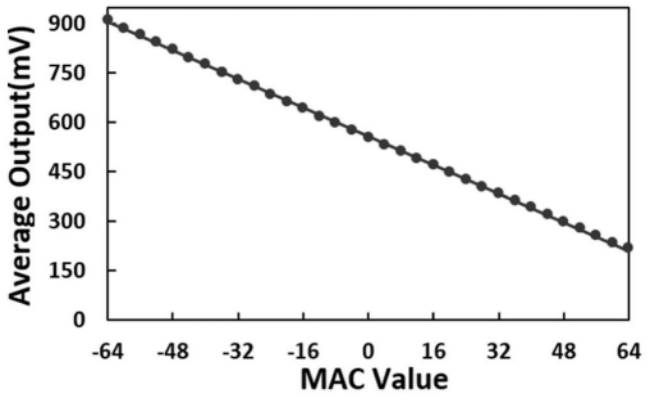

Average Output(mV)
900
750
600
450
300
150
0
-64
-48
-32
-16
0
16
32
48
64
MAC Value

图12

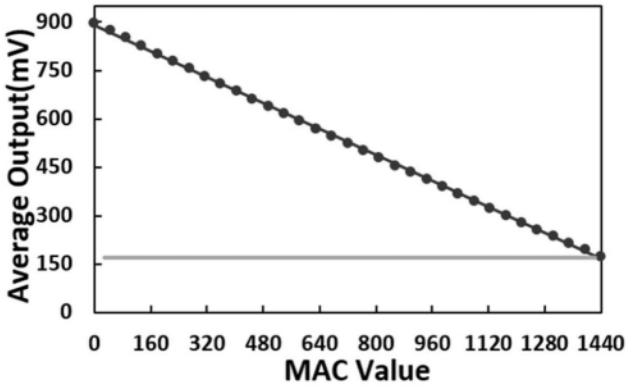

Average Output(mV)
900
750
600
450
300
150
0
0 160 320 480 640 800 960 1120 1280 1440
MAC Value

图13

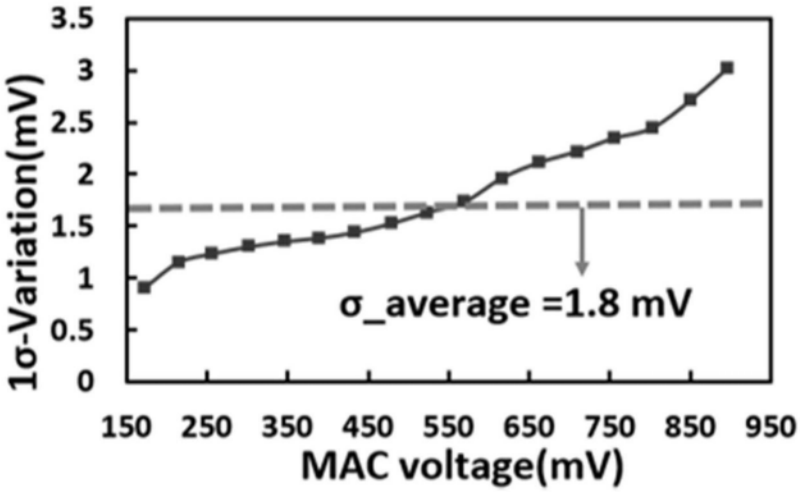

1σ-Variation(mV)
3.5
3
2.5
2
1.5
1
0.5
0
σ_average =1.8 mV
150 250 350 450 550 650 750 850 950
MAC voltage(mV)

图14

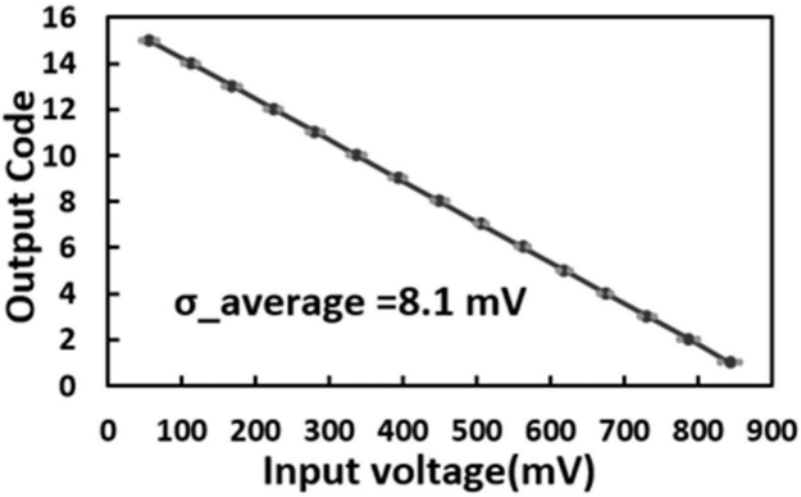

Output Code
16
14
12
10
8
6
4
2
0
σ_average =8.1 mV
0 100 200 300 400 500 600 700 800 900
Input voltage(mV)

图15

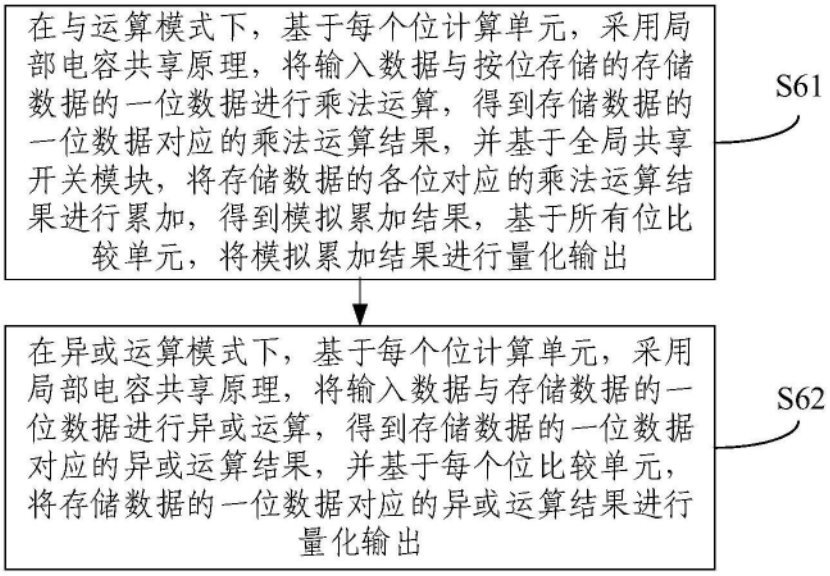

在与运算模式下，基于每个位计算单元，采用局部电容共享原理，将输入数据与按位存储的存储数据的一位数据进行乘法运算，得到存储数据的一位数据对应的乘法运算结果，并基于全局共享开关模块，将存储数据的各位对应的乘法运算结果进行累加，得到模拟累加结果，基于所有位比较单元，将模拟累加结果进行量化输出
S61
在异或运算模式下，基于每个位计算单元，采用局部电容共享原理，将输入数据与存储数据的一位数据进行异或运算，得到存储数据的一位数据对应的异或运算结果，并基于每个位比较单元，将存储数据的一位数据对应的异或运算结果进行量化输出
S62

图16