МАТРИЦА ОТ ИЗКУСТВЕНИ АНАЛОГОВИ НЕВРОНИ ИЗГРАДЕНА ПО СТАНДАРТНА КМОП ТЕХНОЛОГИЯ

инж.Момчил Михайлов Милев
Научен Р-л: доц. д-р. инж. Марин Христов Христов
Технически Университет – София
+1(520) 746-7690 milev_momtchil@ti.com

Milev M.M., Hristov M.H., Artificial Neural Matrix of Analog Neurons Implemented On Standard CMOS technology. *The report concludes a series of findings on the research and design of an artificial neuron model implemented in a conventional CMOS fabrication technology. Shown is the design of the individual elements building the artificial neuron. Attention is focused on the particular choice of these elements, on the overall approach for the design solution and to the operation of the neuron in a system of interconnected neurons. The motivation not to use floating-gate devices is explained. Non-linear relationship in the synapse function due to the specific implementation is briefly discussed. Method of storing synapse weight is discussed. Weight charge degradation due to leakage current is considered and a method for compensation is suggested. Special design of weight- and input-current switches is discussed. A complete system-on-a-chip (SOC) is shown, designed for fingerprint image feature extraction, which is based on the proposed analog neuron design.*

1. Постановка на проблема. Изисквания при моделирането на изкуствени невронни мрежи.

За да бъдат ефективни и приложими в практиката изкуствените невронни мрежи трябва да притежават бързина на обработката на входния сигнал по-бърза от времеконстантата на процеса под наблюдение поне с един порядък. С широкото навлизане на суб-микронните технологични процеси, поради изострящото се противоречие между изискванията за постоянно нарастваща скорост на цифровите интегрални схеми и едновременно изисквания за намаляване на захранващите напрежения и консумация, все повече автори търсят решение за изграждане на изкуствени невронни мрежи чрез аналогови средства. Аналоговото изпълнение позволява не само много по-голяма степен на интеграция на невронните елементи, но и значително предимство при скоростта на обработка на сигнала в сравнение с цифровите решения. Главните предимства на аналоговите решения произтичат от аналоговата обработка на входния сигнал, при която закъсненията са пренебрежими в сравнение с цифровите решения. Главен недостатък на аналоговите решения за невронни елементи, традиционно, се счита тяхната ниска точност при извършване на математическите операции. Това твърдение се оборва в [1] и [2] на база на изследвания доколко значима е "абсолютната точност" при функционирането на невронните мрежи. От друга страна, за да могат да се реализират голям брой изкуствени невронни елементи е необходимо техният размер да бъде намален до минимум. Тъй като основен изграждащ елемент в невроните е синапса, предложено е решение което позволява невронните синапси да се изградят възможно най-просто. Тук възниква противоречието между точността на моделирането и размера на интегралното решение за синапса. Показано е, че чрез алтернативно решение за хардуерното моделиране на синапса [1], може да се постигне значително намаляване на размерите на невронния елемент при

което функционалността на невронната мрежа не се компрометира. Показано е, че квадратичната нелинейност на синаптичната връзка спрямо нейното 'тегло', причинена от специфичното изпълнение не само не оказва негативен ефект върху функционалността на невронната мрежа, но и дори има позитивен ефект върху процеса на адаптиране на невронната мрежа.

Освен намаляването на физическите размери на невронните елементи, за постигането на скорост, необходима за обработка на входната информация в реално време, е необходимо да се конструира ефективен метод за пренасяне на информацията между невронните слоеве. По този показател, цифровите решения са подчинени не само на ограниченията в системната такова честота и необходимостта от поне няколко такта при извършване на математичните операции, но и от необходимостта да използват вътрешно-схемни протоколи за използване на общите информационни ресурси – вътрешни информационни и контролни шини, регистри и други. За да се избягнат тези недостатъци и за да се позволи скорост на обработка необходима за практичното приложение не изкуствените невронни мрежи за обработка на сигнали в реално време, като информационен носител е избран аналогов електрически ток. Това позволява да се избегнат закъсненията и загубите на сигнала поради разпределената резистивност и капацитет на линиите свързващи изходите и входовете на невронните елементи в мрежата (разпределения товарен капацитет и съпротивление на изхода на невронните елементи са значими поради факта, че изхода на всеки неврон от един слой е свързан с поне един вход на всеки неврон от следващия слой).

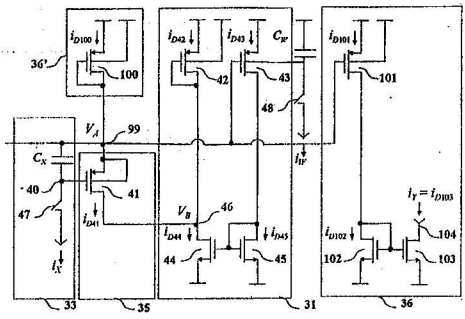
2. Схемно решение на Невронния Синапс

Имайки предвид гореописаните причини, разработено е, подходящо решение за моделирането на невронния синапс чрез използването на електрическите характеристики на обикновен полупроводников прибор – МОП транзистор изграден по стандартна КМОП технология. Това дава предимството настоящия модел на невронен елемент да бъде използван за изграждането на невронни мрежи в ГИС и СГИС като невронната мрежа е интегрирана с друг продукт изграден по стандартна технология върху същият кристал. Това дава възможност за вграждането на невронната мрежа в едночипови системи за събиране и обработка на информацията (data acquisition systems). По тази основна характеристика настоящето моделно решение се различава от съществуващите решения на базата на аналогови памети използващи устройства с плаващ затвор (floating-gate). Представеният модел на синапс използва дрейновия ток на КМОП полеви транзистор за да апроксимира синаптичната активност на неврона като до-синаптичния сигнал се представя от приложеното напрежения на затвора (гейт-сорс), а напрежението дрейн-сорс определя тегловия коефициент на синаптичната връзка. В [1] и [2] вече бяха приведени токовите уравнения и тяхното използване при моделирането на синаптичната връзка. В последващото изложение ще обърнем внимание на някои конкретни проблеми и техните решения позволили окончателното изграждане на универсална невронна матрица от 2176 синаптични елементи.

3. Схемно решение на изкуствения неврон

Изкуствения неврон е изграден на базата на следните модули: синапсен транзистор (41), входно преобразуващо стъпало (33), схема за установяване на синаптичното тегло (31), теглови кондензатор (C_W), общ токов сумиращ нод (99) и изходно стъпало (36 и 36'). Синапсният транзистор както и схемата за установяване на теглото заедно с тегловия кондензатор се повтарят необходимия брой пъти за да реализират многовходов невронен елемент. Схемата на неврона е показана на Фиг.1 За краткост на описанието и поради факта, че по-подробно описание на функционирането може да се намери в [1], по-голямо внимание ще обърнем на преобразуването на входния сигнал и

установяване на тегловия коефициент на синапса.



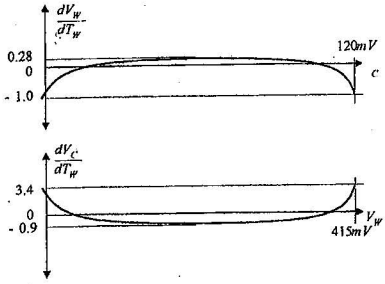
Фигура 1. Схема на основните компоненти на изкуствения нéврон

4. Входно преобразуване

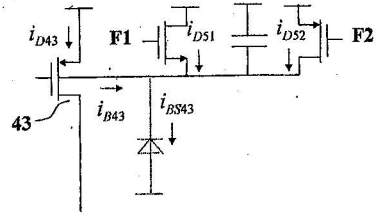
Входния ток към всеки синапс се поддържа постоянен за времето на входното преобразуване през което посредством ключа (47) този ток се интегрира в напрежение върху кондензатора C_X. Поради необходимостта синапсния транзистор да се поддържа постоянно в триоден режим на работа, е необходимо върху входния кондензатор да се подържа напрежение по-голямо от напрежението на насищане. За минимално напрежение на гейта е избрано напрежението от 1.0V, като избрания съответен сигнален обхват е 1.0 - 2.024V. Това съответства на 9-битова разделителна способност при напрежение на най-малкия разряд от 2mV. За да се ускори времето за преобразуване на входния ток, на практика се използва схема с предварително зареждане на C_X. Предварителното зареждане се извършва от фиксиран по продължителност и амплитуда токов импулс (1uA x 100nS) за всички входни кондензатори на всички синапси в матрицата. Чак след като те са заредени до 1.0V, започва същностното преобразуване на токовия сигнал за последващите 65nS (входния сигнал се очаква в обхвата 0-25.6uA). След завършване на зарядния цикъл, ключа (47) се отваря и запазва заряда върху кондензатора C_X. Специални мерки за намаляване на деградацията на заряда, тук, не са приети поради факта, че входното преобразуване се извършва през период от не повече от 6.25us, който не е достатъчно продължителен за да се забележи деградация по-голяма от половината от най-малкия разряд (1.0mV). Този въпрос стои по съвсем друг начин при кондензатора C_W който временно съхранява тегловния коефициент на синапса.

5. Съхранение на тегловния коефициент на синапса

Основно изискване при невронните мрежи е да осигурят съхранението на тегловните коефициенти на синаптичните връзки неопределено дълго време при липса на адаптация (режим на 'тренировка'). Съхранението на тегловните коефициенти при цифровите схемни решения най-често се извършва в някакъв тип програмируема памет (DRAM, SRAM, EEPROM) което прави лесен както процеса на програмиране на невронната мрежа така и интерфейсната комуникация между невронната мрежа и външни цифрови ресурси за обработка и съхранение на информация. При аналоговите решения най-често тегловите коефициенти се съхраняват в динамични аналогови RAM или устройства с плаващ затвор. Обикновено втория метод се предпочита поради независимостта от захранващото напрежение и липсата на необходимост от опресняване на съхранения заряд. Въпреки безспорните предимства на този метод, схемите реализиращи тегловните коефициенти чрез устройства с плаващ затвор притежават някои основни недостатъци, които най-често ограничават тяхната употреба в практиката. Тези схеми изискват първо, специализирана технология за производство която често оскъпява крайния продукт, и второ – изискват повишено работно напрежение позволяващо тунелното пренасяне на заряди към и от плаващия затвор. За да се преодолеят тези недостатъци, както и да се опрости съхранението и използването на тегловните коефициенти от аналоговите синапси, в настоящото решение е избран комбинативен метод за

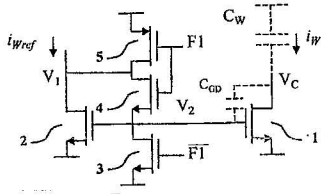


Фигура 2. Криви, определящи динамичния обхват на тегловото напрежение и напрежението върху кондензатора на теглото



Фигура 3 Токове определящи токовете на утечка и деградация на заряда върху тегловия кондензатор

съхранение и използване на заряда. Тегловите коефициенти се съхраняват в цифров вид в статична RAM, след което се преобразуват в аналогов вид при използването им в синапсите под формата на заряд върху кондензатора C_W. Преобразуването се извършва от броячи със зареждане, които едновременно съхраняват теглата в статични регистри и едновременно извършват цифрово-аналоговото преобразуване. Преобразуването се извършва като по управляващ сигнал, в режим на изваждане, съдържанието на броячите определя продължителността на постоянно-токов импулс който зарежда линейно кондензаторите до съответното напрежение. Отново, използвано е предимството на разпространение на сигнала по ток за да се избягнат проблемите с различната дължина на зарядните линии в различните части на невронната матрица. От компютърните симулации на нелинейността на заряда и ефектите причинени от тока на утечка са изведени кривите (Фиг.2) на изменението на напрежението на кондензатора C_W и на напрежението на теглото (напрежението дрейн-сорс на синапсния транзистор). От тези криви е определен динамичния обхват на изменение на напрежението върху кондензатора C_W, и съответното тегловото напрежение на синапса. Максималния теоретичен обхват за двете напрежения е: 120mV за напрежението на теглото и 415mV за напрежението на заряда определящ тегловното напрежение. На практика, избраните максимални напрежения са 100mV и 400mV съответно. Разрядността на теглото е 9-бита. Токовете на утечка, определящи максималната продължителност на съхранение на заряда са показани на Фиг.3. При тънко-окисен кондензатор между поликристален затвор и N-тип джоб, за TSMC 0.35um технология, токът на утечка на заряда е оценен на 5.5pA като определящо значение има тока на утечка на N-ваната към подложка както и утечката през джоба на p-каналния МОП полеви транзистор (43). Деградацията на заряда върху кондензатора се равнява на $0.5U_{LSB}$ след около 8.0us. На практика, е използвано решение 'компенсиращо' тока на утечка в подложката което позволява деградацията на заряда да се забележи чак след 500us. Това позволява, цикъла на презареждане на тегловите коефициенти от SRAM да се извършва много по-рядко, което оптимизира работата на невронните елементи при обработката на входния сигнал. За пример, в системата описана в заключението на доклада, това решение позволява само 160us от 33ms да бъдат използвани за 'опресняване' на тегловите



Фигура 4. Бърз, ниско-токов ключ с минимизация на шума от превключване

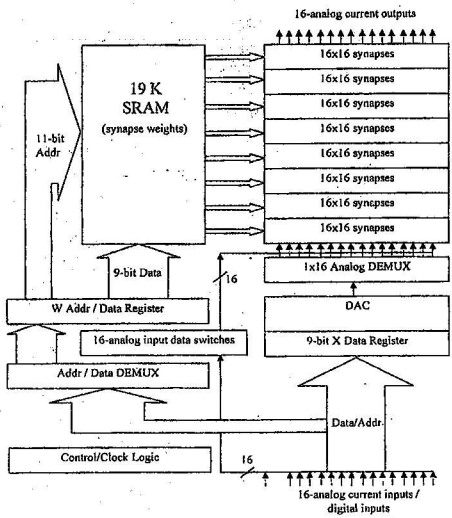
коефициенти при обработката на един пълен кадър CIF формат (320x240 пиксела) и скоростта на обработка да достигне границата на 30 кадъра/сек (на практика показаната система е проектирана за 24 кадъра/сек.)

6. Токови ключове

Както при зареждането на тегловите коефициенти, така и при входно-токовото преобразуване са използвани токови ключове със специална конструкция която позволява както значителна минимизация на шума от превключване (feed-through) така и бързо време на превключване при изключително слаби токове в обхвата 80-100nA. Токовият ключ използва противофазен управляващ сигнал и ускоряващ (pull-up) транзистор (5) който на практика елиминира времето за възстановяване на ключа и добавя ефект на пре-компенсация на шума от превключване при (3) и (4), така че времето за отваряне на ключа да бъде сведено до 380pS.

7. Системно описание

На базата на описания невронен елемент е изградена "Невро-Матрица" влизаща в състава на система предназначена за екстракция на характеристики при обработката на изображения от отпечатъци на пълци. "Невро-Матрицата" се състои от 128 неврона разположени в 8 реда и 16 колони. Всеки от невроните включва 16+1 синапса, или общо за цялата матрица - 2176 синаптични връзки. Матрицата от аналогови неврони има 16 аналогови входове и изходи по ток предназначени за лесно каскадно разширение на невронната мрежата. С цел съпоставка и оценка на бързодействието, системата може да бъде поставена в два основни режима на обработка на входните сигнали – аналогов и цифров. В цифров режим, системата използва вграден 9-битов ЦАП за преобразуване на цифровите входни данни за осветеността на пикселите (512 нива на сивото). 16-те входа, в този режим, се зареждат последователно с данните от поредната сканирана област от 4x4 пиксела, входните вектори се преобразуват в аналогов токов сигнал, зареждат входните кондензатори на синапсите в първия слой след което се обработват от матрицата едновременно. Схема на системно ниво е показана на Фиг.5 Важно е да се посочи че докато цялата система заема площ от 7.3x7.3mm^2, аналоговата матрица от 128 неврона (2176 синапса) заема площ по-малка от 3mm^2 или по-малко от 6% от общата площ на чипа (теоретичната плътност на представените невронни елементи е 1400 синапса на квадратен милиметър). Този факт отново демонстрира, че ограниченията в използваната площ и брой на невронните елементи произтичат от цифровата част на системата.



Фигура 5. Схема на системно ниво на "Невро-Матрица-1"

8. Литература.

1. Момчил Милев, Хардуерен Модел на Невронен Синапс с Нелинейност, десета национална научно-приложна конференция с международно участие "Електронна Техника 2001" Созопол 20-22 Септември.
2. Момчил М. Милев, Марин Х. Христов, Прост Хардуерен Модел на Невронна Синаптична връзка в с единствен полеви-МОП транзистор за конвенционална CMOS технология, девета национална научно-приложна конференция с международно участие "Електронна Техника 2000" Созопол 19-21 Септември.
3. Momchil Mihaylov Milev, Marin Hristov Hristov, Characteristics of Single Perceptron Training, National Scientific and Applied Science Conference, ET-98, Sozopol, September 18-21, 1998.
4. Phillip E. Allen, Douglas R. Holberg, CMOS Analog Circuit Design, Oxford University Press, 1987; pp.98-101, pp.198-207
5. Simon Haykin, Neural Networks – A Comprehensive Foundation, second edition, 1999 Prentice Hall, New Jersey; pp. 3-18, pp.51-53, pp.121-122, pp.128-132
6. Roubik Gregorian, Gabor C. Temes, Analog MOS Integrated Circuits For Signal Processing, A Wiley-Interscience Publication, John Wiley & Sons, New York, 1986; pp.462-483