Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт (государственный университет)»

на правах рукописи

ТЕПЛОВ ГЕОРГИЙ СЕРГЕЕВИЧ

РАЗРАБОТКА МОДЕЛИ ИСКУССТВЕННОГО НЕЙРОНА С ДИНАМИЧЕСКОЙ ФУНКЦИЕЙ АКТИВАЦИИ НА БАЗЕ МЕМРИСТИВНЫХ КОМПОНЕНТОВ

Специальность 05.27.01 –
твердотельная электроника, радиоэлектронные компоненты,
микро- и наноэлектроника, приборы на квантовых эффектах

Диссертация на соискание ученой степени
кандидата физико-математических наук

Научный руководитель
доктор технических наук Горнев Е.С.

Научный консультант
кандидат физико-математических наук Матюшкин И.В.

Москва – 2018

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Развитие науки и технического прогресса в области создания интеллектуальных систем в последние годы является предметом общего интереса производителей, интеграторов и исследователей. Результатом данного интереса, с одной стороны, является получение все большего количества теоретических и эмпирических данных обо все расширяющемся перечне прикладных задач, решаемых с помощью подобного рода систем, с другой стороны, достижения последних десятилетий позволяют говорить о возможности оптимизации архитектуры и компонентной базы при построении подобного рода систем. Примерами подобного рода задач могут служить экспертные системы, системы автоматизированного и автоматического управления, системы обработки информации, плохо поддающейся формализации (изображения, аудио сигналы и т. д.), и многие другие. Примерами подобного рода систем являются различные приложения по обработке фотографий с целью стилизации под известных художников, системы постановки диагноза по симптомам, системы управления различными роботами и многие другие.

Одним из способов создания интеллектуальных систем, в последнее десятилетие все более превалирующий над остальными, является применение в данных целях искусственных нейронных сетей (ИНС). Данная работа направлена, прежде всего, на физико-математические аспекты проектирования и создания ИНС и их составных элементов (искусственных нейронов) с учетом последних достижений в области электронной компонентной базы, а именно мемристивных элементов памяти.

Актуальность работы

Уменьшение минимального топологического размера при производстве микросхем, появление новых электронных компонентов – мемристоров, развитие методов и средств распараллеливания вычислений породили существенный рост интереса ученых и исследователей со всего мира к нейроморфным системам и

аппаратным реализациям искусственных нейронных сетей. Вычислительные системы с фон Неймановской или Гарвардской архитектурой не позволяют достичь требуемой производительности вычислений, что делает актуальными поиск новых вычислительных архитектур, алгоритмов и моделей вычислений.

Отличный от фон Неймановских, принцип организации вычислений в искусственных нейронных сетях позволяет добиться эффективного распараллеливания процессов вычислений, что в будущем заложит основы для создания более производительных вычислительных систем. При решении этих задач важно подчеркнуть взаимообусловленность функционального и физического аспекта, причем не только на уровне архитектуры всей сети, но и в ее элементах. Среди всего многообразия проектов, направленных на исследование аппаратных реализаций нейроморфных систем, как наиболее крупные международные исследовательские проекты могут быть выделены следующие:

FACETS/BrainScaleS (2005-2015) — проекты, направленные на разработку новых базовых технологий, реализацию нейросистем в виде аппаратных средств, разработку архитектуры систем и биологические исследования. Ведущим исследователем выступает Консорциум европейских научных школ под началом Гейдельбергского университета (Германия), возглавляет К. Мейер.

Neurogrid (2006-2018) — проект направлен на повторение достижений Blue Brain Project, моделирующего работу мозга на суперкомпьютере. Среди основных целей сделать это на специализированном устройстве собственной разработки, со значительно меньшей стоимостью, размером и энергопотреблением по сравнению с суперкомпьютером IBM Blue Gene и на изготовленном устройстве проводить моделирование различных функций мозга. Ключевые организации – Brains-In-Silicon и Стенфордский университет. Ведущий исследователь проекта – К. Боэн.

SyNAPSE (2008-2016)а — проект, нацеленный на моделирование мозга млекопитающих животных и человека, создание нейроморфных устройств, разработка архитектуры нейронных систем. Организаторами выступают DARPA совместно с IBM Labs, HRL и рядом научных школ США.

MoNETA/Cog Ex Machina (2008-2016) — проект ориентирован на разработку систем управления роботами и моделирование с применением новых аппаратных архитектур.

Human Brain Project (2014-2023) — самый крупномасштабный проект по количеству участников и финансированию в Европе: среди участников European Commission более 10 научных центров и организаций. В качестве целей проекта выступают моделирование мозга мыши и человека, разработка когнитивных архитектур, теоретические исследования в неврологии, разработка нейроинформационной платформы, высокопроизводительные вычисления, медицинская информатика.

Несмотря на то, что мейнстримом исследований являются сети с глубоким обучением, что связано с изменением архитектуры сети, для нейроморфных систем до сих пор актуальна задача поиска оптимальной с точки зрения аппаратной реализации структурно-функциональной схемы одиночного нейрона. Одним из перспективных направлений развития ЭКБ является мемристивный и резистивный элементы, в частности – на базе тонких пленок нестехиометрических оксидов Ti, Hf, Si, Ta и др.

Среди недавних разработок в области нейронных сетей с мемристивными элементами можно отметить разработку лабораторного прототипа нейропроцессора фирмы IMEC. При реализации нейронной сети на кремниевой фабрике неизбежны вопросы создания специальных библиотек элементов как схемотехнического, так и логического уровня, расширяющих возможности САПР Cadence. Одной из возникающих при этом задач является компактное отображение характеристик мемристора на формализованные математические параметры в Verilog описании.

Разработка нейронов с динамической функцией активации по сравнению со статической функцией активации повысит универсальность системы при тех же целевых параметрах. Несмотря на усложнение в структуре нейрона, данный подход к построению нейронной сети позволит уменьшить общее число нейронов

при реализации вычислений за счет имплементации различных типов функций над входными векторами данных в рамках одного нейрона.

Исследования были поддержаны грантом РФФИ «Исследование и разработка нейросетевых и клеточно-автоматных технологий в проектировании сверхбольших интегральных схем» № 17-07-00570 А (2017-2018гг.).

Цели и задачи

Цель работы – разработка функциональной модели искусственного нейрона позволяющего производить выбор типа функции активации в процессе обучения или работы, его структурной схемы для аппаратной реализации с применением мемристивных компонентов.

Для достижения поставленной цели в работе поставлены и решены следующие научные задачи:

- Синтез абстракций формального нейрона и конечного автомата (модель конечного автомата абстрактного нейрона – КААН) и ее спецификация для мультипликативной и аддитивной функции групповой обработки входных сигналов.
- Разработка и описание высокоуровневой функциональной модели искусственного нейрона с динамической функцией активации.
- Разработка модельных представлений мемристора средствами САПР Cadence на языке высокого уровня предназначенного для описания аппаратуры Verilog-A.
- Анализ разработанных модельных представлений мемристора с помощью САПР Cadence.

- Разработка описания нейрона с динамической функцией активации в виде структурной схемы с учетом физических особенностей мемристивных элементов с возможностью задания многоуровневых дискретных состояний.

Объект исследований

Искусственный нейрон, использующий регистры памяти и мемристивные элементы при аппаратной реализации.

Предмет исследований

Предметом исследований является модель нейрона с динамической функцией активации: математическая и формализованная на языке высокого уровня, включающая элементы схемотехнического описания.

Методы исследования

Для решения поставленных в работе задач использовались методы, основанные на теории множеств, теории автоматов, схемотехнике, теории алгоритмов и прикладного программирования, а так же стандартные методы схемотехнического моделирования САПР Cadence.

Научная новизна

1. Впервые предложена модель нейрона с динамической функцией активации отличающейся тем, что производится выбор функции активации, переключаемой либо в процессе функционирования нейрона, либо в процессе обучения. Показано влияние параметров модели (мощность множества определения функции

активации, мощность множества весовых коэффициентов синапсов, мощность алфавита входных и выходных сигналов) на вычислительную мощность искусственного нейрона.

2. Предложено обобщение в рамках предлагаемой математической модели конечного автомата абстрактного нейрона с динамической функцией активации, агрегационных функций математических моделей суммирующего и мультипликативного нейронов. Идея синтеза основана на задании равномощных множеств определения функций активации. Построение эквивалентных моделей реализуется путем подбора элементов множеств значений весовых коэффициентов с последующим (для обеих моделей нейронов) взаимным повторением порядка и следования элементов множества значений при задании функций активации.

3. Предложена математическая модель с неэквидистантным следованием уровней значений весов синапса, позволяющая в предельном случае увеличить область определения агрегационной функции искусственного нейрона до $I*N^{W}$, где N — количество синапсов, W — мощность множества значений весовых коэффициентов, I — мощность множества значений входных сигналов.

4. Установлено соответствие между физическими параметрами известных мемристивных компонентов (вольт-амперные характеристики процессов переключения, механизм переключения и другие) и формальными параметрами Verilog-A описания в среде САПР Cadence. Модель мемристора позволяет произвести описание множественности состояний проводимости с учетом девиации следующих параметров: напряжений порогов переключения, высокорезистивного и низкорезистивного состояний, количества циклов переключения. Множественность состояний проводимости, в отличие от существующих моделей, реализуется с учетом отклонений параметров и без необходимости отдельного, ограниченного по количеству и преднамеренно задаваемого параметрами модели описания промежуточных состояний проводимости.

5. Предложена структурная схема реализации искусственного нейрона, включающая два блока LUT (LUT – таблица значений функции) и сдвиговый регистр. Преимущество схемы заключается в учете амплитуд агрегированного сигнала в течение всего времени активации в дополнение к учету обобщенного уровня возбуждения нейрона.

Практическая значимость

- Предложенная модель конечного автомата абстрактного нейрона с динамической функцией активации (КААН) позволяет реализовать набор искусственных нейронов с различными функциями активации, что при аппаратной реализации КААН гарантирует относительную универсальность технического решения и широкий набор возможностей при проектировании сети.

- Представленное описание на языке Verilog-A позволяет использовать мемристор в качестве стандартного элемента библиотеки электронных компонентов САПР Cadence, что в свою очередь необходимо для проектирования микросхем с нейроморфной структурой. Предложенное описание может быть модифицировано в зависимости от конкретных параметров создаваемых структур.

- Результаты моделирования демонстрируют предпочтительность снижения разброса параметров высокопроводящего состояния мемристора и параметров разброса пороговых напряжений для изготавливаемых МДМ (металл-диэлектрик-металл) структур. Разброс параметров в высокопроводящем состоянии будет вносить большие искажения в обрабатываемый сигнал в сравнении с влиянием разброса в низкопроводящем состоянии. Разброс параметров пороговых напряжений при переключении от цикла к циклу напрямую влияет на точность получаемых промежуточных состояний проводимости мемристора.

- Следствием результатов моделирования с учетом разброса параметров мемристора является приоритет задания множественности состояний путем

подачи коротких (не более 15 нс) импульсов с малой амплитудой по напряжению (не более 0.15 В выше порога переключения) перед длительными импульсами с малой амплитудой, либо короткими импульсами с большой амплитудой.

Положения, выносимые на защиту

1. Математическая модель искусственного нейрона с динамической функцией активации в виде схемы конечного автомата абстрактного нейрона, реализующая синтез математических моделей искусственных нейронов с агрегирующими функциями сложения и умножения. Динамика функции активации заключается в оперативном изменении функции активации в процессе обучения или работы нейронной сети.

2. Математическая модель синапсов с нелинейной зависимостью между элементами весовых коэффициентов синапсов, позволяющая увеличить максимальную мощность множества определения функции агрегации до $I*N^{w}$, где N – количество синапсов, W – мощность множества значений весовых коэффициентов, I – мощность множества значений входных сигналов.

3. Модель мемристора, описанная на языке Verilog-A и позволяющая моделировать промежуточные состояния проводимости с учетом девиаций параметров при переключении: разброса напряжений порогов переключения между состояниями проводимости, разбросов низкорезистивного и высокорезистивного состояний, разброса количества циклов переключения.

4. Структурная схема реализации искусственного нейрона, включающая два блока LUT и сдвиговый регистр, позволяющая учитывать не только общий уровень активности агрегированных входных сигналов, но и влияние амплитуды агрегированных входных сигналов на выходной сигнал для каждого момента времени активации.

Личный вклад автора

Все теоретические результаты предлагаемой модели искусственного нейрона представлены в разделах 1.2-1.4 Главы 1, получены соискателем лично либо в соавторстве при его непосредственном определяющем или весомом участии. Высокоуровневое модельное описание мемристивного элемента на языке Verilog-A и моделирование его работы в среде САПР Cadence, представленное в разделах 2.3-2.4 Главы 2, а так же разработка структурной схемы модели КААН на базе мемристивных компонентов Главы 3, произведено автором лично под руководством к-та физ.-мат. наук Матюшкина И.В. и д-ра тех. наук Горнева Е.С.

Апробация результатов исследования

Результаты работы были представлены на следующих конференциях и семинарах:

- X научно-техническая конференция молодых специалистов «Высокие технологии атомной отрасли. Молодежь в инновационном процессе», Нижний Новгород, 10-12 сентября 2015
- «Электроника-2015» Международная научно-техническая конференция, г. Зеленоград, 19-20 ноября 2015
- 60-я Научная конференция МФТИ, г. Долгопрудный, 15 октября 2017 г.
- Научный семинар “Нейроморфные системы и их реализация” научного совета РАН “Фундаментальные проблемы элементной базы информационно-вычислительных и управляющих систем и материалов для ее создания”, г. Зеленоград, 24 сентября 2018.
- 11-ый научно-практический семинар "Математическое моделирование в материаловедении электронных наноструктур", ВЦ РАН имени А.А. Дородницына ФИЦ ИУ РАН, г. Москва, 15 мая.

Публикации

По теме диссертации опубликованы 5 работ в научных журналах и сборниках трудов международных и российских конференций, в том числе 2 работы в рецензируемых журналах, входящих в действующий перечень ВАК.

Структура и объем работы

Диссертационная работа состоит из введения, трех глав, заключения, списка терминов и определений, списка сокращений, списка работ, опубликованных по теме диссертации, списка цитируемой литературы из 143 наименований, двух приложений и содержит 156 страниц, в том числе 41 рисунок и 4 таблицы.

Во введении обосновывается актуальность диссертационной работы с научной точки зрения; формулируются цель и задачи исследования. Представлены положения научной новизны, практической значимости и положения, выносимые на защиту.

Первая глава включает аналитический обзор моделей искусственных нейронов и искусственных нейронных сетей. Излагаются и анализируются существующие модели нейронов и их применение в современных моделях нейронных сетей. Отмечается, что ни одна из моделей искусственных нейронов не учитывала возможность динамического изменения функции активации в процессе работы или обучения. Анализ существующих моделей сетей производится с позиций шаблона связности и типа связей с учетом специфики распространения сигналов по сети, а также включает рассмотрение современных типов архитектур и способов их имплементации на основе существующих базовых моделей нейронных сетей. По результатам анализа формулируются требования к математической модели нейрона с динамической функцией активации.

Представлена модель конечного автомата абстрактного нейрона (КААН) учитывающая специфику динамического изменения функции активации. Предлагаемая концепция сформулирована в терминах теории множеств на основе формального описания автомата Мура. В рамках модели осуществляется синтез

моделей искусственных нейронов мультипликативной и суммирующей агрегационной функцией. Производится анализ влияния способа задания весовых коэффициентов синапсов на мощность определения функции агрегации. Отмечается эффективность неэквидистантного подхода к заданию весовых коэффициентов позволяющая повысить мощность области задания функции агрегации без увеличения мощности множества весовых коэффициентов.

На основе предложенной модели КААН производится построении обобщенной схемы аппаратной реализации нейрона включающей два блока LUT, позволяющих задавать функцию активации произвольного вида, что расширяет возможности с позиций относительной универсальности предлагаемого технического решения.

Вторая глава предваряется обзором методов, подходов и средств аппаратных реализаций искусственных нейронов и искусственных нейронных сетей. В рамках обзора исследуются элементы современной компонентной базы, такие как мемристоры, MTJ-элементы, SOT-элементы, PCM-элементы в качестве составных блоков аппаратной реализации искусственного нейрона. Также в рамках литературного обзора производится исследование современных достижений в области нейроморфных систем и имплементации искусственных нейронных сетей на чипе, включающее рассмотрение проектирования указанных вычислителей на основе цифровой схемотехники, ПЛИС и гибридной схемотехники. По результатам обзора производится мотивированный выбор в пользу мемристивных элементов с биполярным механизмом переключения, как наиболее перспективных с позиций занимаемой площади на кристалле (в сравнении с цифровой реализацией) и количества представляемых дискретных состояний (в сравнении с MTJ, SOT и PCM).

На основе известных экспериментальных данных исследований мемристивных компонентов и существующих подходов к их моделированию формулируется описание мемристора средствами языка Verilog-A. В отличие от существующих моделей описания, предлагаемое представления позволяют учитывать влияние разброса параметров высокорезистивного и низкорезистивного

состояний, а также разброса параметров порогов переключений на количество циклов переключения, разброса переключений между состояниями промежуточной проводимости элемента и типа отказа мемристора в высокорезистивном, низкорезистивном или промежуточном состоянии. Указанные возможности позволяют производить проектирование последующих схем обработки агрегированного от входов сигнала с учетом его дисперсии, что позволяет учитывать физические особенности мемристивных структур.

Третья глава посвящена конкретизации обобщенной схемы реализации искусственного нейрона с динамической функцией активации. Вначале производится уточнение схемы с учетом применения двух блоков LUT. Отмечается позитивный эффект данного технического решения, заключающийся в возможности учета не только общего уровня активности, что справедливо для интегрирующих и связывающего типов нейрона, но учета амплитуды уровня активности каждого агрегированного сигнала в период активации искусственного нейрона. На основе полученного уточнения производится конкретизация ряда блоков структурной схемы с учетом применения в аппаратной реализации мемристивных компонентов.

В заключении представлены основные научные и практически значимые результаты диссертационной работы.

Приложение № 1 содержит Verilog-A описание мемристивного компонента.

Приложение № 2 содержит дополнительные графические результаты моделирования мемристивного компонента в среде Cadence.

1 ГЛАВА. Теоретические основы искусственных нейронов.

В данной главе рассматриваются математические модели искусственных нейронов и сетей на их основе. В начале главы представлен краткий обзор литературы, далее авторская модель искусственного нейрона, авторский алгоритм обучения и верификация модели с окончательными выводами по главе.

В обзоре литературы рассматриваются основные направления развития теории и текущий уровень достижений. Результатом обзора являются выводы относительно перспектив и возможностей последующего развития теории искусственных нейронных сетей, включая алгоритмы их функционирования. Акцент в авторской математической модели нейрона смещен в сторону аппаратных реализаций. При описании формализма использовались аксиоматики теории множеств и теории конечных автоматов. Предлагаемый формализм позволяет моделировать основные типы существующих и моделей искусственных нейронов и нейронных сетей.

В завершении главы представлены данные по результатам верификации модели нейрона и алгоритма обучения. На основе данных моделирования предлагаются основные подходы к применению результатов исследования и дальнейшие направления в развитии теории искусственных нейронных сетей.

1.1 Литературный обзор математических моделей искусственных нейронов и искусственных нейронных сетей

В теории нейронных сетей, если абстрагироваться от способов имплементации нейроморфных систем и всего спектра, относящихся к этой научно-технической области вопросов, (архитектура вычислительной системы, компонентная база, материалы и т. д.), может быть выделено три превалирующих взаимозависимых направления: модели искусственных нейронов, архитектуры искусственных нейронных сетей и алгоритмы обучения сетей.

Согласно существующим представлениям в истории науки развитие современной теории искусственных нейронных сетей (далее по тексту ИНС) началось с работы McCulloch W.S. и Pitts W.A. [1]. В рамках работы авторами впервые был предложен подход, позволяющий описывать биологические нейронные сети, приведена методика для формализации не только нейронных сетей, но и процессов в них. Стоит отметить, что предлагаемые нейроны могли выполнять все классические операции двоичной логики и включали исследование таких вопросов, как наличие петель обратной связи и «тормозящие» связи у нейронов. В качестве функций активации, как правило, выступала некоторая комбинация логических функций от входов, исчислявшаяся только в случае преодоления некоторого порога при суммировании входных сигналов от других нейронов. Сложно переоценить данную работу, так как она не только продемонстрировала возможность описания математическими формализмами нейронов, но поставила вопрос об оптимальных методах и моделях данного описания.

В качестве следующей значительной работы в данной области следует отметить разработку ИНС Perceptron, названной так ее автором F. Rosenblatt'ом. Концепция Perceptron как ИНС и метода построения нейронных сетей явилась первой удачной попыткой создания параметризованной сети с возможностью обучения. Предложенная, например, в статье [2] модель представляла сеть

прямого распространения сигнала и состояла из трех слоев нейронов. В рамках работы автор приводит пример «фотоперсептрона» (photoperceptron англ.), сети состоящей из 3-х слоев. Первый слой – это сетчатка (слой афферентных нейронов с функций «активации все или ничего» (all-or-nothing – англ.) и зрительными рецепторами в качестве входов). Второй слой – это набор интернейронов, связанных случайным образом с первым и третьим слоями. Третий слой – это слой эфферентных нейронов имеющих рекуррентные ингибиторные связи со вторым слоем. Предлагаемый подход позволил получить качественное совпадение между кривыми обучения и переменными описания сети, а также произвести обратное сопоставление, что явилось прорывом в понимании и описании механизмов функционирования реальных биологических нейронных сетей.

Впервые обобщенные правила обучения персептронов и «Adaline» нейронов (Adaptive linear neuron) были представлены в совместной работе Widrow B. и Hoff M.E. [3]. Предложенное обобщенное «дельта правило» легло в основу целого класса алгоритмов обучения и ознаменовало формализацию и выделение данной области в отдельное направление по изучению алгоритмов обучения таких, как «обучение с учителем» и алгоритмов обучения в целом.

Модели искусственных нейронов. Обзор основных моделей. Как было упомянуто ранее, согласно общепринятому на текущий момент представлению, первой моделью искусственного нейрона был нейрон, описанный в работе [1]. Модель разрабатывалась для описания процессов нейронной активности, происходящих в мозге человека. В работе продемонстрирована возможность вычисления сложных логических выражений на наборе нейронов. В качестве составных частей сети использовались нейроны с функцией активации в виде логического выражения от значений на входах. В логических выражениях использовались такие операции, как конъюнкция, дизъюнкция и отрицание.

«$N_3(t) = N_1(t-1) \vee N_2(t-1)$ Figure 1b

$N_3(t) = N_1(t-1) \wedge N_2(t-1)$ Figure 1c

$N_3(t) = N_1(t-1) \vee \overline{N_2(t-1)}$ Figure 1d», работа [1] стр. 19.

Указанный набор операций позволяет заключить о выполнении критерия «полноты» по Посту, что подразумевает возможность построения (вычисления) любой сложной зависимости в рамках двоичной логики. Вторым отличительным свойством является исследование вопроса построения сетей с обратными связями и без обратных связей. Все рассматриваемые элементы работают синхронно, то есть каждый такт срабатывает каждый элемент.

Модель искусственного нейрона, представленная в работе [2], представляет собой генератор сигнала, обрабатывающий входной вектор, каждая компонента которого имеет собственный вес. Обработка производится путем сравнения с некоторым пороговым значением функции активации искусственного нейрона. Отличием от представленного в работе [1] нейрона заключается в замене функции активации с логического выражения на алгебраическую (или пороговую) функцию, учет входных сигналов также не является логическим выражением конъюнкции или дизъюнкции, а представляется в виде суммы покомпонентно перемноженных векторов входного сигнала и весовых коэффициентов. Преимуществом данного подхода является возможность применения численных методов поиска весовых коэффициентов для решения задач распознавания образов, классификации прогнозирования. Условное разделение нейронов на сенсорные, реагирующие и нейроны ассоциативного слоя непринципиально, если рассматривать обобщенное строение нейронов (1).

$$y = f_A(w_1 * x_1 + w_2 * x_2 + \cdots + w_n * x_n)$$

где y – выходной сигнал, w – весовой коэффициент, x – входной сигнал, f_A – функция активации (пороговая, линейная, нелинейная и т. д.). Представленное уравнение описывает сенсорные и реагирующие нейроны с функцией пороговой активации, а так же нейроны ассоциативного слоя. Немаловажным отличием от нейрона McCulloch-Pitts`a является отсутствие переменной времени (итерации, шага) в описании работы нейрона, что возможно только для синхронного режима работы сети нейронов.

«$P_a = \sum_{e=0}^{x} \sum_{i=0}^{\min(y,e-\Theta)} P(e,i)$ (1)» работа [2] стр. 392.

Последующее исследование было направлено на изучение влияния на возможности классификации и обработки информации таких аспектов как: функция активации искусственного нейрона, увеличение количества ассоциативных слоев, введение рекуррентных связей в сеть.

Модель искусственного нейрона Widrow [4] a Adaline (ADAptive LInear Neuron) включает расширенный диапазон обрабатываемых сигналов [-1, 1] вместо [0, 1] у искусственного нейрона Rosenblatt`a. Результатом данного подхода является возможность использовать модифицированные алгоритмы обучения и повысить вычислительные возможности дифференциации образов нейронными сетями. Модели были удачно применены для решения задач распознавания речи и изображений, диагностирования на основе кардиограмм, прогнозирования, управления.

Комплекснозначный нейрон является логическим продолжением подхода к расширению диапазона значений обрабатываемых сигналами искусственных нейронных сетей и их нейронов. В работе [5] опубликован анализ влияния комлекснозначной функции активации нейрона на выход единичного нейрона. Как отмечают сами авторы, применение комплексных сигналов в ИНС позволяет моделировать и учитывать в вычислениях не только амплитуду, но и фазу сигнала активности нейрона.

Концепции нейронов, предложенные Grossberg`ом [6, 7], направлены на модификацию алгоритмов обучения отдельно взятых искусственных нейронов. Принципиальным отличием от предшественников являются разработанные автором правила обучения, позволяющие реализовать алгоритмы обучения без учителя и самоорганизацию. Структура изучавшихся нейронов описывается уравнением (1).

Математическое описание WTA-нейрона (winner take all – англ.) [8], вероятностного нейрона [9], RBF-нейрона [10] и нейрона с сигмоидальной

функцией активации отличается от рассмотренных ранее моделей используемой функцией активации. Указанные подходы имеют более высокий коэффициент эффективности с точки зрения времени обучения сети для решения задач классификации и прогнозирования, а также позволяют снизить количество нейронов в сети. Изначально предусматривался синхронный режим работы нейронов.

Помимо работ, направленных на определения степени влияния функции активации на вычислительные возможности реализуемой сети, различные исследователи проявляли интерес к влиянию функции или же n-мерной операции учета взвешенных входных сигналов. Согласно сложившемуся подходу далее для наименования данной функции будет использоваться термин агрегирующая функция (aggregation function – англ.). Результатом данного интереса стало применение в качестве указанной агрегирующей функции n-мерной операции умножения [11], что в свою очередь привело к уложению процесса учета веса синапса из-за введения в него дополнительной константы смещения. Общая формула структуры нейрона приняла вид:

$$y = f_A\big((w_1 * x_1 + c_1) * (w_2 * x_2 + c_2) * \cdots * (w_n * x_n + c_n)\big)$$

Введение неравных нулю констант позволяет избежать потери информации в процессе учета независимых компонент входного вектора в случае наличия нулевого уровня сигнала на входе. Срабатывание элементов синхронно.

Модель, предназначенная для учета особенностей связности искусственных нейронов в сети, путем обобщения входных сигналов в группы реализуется математическим аппаратом $\sum\prod$-нейронов [12]. Подход, как указано в работе [13], позволяет учитывать взаимное влияние между сигналами группы синапсов. В наиболее общем случае структура нейрона принимает вид:

$$«y = f_A\big(\sum_{k=1}^{2n} \prod_{i \in I_k} x_i\big)»$$

Вторым возможным применением концепции является определение количества групп во входном векторе: например, векторы 1100011100 и 1011100111 имеют две и три группы соответственно» [13] стр. 233. Режим работы нейронов синхронный.

Формализм комбинированного нейрона предполагает включение в структуру элемента двух агрегирующих функций исчисляющихся параллельно, после чего производится суммирование либо перемножение сигналов и вычисление результирующего сигнала в соответствии с применяемой активационной функцией. В работе [14] представлено два типа подобного рода нейронов включающих в структуру нейрона сложение и произведение результатов агрегации входных сигналов соответственно.

«$y = f_A\left(\sum_{i=1}^{k} w_i * x_i + \prod_{j=n-k}^{n} w_j * x_j\right)$» – сложение, работа [14] С. 93.

«$y = f_A\left(\sum_{i=1}^{k} w_i * x_i * \prod_{j=n-k}^{n} w_j * x_j\right)$» – произведение, работа [14] С. 93.

Режим работы сети синхронный. В работе отмечено преимущество размещения в разных слоях сети искусственных нейронов с разными типами агрегирующих функций, а также приводятся данные по результатам ускорения обучения подобного рода сетей полученные в ходе моделирования. Обобщение концепций, использующих синапсы с функциями агрегации, можно найти в работе [15]. Изначально предполагался синхронный режим работы нейрона.

Концепция адаптивного нейрона, представленная в работе [16], в дополнение к процессу параллельного вычисления на нейроне двух агрегирующих функций предусматривает реализацию некоторого набора функций активации. Данный подход позволяет добиться более высокой скорости перенастройки сети в случае изменения внешних условий и требований к способам обработки входных сигналов. Несмотря на увеличение времени обучения, метод может найти применение в системах с повышенными требованиями к отказоустойчивости. Вторым преимуществом подхода выступает снижение требований к аппаратным ресурсам нейроморфной системы в случае отсутствия необходимости в параллельной реализации вычислительных

процессов выполняемых сетью. Механизм срабатывания нейрона жестко не задан, что предполагает возможность использования асинхронного режима работы.

Модель нечеткого нейрона базируется на математическом формализме нечетких множеств. Базовая модель нечеткого нейрона [17] предусматривает использование нечетких операций агрегации над входными данными, представленными в виде элементов нечетких множеств или четких значений, или численных значений. Выход нейрона может осуществляться также в виде элемента нечеткого множества или четкого или численного значений. Математическое описание имеет вид:

«If X_{1i} and X_{2i} and … X_{ni} then Y_i нейрон с нечеткими или четкими входными значениями и логическим выражением функции активации;

$\mu(x_1, x_2, \dots, x_n) = \mu_1(x_1) \otimes \mu_2(x_1) \otimes \dots \otimes \mu_i(x_i) \otimes \dots \otimes \mu_n(x_n)$, где $\mu_i(x_i)$ – операция взвешивания входного сигнала, представляет описание нечеткого нейрона с четкими значениями входов;

$Y = X_1 \otimes X_2 \otimes \dots \otimes X_i \otimes \dots X_n$, $X_i = G_i(x_i)$ – операция взвешивания i-го нечеткого входного сигнала представляет описание нейрона с нечеткими входными сигналами. В определениях рассматриваемой модели искусственный нейрон являлся синхронным вычисляющим элементом.

Стохастический нейрон является методом описания элемента ИНС со связями каждый с каждым. Выходы нейрона связаны со всеми нейронами цепи, за исключением самого себя [18]. Данный метод позволяет путем изменения функции активации за счет введения расчета общего уровня возбуждения сети и заранее заданного параметра шума реализовать память, адресуемую по содержимому или ассоциативную память.

« $V_i \rightarrow 1$ if $\sum_{j \neq 0} T_{ij} V_j > U_i$

$Vi \rightarrow 0$ if $\sum_{j \neq 0} T_{ij} V_j < U_i$ (1)»

Режим работы нейронов сети асинхронный. В случае использования синхронного режима работы сети сходимость может не наступать, что приводит к некорректной работе.

Рассмотренные выше модели изначально предусматривали синхронный режим срабатывания нейронов; впоследствии ряд моделей были рассмотрены с позиций асинхронного режима работы сети (см. раздел «Термины и определения»). Рассматривая принципы построения указанных формализмов, перечисленные выше подходы вне зависимости от режима работы сети могут быть отнесены к классу нейронов, *срабатывающих по совпадению*. Последующие описываемые модели искусственных нейронов могут быть отнесены к классу *динамических искусственных нейронов* (см. раздел «Термины и определения»).

Модель «integrate-and-fire» впервые была предложена для описания биологического нейрона в 1907 г. [19]. В простейшем ее варианте модель связывает возбуждение нейрона с электрохимическими процессами и мембранным потенциалом нейрона. В базовом текущем варианте физическая модель определяющая уровень возбуждения через электрический ток может быть описана следующим уравнением:

$$I(t) = C * \frac{dV}{dt}$$

Срабатывание нейрона происходит в случае превышения заранее определенного порога, после чего происходит сброс в начальное состояние. Критика модели «integrate-and-fire», заключающаяся в постоянном хранении уровня текущего возбуждения с момента начала возбуждения до момента срабатывания привела к модифицированию указанной модели в модель the leaky integrate-and-fire:

«$I(t) = C * \frac{dV}{dt} + \frac{V}{R}, t > 0, V < \Theta, V(0) = 0$ (9.1) » работа [20] стр. 111.

При больших уровнях сигнала и низких уровнях порога срабатывание нейронов обеих рассмотренных моделей может происходить слишком часто, что в свою очередь может быть исправлено путем искусственного введения константы

задержки семантически отождествляемой с рефракторным периодом биологического нейрона.

Связывающий нейрон, представленный в работе [21], обобщает классы нейронов, *срабатывающих по совпадению* и *динамические нейроны*. Обобщение производится путем ведения в функцию активации дополнительной константы корректировки времени срабатывания нейрона:

« $I(t) = -mC * \frac{dCompEPSP(t)}{dt}$ (9)

$CompEPSP(t) = \sum_{k=1}^{N} EPSP(t - t_k)$ (10) » работа [21] стр. 4.

При $t\text{-}t_k \rightarrow 0$ и больших значениях агрегированного сигнала модель переходит в *нейрон, срабатывающий по совпадению*, при противоположных условиях модель модифицируется в integrate-and-fire модель. Данная модель завершает обзор основных моделей искусственных нейронов, что позволяет перейти к рассмотрению математических моделей ИНС.

Архитектуры искусственных нейронных сетей. Обзор основных архитектур. Прежде чем перейти к рассмотрению типов архитектур ИНС укажем аспекты, по которым будет производиться рассмотрение:

- *Распространение сигнала*. По типу распространения сигнала все сети могут быть разделены на сети прямого распространения FFNN (feed forward neural networks – англ.), рекуррентные сети RNN (recurrent neural networks – англ.), сети двунаправленного распространения сигнала BNN (bidirectional neural networks – англ.) и сети с распространением сигнала во всех направлениях (Hopfield). Классическим примером сети прямого распространения сигнала является «Perceptron» [2]. В данном типе сетей распространение сигнала происходит от слоя к слою, от входов нейронов сети к выходам нейронов последнего слоя. В указанном варианте обратные (рекуррентные) связи отсутствуют, что означает, что результат обработки сигналов последующими (по направлению распространения сигнала) слоями не поступает для обработки на предыдущие слои. В рекуррентных ИНС отличие от

рассмотренных выше сетей заключается в существовании обратных связей между слоями сети, т. е. наличие связей, обеспечивающих передачу обработанной информации от последующего по направлению передачи информации слоя к предыдущему. Характерным примером может служить Recurrent multilayer perceptron, представленная в работе [22]. В двунаправленных нейронных сетях нейроны предыдущего и последующего слоев обмениваются результатом своих вычислений каждую текущую итерацию (под итерацией подразумевается синхронное вычисление функции активации на нейронах). Результатом данного подхода является итеративный подбор минимума возбуждения на нейронах сети, который и является результатом вычислений, либо наличие обратной связи для корректного учета временной корреляции сигналов. В качестве примера указанных сетей выступает работа [23]. Примером сети с распространением сигналов по всем направлениям выступает сеть Hopfield`a [18].

- *Шаблон связность сети*. Под типом связности сети (шаблоном связности) подразумевается способ задания связей между нейронами. Возможны два варианта: случайный заданный набор связей и набор связей, задаваемый преднамеренно исследователем или алгоритмом (параметрический шаблон связности).

- *Связи сети*. С позиций применяемых связей архитектуры могут быть разделены на архитектуры с двунаправленными связями и архитектуры с однонаправленными связями.

Первой предложенной работоспособной архитектурой ИНС является «Perceptron» [2] данная архитектура рассматривалась ранее. Является сетью прямого распространения сигнала со случайным шаблоном связности между слоями и однонаправленными связями.

Сеть Madaline помимо входного слоя нейронов имеет один или несколько слоев из нейронов Adaline [4]. Первая сеть, получившая аппаратную реализацию. Первоначальный вариант построения архитектуры можно характеризовать как

сеть прямого распространения сигнала со случайным шаблоном связности сети и однонаправленными связями. Решает задачу распознавания и классификации образов, прогнозирования. Вычислительные возможности разработанной архитектуры позволяют реализовать с приемлемой точностью: прогнозирование погоды, распознавание речи, диагностирование по кардиограммам и системы адаптивного управления.

Архитектура Multilayer Perceptron [24] является структурной модификацией Perceptron с более чем одним слоем ассоциативных нейронов. Введение более одного слоя нейронов было обусловлено ограниченностью модели в классификации линейно неразделимых образов. Примером задачи может служить невозможность выполнения XOR на одном ассоциативном слое Perceptron. Введение дополнительных ассоциативных слоев позволяет решить указанную проблему, но существенно увеличивает время обучения сети. Изначально относится к сетям прямого распространения сигнала с однонаправленными связями и случайным шаблоном связности. Существуют модификации сети с рекуррентным распространением сигнала [23] и параметрическим шаблоном связности.

Самоорганизующиеся карты Кохонена SOM (self-organizing Map – англ.) в элементарном варианте исполнения представляет собой однослойную нейронную сеть [25]. В элементарном варианте сеть использует случайный шаблон связности между входами сети и слоем нейронов, сеть относится к сетям прямого распространения сигнала с однонаправленными связями. Сеть решает параметрическую задачу кластеризации векторов многомерных входных данных исходя из заранее заданного параметра ошибки. Возможны модификации архитектуры с более чем одним слоем нейронов [26].

Архитектура сети Хопфила [18] относится сети распространения сигнала по всем направлениям и состоит из одного слоя нейронов. Схема организации шаблона связности является параметрически задаваемой. В соответствии с алгоритмом задания связей выход каждого нейрона H1 подается на вход всех

нейронов сети, кроме самого нейрона H1. В качестве выхода сети определяется текущее состояние нейронов, вычисление считается завершенным при достижении сети устойчивого неизменного состояния возбуждения на нейронах. Механизм работы нейронов асинхронный, применение синхронного механизма может приводить к поочередной сменен состояний сети с двумя уровнями возбуждения, что расценивается как некорректная работа. Тип применяемых связей сети двунаправленные. На базе указанной архитектуры реализуются память, адресуемая по содержанию CAM [27] (content addressable memory – англ.), автоассоциативная память [23] и многие другие решения, базирующиеся на кластеризации и классификации.

Машина Больцмана BM [28] (Boltzmann machines – англ.). Архитектура является модификацией сети Хопфилда и относится к сетям двунаправленного распространения сигнала. Функционально нейроны сети делятся на входные (они же выходные) и скрытые нейроны. Шаблон связности сети является параметрически задаваемым и повторяет шаблон связности сети Hopfield`a. Тип используемых связей сети двунаправленные связи. Функция активации нейронов предполагает включение константы теплового шума сети, что позволяет в процессе обучения сети избегать локальных минимумов. Рассматриваемая архитектура имеет приложения в областях комбинаторики [29], распознавание речи [30] и многих других [31]. К недостаткам архитектуры обычно приводят временные затраты на обучение сетей с большим количеством нейронов.

Ограниченная Машина Больцмана RBM [32] (Restricted Boltzmann Machines – англ.). Предлагаемый в архитектуре подход был направлен на снижение затрат на обучение сети и модификацию алгоритма обучения путем изменения шаблона связности BM. Все нейроны, так же как и в Машине Больцмана, разбиваются на скрытые и входные. Шаблон связности сети задается параметрически по правилу: все нейроны, скрытые или входные имеют связи только с нейронами противоположного типа, т. е. каждый скрытый нейрон связан только со всеми входными, а каждый входной связан только со скрытыми нейронами.

Используемые связи двунаправленные. Схема распространения сигнала соответствует двунаправленным нейронным сетям. Сеть нашла применение в задачах выявления признаков, снижение размерности данных, фильтрация данных, классификации и многих других [33].

Рассмотренные выше сети могут считаться базовыми примерами построения ИНС. Представленные далее сети будут базироваться на указанных архитектурах с блочной структурой построения обработки информации, и иметь FFNN схему обработки сигнала. Особо стоит отметить, что различные блоки могут иметь одну и ту же базовую архитектуру с различиями в функциях, реализуемых на разных блоках ИНС. Например, архитектура Perceptron или Multi Layer Perceptron использующая нейроны ассоциативного слоя с RBF (Radial Basis Function – англ.) функциями активации [34]. Как правило, концепция воплощается различным параметрическим заданием шаблона связности и различными функциями активации для нейронов разных блоков, также возможен подход, предусматривающий отличные друг от друга алгоритмы обучения для разных блоков сети.

Иерархические нейронные сети HNN (Hierarchical neural networks – англ.) [35]. Сети прямого распространения сигнала FFNN. Отличительной особенностью является параметрический шаблон связности. В рамках шаблона сигналы от входов ИНС поступают на набор подсетей, при этом напрямую запрещено использование полной связности (full connection – англ.) между всеми слоями, т. е. существует хотя бы один слой, в котором сигналы от всех нейронов предыдущего слоя не подаются на все нейроны последующего. В слое нейронов, состоящем из различных подсетей, нейроны одной подсети могут иметь отличное время срабатывание от нейронов другой подсети. Время срабатывания нейронов в одной подсети всегда одинаково. Сети демонстрируют меньшую ошибку при обработке входных сигналов и требуют меньше ресурсов для обучения сети.

Neocognitron [36]. Сеть относится к сетям прямого распространения сигнала HNN. Используемый тип связей – однонаправленные связи. Шаблон связности

является параметрическим. Каждый блок сети, кроме входного, состоит из двух слоев: входного и ассоциативного. Первый слой каждого блока, кроме входного блока сети, выполняет операцию свертки (Convolutions – англ.) обобщая входной вектор от предыдущего слоя на одном из нейронов. Второй слой всех блоков кроме входного и последнего блока осуществляет операцию выборки из результатов свертки (Subsampling – англ.), результатом которой служит максимальное возбуждение группы нейронов подвыборки с наиболее релевантными входному сигналу весовыми коэффициентами. Последний слой последнего блока сети является ассоциативным и предназначен для операции субдискретизации (downsampling – англ.), позволяющей классифицировать релевантный текущему входному сигналу образ. Изменению весовых коэффициентов подвержены только входные связи первого слоя каждого блока. Существуют отличные от Neocognitron модификации, использующие для операций субдискретизации WTA нейроны или другие локальные правила [37].

Cresceptron [38], HNN. Является модификацией Neocognitron. Отличие состоит в использовании двух новых типов блоков. Первый блок – Max-Pooling, выполняет операцию субдискретизации путем выбора наибольших значений из сигналов на выходе слоя свертки. Второй тип блока – Blurring [39] осуществляет развертку входного сигнала от предыдущего слоя на последующий слой, позволяя повысить количество активных откликов от Subsampling перед очередной сверткой или субдискретизацией.

Клеточные нейронные сети (Cellular Neural Networks – англ.). Концепция построения сети переставляет из себя объединение таких направлений обработки информации, как клеточные автоматы КА и HNN. Тип распространения сигнала в сети – распространение по всем направлениям. В качестве вычисляющего элемента ВЭ выступает нейрон с наперед задаваемой алгебраической (не логической, в отличие от КА) функцией активации. ВЭ располагаются в узлах решетки, аналогично ВЭ клеточного автомата. Шаблон связности является параметрически задаваемым и фактически определяет собой тип решетки. Тип

используемых связей в основном однонаправленные, но возможно применение и двунаправленных связей [40]. Архитектура имеет широкое применение в области обработки изображений в задачах детектирования границ, поиска связных областей, распознавания зашумленных образов и их коррекции [41].

Нейронные сети кодирования информации. Относятся к сетям FFNN. Autoencoder [42], Sparse Autoencoder [43], Denoising Variational Auto-Encoding [44]. Являются сетями прямого распространения сигнала FFNN. Указанные архитектуры строятся из трех типов слоев: входной слой, слой обработки информации и выходной слой. Шаблон связности задается параметрически и является, как правило, полносвязным между слоями. Тип связей однонаправленные. Основное отличие архитектур выражается в слое обработки информации и выходном слое. В архитектуре Autoencoder слой обработки слой осуществляет представление входного вектора данных вектором меньшей размерности, выходной слой предназначен для восстановления исходного вектора. В архитектуре Sparse Autoencoder слой обработки информации представляет входной вектор в виде активности на группе своих нейронов, при этом на нейронах, не относящихся к группе, в процессе обучения сети активность сводится к минимуму. Выходной вектор представляет набор выявленных признаков. Denoising Variational Auto-Encoding получает на вход данные с ошибкой, после чего слой обработки информации корректирует образ и выдает на выход не зашумлённые обобщенные данные образа.

Глубинные ИНС доверия DBN (Deep Belief Networks – англ.) [45]. Архитектура относится к сетям двунаправленного распространения сигнала с параметрически задаваемым шаблоном связности и двунаправленными связями. Структура сети переставляет собой последовательно соединенные друг с другом блоки RBM. Обучение сети производится послойно. Вначале обучается первый скрытый слой нейронов до приемлемого показателя ошибки, после обучения слоя его веса замораживаются и происходит обучение последующего слоя. Сети нашли применение в задачах распознавания изображений [46], акустического

моделирования [47] и многих других задач классификации, распознавания и кластеризации. Данная архитектура завершает рассмотрение основных моделей нейронных сетей. Широко применяемые на текущий момент сверточные сети CNN (Convolutional Neural Networks – англ.) или развертывающие нейронные сети DN [48] (Deconvolutional Network – англ.) представляют собой комбинацию из рассмотренных в обзоре блоков. В качестве примера построения архитектуры CNN может выступать рассмотренный ранее Cresceptron.

Отказоустойчивость архитектуры ИНС. В качестве одного из положительных качеств параллельной обработки информации искусственными нейронными сетями традиционно приводят их высокую отказоустойчивость. Тем не менее, данный фактор требует дополнительного рассмотрения. Причиной рассмотрения данного вопроса является возможность выхода из строя элемента выходного слоя WTA нейронов, что в свою очередь, несомненно, приведет к некорректной работе сети в целом [49]. Аналогично и для архитектуры Cellular Neural Networks будут справедливы варианты некорректной обработки информации свойственные клеточным автоматам [50, 51]. Сложность построения отказоустойчивых архитектур на основе рассматриваемых моделей напрямую следует из их структуры и решаемых системами задач. Специфичность алгоритмов построения указанных систем не позволяет напрямую определить количество требуемых элементов для сохранения заданных показателей надежности на должном уровне, что приводит к сегментации не только по классам алгоритмов, но и по объему и однородности используемой информации [52]. Результатом является необходимость решения сложной многопараметрической задачи оптимизации архитектуры, включающей модификацию алгоритмов обучения и обработки данных, шаблон связности и представление информации в сети.

Модификация алгоритмов обучения ИНС рассматривается в работах [49, 53]. В частности, приводится пример модификации правила изменения весовых коэффициентов:

« $\Delta w^{total}_{ij} = \Delta w^{normal\_terms}_{ij} + \alpha * \Delta w^{extra\_terms}_{ij}$, где $\Delta w^{normal\_terms}_{ij}$ – приращение связи нормальных условий, $\Delta w^{extra\_terms}_{ij}$ – приращение связи из необходимости отказоустойчивости»

Преимуществом модификации алгоритмов является возможности их последующего применения к другим типам сетей использующих данные алгоритмы для обучения.

В работе [54] исследуется вопрос отказоустойчивости ИНС при их аппаратной реализации. В статье представлен анализ, показывающий невозможность достижения необходимого уровня отказоустойчивости систем ИНС на базе VLSI без внесения значительного количества дополнительных элементов; тем не менее, данных подход позволяет избежать подхода с полным резервированием блоков. К существующим методам оптимизации архитектуры для целей повышении наработки до отказа можно отнести использование при проектировании нейроморфных систем, генетических алгоритмов [55] и специализированного ПО [56].

Выводы из обзора литературы моделей нейронов и архитектур искусственных нейронных сетей. Тенденции изучения ИНС последних лет сместили фокус исследователей от поиска новых моделей нейронов и построения небольших (до сотни нейронов) архитектур на их основе в область построения смешанных сложных сетей. Как правило, указанные сети строятся из нескольких видов базовых архитектур, функционально реализующих различные блоки. Данный подход позволяет чередовать (DBN), повторять (CNN) различные типы блоков для получения более абстрактного представления или же выделения признаков обрабатываемых данных (DN). В качестве причин указанного перехода могут служить рост вычислительной мощности систем доступных для исследователей, а также создание новых алгоритмов обучения сетей требующих существенно меньше ресурсов для обучения сети. На текущем уровне развития теория ИНС, включая методы и подходы к их имплементации, позволяет заключить следующее:

1. В ходе анализа литературы установлено, что модель нейрона с динамической функцией активации, т. е. функции активации изменяющейся в процессе работы или обучения сети, исследователями ранее не предлагалась и не рассматривалась.
2. В концепциях существующих моделей искусственных нейронов можно выделить три функционально различных блока обработки информации: синапс (блок учета вклада сигнала от входа нейрона в общий уровень выходного сигнала), функцию агрегации (блок группового учета взвешенных входных сигналов) и функцию активации (блок генерации выходного сигнала). Аксон нейрона – блок подстройки выходного сигнала под последующий нейрон, см. концепцию Outstar Grossberg, может быть объединен с синапсом в единый блок, учитывающий вклады обоих блоков.
3. Функциональные блоки искусственных нейронов могут иметь сложную структуру. Например, синапсы ΣΠ-нейронов, агрегационная функция комбинированного нейрона, функция активации интегрирующих нейронов и связывающего нейронов.
4. Современная концепция построения искусственных нейронных сетей предполагает поблочную обработку информации, в которой блоки выполнены в виде слоев или подсетей различной архитектуры. Типы применяемых в блоках нейронов могут различаться от блока к блоку и иметь различные функции активации, функции агрегации и строение синапсов. Данный факт приводит к необходимости обобщения ранее разработанных моделей искусственных нейронов в каждой последующей предлагаемой модели. Например, комбинированный нейрон обобщается адаптивным нейроном, связывающий нейрон обобщает интегрирующие нейроны и срабатывающие по совпадению нейроны.
5. Несмотря на обилие методов моделирования и имплементации ИНС программными средствами, подход ориентированный на воплощение ИНС аппаратными средствами должен иметь большие показатели эффективности

с точки зрения затрат ресурсов, в сравнении с программными концепциями. В данном подходе стоит предусмотреть избыточность состояний, при обучении сетей, для повышения работоспособности систем в случае отказа вычислительных элементов.

Все вышеперечисленные пункты должны быть учтены при разработке модели искусственного нейрона с динамической функцией активации.

1.2 Математическая модель конечного автомата абстрактного нейрона

В данном разделе демонстрируется авторская модель нейрона с динамической функцией активации. С учетом пункта 5 *Выводов из обзора литературы моделей нейронов и архитектур искусственных нейронных сетей* математический формализм ориентирован на последующее описание высокоуровневой модели искусственного нейрона средствами схемотехническими моделирования.

Обобщенное абстрактное представление нейрона согласно пунктам 2, 3 *Выводов из обзора литературы моделей нейронов и архитектур искусственных нейронных сетей* отображено на схеме Рисунок 1. Блок учета вклада сигнала предполагает обобщенное представление функций синапса и аксона единым блоком учитывающим вклад аксона и синапса в уровень активации нейрона. Блок агрегации входных сигналов реализует функцию агрегации. Блок генерации выходного сигнала обеспечивает реализацию функции активации нейрона.

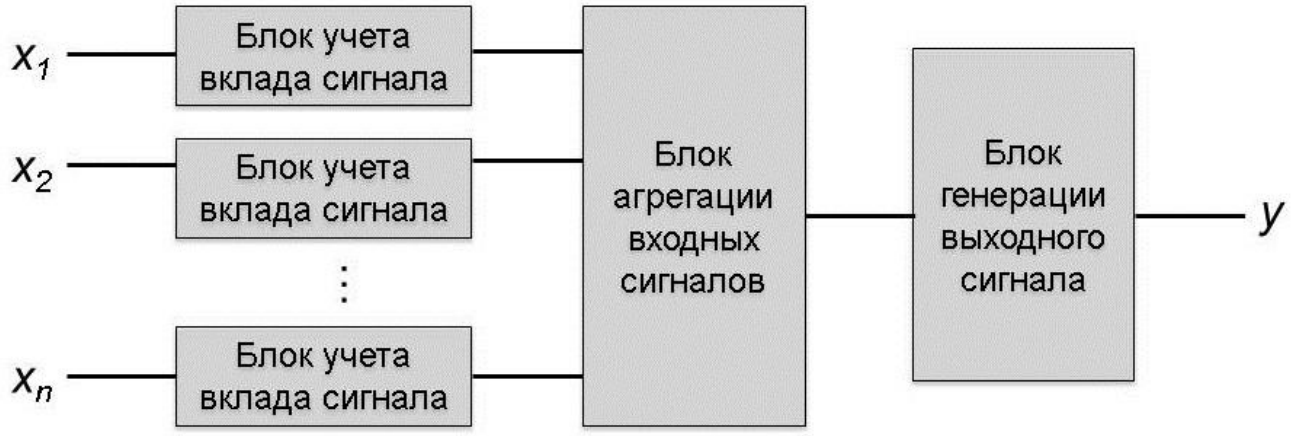


Рисунок 1. Обобщенное представление искусственного нейрона. x1-xn – входные сигналы, y – выходной сигнал нейрона.

Входные сигналы x_1-x_n, пройдя последовательную обработку на нейроне, приводят к генерации выходного сигнала y, после чего сигнал с выхода нейрона подается на входы блоков учета входного сигнала последующих нейронов.

Обобщение синапса и аксона. Представление информации в математических моделях ИНС осуществляется путем постановки в соответствие им числовых значений. При этом могут использоваться все существующие множества чисел: бинарное логическое множество {0, 1} (в нейронах McCulloch-Pitts), множество целых чисел $\mathbb{Z}$ (в сетях Madaline, Perceptron), множество вещественных чисел $\mathbb{R}$, (стохастический нейрон) и множество комплексных чисел $\mathbb{C}$. Превалирующим подходом к математическому выражению учета вклада сигнала на аксоне выступает операция умножения на весовой коэффициент. Для синапсов возможны различные варианты, например: помимо операции умножения на весовой коэффициент связи, синапсы могут иметь более сложное строение включающее константу смещения (мультипликативный нейрон). В случае применения бинарного множества сигналов между нейронами так же используется операция умножения, что приводит к умножению бинарного сигнала на целочисленный весовой коэффициент.

Поскольку для всех рассматриваемых множеств чисел заданы операции умножения и сложения, относительно которых справедливы аксиомы дистрибутивности, коммутативности и дистрибутивности умножения относительно сложения, возможно задание обобщенного коэффициента для учета вклада сигнала, что является тривиальным. Покажем это, пусть x – численное значение сигнала, w_a – численное значение коэффициента вклада аксона для данной связи, w_s – численное значение коэффициента вклада синапса для данной связи, c_s – численное значение константы смещения для учета вклада синапса данной связи, c_a – численное значение константы смещения для учета вклада аксона данной связи. Для случая с простыми синапсами и простым аксоном справедливо: $w_s(xw_a) = w_{sa}x$, где w_{sa} – результат перемножения коэффициентов w_s и w_a. Для случая с простым аксоном и сложным синапсом справедливо: $w_s(xw_a) + c_s = w_{sa}x + c_s$. Для полноты описания рассмотрим также случаи применения сложных аксонов, несмотря на то, что в моделях данные методы не применялись. Для случая со сложным аксоном и простым синапсом справедливо: $w_s(xw_a + c_a) = w_{sa}x + c_{aw}$, где c_{aw} – константа смещения, помноженная на

весовой коэффициент синапса. В случае сложных синапсов и аксонов зависимость выражается следующим соотношением: $w_s(xw_a + c_a) + c_s = w_{sa}x + c_{aws}$, где c_{aws} – константа смещения, $c_{aws} = c_{aw} + c_s$. Таким образом, можно заключить, что представление аксона и синапса нейрона может быть сведено к применению единого блока учета вклада сигнала для сигналов, представляемых на любом множестве чисел. Данный блок будет иметь сложное строение для множеств целых и комплексных чисел. Для комплексных чисел это обуславливается невозможностью сравнения комплексных чисел, а для целых, спецификой операции умножения. В случае использования для обозначения сигналов множества вещественных чисел, блок учета вклада входного сигнала может быть сведен к операции умножения на коэффициент путем прямого запрета использования в качестве сигналов 0, что тривиально и следует из аксиомы непрерывности $\mathbb{R}$: $w_s(xw_a + c_a) + c_s = w_{sa}x + c_{aws} = w_{sac}x, x \neq 0$, где w_{sac} – искомая константа.

Функции агрегации искусственных нейронов. Из существующих моделей искусственных нейронов по функциям агрегации входных сигналов могут быть выделены следующие типы: суммирующий, мультипликативный, ΣΠ-нейрон, агрегирующий, комбинированный, адаптивный и нейрон McCulloch-Pitts`a.

ΣΠ-нейроны используют сложную функцию агрегации, позволяющую выявить наличие сигналов на соседних синапсах, что часто реализуется путем попарного перемножения сигналов между собой. Аппаратная реализация указанной концепции потребует резкого увеличения площади на кристалле, как за счет реализации блока умножения сигналов, так и за счет увеличения количества складываемых перемноженных между собой сигналов, что приведет к увеличению площади занимаемой сумматором. Указанный недостаток справедлив для нейронов McCulloch-Pitts`a и агрегирующего типа. Для нейрона McCulloch-Pitts`a агрегация выполняется произвольной логической функцией от входов, что потребует значительного использования площади кристалла. Поскольку нейрон агрегирующего типа является обобщением радиальных, ΣΠ и Розенблаттовских

нейронов требуемая для имплементации блока агрегации площадь будет значительно больше.

Комбинирующий нейрон и адаптивный нейрон потребуют удвоения занимаемой на кристалле площади, так как функция агрегации нейронов обоих типов предусматривает параллельное выполнение перемножения и сложения сигналов от всех входов. Мультипликативные нейроны обладают сравнительно низкими требованиями к реализации блока агрегации входных сигналов и имеют широкое распространение для решения ряда комбинаторных и оптимизационных задач. Суммирующие нейроны – это нейроны, использующие для агрегации n-мерную операцию сложения, обладают самыми низкими требованиями к площади на кристалле и являются превалирующей моделью среди всех типов.

Из рассмотренного выше следует, что при построении модели искусственного нейрона с динамической функцией активации следует обобщить функции агрегации сложения и умножения в рамках одного блока агрегации. Реализация подхода должна исключать дублирование вычислительных блоков быть реализована в рамках математической модели.

Описание искусственного нейрона в терминах теории множеств. Для осуществления последующего анализа введем описание нейрона наборами числовых множеств. Искусственный нейрон представляет собой элемент обработки информации и имеет не менее одного входа и одного выхода. Количество связей искусственных нейронов, но при этом не может иметь не целое или отрицательное количество и очевидно менее чем счетно, что в свою очередь ведет к его описанию натуральным числом. Поскольку использование комплексных чисел не позволяет применять операцию сравнения, последующее описание искусственных нейронов будет осуществляться множеством вещественных и целых чисел. Множество выходных и входных сигналов нейрона при аппаратной реализации будет представляться некоторым диапазоном физических величин, что может быть описано на множестве $\mathbb{R}$. Запрет на представление входных и выходных сигналов «0», как отмечалось ранее, имеет преимущества при описании блока учета вклада сигнала с помощью операции

умножения над множеством чисел. Поскольку базовая модель нейрона предусматривает использование блока учета входного сигнала, возможно заменить математическую операцию функцией учета входного сигнала δ. Множество весовых коэффициентов синапса удобно представлять на множестве $\mathbb{R}$. Помимо весовых коэффициентов синапса необходимо ввести множество подгоночных коэффициентов функции активации, представляемом на множестве $\mathbb{R}$, а также функцию агрегации Δ и функцию активации λ. Из всего вышеперечисленного один из способов описания абстрактного искусственного нейрона представляется набором вида (n, Y, X, W, C, δ, Δ, λ), где:

$n \in \mathbb{N}$ – число входов (синапсов);

$Y \subseteq \mathbb{R}$ – множество выходных сигналов нейрона;

$X = \bigcup_{i=0}^{n} X_i, X \subseteq \mathbb{R}$ – множество входных сигналов, подаваемых на синапсы искусственного нейрона;

$W = \bigcup_{i=0}^{n} W_i, W \subseteq \mathbb{R}$ – множество всех возможных значений весовых коэффициентов синапса;

$C \subseteq \mathbb{R}$ – множество подгоночных коэффициентов функции активации;

δ – функция учета входного сигнала, традиционно применяется операция умножения $(w_i, x_i) = w_i * x_i, w_i \in W, x_i \in X$;

Δ – функция агрегации входных сигналов $\sum_{i=1}^{n}(w_i * x_i)$ или $\prod_{i=1}^{n}(w_i * x_i + c_i), c_i \in C, w_i \in W, x_i \in X$;

λ – функция активации искусственного нейрона.

Здесь и далее $\mathbb{N} \cap 0 = \emptyset$.

Анализ суммирующего и мультипликативного нейронов

Зависимость сигнала на выходе для суммирующего нейрона имеет вид

$$y = \lambda(\sum_{i=1}^{n} w_i * x_i + c), c \in C \quad (1)$$

а для мультипликативного нейрона описывается

$$y = \lambda(\prod_{i=1}^{n}(w_i * x_i + c_i)), c_i \in C \quad (2)$$

С учетом введенного ранее описания, имеем следующие зависимости для абстрактного, суммирующего и мультипликативного нейронов соответственно:

$$y = \lambda_a\left(\Delta_{i=1}^{n}\left(\delta(w_i, x_i)\right)\right), \delta\colon \mathbb{R}^2 \to \mathbb{R}, \Delta\colon \mathbb{R}^n \to \mathbb{R}, \lambda_a\colon \mathbb{R} \to Y \quad (3)$$

$$y = \lambda_s(\sum_{i=1}^{n} w_i * x_i + c), \delta\colon \mathbb{R}^2 \to \mathbb{R}, \Sigma\colon \mathbb{R}^n \to \mathbb{R}, \lambda_s\colon \mathbb{R}^2 \to Y \quad (4)$$

$$y = \lambda_m(\prod_{i=1}^{n}(w_i * x_i + c_i)), \delta\colon \mathbb{R}^3 \to \mathbb{R}, \Pi\colon \mathbb{R}^n \to \mathbb{R}, \lambda_m\colon \mathbb{R} \to Y \quad (5)$$

Первым условием эквивалентности автоматов M_1, M_2 является равенство алфавитов входных $X_1 = X_2$ и выходных символов $Y_1 = Y_2$. Вторым условием эквивалентности является получение одинаковых выходных символов $y_1 = y_2, y_1 \in Y_1, y_2 \in Y_2$ при задании на входе автомата одинаковых входных символов $x_1 = x_2, x_1 \in X_1, x_2 \in X_2$ для всех элементов множества входных символов.

Аналогично будем считать эквивалентными модели искусственных нейронов, для которых выполняются следующие условия: множества входных символов (сигналов, элементов) равны; множества выходных символов (сигналов, элементов) равны; при подаче на два эквивалентных нейрона любого входного символа, их выходные символы равны.

Приведенное уравнение (2) не совпадает с описанием абстрактного нейрона (3) из–за наличия в нем констант, приведем его к требуемому виду.

Первым подходом является запрет обозначения входных и выходных сигналов «0», что по аксиоме о полноте $\mathbb{R}$ приводит к тривиальному подбору весового коэффициента синапса:

$$\left(X \subseteq (\mathbb{R}\backslash 0)\right) \wedge \left(Y \subseteq (\mathbb{R}\backslash 0)\right) \wedge (\forall x_i \neq 0) \Rightarrow \exists \widetilde{w}_i \in \mathbb{R}(\widetilde{w}_i * x_i = w_i * x_i + c_i) \Longrightarrow$$

$$\Longrightarrow \prod_{i=1}^{n}(w_i * x_i + c_i) = \prod_{i=1}^{n} \widetilde{w}_i * x_i = y, y \in Y$$

Второй подход заключается в задании специальной функции учета входного вклада сигнала. Поскольку $\mathbb{R}$ непрерывно (аксиома о полноте $\mathbb{R}$), из (2) для случая $\forall x_i \neq 0$ справедливо:

$$\exists \widetilde{w}_i \in \mathbb{R}(\widetilde{w}_i * x_i = w_i * x_i + c_i) \Longrightarrow$$

$$\Longrightarrow \prod_{i=1}^{n} \widetilde{w}_i * x_i = \prod_{i=1}^{n}(w_i * x_i + c_i) = y_k, \forall y_k \in Y_k, Y_k \subseteq Y$$

Тогда для выполнения условий равенства функций активации

$$\prod_{i=1}^{n} \widetilde{w}_i * x_i = \prod_{i=1}^{n}(w_i * x_i + c_i)$$

при $\exists x_i = 0$, перезапишем функции учета входного сигнала δ в ином виде

$$y = \lambda_{m1}\left(\prod_{i=1}^{n}(w_i * x_i + c_i)\right) = \lambda_{m1}\left(\prod_{i=1}^{n}\delta(\widetilde{w}_i, x_i)\right), \delta(\widetilde{w}_i, x_i) = \begin{cases}\widetilde{w}_i * x_i, x_i \neq 0 \\ c_i, x_i = 0\end{cases}$$

Третий метод состоит в построении специальной функции агрегации и модификации функции активации нейрона. Условиями равенства функций f_1 , f_2 являются:

$$Dom\ f_1 = Dom\ f_2$$

$$ran\ f_1 = ran\ f_2$$

$$\forall x_i \in Dom\ f_1\left(\left(f_1(x_i) = f_2(x_i)\right) \wedge \left(\left(\bigcup_i x_i\right) \cap Dom f_1 = \emptyset\right)\right).$$

Представим входной сигнал в виде многокомпонентного вектора $\mathbf{x_{in}}$, а весь набор сигналов в виде множества $\mathbf{X_{in}} = \cup\, \mathbf{x_{in}}$. Весовые коэффициенты синапсов обозначим в виде равного по размерности вектору $\mathbf{x_{in}}$ многокомпонентного вектора $\widetilde{\mathbf{w}}_{\boldsymbol{in}}$, образующего множество всех возможных векторов весовых коэффициентов $\widetilde{\mathbf{W}}_{\boldsymbol{in}} = \cup\, \widetilde{\mathbf{w}}_{\boldsymbol{in}}$ такое, что будет выполняться условие $|X \times W \times C| = \left|\widetilde{\mathbf{W}}_{\boldsymbol{in}} \times \mathbf{X_{in}}\right|$ и порядок на множествах будет совпадать. Тогда возможно задание функции агрегации $\Delta: \left|\widetilde{\mathbf{W}}_{\boldsymbol{in}} \times \mathbf{X_{in}}\right| \to \mathbb{R}$. Поскольку порядок для множеств совпадает и функция активации не имеет специальных ограничений к своему виду (гладкость, непрерывность и т.д.), можно ее задать в другом виде λ_{m2}, таком что:

$$\forall x_{in}, \exists y_j\left(y_j \middle| \left(y_j = \lambda_{m1}\left(\prod_{i=1}^{n} w_i * x_i + c_i\right)\right) \wedge \left(\bigcup_j y_j \subseteq Y\right) \wedge (\exists x_i = 0)\right) \Rightarrow$$

$$\Rightarrow \lambda_{m1}\left(\prod_{i=1}^{n}(w_i * x_i + c_i)\right) = \lambda_{m2}(\Delta\langle \widetilde{\mathbf{w}}_{\boldsymbol{in}}, \mathbf{x_{in}}\rangle)$$

поставив в соответствие значению функции $\lambda_{m2}(\Delta\langle \widetilde{\mathbf{w}}_{\boldsymbol{in}}, \mathbf{x_{in}}\rangle)$ соответствующее значение функции λ_{m1}, для каждого $\mathbf{x_{in}}$ при $\forall x_i \neq 0$ получим функцию вида

$$y = \lambda_{m2}(\Delta\langle \widetilde{\mathbf{w}}_{\boldsymbol{in}}, \mathbf{x_{in}}\rangle) = \begin{cases} y_j, \exists x_i = 0, \forall y_j \in Y_j, Y_j \subseteq Y \\ y_k, \forall x_i \neq 0, \forall y_k \in Y_k, Y_k \subseteq Y \end{cases}, \left(Y_k \cup Y_j\right) \cap Y = \emptyset$$

Сведем, таким образом, описание (5) к (3).

Приведенное уравнение (1) не совпадает с описанием абстрактного нейрона (3) из–за наличия в нем константы, приведем его к требуемому виду.

Теорема 1. *Для любой модели суммирующего искусственного нейрона с определенными на* $\mathbb{R}$ *множествами входных* X *, выходных* Y *и внутренних сигналов может быть построена эквивалентная ей модель суммирующего нейрона без использования константы смещения.*

Доказательство. По условию $n \geq 1$, следовательно:

$$y = \lambda_{s1}\left(\sum_{i=1}^{n} w_i * x_i + c\right) = \lambda_{s1}\left(\sum_{i=1}^{n}\left(w_i * x_i + \frac{c}{n}\right)\right)$$

Поскольку $\mathbb{R}$ непрерывно (аксиома о полноте $\mathbb{R}$), из (2) для случая $\forall x_i \neq 0$ справедливо:

$$\forall x_i \neq 0, \exists \widetilde{w}_i \in \mathbb{R}\left(\widetilde{w}_i * x_i = w_i * x_i + \frac{c}{n}\right) \Longrightarrow$$

$$\Rightarrow \lambda_{s1}\left(\sum_{i=1}^{n}(\widetilde{w_i} * x_i)\right) = \lambda_{s1}\left(\sum_{i=1}^{n}\left(w_i * x_i + \frac{c}{n}\right)\right) = y_k, \forall x_i \neq 0, \forall y_k \in Y_k, Y_k \subseteq Y$$

Задав функцию λ_{s2} ставящую в соответствие значения функции λ_{s1} при $\exists x_i = 0$ и $\forall x_i \neq 0$, аналогично рассмотренному выше способу, получим искомую модель и сведем описание (4) к описанию (3):

$$y = \lambda_{s1}\left(\sum_{i=1}^{n}\left(w_i * x_i + \frac{c}{n}\right)\right) = \lambda_{s2}\left(\sum_{i=1}^{n}(\widetilde{w_i} * x_i)\right)$$

$$y = \lambda_{s2}\left(\sum_{i=1}^{n}(\widetilde{w_i} * x_i)\right) = \begin{cases} y_j, \exists x_i = 0, \forall y_j \in Y_j, Y_j \subseteq Y \\ y_k, \forall x_i \neq 0, \forall y_k \in Y_k, Y_k \subseteq Y \end{cases}, \left(Y_k \cup Y_j\right) \cap Y = \emptyset$$

Теорема доказана.

Докажем эквивалентность мультипликативного и суммирующего нейронов.

Теорема 2. *Для любой модели мультипликативного искусственного нейрона с определенными на* $\mathbb{R}$ *множествами входных* X*, выходных* Y *и весовых коэффициентов может быть построена эквивалентная ей модель суммирующего нейрона.*

Доказательство. По определению эквивалентности, множества входных сигналов $X_s = X_m$ и выходных сигналов $Y_s = Y_m$ искусственных нейронов равны, где индексы s, m — индексы множеств суммирующего и мультипликативного нейронов соответственно.

Из рассмотренного выше уравнения моделей искусственных нейронов можно переписать в следующем виде:

$y = \lambda_s(\sum_{i=1}^{n}(w_i * x_i))$ (6) $y = \lambda_m(\Delta\langle \widetilde{\mathbf{w}}_{\boldsymbol{in}}, \mathbf{x}_{\boldsymbol{in}} \rangle)$. (7)

Из плотности $\mathbb{R}$ при $\forall x_i \neq 0$ выполняется

$$\exists \widetilde{w_i} \in \mathbb{R}, \left(\lambda_s \left(\sum_{i=1}^{n} \widetilde{w_i} * x_i \right) = \lambda_m \left(\prod_{i=1}^{n} w_i * x_i \right) \right)$$

Задав функцию λ_s и поставив в соответствие значения функции при $\exists x_i = 0$ значениям функции активации λ_m при $\exists x_i = 0$, получим искомую модель

$$y = \lambda_s \left(\sum_{i=1}^{n} (\widetilde{w_i} * x_i) \right) = \lambda_m \left(\prod_{i=1}^{n} w_i * x_i \right)$$

$$y = \lambda_s \left(\sum_{i=1}^{n} (\widetilde{w_i} * x_i) \right) = \begin{cases} y_j, \exists x_i = 0, \forall y_j \in Y_j, Y_j \subseteq Y \\ y_k, \forall x_i \neq 0, \forall y_k \in Y_k, Y_k \subseteq Y \end{cases}, (Y_k \cup Y_j) \cap Y = \emptyset$$

Теорема доказана.

Обратное построение эквивалентной модели мультипликативного нейрона для модели суммирующего также возможно, ввиду отсутствия ограничений на тип функции и равенства множеств входных и выходных сигналов.

Приведенные уравнения (6) и (7) могут рассматриваться как частные случаи модели абстрактного нейрона (3). Эквивалентность моделей искусственных нейронов, включающих переменные определенные на $\mathbb{R}$, не гарантирует их эквиватентности при определении всех их переменных на $\mathbb{N}$ ввиду отсутствия у $\mathbb{N}$ непрерывности. Учитывая условие ограниченности любых аппаратных реализаций, докажем эквивалентность моделей определенных на конечных множествах элементов из $\mathbb{N}$.

Теорема 3. *Для любой модели мультипликативного искусственного нейрона с определенными на* $\mathbb{N}$ *конечными множествами входных* X*, выходных* Y *и внутренних переменных* (W, C)*, может быть построена эквивалентная ей модель суммирующего нейрона.*

Доказательство. Множества $X_s = X_m = X$, $Y_s = Y_m = Y$ равны по определению, где s, m индексы суммирующего и мультипликативного нейронов. Результатом функции учета веса синапса является конечное множество ввиду конечности

множеств W_m, X_m, C_m. Результатом конечной n-мерной операции умножения для учета взвешенных входных сигналов является конечное линейно упорядоченное множество $Dom\,\lambda_m$, поскольку все элементы операции принадлежат $\mathbb{N}$. Для эквивалентности моделей требуется равенство выходных сигналов (символов) обеих моделей в зависимости от одинаковых входных сигналов (символов) на всем множестве возможных входных сигналов. Поскольку отсутствуют ограничения на величину константы «c» в модели суммирующего нейрона, модель может быть перезаписана в следующем виде:

$$y = \lambda_s\left(\sum_{i=1}^{n} w_{is} * x_i + c\right) = \lambda_s\left(\sum_{i=1}^{n} (w_{is} * x_i + c_{is})\right), c = \sum_{i=1}^{n} c_{is}$$

Поскольку множества W_m, C_m конечны, то могут быть подобраны такие элементы w_{is}, c_{is}, которые обеспечат линейно упорядоченное множество $Dom\,\lambda_s$ как результат n-мерной операции сложения для модели суммирующего нейрона, имеющее ровно тот же порядок и мощность что и множество $Dom\,\lambda_m$ в зависимости от входных сигналов X. Так как $|Dom\,\lambda_s| = |Dom\,\lambda_m|$ и их элементы имеют одинаковый порядок в зависимости от входных сигналов, то можно задать функцию λ_s, поставив в соответствие выходные элементы Y элементам множества $Dom\,\lambda_s$, ровно таким же способом, как и функция λ_m. Тоесть: возьмем наименьший элемент $\min(\mathrm{Dom}\,\lambda_s)$ из $\mathrm{Dom}\,\lambda_s$ и поставим ему в соответствие элемент из Y соответствующий элементу $\min(\mathrm{Dom}\,\lambda_m)$ по функции λ_m. Возьмем следующий за ним наименьший элемент $\min\big(\mathrm{Dom}\,\lambda_s \setminus \min(\mathrm{Dom}\,\lambda_s)\big)$ из $Dom\,\lambda_s$ и поставим ему в соответствие элемент из Y соответствующий элементу $\min\big(\mathrm{Dom}\,\lambda_m \setminus \min(\mathrm{Dom}\,\lambda_m)\big)$ по функции λ_m. Будем повторять указанные действия для каждого последующего наименьшего, пока не поставим в соответствие всем элементам из $\mathrm{Dom}\,\lambda_s$ все элементы из Y. Полученная модель эквивалентна модели мультипликативного нейрона. Теорема доказана.

В случае необходимости более сложные модели нейронов, такие как ΣΠ-нейроны или же агрегирующие нейроны, а также многие другие могут быть сведены к абстрактному. Анализ данных нейронов не производится по причине необходимости использовать большее количество ячеек памяти при аппаратной

или программной реализации этих моделей (память для хранения дополнительных переменных), что явным образом увеличивает затраты на реализацию. Введение дополнительных операций (агрегирующие и ΣΠ-нейроны) явно (в общем случае) увеличивает количество реализуемых операций для вычисления выходного сигнала, что ставит вопрос о возможности применения данных моделей для широкого круга задач.

Модель конечного автомата абстрактного нейрона (КААН, FAAN – англ.)**.** Все существующие основные подходы к аппаратной реализации вычислительной техники вне зависимости от схемотехнических решений могут быть проанализированы с применением дискретных конечных множеств. Дискретность цифровой схемотехники не вызывает сомнений. Дискретность гибридной и аналоговой схемотехники определяется соотношением сигнал–шум, в результате которого все сигналы ниже определенного порога являются неразличимыми между собой, что в свою очередь влечет дискретизацию множества их значений. Конечность множества обрабатываемых сигналов обуславливается ограниченностью любой аппаратной реализации вычислительной системы.

Таким образом, описание аппаратной реализации абстрактного нейрона как вычислительного элемента удобно проводить в терминах конечных автоматов. С учетом математического описания рассматриваемых моделей, а именно зависимости результата функции активации от результата функции агрегации, наиболее подходящим является использование автомата Мура [57]. Далее приводится формализм для описания абстрактного нейрона. Ввиду конечности любой аппаратной реализации описание проводится в терминах теории конечных множеств. Поскольку основной абстракцией при построении вычислительных систем является автомат, формализм строится на основе конечного автомата Мура. Автомат Мура описывается упорядоченным множеством $A = (S, S_0, X, Y, \delta, \lambda)$, где S — множество внутренних состояния автомата; $S_0 \in S$ —начальное состояние автомата; X и Y — множества входных символов и множество выходных символов соответственно; $\Delta : S \times X \to S$ — функция перехода в

новое состояние; $\lambda : S \to Y$ — функция вывода символа. Поведение автомата определяется как $s^{t+1} = \Delta(s^t, x^t)$ и $y^t = \lambda(s^t)$.

Введем формальное описание абстрактного нейрона в терминах конечных множеств. Конечный автомат абстрактного нейрона (КААН) может быть описан набором конечных множеств $(N, E, W, Q, \Delta, \Lambda, T)$, где:

N – множество индексов входов (управляющие входы и синапсы), $N \in \mathbb{N}$;

E – алфавит, включающий алфавит всех входных символов E_{input} и алфавит всех выходных символов, E_{output} и $E = E_{input} \cup E_{output}, E \subset \mathbb{N}$;

W – линейно упорядоченное конечное множество всех возможных значений весовых коэффициентов информационных входов (синапсов), $W \subset \mathbb{N}$;

Q –множество определения функции активации и множество значений функции агрегации, $Dom\ \Lambda = Q, ran\ \Delta = Q, Q \subset \mathbb{N}$;

Δ – агрегирующая функция входных сигналов $\Delta_{i=1}^{n}\left(\delta_i\left(w_i\ , E_j\right)\right), i \in N$, зависящая от операции учета весового коэффициента сигнала каждого входа (синапса) $\delta_i\left(w_i, E_j\right), E_j \in E$;

Λ – множество функций активации, реализуемых на нейроне $|\Lambda| \in \mathbb{N}$;

T – множество функций изменения параметров функций Δ, Λ и $\tau_i \in T, |T| \in \mathbb{N}$ искусственного нейрона.

На основе E_{input} формируются подмножества входных информационных сигналов E_{in} и множества управляющих сигналов E_r, такие что $E_{input} = E_{in} \cup E_r, \bigcup_i e_i = E_{in}$. В общем случае при аппаратной реализации не всегда имеется возможность разделить входящие сигналы на управляющие и информационные. С учетом того, что сигнал выхода искусственного нейрона может подаваться на его вход или содержать управляющие символы, подаваемые на другие нейроны, имеем следующие соотношения: $E_{input} \cap E_{output} \neq \emptyset, E_{input} \cap E_{output} = \emptyset$, и $E_{input} = E_{output}$. Множество входных символов каждого i–го входа может быть описано

$$I_i = \left\{ (E_{in}, E_r) \middle| \begin{matrix} \big((E_{in} = \emptyset) \wedge (E_r \neq \emptyset)\big) \vee \\ \vee \big((E_{in} \neq \emptyset) \wedge (E_r = \emptyset)\big) \vee \\ \vee \big((E_{in} = \emptyset) \wedge (E_r = \emptyset)\big) \end{matrix} \right\}, i \in N$$

Например: подача сигнала выше определенного порога может не только переводить мемристор в новое состояние проводимости, но и распознаваться как информационный сигнал. Множества символов на различных входах образуют алфавит входных векторов нейрона

$$I_{in} = \left\{ \big(i_j, \cdots, i_k\big) \middle| i_j \in I_1, \cdots, i_k \in I_n, \bigcup\nolimits_j i_j = I_1, \cdots, \bigcup\nolimits_k i_k = I_n \right\}, n = N$$

Алфавит выходных символов имеет вид:

$$O = \{o_i | o_i = (e_{out}, e_r), e_{out} \in E_{out}, e_r \in E_r, E_{out} \cup E_r = E_{output}\}$$

где E_{out} — множество символов информационных выходных сигналов.

Резюмируя все выше перечисленное, множествам входных и выходных символов X и Y автомата Мура могут быть поставлены в соответствие алфавиты I_{in} и O.

С учетом возможности реализации нескольких функций на одном нейроне функция активации КААН $\lambda_i \in \Lambda, i \in \mathbb{N}$ и определяется зависимостями:

$$\lambda_i : Q \to O \tag{8}$$

$$o^t = \lambda_i (q^t)^t \tag{9}$$

что позволяет поставить в соответствие данную функцию — функции вывода автомата Мура.

Все возможные комбинации весовых коэффициентов на всех входах КААН могут быть представлены в виде вектора:

$$V = \{\mathbf{v}_i | \mathbf{v}_i = (w_j, \cdots, w_k), w_j \in W_1, \cdots, w_k \in W_n, \cup_j w_j = W_1, \cdots, \cup_k w_k = W_n\}, n \in N$$

где W_n — подмножество весовых коэффициентов каждого i–го синапса.

Внутренние состояния абстрактного нейрона описываются множеством

$$V \times \Delta \times \Lambda \times T$$

Функция агрегации информационных входных сигналов Δ является заданной на множестве $I_{in} \times V$ и имеет вид:

$$\Delta : I_{in} \times V \to Q \tag{10}$$

$$q^{t+1} = \Delta_{i=1}^{n}(\delta(\mathbf{v}_i^t, i_i^t)^t)^t \quad (11)$$

Функция изменения параметров КААН $\tau_i \in T, i \in \mathbb{N}$ в наиболее общем виде описывается следующим отображением:

$$\tau_i : E_r \rightarrow V \times \Delta \times \Lambda \times T \quad (12)$$

Динамика изменения параметров имеет вид:

$$\left(\mathbf{v}_i^{t+1}, \Delta^{t+1}, \lambda_i^{t+1}, \tau_i^{t+1}\right) = \tau_j(i_i)^t, \tau_i^{t+1} \in T, \tau_j \in T \quad (13)$$

Функция перехода КААН в новое состояние является результатом двух независимых друг от друга функций (11) и (12), что может быть описано следующим образом:

$$\left(q^{t+1}, \mathbf{v}_i^{t+1}, \Delta^{t+1}, \lambda_i^{t+1}, \tau_i^{t+1}\right) = \left(\Delta_{i=1}^{n}(\delta(\mathbf{v}_i^t, i_i^t)^t)^t, \tau_j(i_i)^t\right) \quad (14)$$

и позволяет сопоставить данную зависимость функции перехода автомату Мура.

Начальное состояние КААН выражается через (11) и (13) с пустым множеством сигналов на входе:

$$\left(\Delta_{i=1}^{n}\left(\delta\left(\mathbf{v}_i^{t=0}, \emptyset\right)^{t=0}\right)^{t=0}, \tau_j(\emptyset)^{t=0}\right)$$

где $t = 0$ — значения в начальный момент времени.

Из всего перечисленного выше следует, что автомату Мура может быть поставлен в соответствие КААН (таб. 1).

Таблица 1. Соответствие между автоматами Мура и КААН

Объект соответствия	Автомат Мура $A = (S, S_0, X, Y, \delta, \lambda)$	КААН $FAAN = (N, E, W, Q, \Delta, \Lambda, T)$
Множество внутренних состояний	S	$V \times \Delta \times \Lambda \times T$
Алфавит входных символов	X	I_{in}
Алфавит выходных символов	Y	O

Начальное состояние	S_0	$\Delta_{i=1}^{n}\left(\delta\left(v_i^{t=0},\emptyset\right)^{t=0}\right)^{t=0}$ $\tau_j(\emptyset)^{t=0}$
Функция переходов в новое состояние	$\Delta: S \times X \to S$	$\Delta: I_{in} \times V \to Q$ $\tau_i: E_r \to V \times \Delta \times \Lambda \times T$
Функция вывода символов	$\lambda: S \to Y$	$\lambda_i: Q \to O$
Динамика переходов	$s^{t+1} = \Delta(s^t, x^t)$	$q^{t+1} = \Delta_{i=1}^{n}(\delta(\mathbf{v}_i^t, i_i^t)^t)^t$ $\left(\mathbf{v}_i^{t+1}, \Delta^{t+1}, \lambda_i^{t+1}, \tau_i^{t+1}\right) = \tau_j(i_i)^t$
Динамика вывода символов	$y^t = \lambda(s^t)$	$o^t = \lambda_i(q^t)^t$

Свойства синапсов КААН. Возможны следующие ситуации при задании множества E_{in} и его подмножеств значений входных сигналов i–го синапса $E_{in} = \bigcup_{i=1}^{n} E_i, E_i = \{e_i | e_i \in E_{in}\}$, n — количество связей, для которых справедливо $w_i \neq 0$ при $\delta(w_i, e_i) = w_i * e_i$. Для информационных сигналов i–х входов (синапсов) возможны различные ситуации при их задании.

Элементы различных множеств E_{in} равны между собой

$$E_i \cup E_j = E_{in}, \nexists e_i \in E_i, \nexists e_j \in E_j \left(\left(e_i > e_j\right) \vee \left(e_i < e_j\right)\right) \Rightarrow$$

$$\Rightarrow \forall e_i \in E_i, \forall e_j \in E_j \left(e_i = e_j\right).$$

Например: на 1–й вход подается информация о цвете (красный, зеленый и т.д.), на 2–й вход подается информация о форме (шар, параллелепипед и т.д.).

Элементы не всех подмножеств сравнимы между собой: $E_i \cup E_j = E_{in}, \exists e_i \in E_i, e_j \in E_j \left(\left(e_i \succ e_j\right) \vee \left(e_i \neq e_j\right) \vee \left(e_i \prec e_j\right)\right)$.

На один из входов подаются сигналы, несравнимые с другими.

На элементах подмножеств может быть задан нестрогий порядок

$$\forall e_i \in E_{in}, \exists e_j \in E_{in} \left(\left(e_i > e_j\right) \vee \left(e_i < e_j\right)\right) \Rightarrow \left(E_{in} \cap \left(\bigcup\nolimits_j e_j\right) = \emptyset\right)$$

Все подаваемые на входы сигналы сравнимы друг с другом. Выход хотя бы одного нейрона (входной нейрон) связан с входом хотя бы еще одного нейрона (для минимальной сети необходимо более одного нейрона). В общем случае множество выходных информационных сигналов нейрона E_{out} может иметь мощность, отличную от мощности входных сигналов последующего нейрона E_{in}, например диапазон напряжений на выходе 1…3 В 1–го нейрона подается на вход второго нейрона, принимающий сигналы в диапазоне 2…3 В. Тогда очевидно, что входные сигналы в диапазоне [1, 2) не должны рассматриваться как информационные.

Данный случай формально описывается применением характеристической функции к множеству входных информационных сигналов каждого синапса: $\chi E_i(e_j), e_j \in E_i, \bigcup_{i=1}^{n} E_i = E_{in}$, где значения функции $\chi E_i(e_i) = 0$, не учитываются на данном нейроне.

На множествах информационных сигналов синапсов $\chi E_i(e_i) \neq 0$, учитываемых принимающим их нейроном $\bigcup_{i=1}^{n} E_i = E_{in}, |E_i| > 1$, может быть задан линейный порядок. В случае если порядок не задан, все сигналы (или их часть, в случае задания нестрогого порядка) будут неразличимы между собой.

Резюмируя все выше описанное можно заключить следующее: входные информационные сигналы на разных синапсах, с одной стороны, могут быть несравнимы между собой, с другой стороны, множество сигналов каждого синапса может описываться характеристической функцией. Результатом чего является определение оптимальной мощности множеств информационных сигналов для общего случая:

$$\bigcup_{i=1}^{n} E_i = E_{in}, |E_{in}| = |E_{out}|$$

Отсутствие связи может быть задано через значение весовых коэффициентов синапсов $w_i = 0$ при $\delta_i(w_i, i_i) = w_i * i_i$. Поскольку множество весовых коэффициентов W линейно упорядочено и для всех несуществующих связей $w_i = 0$, то элементы множества принимают следующие значения:

$w_i = 0$ для отсутствующих связей,

$\min(W_i) = k,$ где $k \in \mathbb{N}$ некоторая константа.

$\min\big(W_i \backslash \min(W_i)\big) = k_w + h,$ где $h \geq k, h \in \mathbb{N}$ некоторая константа и $W_i = \{w_j | w_{j+1} - w_j = h\}$.

Традиционным подходом к реализации учета веса входных информационных сигналов каждого i–го входа является умножение на его весовой коэффициент

$$\delta: w_i * e_i \to \mathbb{N}, e_i \in E_i, \bigcup_{i=1}^{n} E_i = E_{in}, w_i \in W_i, \bigcup_{i=1}^{n} W_i = W$$

При соотношении элементов множества $e_i \in E_{in}, w_j \in W,$ задаваемых константой, одинаковой для обоих множеств, максимальная возможная мощность множеств в предельном случае определяется следующим образом:

$$\forall e_{i+1} \in E_{in}, \forall e_i \in E_{in}, \forall w_{j+1} \in W, \forall w_j \in W\big((e_{i+1} - e_i = h) \wedge (w_{i+1} - w_i = h)\big)$$
$$\Rightarrow |Dom\ \lambda| = |E_{in}| * |W|$$

Увеличение мощности множества области определения $Dom\ \Delta$ КААН при сохранении мощности множеств E_{in}, W возможно за счет задания нелинейной зависимости между элементами $w_i \in W$. Первый из возможных способов поставить в соответствие элементам множества W элементы линейно упорядоченного множества с нелинейной зависимостью, например:

$$W = \{0, w_1 = 1, w_2 = n + 1, w_3 = n * w_2 + 1, \cdots, w_j = n * w_{j-1} + 1\}$$

где n – количество связей, для которых справедливо $w_j \neq 0$. Указанный метод максимально увеличивает мощность области определения $\mathrm{Dom}\ \Delta$ КААН при аппаратной реализации до

$$|\mathrm{Dom}\ \Delta| = |E_{in}| * |W|^{|n|}$$

К недостаткам подхода можно отнести полное отсутствие возможности задания линейной зависимости между, входными сигналами на синапсах, что ограничивает его применение при аппаратных реализациях, предназначенных для построения сетей на основе КААН.

Второй метод может быть определен как внесение нелинейности между подмножествами множества W и имеет следующую схему задания:

$$W = \left\{ \begin{array}{c} 0, \\ w_1 = 1, w_2 = 2 * w_1, \cdots, w_i = i * w_1, \\ w_{i+1} = n * w_i + h, w_{i+2} = 2 * w_{i+1}, \cdots, w_{i+j} = j * w_{i+1}, \\ \cdots, \\ w_{k+1} = n * w_k + h, \cdots, w_{k+l} = l * w_{k+1} \end{array} \right\},$$

$$\bigcup_{n=1}^{i} w_i = W_i, \bigcup_{i+1}^{j} w_j = W_j, \cdots, \bigcup_{k+1}^{l} w_l = W_l, \bigcup_p W_p = W,$$

где n – число связей, для которых справедливо $w_i \neq 0$.

Подход является компромиссом между возможностью задавать линейные зависимости при учете входных сигналов на синапсах и увеличением мощности множества $Dom\ \Delta$ КААН. Нелинейная зависимость между входными сигналами E_{in}, также приводит к увеличению множества $Dom\ \Delta$ и может быть описана одним из двух вариантов рассмотренных выше для множества W.

Формализм искусственных нейронных сетей на основе КААН. Формализм описания сети на основе КААН определяется кортежем $(E, W, Q, \Delta, \Lambda, T)$ без использования N. Покажем это. Используемые понятия входных нейронов могут быть распространены не только на информационные входы (синапсы), но и на управляющие входы. Используя данное условие, получим единое описание управляющих и информационных входов на случай аппаратной реализации гарвардской архитектуры построения вычислительных систем. Определив, что выходной сигнал каждого КААН подается на вход всех остальных КААН сети, получим полносвязную сеть. Задав для не существующих связей значение весовых коэффициентов синапсов $w_i = 0$, получим описание искусственнной нейронной сети (ИНС) без использования множества N.

Классы реализуемых на КААН функций напрямую зависят от мощности множества Q. Множество Q является линейно упорядоченным множеством $Dom\ \Lambda$ с одной стороны и областью значений Δ с другой стороны.

Для случая $|E_{in}| = 1$ возможны следующие ситуации. Поскольку $\mathrm{Dom}\,\Delta$ линейно упорядоченно и $\forall a \in Dom\Delta\big(\Delta(a_{i+1}) > \Delta(a_i)\big)$, то при $|Q| = 1$ возможна реализация только пороговой функции. При $|Q| = 2$ возможны реализации пороговой, линейной и дельта-функции. Задание пороговой функции возможно четырьмя способами, первые два:

$$\forall\lambda \in \Lambda\big((\lambda(maxQ) = 1) \wedge (\lambda(q_i) = 0, q_i \neq maxQ)\big)$$

$$\forall\lambda \in \Lambda\big((\lambda(minQ) = 0) \wedge (\lambda(q_i) = 1, q_i \neq minQ)\big)$$

Инвертирование значений функции λ с 0 на 1 дает два других способа.

Линейная функция задается следующим образом:

$$\forall q \in Q, \forall\lambda \in \Lambda\begin{pmatrix}\big((\lambda(q_{i+1}) > \lambda(q_i)) \wedge (q_{i+1} > q_i)\big) \vee \\ \vee\big((\lambda(q_{i+1}) < \lambda(q_i)) \wedge (q_{i+1} > q_i)\big)\end{pmatrix}$$

дельта-функция задается следующим образом:

$$\forall q \in Q, \forall\lambda \in \Lambda\begin{pmatrix}\big((\lambda(q_1) > \lambda(q_2)) \wedge (\lambda(q_1) > \lambda(\emptyset))\big) \vee \\ \vee\big((\lambda(q_1) < \lambda(q_2)) \wedge (\lambda(q_1) < \lambda(\emptyset))\big)\end{pmatrix}$$

Минимальная необходимая мощность для задания периодических функций $|Q| = 3$. Аппаратная реализация задания периодических функций может быть реализована путем модификации функции агрегации входных сигналов с применением операции взятия по модулю над суммой или произведением взвешенных входных сигналов $\Delta_i\big(\delta(w_i, i_i)\big) = mod_d(\sum_i \delta(w_i, i_i))$, где $d -$ размер периода на $Dom\,\Delta$. Результатом применения модифицированной функции будет значение от начала текущего периода. Путем задания соответствующих значений E_{out} функция активации может принимать вид RBF, линейной или пороговой функции на текущем периоде. В предельном случае $|Dom\,\Delta| = |Q|$, что позволяет задавать не только линейные, пороговые, нелинейные, RBF или периодические функции, но и почти периодические функции.

Включение в модель КААН множества функций изменения параметров T позволяет учитывать в алгоритмах обучения сети КААН изменения функции активации в процессе обучения и работы сети. Например, производить переход от пороговых функций к линейным функциям. Осуществлять переход от линейных функций к нелинейным (сигмоидальная функция, заданная на дискретном множестве) и RBF функциям. Обеспечивать переход от RBF и нелинейных функций к периодическим, а также от периодических функций к почти периодическим функциям во время тренировки ИНС.

Имеющееся описание КААН применимо для описания работы синхронных искусственных нейронных сетей – сетей, в которых все нейроны срабатывают (либо не срабатывают) в один момент времени, после чего данные переменных обнуляются и начинается новый цикл работы сети. Попытка асинхронного описания работы одиночного КААН невозможна ввиду отсутствия объекта сравнения.

В отличие от одиночных КААН, сети на их основе могут иметь асинхронный механизм работы. Обобщение сетей КААН на случай асинхронной работы сети требует внесения изменений в выражения (9), (11). Вносимые изменения должны учитывать возможность настройки времени срабатывания функции активации КААН t_A и/или возможность учета предыдущего состояния функцией Δ. Уравнение изменения состояний КААН описываемых уравнением (11) и уравнение динамики вывода символов (9) для асинхронных сетей КААН принимают вид:

$$o^{t_A} = \lambda_i(q^t)^{t_A} \quad (15) \qquad q^{t+1} = \Delta_{i=1}^{n}(\delta(v_i^t, i_i^t)^t, q^t)^t \quad (16)$$

Представленное математическое описание конечного автомата абстрактного нейрона позволяет, как было показано, путем задания соответствующих весовых коэффициентов осуществлять агрегацию входных сигналов как для модели суммирующего, так и для модели мультипликативного нейронов на едином блоке. В качестве данного блока может быть использован сумматор, что в свою очередь существенно снизит затраты на аппаратную реализацию функции агрегации в сравнении с блоком умножения. Предлагаемый формализм позволяет реализовать в рамках модели нейрон, срабатывающий по совпадению, интегрирующие нейроны и связывающий нейрон. В рамках рассматриваемой дискретной модели показан метод увеличения множества задания функции активации, без увеличения множества значений весовых коэффициентов связи.

1.3 Обобщенная схема реализации КААН

Базовая схема искусственного нейрона, представленная на рис. 1, имеет высокий уровень абстракции и предназначена обобщения основных типов функций и блоков, позволяющих концептуально описать вычислительный элемент сети. Предложенный формализм КААН позволяет произвести уточнение структуры искусственного нейрона.

Математическая модель процесса вычисления для срабатывающих по совпадению нейронов реализуемых с применением КААН может быть схематично представлена в виде отображения на элементах множествах нейрона рис. 2. Из математического формализма явно следует введение дополнительного множества *Q*, представленного на схеме и позволяющего имплементировать математическое представление внутреннего состояния нейрона.

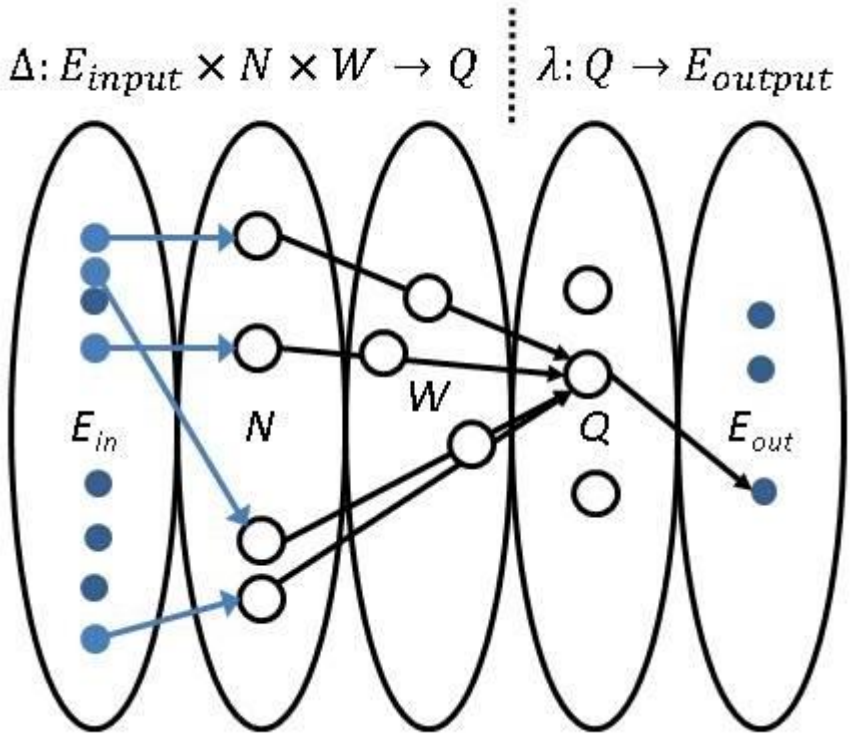


Рисунок 2. Схема вычислений срабатывающего по совпадениям нейрона на основе КААН.

С учетом аппаратной реализации представленного концепта КААН, данное множество должно включать функции хранения во времени (от срабатывания к срабатыванию) заданных переменных функции агрегации и текущей функции

активации нейрона, что приводит к необходимости включения в модель элементов памяти.

Процесс вычислений интегрирующим нейроном, связывающим нейроном и интегрирующим с утечками нейроном описываемых в терминах КААН схематично представлен на рис. 3. На схеме не отображены временные параметры вычисления функции активации, что вызвано нецелесообразностью рассмотрения динамики функции активации единичного нейрона.

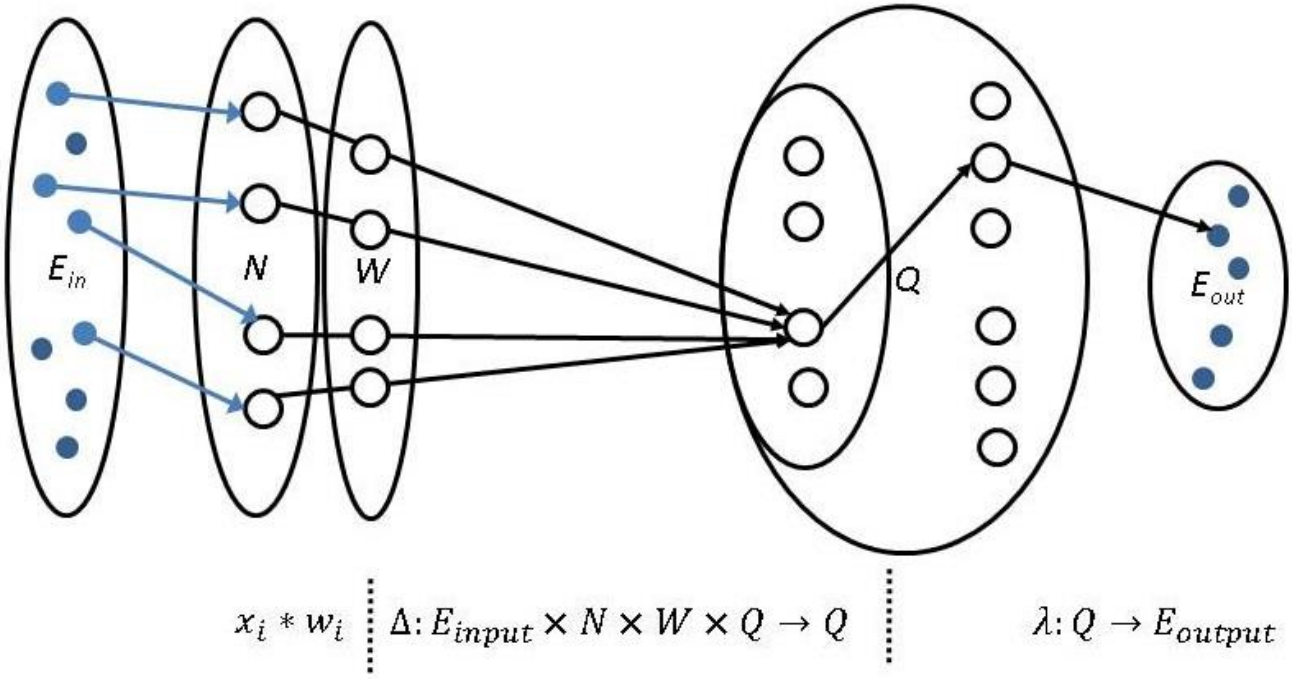


Рисунок 3. Схема вычислений интегрирующих нейронов на основе КААН

Исходя из необходимости обобщения единым представлением различных типов нейронов и способов описания их средствами КААН, логично рассматривать в качестве базовой концепции блочного представления КААН имплементацию связывающего нейрона. Поскольку в рамках концепции связывающего нейрона функция активации в явном виде содержит параметр времени учета данных поступающих от функции агрегации, схема обработки информации может быть модифицирована с учетом отображения временной зависимости. Модифицированная схема с поблочным представлением показана на рис. 4. Время рефрактерного состояния задается как параметр работы входа блока функции активации. Время удержания сигнала параметр выхода указанного блока.

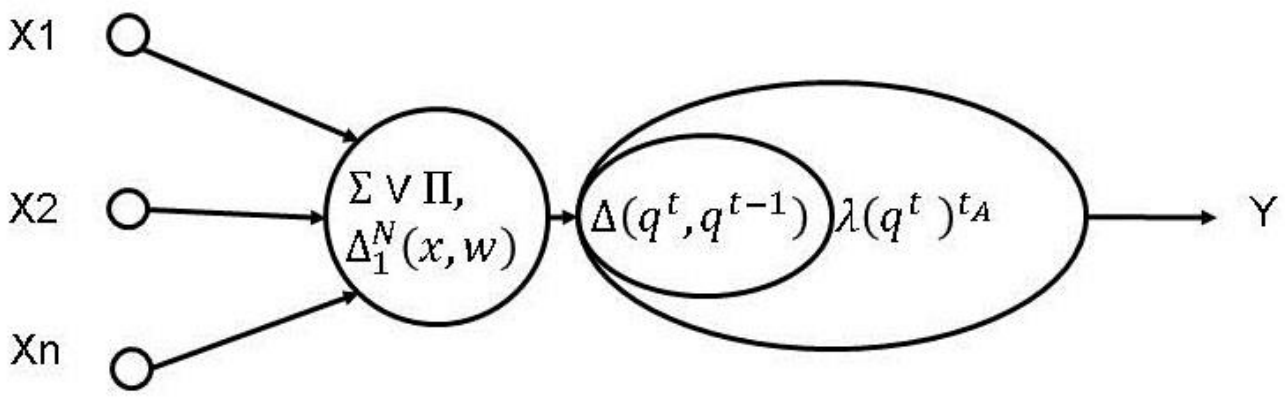


Рисунок 4. Схема с поблочным представлением обработки информации, связывающим нейроном на основе КААН.

Помимо реализации рефрактерного периода работы нейрона параметром входа блока функции активации существует возможность задания данного параметра на блоке функции агрегации или блоках синапсов. Оптимальное техническое решение имплементации рефрактерного периода в качестве функции обработки информации одним из блоков будет зависеть от конкретных технических требований к системе и схемотехнического решения обработки информации. Например, требования к низкому энергопотребления и применению средств цифровой схемотехники приведут к эффективности применения рассматриваемой функции на входах нейрона, что позволит избежать излишних переключений для логических ячеек и вентилей в период рефрактерного периода, но увеличит необходимую для реализации каждого синапса и площадь на кристалле.

Реализация динамической природы КААН предполагает задание произвольной функции активации на блоке формирования выходного сигнала. Исходя из ограниченности любой аппаратной реализации, данное требование может быть снижено, что является вынужденной мерой, до реализации любой произвольной функции на множестве внутренних состояний нейрона *Q*. Возможные методики построения требуемого блока лежат в областях аналоговой, гибридной и цифровой схемотехники, и в общем случае не гарантируют оптимального метода решения задачи.

Применение аналоговой схемотехники будет сопровождаться сниженной устойчивостью к помехам, значительными размерами площади кристалла и

сложностью реализации произвольной функции активации. Например, в случае имплементации функции активации с применением аналогового схемотехнического решения для задания функции гиперболического арктангенса с четырьмя настраиваемыми параметрами, возможно задание функции стремящейся к линейной, пороговой или сигмоидальной функциям рис. 5-7.

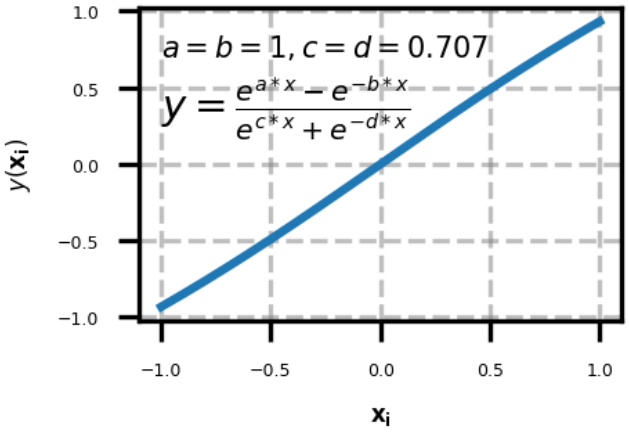


Рисунок 5. Линейная функция.

На рисунке 5 представлен пример задания функции гиперболического арктангенса стремящейся к линейной функции с четырьмя параметрами настройки. На рисунке 6 представлен пример задания функции активации гиперболического арктангенса с подгоночными коэффициентами, приближающими ее вид к пороговой функции активации. Параметры функции, приводящие ее вид к сигмоидальному типу, представлены на рис. 7. На рисунках 5-7 $\mathbf{x_i}$ – результат функции агрегации входных сигналов. Рассматриваемые коэффициенты, несмотря на широкие возможности по подстройке функции к требуемому виду, не позволяют задавать RBF-функции или взятие по модулю 2 в качестве функции активации искусственного нейрона.

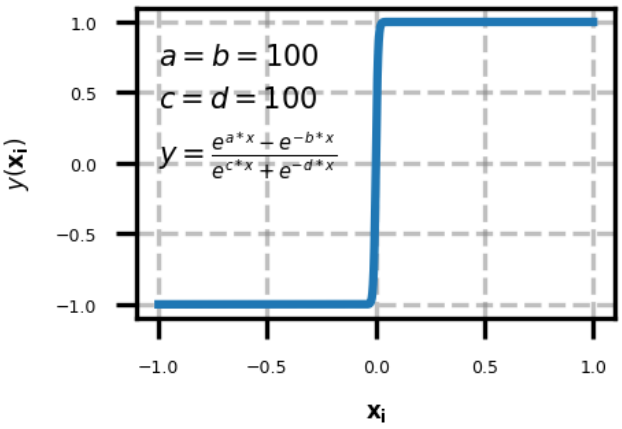


Рисунок 6. Пороговая функция.

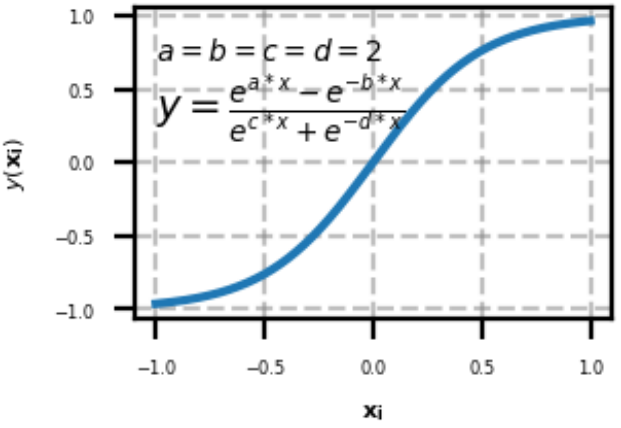


Рисунок 7. Сигмоидальная функция.

Имплементация функции активации применением вычислителя на основе цифрового схемотехнического решения в общем случае приведет либо к снижению скорости обработки входного сигнала (за счет необходимости последовательного выполнения элементарных операций), либо к увеличению площади занимаемой вычислителем на кристалле.

Исходя из всех рассмотренных факторов, оптимальным методом, позволяющим реализовать произвольную функцию генерации выходного сигнала нейрона, будет табличный метод задания соответствия значений функции агрегации значениям функции активации LUT (Look Up Table – англ.). Обобщённая схема искусственного нейрона на основе КААН представлена на рис. 8. Схема предполагает использование двух таблиц LUT для блока генерации.

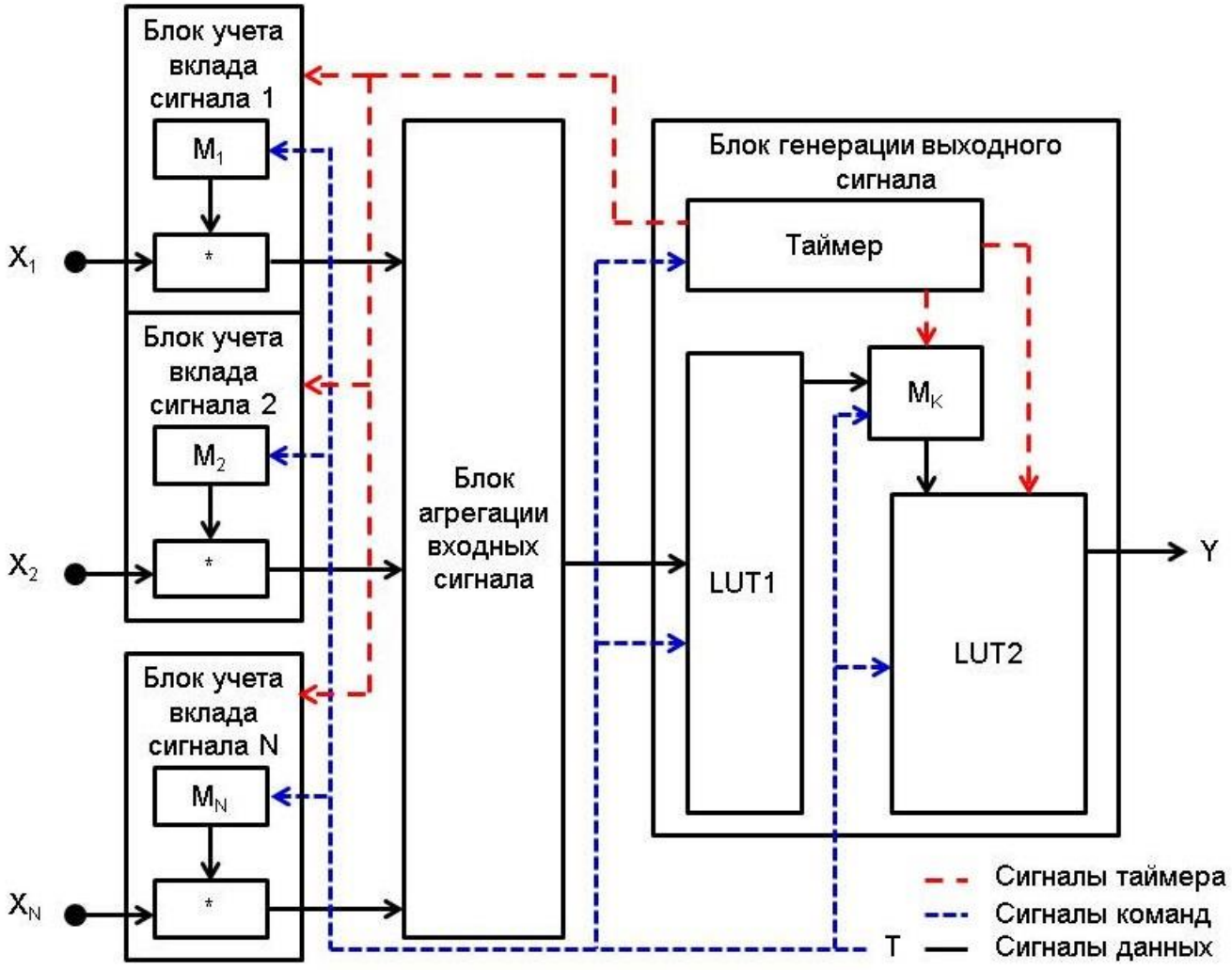


Рисунок 8. Обобщённая схема КААН.

Первый блок LUT1 предназначен для учета амплитуды сигнала в текущий момент времени, второй блок LUT2 реализует учет последовательности амплитуд во времени.

Блоки учета входного сигнала состоят из регистров хранения весовых коэффициентов $\mathbf{M_1}$, $\mathbf{M_2}$, ..., $\mathbf{M_N}$ и схем умножения данных регистра на входной сигнал. Блок генерации выходного сигнала состоит из таймера, регистра хранения текущего состояния и двух таблиц значений. Первая таблица значений реализует отображение текущей активности на входах нейрона. Вторая таблица предназначена для отображения текущего состояния нейрона выходным сигналом. Регистр $\mathbf{M_K}$ хранит текущее состояние нейрона. Рефрактерный механизм реализуется путем контроля сигналами таймера блоков учета входных сигналов или же регистра хранения текущего состояния. Время активации и время удержания выходного сигнала управляются командами от таймера, передающимися на регистр текущего состояния нейрона и LUT2 соответственно.

Обработка информации осуществляется по следующему алгоритму. Таймер начинает отсчет времени активации нейрона. Входной вектор поступает на блоки учета входного сигнала, где происходит его покомпонентное умножение на хранящийся в регистрах весовых коэффициентов вектор. После перемножения происходит агрегация компонент вектора. Сигнал от блока агрегации поступает в LUT1, на выходе которого генерируется значение выходного сигнала, поставленное в соответствие входному сигналу. Сигнал с выхода LUT1 поступает в регистр хранения текущего состояния нейрона. Процесс повторяется до завершения времени активации нейрона. После завершения времени активации нейрона начинается отсчет таймером времени удержания сигнала на выходе. На блоки учета вклада сигнала и на регистр учета текущего состояния нейрона подается сигнал препятствующий изменению состояния. На LUT2 подается сигнал генерации выходного сигнала и генерируется сигнал. По окончании времени удержания выходного сигнала таймер подает команду сброса [либо частичного изменения данных (интегрирующий нейрон с утечками)] текущего состояния и команду прекращения генерации выходного сигнала. Таймер

начинает отсчет рефрактерного периода. По завершении отсчета процесс повторяется.

1.4 Выводы по главе 1

Обобщенная модель КААН позволяет моделировать широкий набор искусственных нейронов с различными функциями активации, что при аппаратной реализации КААН гарантирует относительную универсальность технического решения и широкий набор возможностей при проектировании сети.

На основании предложенной модели могут быть построены сети, реализующие помимо распознавания изображений, математические операции сложения, вычитания, умножения и др. Реализация математических операций позволяет применить часть ресурсов сети (реализовав на подсети алгоритм обучения) для обучения остальной части сети без применения классических (Неймановских и Гарвардских) архитектур.

Предлагаемая модель КААН позволяет учесть в алгоритмах обучения изменение активационной функции КААН и построить алгоритм, автоматически генерирующий сеть с оптимальными активационными функциями нейронов.

Помимо генерации сетей КААН (путем применения алгоритмов обучения) появляется возможность задания алгоритма работы сети с динамически изменяющимися функциями активации.

Формализм КААН позволяет проанализировать применение неэквидистантного подхода к аппаратной реализации множества весовых коэффициентов при его построении и добиться увеличения мощности задания функции агрегации без увеличения мощности множества весовых коэффициентов.

2 ГЛАВА. Моделирование элементов нейрона.

Предметом рассмотрения данной главы выступают методы и способы реализации искусственных нейронных сетей аппаратными средствами, а так же непосредственно сами реализации. Глава предваряется обзором литературы, по результатам которого производится обоснованный выбор в пользу мемристивных компонентов как основного элемента нейрона. Производится Verilog описание мемристивных компонентов.

Вначале главы представлен обзор реализаций ИНС и обзор реализаций искусственного нейрона. После чего рассматриваются различные элементы электронной компонентной базы для применения в искусственном нейроне. Обзор завершается исследованием литературы, которое посвящено методам и средствам описания составных блоков искусственных нейронов. На основе материалов обзора литературы производится обоснование применения мемристивных компонентов в аппаратной реализации искусственного нейрона, рассматриваются положительные и отрицательные стороны данного подхода. На основе известных теорий о механизмах переключения проводимости в мемристивных компонентах, производится построение Verilog описания для биполярного механизма.

Глава завершается представлением результатов моделирования в среде Cadence разработанного Verilog описания.

2.1 Литературный обзор аппаратных реализаций искусственных нейронных сетей

Парадигма коннекционизма. Аппаратные реализации искусственных нейронных сетей, равно как и их математические модели, берут свое начало в концепции коннекционизма. Исторически [58] коннекционизм (=коннективизм) происходит от попыток понять, как осуществляется обработка информации мозгом. Известный психолог Э.Торндайк полагал (1910):

««Возникновение связей является результатом, как состояния мозга, так и действия внешних ситуаций. Часто связи приобретают вид длинных последовательностей, в которых реакция на одну ситуацию становится новой ситуацией, вызывающей следующую реакцию, и т. д. Связи могут создаваться как частями, элементами или особенностями отдельной ситуации, так и всей ситуаций в целом. Связываться могут едва различимые отношения или смутные аттитюды и интенции».

Заслуга Торндайка состоит в выделении элементов и элементарных актов (identical-elements theory). Современный коннекционизм имеет начало в работах МакКаллока по искусственным нейронным сетям и Тьюринга, рассматривавшего случайную сеть ассоциатов логических вентилей (1948). Как пишет Янковская Е.А., акцентируя внимание на децентрализованности, «понятие гетерархия, предложенное МакКаллоком, становится основой для концептуальных и/или формализованных моделей сложных систем является контингентным и отчасти альтернативным по отношению к понятию иерархии» [59].

Главный принцип коннекционизма состоит в описании процесса обработки информации сетями из взаимосвязанных простых элементов. Форма связей и элементов может меняться от модели к модели. Например, элементы в сети могут представлять нейроны, а связи — синапсы. Другая модель может считать каждый элемент в сети словом, а каждую связь признаком семантического подобия и т.п. Лозунг радикального коннекционизма – «связи – все, элементы – ничто», т.е. для результата вычислений более значимы связи, а не особенности структуры

элементов, чем бы те ни были. Под коннекционизмом применительно к вычислительным системам обычно понимается подход, ориентированный на максимальное распараллеливание обработки данных и терминологически тождественный PDP (Parallel Distributed Processing).

Основные принципы модели PDP, не без влияния терминов искусственных нейронных сетей, были сформулированы [60] Д.Е. Румельхартом и др. в 1987:

1. Набор процессорных элементов (units)
2. Состояние активации
3. Выходная функция для каждого элемента
4. Шаблон связности между элементами, включающий веса и локализацию связей (нагруженный граф)
5. Правило распространения активностей по сети
6. Правило активации, порождающее из совокупности «входов» и состояния элемента новый уровень активации элемента
7. Правило обучения, т.е. возможность изменения шаблона связности под действием опыта.
8. Окружение, в котором должен работать вычислитель.

Первоначально среди нейрофизиологов предпочтительной считалась концепция последовательной обработки данных, в ходе которой активизируются следующие друг за другом узлы в цепи связей в сети. Исходя из физических принципов строения мозга, произошла смена парадигмы на параллельную обработку, где предположительно происходит независимая активизация двух или более взаимозависимых цепей. Стремясь обобщить PDP, Бештел и Абрахамсен в 1991г. предложили 4 принципа коннекционизма [61]; три из них совпадают с P1+P4, P2 и P7, а четвертый требует дать семантическую интерпретацию сети (например, данные или результаты могут храниться в одном элементе или же быть распределенными между элементами). Под понятием «вес» можно понимать любое предикативное свойство связи (например, двунаправленность).

В настоящее время можно выделить пять коннекционистских имплементаций рис. 9:

- Клеточные автоматы (включая специализированные мультипроцессорные системы типа CAM-8);
- Искусственные нейронные сети;
- Клеточные нейронные сети;
- Кластерные вычислительные системы (от многоядерных персональных ЭВМ до нескольких серверов в пределах одного помещения);
- Облачные и GRID-проекты.

Рисунок 9. Модели коннекционизма в современной вычислительной технике.

Отдельно стоит выделить концепцию клеточных автоматов, так как данный подход может быть описан нейронными сетями с логистической функцией активации и ограниченным шаблоном связности. Создателями концепции клеточных автоматов считаются Конрад Цузе (1969) и Джон фон Нейман (1952). Начало исследований датируется 40ми годами XX века и инспирировано ранними работами по нейронным сетям [62]. Доказывая идею возможности существования самовоспроизводящегося автомата, Нейман столкнулся с рядом технических проблем, обусловленных инженерной сложностью такой системы. Станислав Улам [63] предложил абстрагироваться до математической модели и использовать метод, схожий с методом объяснения роста кристаллов. С этой целью пришлось объединить вычислительное устройство и данные, результатом этого стал первый

клеточный автомат (СА). В это же десятилетие основоположники кибернетики Норберт Винер и Артуро Розенблют публикуют работу, где рассматривается распространение нервного импульса вместе с изложением нового математического аппарата - клеточных автоматов [64]. В модели использовались три состояния нервных клеток (возбужденное - или активное, рефрактерное - или расслабленное, и покоя). Рефрактерное состояние – состояние, в котором клетка не может быть возбуждена и передавать возбуждающие импульсы. В 1969 г. Конрад Цузе опубликовал книгу (Rechnen der Raum) [65], в которой Вселенная рассматривалась с позиций огромного клеточного автомата, реализующего природные вычисления.

Математическое основание использования нейронных сетей базируется на теореме о представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одного переменного и сложения Колмогорова-Арнольда (1957). Теорема о представлении непрерывных функций нескольких переменных в виде суперпозиций непрерывных функций одного переменного и сложения в 1987 году была переложена Хехт-Нильсеном для нейронных сетей [66]. Теорема Хехт-Нильсена доказывает представимость функции многих переменных достаточно общего вида с помощью двухслойной нейронной сети с прямыми полными связями с n нейронами входного слоя, (2n+1) нейронами скрытого слоя с заранее известными ограниченными функциями активации (например, сигмоидальными) и нейронами выходного слоя с неизвестными функциями активации. Исследовав истоки зарождения концепции и предпосылки к формированию аппаратных и модельных представлений описания ИНС, перейдем к рассмотрению современного уровня технических достижений в области аппаратной имплементации нейроморфных систем и ИНС в целом.

Современные имплементации ИНС в интегральных схемах. Поскольку реализация программными средствами на фон Неймановской архитектуре не позволяет оптимально использовать возможности искусственных нейронных сетей. Данное обстоятельство, а так же все возрастающая потребность в

повышении скорости обработки информации и постоянно увеличивающиеся объемы обрабатываемой информации привели к аппаратным реализациям ИНС.

На текущий момент в мире ведется множество проектов с применением искусственных нейронных сетей. Основными направлениями исследований можно выделить: информатику (реализация когнитивных функций от классификации изображений до автоматического перевода), нейрофизиологию (применение моделей нейронных сетей для исследования работы мозга и нейроморфинг [67]) и системы управления (как правило, системы принятия решений и управления для роботизированных систем). Под нейроморфингом обычно понимается процесс создания технических систем (как алгоритмический, то есть программный, так и аппаратно-программный) осуществляющих обработку сигналов на основе математических моделей обработки сигналов в биологических нейронных сетях. Подходы в реализации ИНС выбираемые группами ученых условно можно разделить на алгоритмический подход (реализуется применением СБИС стандартной архитектуры, оптимизированными для параллельных вычислений) и аппаратный (реализуется созданием архитектуры, ориентированной на моделирование нейронных сетей). Оба подхода нашли отражение в классификации аппаратных реализаций представленных в работе [68], в результате чего она приобрела следующий вид: рис. 10.

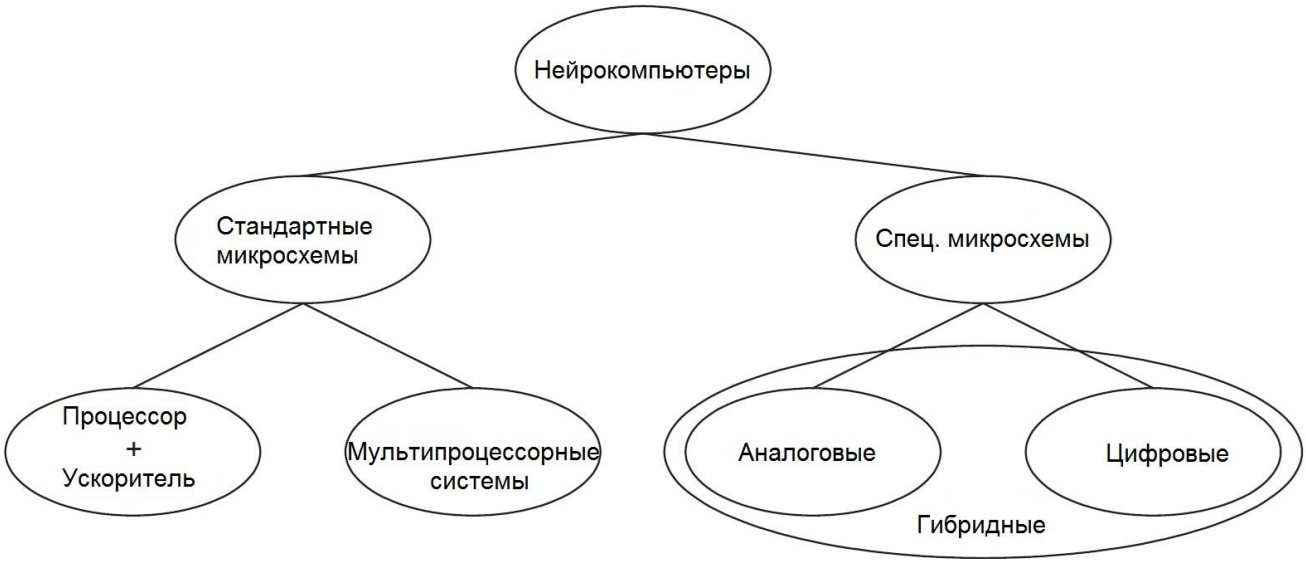


Рисунок 10. Классификация аппаратных реализаций ИНС [68].

Интегральные схемы (ИС), реализующие ИНС путем создания специальной архитектуры для типовых аналоговых и цифровых элементов, представлены на классификации слева. В основе подхода лежат схемотехнические решения, определяющие необходимое количество блоков обработки, размер стека команд и их длину, а так же маршруты прохождения сигналов в схеме. К преимуществам ИС этого типа относятся легкая интегрируемость с существующими системами, отработанная система разработки технических решений и возможность оперативной модернизации или изменения сети и ее элементов.

Аппаратные реализации искусственных нейронов и их элементов породили второй подход к имплементации ИНС (рис. 10 справа). Основной отличительной особенностью данного подхода является применение специально разработанной схемы обработки сигналов на основе элементной базы, позволяющей реализовать искусственный нейрон и схемотехнические решения по построению ИНС. Преимуществами подхода является более высокое быстродействие и низкое энергопотребление в сравнении с алгоритмическим подходом. К недостаткам можно отнести ограничения в модификации, свойственные любой аппаратной реализации.

Мировые производители ИС и исследовательские центры создали ряд аппаратных реализаций ИНС (представлены в таблице 2).

Таблица 2. Номенклатура коммерческих ИС для ИНС на 2004 г. [68].

Фирма/ Название	Архитектура	Алгоритмы обучения	Точность	Кол-во нейронов	Кол-во Синапсов	Скорость вычислений
Slice architectures						
Micro devices MD-1220	Feedforward, ML	No	1x16 bits	8	8	1.9 MCPS
Nuralogix NLX-420	Feedforward, ML	No	1-16 bits	16	off chip	300 CPS
Philips Lneuro-1	Feedforward, ML	No	1-16 bits	16 PE	64	26 MCPS
Philips Lneuro-1	N.A.	No	16-32 bits	12 PE	N.A.	720 MCPS

SIMD						
Inova N64000	GP,SIMD, FP	Program	1-16 bits	64PE	256k	
Hecht-Nielson 100NAP	GP,SIMD, FP	Program	32x32,FP	4PE	512k Off chip	250 MCPS 64 MCUPS
Hitachi WSI	Wafer, SIMD	Hopfield	9x8 bits	576	32k	138 MCPS
Hitachi WSI	Wafer, SIMD	BP	9x8 bits	144	N.A.	300 MCUPS
Neuricam NC3001 TOTEM	Feedforward, ML, SIMD	No	32 bits	1-32	32k	1 GCPS
Neuricam NC3003 TOTEM	Feedforward, ML, SIMD	No	32 bits	1-32	64k	750 MCPS
RC Module NM6403	Feedforward, ML	Program	1-64 x 1-64 bits	1-64	1-64	1200 MCPS
Systolic arrays						
Siemens MA-16	Matrix ops	No	16 bits	16 PE	16x16	400 MCPS
Radial basis functions						
Nestor/Intel NI1000	RBF	RCE, PNN, program	5 bits	1 PE	256x1024	40kpat/s
IBM ZISC036	RBF	ROI	8 bits	36	64x36	250kpat/s
IBM Silicon recognition ZISC78	RBF	KNN, L1, LSUP	N.A.	78	N.A.	1 Mpat/s
Hybrid hardware implementations						
AT&T ANNA	Feedforward, ML	No	3x6 bits	16-256	4096	2.1 GCPS
Bellcore CLNN-32	FCR	Boltzmann	6x5 bits	32	992	100 MCPS 100 MCUPS
Mesa Research	Feedforward	No	6x5 bits	6	426	21 GCPS

Neuralclassifier	, ML					
Ricon RN-200	Feedforward , ML	BP	N.A.	16	256	3.0 GCPS
Other chips						
SAND/1	Feedforward, ML, RBF, Kohonen	No	40 bits	4 PE	Off-chip	200 MCPS
MCE MT19003	Feedforward, ML	No	13 bits	8	Off-chip	32 MCPS

ML – multilayer; GP – General purpose architecture; FP – float point; PE – indicates a processing unit; BP – back propagation; RBF – архитектура сети, предусматривает применение радиально базисных функций; N.A. – данные не доступны; WSI Wafer – специализированная архитектура с программированием SIMD; No – не предусматривается обучение во время работы;

Одновременно с разработкой коммерческих реализаций были определены основные критерии для сравнения технических решений [68].

1. CPS (connect per second – англ.) – показатель, определяющий скорость обработки информации нейронов от момента приема сигнала синапсом, до момента генерации выходного сигнала на выходе.
2. CUPS (connect update per second – англ.) – показатель скорости обучения сети, представляет отношение количества изменяемых значений весовых коэффициентов к единице времени.
3. WCPS (watt per connect update per second – англ.) – количество энергии, затрачиваемое на вычисления и изменение значения каждого синапса.

В работе [69] исследуется список применений указанных аппаратных реализаций в различных областях науки и техники таб. 3

Таблица 3. Научно-технические области применения аппаратных ИНС [69]

Приложение	**Способ реализации**
Физика высоких энергий	Цифровой нейрочип
Распознавание образов	FPGA, Цифровые
Распознавание объектов, картинок	RAM based, Оптические

Классификация изображений	FPGA, Цифровые
Обработка видео	RAM based, Оптические, Аналоговые, FPGA
Интеллектуальный видео анализ	Гибридные, FPGA
Извлечение дактилоскопических данных	Аналоговые
Управление с обратной связью	Аналоговые
АСУ автономными роботами	Цифровые, FPGA, Гибридные, DSP
Управляющие системы	FPGA
Распознавание почерка	Цифровые
Распознавание звука	DSP
АСУ реального времени	Цифровые
Генерация звука	Аналоговые
Планирование задач	Аналоговые, Цифровые
Анализ газа	Аналоговые

Достижения в развитии программных алгоритмов ИНС и элементной базы привели к запуску целого ряда исследовательских программ и проектов по аппаратной реализации (см. актуальность работы). В процессе реализации проектов помимо расширения теоретической базы построения аппаратных реализаций ИНС были разработаны и реализованы в законченном виде новые технологии (PCM) и архитектуры (применение кроссбаров на мемристорах). Отдельно стоит отметить значительные успехи перечисленных ранее проектов в области аппаратной реализации.

Технологические успехи Human Brain Project [70] заключаются в реализации на 8-ми дюймовой кремниевой пластине физической модели эмулирующей 200000 биологически реалистичных нейронов и 50*106 пластичных синапсов. Платформа имеет специально разработанный язык описания сети PyNN. Технологическая платформа предполагает многоядерную реализацию на основе архитектуры ARM, что позволяет произвести масштабирование системы. Каждый блок системы имеет 8 ядер и 128мбайт общей оперативной памяти, что позволяет имитировать до 16000 нейронов и 8000000 синапсов в режиме реального времени

с энергопотреблением в 1 Вт. Используется планарная кремниевая технология 180 нм.

Экспериментальные разработки IBM “True North” представляют собой размещенные на одном чипе 64x64 микропроцессора (ядра). Кристалл изготовлен по 28 нм кремниевой планарной технологии на заводе Samsung. Каждое ядро включает в свой блок SRAM, Router, Neuron, Token Control и Scheduler. Разработанная микросхема позволяет моделировать работу 1 миллиона нейронов и 250 миллионов синапсов. Архитектура позволяет производить масштабирование, на текущий момент построены системы из 16 миллионов нейронов с 4-мя миллиардами синапсов. Энергопотребление чипа составляет 100 мВт, плотность энергопотребления 20 мВт/см^2. Для упрощения программирования чипа разработаны специальные библиотеки и переработан ряд существующих алгоритмов [71]. Рассмотрев основные достижения в области создания ИНС и нейроморфных архитектур, перейдем к обзору основных концепций и подходов к имплементации искусственных нейронов.

Физические реализации искусственного нейрона. Одним из ключевых факторов, влияющих на возможности аппаратной реализации ИНС, является физическая реализация искусственного нейрона, в связи с чем, целесообразно рассмотреть технические решения искусственных нейронов.

Применение электромагнитных волн для передачи и обработки информации продолжает занимать исследователей во всем мире. Преимущества, получаемые при применении данного подхода, нашли свое отражение и в аппаратной реализации ИНС [72]. Физические реализации искусственных нейронов на оптоэлектронной элементной базе имеют внушительные габариты. Реализация представляет оптическую матрицу с применением VSCEL лазеров, линз и фотодетекторов и т. д., подробнее см. [73]. Нейроны обладают высокой скоростью обработки информации, но для достижения высокой скорости обучения сети приходится использовать ИНС с неизменяемыми весовыми коэффициентами (Fixed-weight Learned), что в свою очередь дополнительно осложняет процесс обучения. Масштабирование связности сети ограничивается размерами

пространственных модуляторов света и на текущий момент составляет 256x256 нейронов по принципу «каждый с каждым» рис 11. Для обработки выходных сигналов требуется реализация гибридных схемотехнических решений.

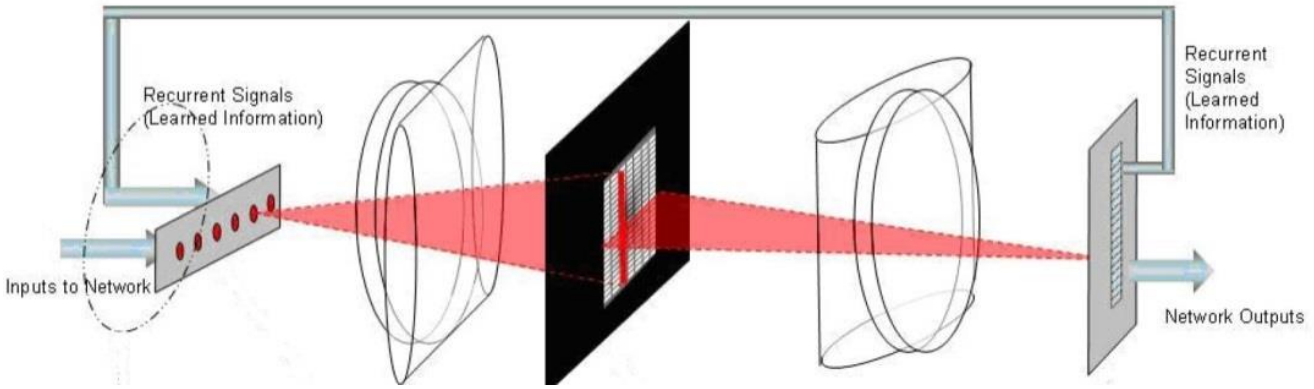


Рисунок 11. Пример реализации оптических искусственных нейронов [73]

Другой подход к вычислениям с применением оптоэлектронных технологий рассматривается в статье [74]. Техническое решение позволяет избежать применения линз для распространения сигналов от источников излучения до приемников, но тем не менее, все еще требует значительных габаритов при реализации. Пример построения многослойного персептрона с модуляцией оптического сигнала на задержках представлен в работе [75]. Полученная точность работы ИНС составляет 99,7%. Пример использования технологии WDM для анализа главных компонент рассматривается в исследовании [76]. Авторы указывают потенциальную возможность применения подхода в обработке спайковых сигналов в фотонных системах и процессорах. Проект решения ориентированного на более низкое энергопотребление и снижение количества активных элементов в схеме вычислений рассмотрен в публикации [77]. Основным недостатком подхода являются линейные размеры элементной базы для создания схем вычислений.

Одним из доминирующих направлений создания искусственных нейронов на кристалле является технологии на основе мемристоров. Применение мемристоров позволяет реализовать схемотехнические решения воплощающие синапсы более компактно в сравнении с традиционными подходами рис. 12. Отдельно стоит отметить, что применение мемристоров в обработке сигналов,

сопровождается переходом к аналоговому способу вычисления выходных сигналов или же нечетким множествам.

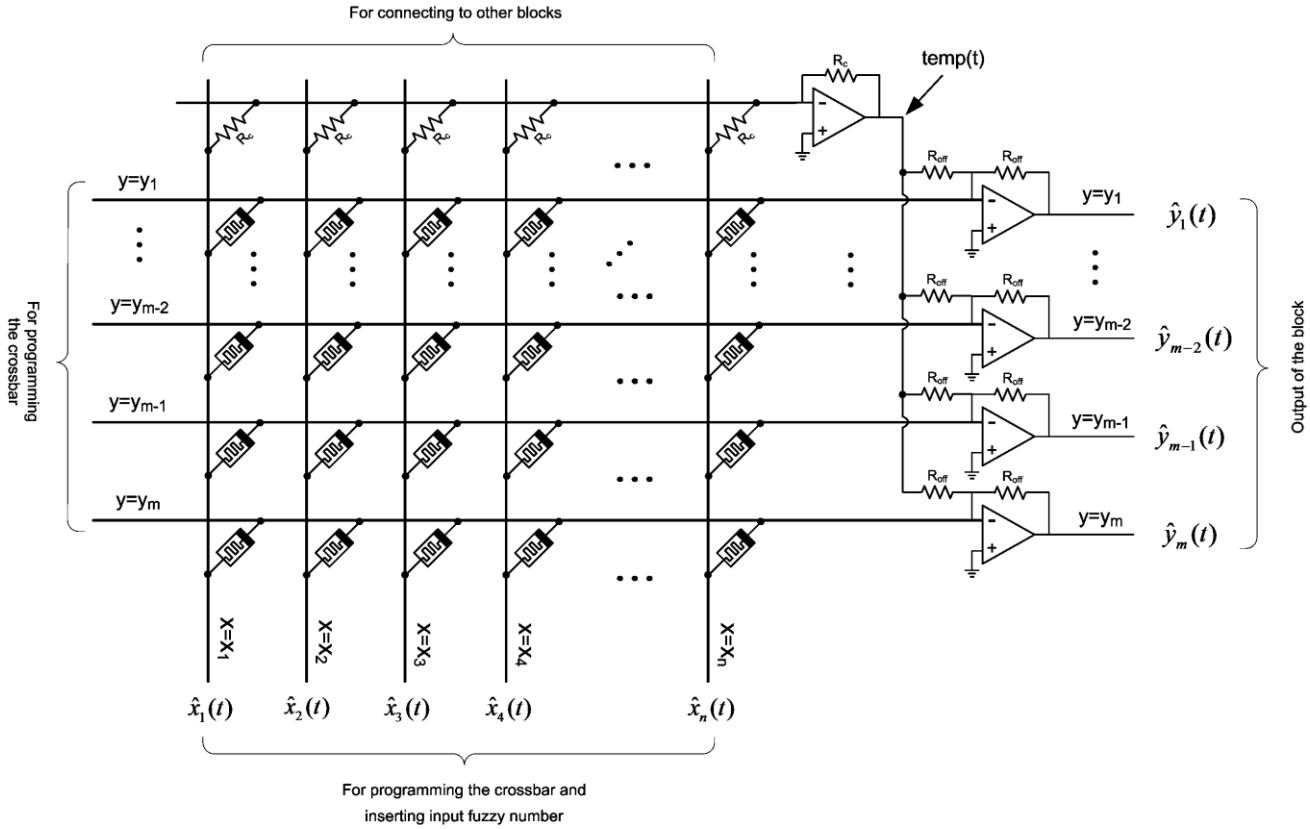


Рисунок 12. Искусственные нейроны с применением мемристоров. [78]

Представленная схема рис. 12 имеет неоднозначность задания весовых коэффициентов связей при применении аналоговых сигналов передачи информации и низкую плотность представления информации на синапсах при реализации цифровых сигналов. В качестве преимуществ решения стоит выделить высокую связность нейронов, за счет применения мемристорных кросс-баров. Пример вариации нейрона с постсинаптической обработкой предложен в работе [79]. Активационная функция искусственного нейрона в обоих предложениях представлена операционными усилителями и компараторами создаваемых по CMOS технологии.

Техническое решение использующее эффект изменения сопротивления на основе вращения магнитного домена для задания коэффициентов связности представлено в работе [80]. Связи между элементами сети реализуются посредством DWM (Domain Wall Magnet – англ.) синапсов, представляющих из

себя два магнитных домена разделенных диамагнетиком. Распространение сигнала происходит от одного домена к другому, что позволяет управлять проводимостью канала, изменяя поляризацию одного из доменов от сонаправленной к противоположно направленной. Функции активации реализуется MTJ (magnetic tunnel junctions – англ.) логическим вентилем, рис. 13.

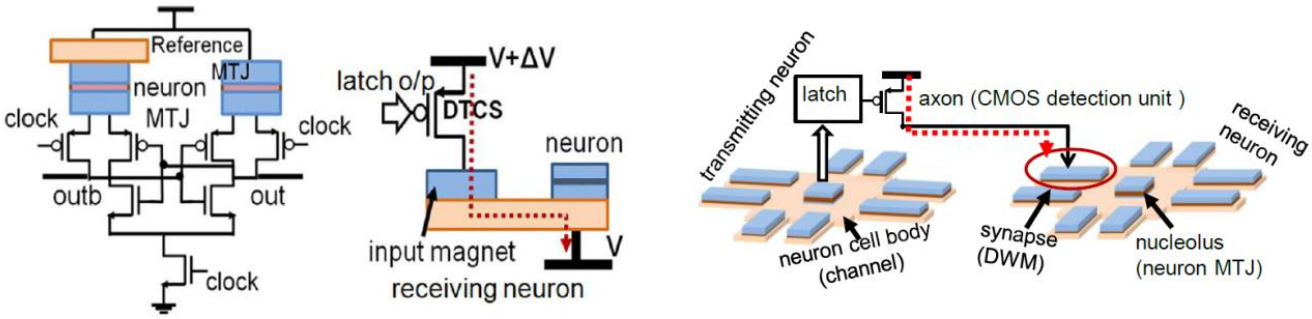


Рисунок 13. Искусственный нейрон MTJ [80].

К недостаткам предлагаемой архитектуры с точки зрения ИНС относится сравнительно низкий уровень связности сети, реализуемый на подобных нейронах. Данный подход наиболее перспективен для реализации клеточных автоматов (CA – cellular automata, англ.) и клеточных нейронных сетей (CNN – cellular neuron networks, англ.) с не большим (до 20 связей на элемент) уровнем связности.

Использование технологий на основе транспорта заряда ионами в гранулированном SiO_2 рассмотрено в работе [81]. В предлагаемой конструкции стоит отдельно выделить отсутствие жестко заданных связей между нейронами. Искусственный нейрон состоит из контактных площадок IZO создаваемых на гранулированном P-допированном оксиде SiO_2, нанесенном на подложку из Si. В процессе подачи импульсов, между контактными площадками аксонов и синапсов в объеме P-допированного оксида кремния образуются проводящие каналы. Схематическое изображение процесса образования канала между контактными площадками, представляющими предсинаптическую и постсинаптическую коммутацию синапса, показано на рис. 14.

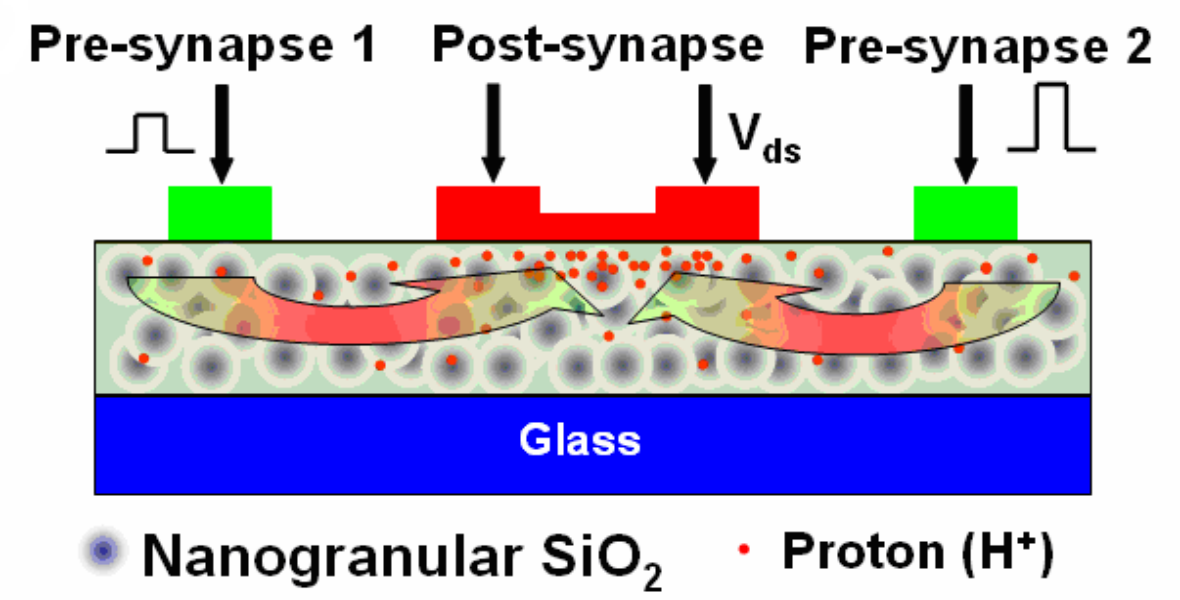


Рисунок 14. Механизм обработки сигналов [81].

В рамках работы рассматривается только обработка сигналов в синапсах, тем не менее, за счет использования дополнительных площадок возможно задание смещения относительно суммы входных сигналов от синапсов. Предлагаемому подходу свойственны ограничения связности порядка 10-30 связей между нейронами, в связи с планарным размещением структуры на кристалле. Вторым недостатком является низкая скорость работы, измеряемая в миллисекундах. Для увеличения связности требуется организация иерархических структур.

Предложенное в работе [82] техническое решение по созданию искусственных нейронов объединяет применение мемристоров и MTJ. Для задания коэффициентов на входных сигналах предлагается использовать матрицу из спинтронных мемристоров реализованную на кроссбаре. Суммирование входных сигналов (в данном случае токов) осуществляется на верхних или нижних проводниках кроссбара. В качестве вычисляющей активационной функции используется пороговая функция, реализованная с применением MTJ нейрона рис. 15. Моделирование проектного решения показывает улучшение параметров энергопотребления на два три порядка в сравнении с реализацией активационной функции искусственного нейрона по CMOS технологии. Авторы, отмечают возможность применения на кроссбаре, в качестве устройств выполняющих умножение входного сигнала синапса на коэффициент, PCM (Phase

change memory – англ., память на основе изменения фазового состояния) устройств.

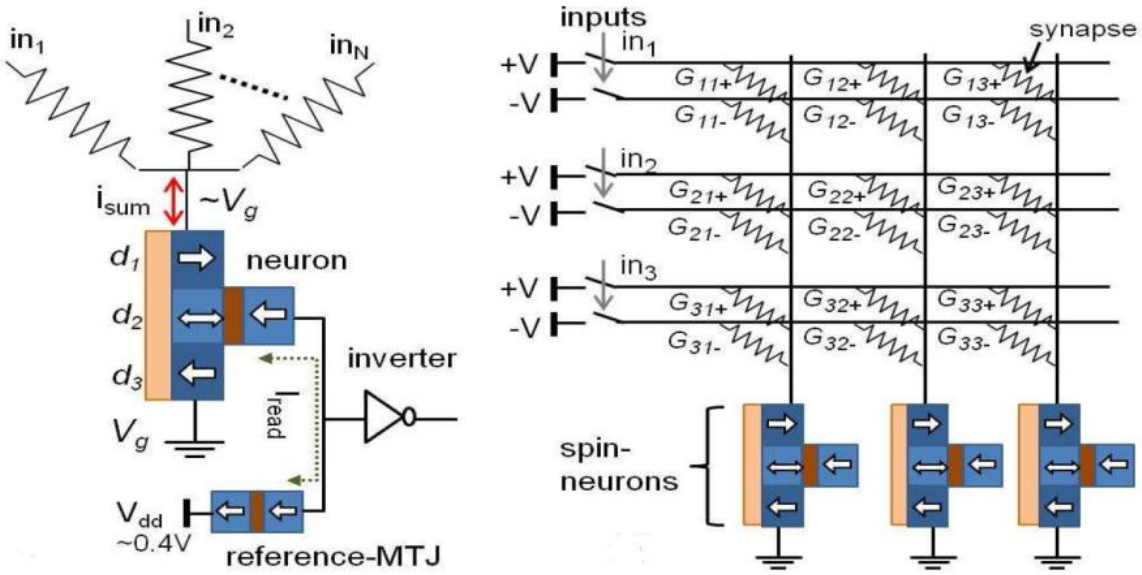


Рисунок 15. Искусственный нейрон на мемристорах и MTJ логике [82]

Общим недостатком применения мемристоров и PCM устройств является их ограниченное количество циклов переключения, что в свою очередь препятствует встраиванию алгоритмов обучения сети в устройство. Таким образом, требуется предварительное моделирование обученной сети с определением задаваемых коэффициентов для синапсов, после чего производятся программирование сети на кристалле. Немаловажным преимуществом рассматриваемой технологии, является полярность сигналов синапсов, что при должной проработке позволит реализовать сигналы торможения активности нейронов. Недостатком применения обоих полярностей в данном решении является, удваивание реализуемых на кроссбаре устройств задания коэффициентов и разрядность в один бит на выходе нейрона.

Модель аппаратной реализации нейрона, основанная на технологии SOT (Spin-orbit torque – англ.) представлена в работе [83]. Разработанное схемотехническое решение демонстрирует в снижение энергопотребления в 3 раза по сравнению с 45nm CMOS технологией. Весовые коэффициенты синапсов задаются на кроссбаре мемристоров или PCM. В качестве активационной

функции нейрона используется пороговая функция, реализованная аппаратными средствами. Техническое решение представлено на рис. 16.

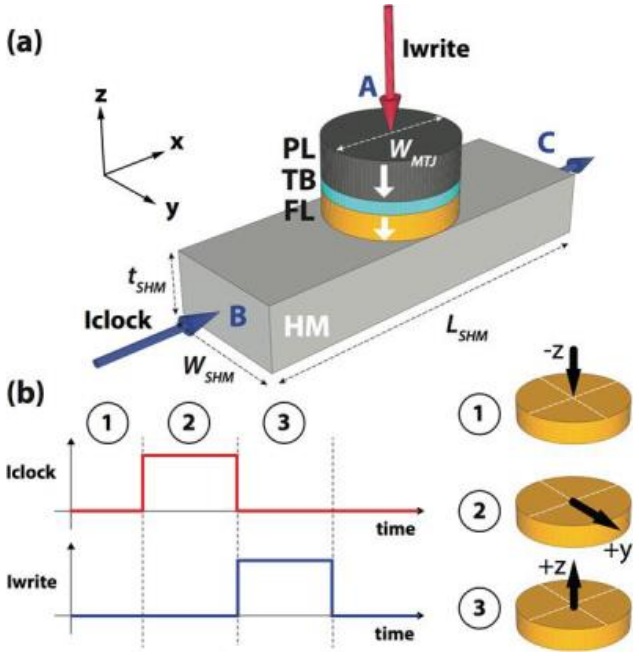


Рисунок 16. Пороговая функция SOT нейрона [83]

Спроектированная конструкция вычисляющей функции нейрона включает: проводник из тяжелого металла с высоким показателем орбитально-спинового взаимодействия (HM), слой с перпендикулярной магнитной анизотропией (PL), диэлектрический слой (TB, выполняет функцию туннельного барьера) и свободный слой (FL, используется для регистрации суммы коэффициентов от синапсов), рис. 16a. Используемые для вычислений дискретные состояния намагниченности свободного слоя рис. 16b. Процесс регистрации изменений состояния свободного слоя осуществляется по двухтактной схеме, механизм аналогичен MQCA (magnetic quantum-dot cellular automata). На первом этапе ток $\boldsymbol{I_{clock}}$ проходящий по проводнику из тяжелого металла переводит свободный слой в неустойчивое состояние намагниченности рис. 16b2. На втором этапе полярность инжектируемого тока $\boldsymbol{I_{write}}$ проходящего через свободный слой производит переключение свободного слоя в одно из двух устойчивых состояний намагниченности. Недостатки предлагаемого подхода удваивание устройств задания коэффициентов и разрядность в один бит на выходе нейрона.

Другой метод построения спиновых нейронов на основе наномагнитов рассматривается в работе [84]. Свободный слой создается из мультиферроика или же магнитострикционного материала. Переключение состояния осуществляется под действием приложенного напряжения. Конструкция состоит из двух электродов (A, A’) и находящегося между ними вычисляющего элемента (MTJ) размещенных на тонком слое пьезоэлектрической пленки рис. 17. Вычисляющий элемент состоит из слоя наномагнетика с неизменяемой поляризацией, промежуточного слоя и магнитострикционного слоя контактирующего с пьезоэлектрической пленкой. На A и A’ подается напряжение от входов нейрона, в качестве весовых коэффициентов используются сопротивления (r_1, r_2,…, r_n). Смещение пороговой функции задается током от источника I. Смешение магнитного поля B направлено вдоль оси намагниченности магнитострикционного слоя. Под действием приложенного к электродам напряжения происходят механические деформации в пьезоэлектрическом слое, которые передаются на магнитострикционный слой вычисляющего элемента, что приводит к изменению сопротивления MTJ структуры. Изменение сопротивления MTJ структуры приводит к изменению напряжения на выходе.

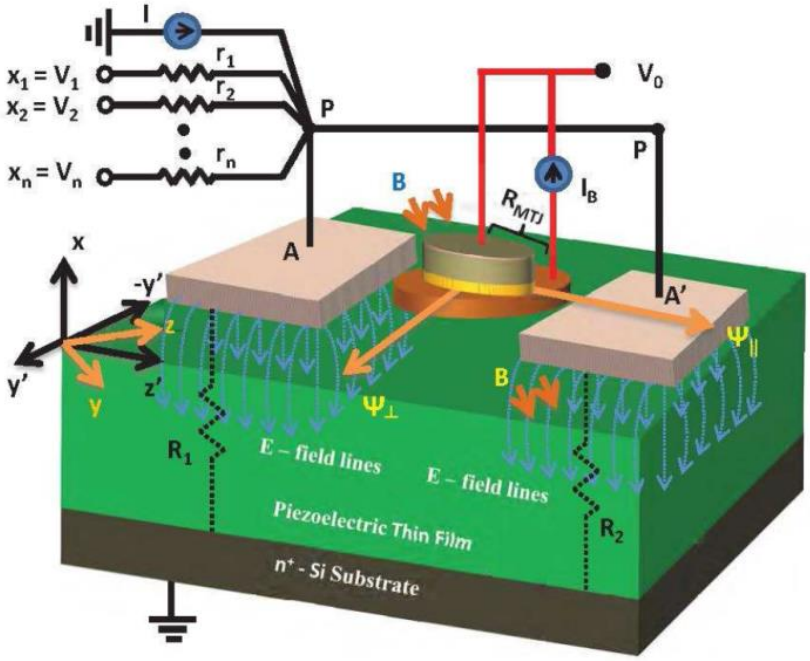


Рисунок 17. Схема функционирования Straintronic spin-neuron [84]

Спроектированное решение показывает большую энергетическую эффективность работы при комнатной температуре в сравнении с ранее рассмотренными подходами. Недостатками структуры является большее число компонентов и увеличение площади на кристалле.

Применение мемристинвых конденсаторных структур для реализации синапсов и вычисляющей пороговой функции нейрона рассматривается в статье [85]. Схемотехнические решения реализации синапсов и вычисляющей функции, а так же диаграмма работы представлены на рис. 9. В процессе моделирования SPICE модели сборки на вход N_1 подавались 100 нс сигналы 5В, динамика выходного сигнала V_{out} показана на рис. 18а.

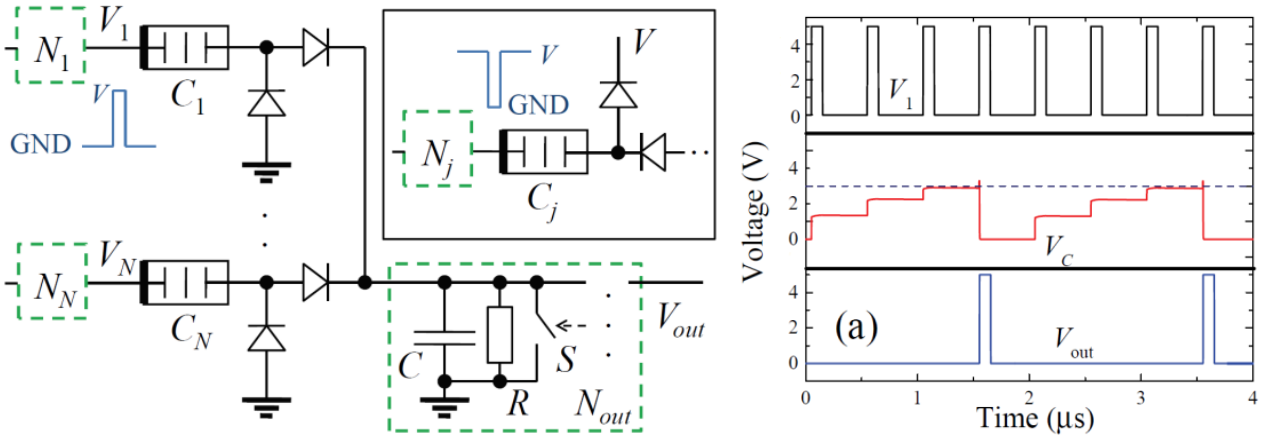


Рисунок 18. Схемотехнические решения позитивной связи NN и N1, Nj схемотехническое решение тормозящей связи, вычисляющая функция нейрона в пунктире. а) Диаграмма работы [85]

К недостаткам подхода реализации относится необходимость использования диодов при реализации синапсов, что в свою очередь существенно увеличивает площадь на кристалле. Преимуществом подхода является более низкое энергопотребление энергии, чем при реализации мемристорных кроссбаров.

Реализация искусственных нейронов только на основе стандартной CMOS технологии, как правило, является схемотехническим решением на основе аналогово-цифровой обработки сигналов. Решения заключаются в специализации архитектуры СБИС для нейровычислений, без предварительной разработки самой элементной базы.

CMOS технологии с гибридной обработкой (аналогово-цифровой) сигналов позволяют реализовывать ИНС с большим набором настраиваемых параметров сети. Так, например, ETANN 80170NX позволяет менять наклон активационной сигмоидальной функции. Чип содержит 64 нейрона и 64 входных порта соединенных с каждым нейроном. Предусмотрены аналоговый и цифровой режимы работы. Веса лежат в диапазоне от -2.5 до 2.5, скорость прохождения сигнала по слою 1.5 мкс, с точностью 6 бит.

В качестве второго примера гибридной реализации может быть представлен чип CLNN64. На кристалле размещены 64 нейрона связанные между собой 1024 двунаправленными синапсами. Обработка сигналов полностью реализована на аналоговой схемотехнике, однако веса синапсов хранятся в цифровой форме с точностью 5 бит. Структура сети RBF, алгоритм обучения встроен.

Таким образом, аппаратная реализация искусственных нейронов накладывает ограничения на создаваемые на их основе искусственные нейронные сети. К ограничениям такого рода относятся: реализуемые функции активации, реализуемые функции обработки входящих на нейрон сигналов (суммирование или же умножение), наличие сигналов разной полярности (которые могут выступать как тормозящие и активационные), тип представления сигналов в синапсах и множество выходных значений на выходе нейрона. С точки зрения вычислительной способности искусственного нейрона ограничения могут быть представлены минимальным порогом срабатывания функции активации, что может накладывать ограничения на масштабирование систем.

Аппаратные решения искусственных нейронных сетей. Помимо ограничений, накладываемых применяемыми на кристалле искусственными нейронами, свои ограничения в реализуемые ИНС вносят и схемотехнические решения по трассировке сигналов между нейронами. Так, например, наличие обратных связей в сети позволяет моделировать временную динамику в случаях, когда не используется STDP механизм. Режимы работы сети асинхронный и синхронный определяют возможности по реализации алгоритмов вычислений и

скорости их исполнения на той или иной платформе. Так же ранее, исследователями рассматривались [86] и такие аспекты реализации ИНС систем как параллелизм обработки данных. В работе ученые выделили четыре уровня параллелизма: параллелизм работы слоев, параллелизм работы нейронов, параллелизм прохождения сигналов по синапсам и параллелизм обработки битов сигнала. Обзор эффективности программного построения нейронных сетей на вычислительных кластерах со специальной архитектурой процессоров представлены в статье [87].

На сегодняшний день самым экономически выгодным решением по аппаратной реализации ИНС с малым количеством нейронов являются реализации на ПЛИС. Возможность реконфигурации схемы обработки сигналов и коммутации внутри микросхемы позволяет ускорить обработку данных на аппаратном уровне и предоставляет более гибкие механизмы по перенастройке уже готовых решений в сравнении со спроектированными чипами. Примером может служить работа [88], в которой осуществлен анализ ускорения выполнения вычислений сверточной сетью ИНС программируемой на FPGA. Другими немаловажными преимуществами являются скорость разработки прототипов на данной аппаратной основе и возможность динамически перенастраивать оборудование для обработки разных участков алгоритма [89], что позволяет, повысить количество эффективной функциональности на единицу пощади кристалла FPGA. Максимальная эффективность данного подхода выражается формулой:

$$q = \frac{r}{(s-1)}$$

где, q – точка безубыточности, r – время в циклах для перенастройки ПЛИС S – общее время вычислений после каждой реконфигурации. Необходимый диапазон весов для решения задачи классификации определяется как [-p,p] и оценивается через минимальное расстояние между классами D по формуле:

$$d = \frac{\sqrt{n}}{2p}$$

где n размерность входного вектора, p – целое число. Пример создания ИНС на ПЛИС представлен в работе [90]. Помимо проблемы снижения эффективности вычислений при переходе к дискретным моделям вычислений на ПЛИС коллектив сотрудников обозначил немаловажную проблему конечного размера внутренней сети линий в работе [91], что так же приводит к снижению производительности вычислений и ограничению на реализацию проектов содержащих большое количество нейронов.

Один из вариантов алгоритма реализации ИНС на ПЛИС предложен в [92]. Исследователи выделяют три основных последовательных стадии: дискретизацию модели одного искусственного нейрона, преобразование модели в структуру логических вентилей и оптимизацию структуры путем исключения избыточности. Дискретизация модели производится в два этапа. На первом этапе аналоговые входные связи нейронов представляются в виде битовых групп входных диапазонов сигнала, за счет вычисления весовых коэффициентов для каждой входной цепи нейрона. На втором этапе нейроны с отрицательными весовыми коэффициентами заменяются эквивалентными нейронами с положительными значениями весовых коэффициентов. Перевод в логическую структуру начинается с задания в убывающем порядке массива весовых коэффициентов. Далее, путем последовательных итераций создается функция логического затвора на основе анализа весов. В случае высокой сложности нейрона функция может быть разбита на подгруппы нейронов. На последней стадии из конфигурационного файла исключаются избыточные элементы.

Полное распараллеливание архитектуры ИНС при решении вычислительной задачи не является обязательным условием для ее эффективного решения. В работе [93] представлен анализ влияния параллельности работы вычисляющих блоков на FPGA демонстрирующих совпадающую по точности производительность для смешанного (параллельно последовательного подхода) в

ряде случаев. Для других случаев точность вычислений не значительно снижается. Учитывая ограниченность ресурсов FPGA, данный представляет интерес при использовании аппаратных платформ со схожими ограничениями и возможностями.

Тенденции последних лет позволяют говорить о смене подхода при реализации аппаратной реализации ИНС на кристалле или плате. Основная суть применяемых технологических особенностей заключается в реализации некоторого количества «нейроядер» (neurocore) в виде блоков на кристалле. Примером подобного подхода служит работа коллектива Johannes Schemmel [94]. В статье рассматривается технология создания ИНС разработанная в рамках проекта FACETS. На каждый нейрон предполагается более чем 10 000 связей. Для реализации ядра использована технология 180 нм, скорость работы схемы превышает скорость работы биологического аналога в 10^3-10^5 раз. Энергопотребление большинства цепей при реализации искусственных нейронов лежит в диапазоне 100 нА – 1 мкА.

Биологически эквивалентная модель нейрона на основе мембранных потенциалов была дополнена и объединена с моделью «Integrate and Fire», сохранив все шесть известных режимов функционирования сети. Схема модели искусственного нейрона на рис. 19.

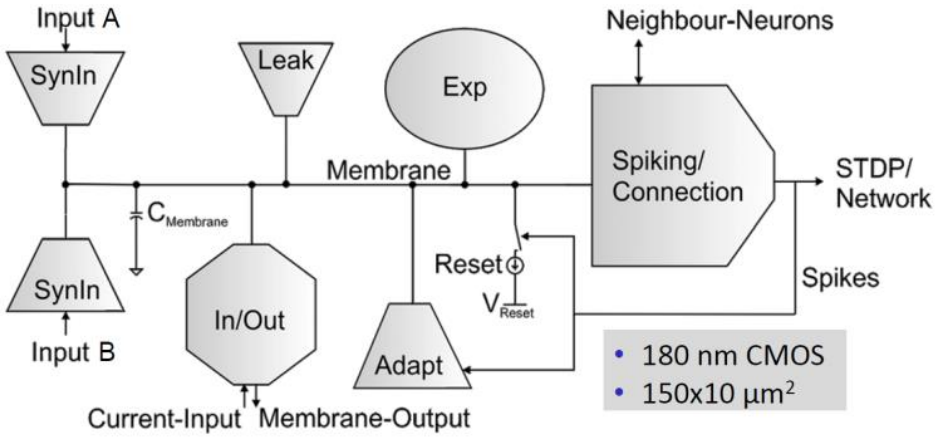


Рисунок 19. Модель нейрона на основе мембранных потенциалов [94].

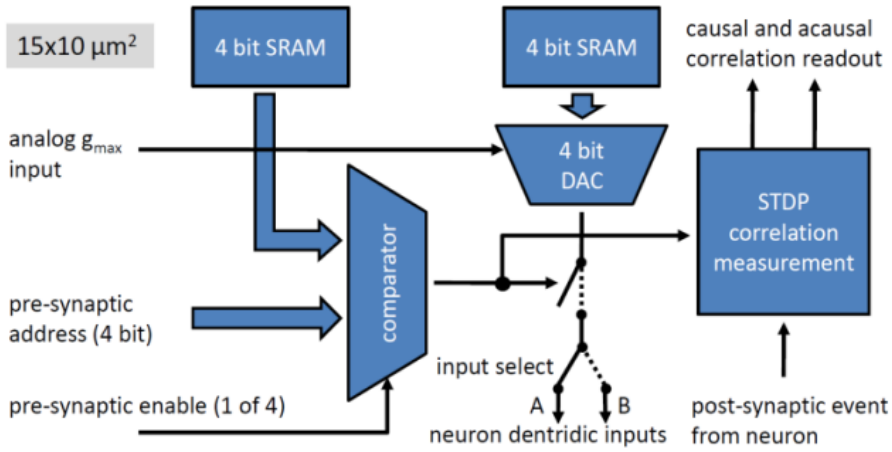


Рисунок 20. Схема синапса модели нейрона на основе мембранных потенциалов [94].

В основе нейрона лежат аналоговые ядра сети (ANC – англ.) к которым могут присоединяться цепи дендритных мембран (DenMen). Каждая цепь подключена к 2241 синапсу и программируется 23 входными аналоговыми параметрами на основе ячеек памяти с плавающим затвором, что в сою очередь позволяет создавать нейроны с переменным количеством синапсов. Ячейки расположены между DenMen цепями и блоком построения нейрона. Схема модели синапса представлена на рис. 20. Синапс имеет от 6 до 8 бит разрядности программируемых SRAM. Реализованы два механизмами пластичности. Взаимодействие ANC представлено на рис. 21.

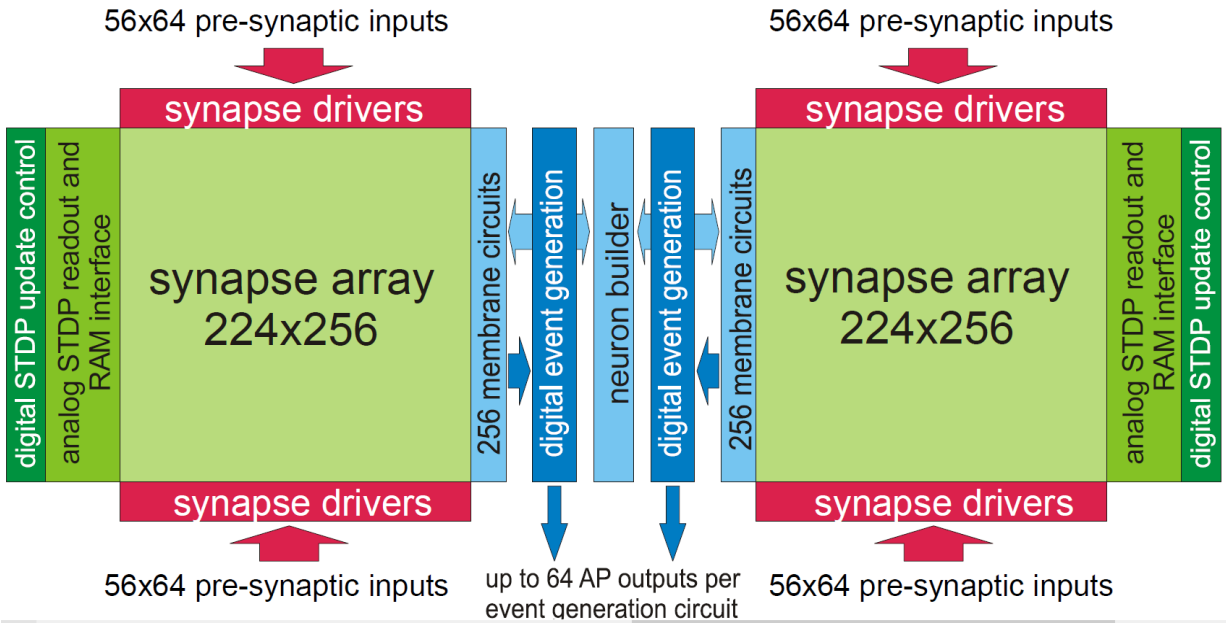


Рисунок 21. Схема работы ANC [94].

Помимо реализации архитектуры сети на кристалле так же схемотехническими решениями могут обеспечиваться и механизмы обучения. В качестве примера можно проектное решение предлагаемое авторами в работе [95]. В исследовании анализируется два механизма обучения самоорганизующихся карт Кохонена. Схемотехническое решение включает 4 основных блока рис. 22.

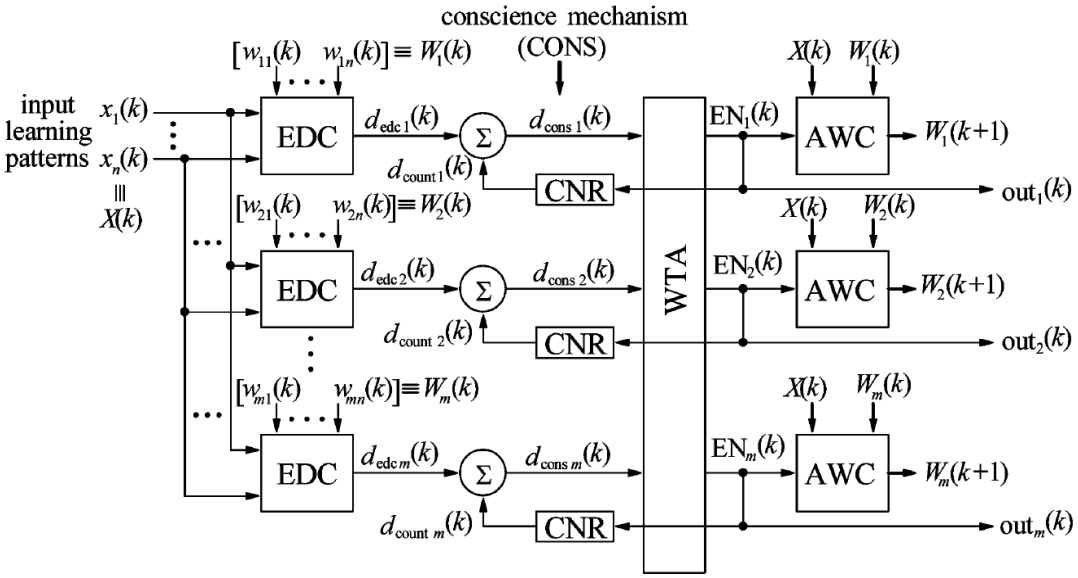


Рисунок 22. Блок схема реализации модифицированного механизма обучения самоорганизующихся карт Кохонена. EDC – блок вычисления эвклидова расстояния, WTA – блок определения победившего нейрона, AWC – блок адаптивного изменения веса, CNR – аналоговый счетчик [95].

Расчетные данные по реализации данного подхода продемонстрировали энергопотребление в 1 мВт с активным режимом обучения и 0,7 мкВт в режиме работы сети без обучения.

Применение ANC (аналоговых нейронных ядер) в рамках которых каждый нейрон связан с каждым, поднимает вопрос о типах возможных соединений между самими блоками. В работе [96] продемонстрировано, что для задачи воспроизведения образа моделью ассоциативной памяти на основе ИНС Хопфилда, наличие регулярно заданных связей между блоками (например, в соответствии с шаблоном связности как в КА) повлекло за собой более низкую производительность в сравнении с нерегулярной связностью между блоками.

Исследование [97] демонстрирует пример построения RBF сети из 3-х слоев, где нейроны скрытого слоя и нейроны выходного слоя имеют

стохастические функции активации. В сравнении с решениями на основе аналогичных сетей с детерминированным выходным слоем, схемотехнический подход имеет преимущества в площади, занимаемой на кристалле и короткой спецификации применяемых логических блоков. Тем не менее, спроектированное решение потребляет больше энергии, имеет большее время вычислений и в задачах распознавания изображений показало худшую точность. Эффективность распознавания снизилась на 1.3%. Увеличение длины входного битового потока приводит к значительному снижению производительности при данной реализации. Структура сети представлена на рис.12.

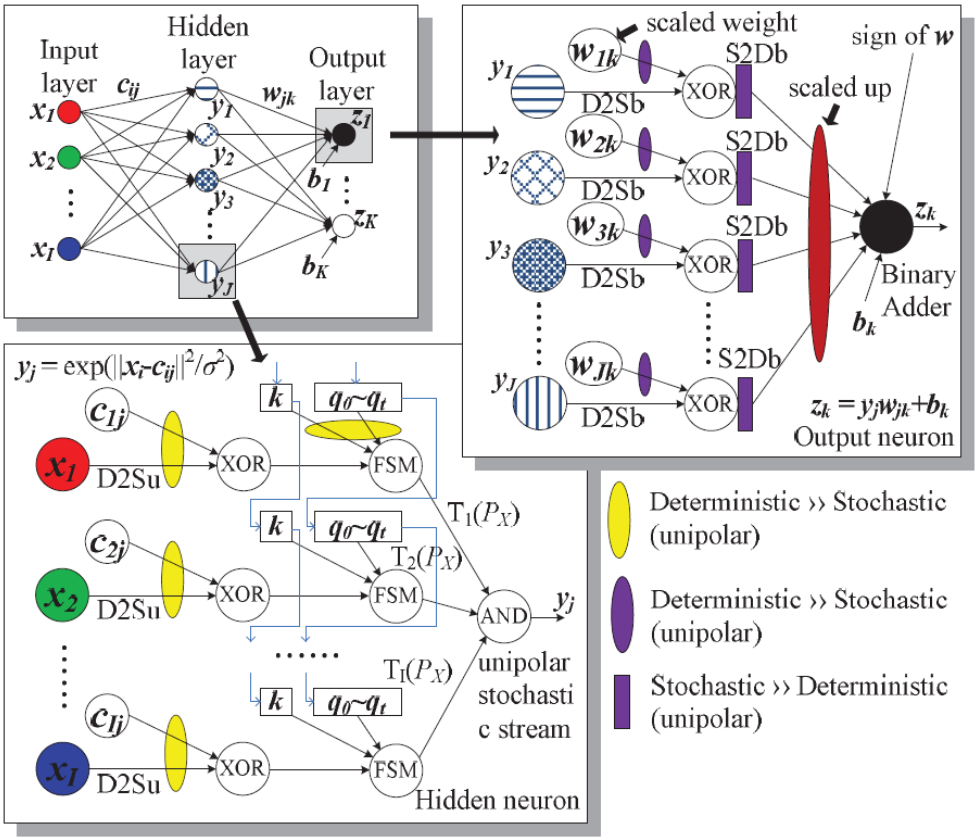
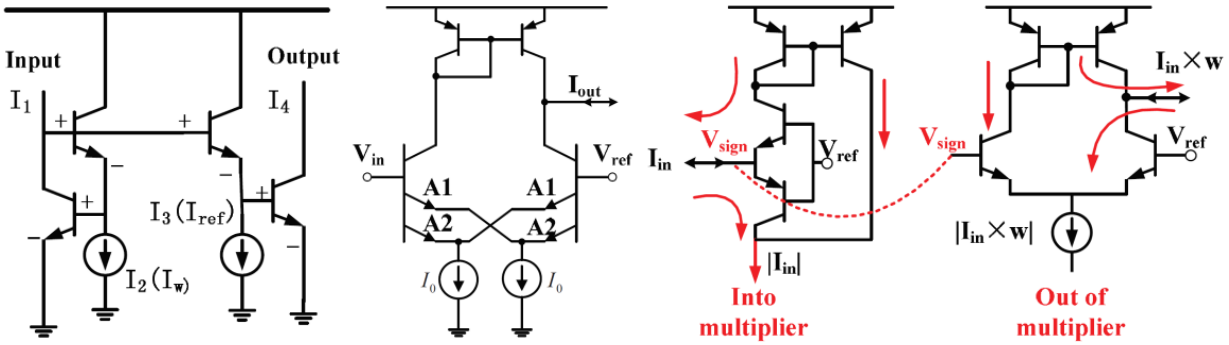


Рисунок 23 Реализация RBF сети с применением стохастической и детерминированной логики [97].

Техническое решение построения ИНС с применением биполярных транзисторов представлено в статье [98]. Авторы рассматривают два схемотехнических решения, одно из которых ориентировано на точность, второе на более простую реализацию. Рис. 24.

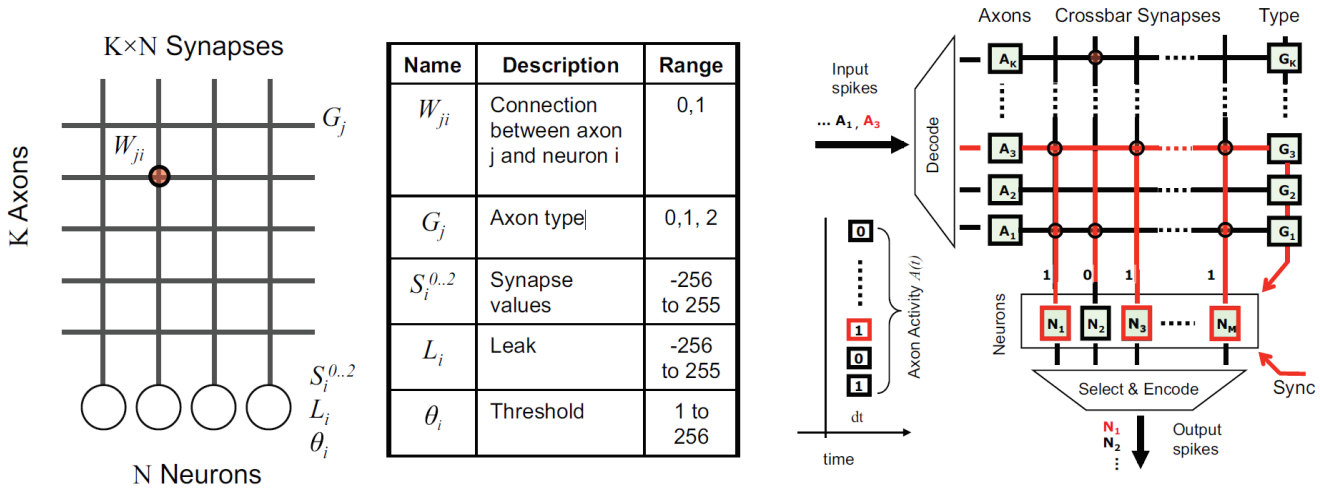


а б в г

Рисунок 24. Схемотехнические решения на биполярных транзисторах. а – решение для высокой точности; б – предложенное решение на основе дифференциальной пары; в,г – схемы учета знака при умножении на коэффициент.

Примечательно, что более точное решение позволяет выполнять умножение на весовой коэффициент только со скоростью, не превышающей 1 ГГц. Вторым преимуществом подхода помимо простоты реализации является широкий линейный диапазон напряжений, образующийся за счет суперпозиции перемножения тангенсальных гиперболических функций.

Вторым подходом при построении искусственных нейронных сетей на основе нейроядер является применение в качестве строительных блоков схемотехнических решений на основе цифровой логики цифровых нейронных ядер (DNC). В качестве примера можно привести работу [99]. Архитектура рассматриваемого блока представлена на рис. 25.



Name	Description	Range
W_{ji}	Connection between axon j and neuron i	0,1
G_j	Axon type	0,1, 2
$S_i^{0..2}$	Synapse values	-256 to 255
L_i	Leak	-256 to 255
θ_i	Threshold	1 to 256

Рисунок 25. Архитектура блока DNC.

Нейросинаптическое цифровое ядро состоит из 256 нейронов, 1024 аксонов и 2^{18} синаптических двоичных связей. Модель работы нейрона «integrate and fire». Алгоритм предусматривает последовательную подачу сигналов состоящих из типа сигнала активации (возбуждение, торможение или игнорирование) на декодер. Сигнал вызывает активацию определенной строки, после чего входные сигналы суммируются на нейроне. Порог функции задается в диапазоне от 1 до 255. Вывод данных осуществляется последовательно.

Проектирование объединения цифровых нейросинаптических ядер DNC на кристалле продемонстрированно в статье [100]. Каждое цифровое ядро оснащается маршрутизаторами, создающими двумерную ячеистую сеть из ядер размерностью 64x64. Маршрутизаторы оснащены 5 портами (север, юг, восток, запад и локальный). Задержка спайка осуществляется 4 битами. Для маршрутизации спайка на выходе нейрона используется 8 битный абсолютный адрес и 9 битный относительный адрес аксона синаптического ядра рис. 26. Сигналы от ядра к ядру передаются по мультиплексированным каналам с разнесением по времени.

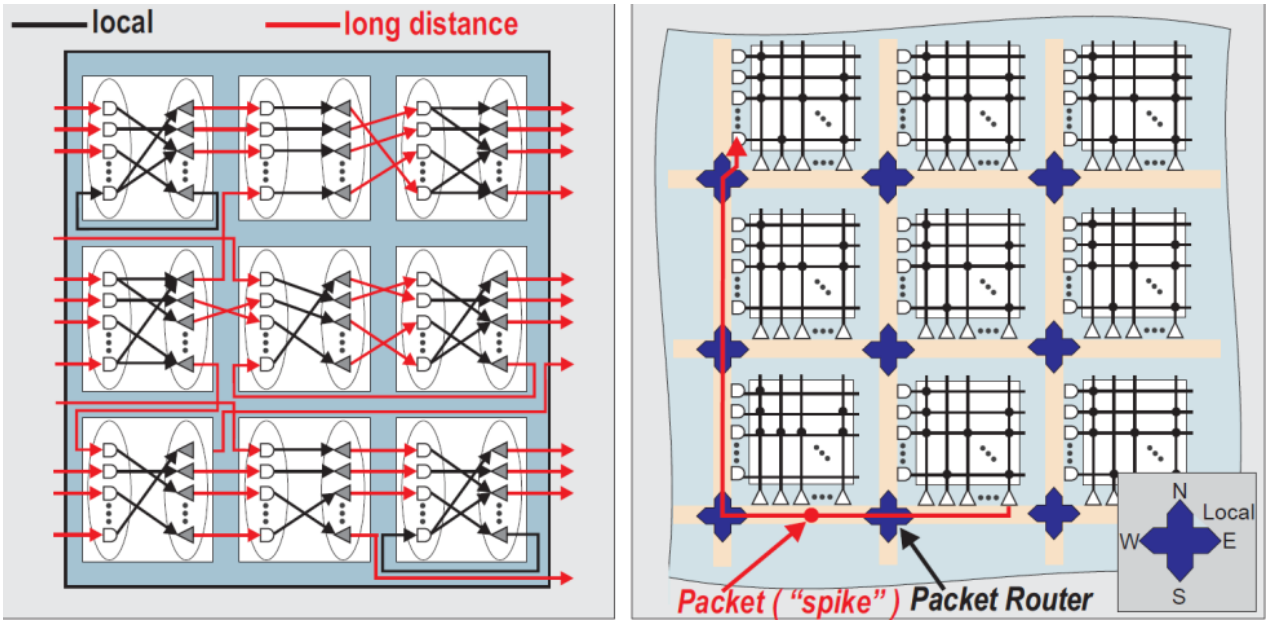


Рисунок 26. Схема маршрутизации спайка.

Как следует из рассмотренного материала, основным направлением развития архитектуры при реализации систем на кристалле вне зависимости от

реализации способа вычислений (аналоговая схемотехника или цифровая) является имплементация крупных схемотехнических блоков в виде нейросинаптических ядер. Отдельно в данном контексте стоят решения по воплощению ИНС средствами FPGA, где исходя из особенностей архитектуры, применяется перепрограммирование связности участков сетей в процессе вычисления, что дополнительно позволяет увеличить скорость обработки информации на одном кристалле. К основным недостаткам реализации на ПЛИС относится крайне низкая эффективность использования площади кристалла и сравнительно малое количество имплементируемых нейронов.

Выводы из обзора литературы аппаратных реализаций искусственных нейронных сетей. Построение искусственных нейронных сетей базируется преимущественно на двух подходах совпадающих с областями схемотехники. Первый подход обобщает цифровые технические решения имплементации ИНС и включает в себя, помимо проектирования специализированных вычислителей, таких как True North, программно-аппаратные методы ПЛИС. Второй подход представлен гибридной и аналоговой схемотехникой специализированных архитектур и, как правило, не только учитывает особенности ИНС при проектировании схем обработки информации, но и ориентирован на применение новой компонентной базы, такой как мемристинвые элементы, PCM, SOT-элементы и многие другие.

Обобщая все положительные и отрицательные стороны имплементации искусственных нейронов и архитектур ИНС можно заключить следующее:

1. Наиболее перспективным подходом является гибридное схемотехническое решение реализации ИНС в виде ANC с организацией схемы цифровой передачи и коммутации сигналов между ядрами. Указанный метод в теории позволит добиться ускоренной обработки сигналов путем применения современной компонентной базы и избежать проблем с организацией маршрутизации сигналов между ядрами на кристалле.
2. Наиболее перспективным блоком для организации связей между искусственными нейронами выступает кроссбар. Данный подход позволяет

оптимизировать площадь, занимаемую на кристалле и упростить реализацию блока агрегации нейрона.

3. Несмотря на ограниченное количество циклов перезаписи, в качестве оптимального подхода к реализации блока учета входных сигналов выступают мемристивные элементы. Данный метод позволяет реализовать не только блок хранения весового коэффициента, но и блок умножения входного сигнала. Возможность реализации многоуровневых состояний дает существенное преимущество мемристивным элементам перед такими элементами как PCM, SOT и MTJ.
4. Исходя из необходимости реализации цифровых схемотехнических решений передачи данных и имплементации в структуру нейрона кроссбара, а так же применения концепции ANC, техническое решение разработки искусственного нейрона должно базироваться на гибридной схемотехнике.

2.2 Физика мемристоров

Описание первого мемристивного элемента впервые предложенное Chua [101] во второй половине прошлого века, формализует его как элемент памяти, меняющий свое состояние в зависимости от его текущего состояния и протекающий сквозь него заряд. Впервые указанные структуры были получены экспериментально в 2000-х г. в лаборатории HP [102]. В основе работы элемента лежит механизм обратимого изменения электрической проводимости в тонких пленках. К первым работам, направленным на изучение указанного эффекта можно отнести статьи коллективов Simmons`a и Dearnaley конца 1960-х годов.

Общее описание мемристора как элемента изменяющего внутреннее состояние от приложенного напряжения или же от проходящего сквозь структуру заряда позволило расширить понятие мемристивных элементов не только на элементы с изменяемой проводимостью. Так, например, в работе [103], помимо описания мемристора, приводится описание таких элементов как мемконденсаторы и меминдукторы. Поскольку, исследование свойств меминдуктивных и мемконденсаторных элементов плохо изучено, а технические реализации остаются под вопросом, данные элементы не подходят в качестве элементов искусственного нейрона и далее не рассматриваются.

Имплементация мемристоров на основе эффекта обратимого переключения электрической проводимости осуществляется в виде МДМ структуры рис .27. Металлы структуры выступают в качестве электродов. Диэлектрик выступает в качестве активного слоя, в котором и осуществляется функция памяти. Функция памяти активного слоя реализуется путем обратимого изменения его проводимости под действием тока или напряжения. В зависимости от материала активного слоя эффект обратимого переключения электрической проводимости может различаться в части превалирующих физико-химических процессов приводящих к изменению сопротивления. Для ряда процессов структура МДМ может быть дополнена дополнительным интерфейсным слоем улучшающим процесс переключения электрической проводимости и параметры структуры.

Рисунок 27. Два типа МДМ структур. Слева МДМ структура без дополнительного слоя, справа структура с интерфейсным слоем, упрощающим переключение состояний проводимости.

Особо стоит отметить, что формирование мемристора как элемента осуществляется проведением специальной операции – операции электроформовки. Указанная подготовительная операция представляет собой подачу импульсов тока или напряжения обеспечивающих первичное изменение проводимости. В результате подачи импульса непроводящее состояние диэлектрической пленки изменяется на низкорезистивное. Отличие операции электроформовки от операции перевода мемристора в высокоомное состояние заключается в недостижении состоянием активного слоя уровня сопротивления соответствующего мемристору до электроформовки. Операции переключения из высокорезистивного состояния в низкорезистивное состояние проводящиеся в последующем на МДМ структуре происходят под действием импульсов тока или напряжения.

Одним из подходов к объяснению процессов электроформовки является модель управляемого электрического пробоя. Согласно данным модельным представлениям в процессе формовки локально формируется область с низким сопротивлением, через которую и проходит ток. Последующие операции перевода мемристора в низкорезистивное состояние уменьшают локальную область с низким сопротивлением, в результате чего сопротивление возрастает. Применение операции перевода мемристора в высокорезистивное состояние увеличивает локальную область с низким сопротивлением рис. 28.

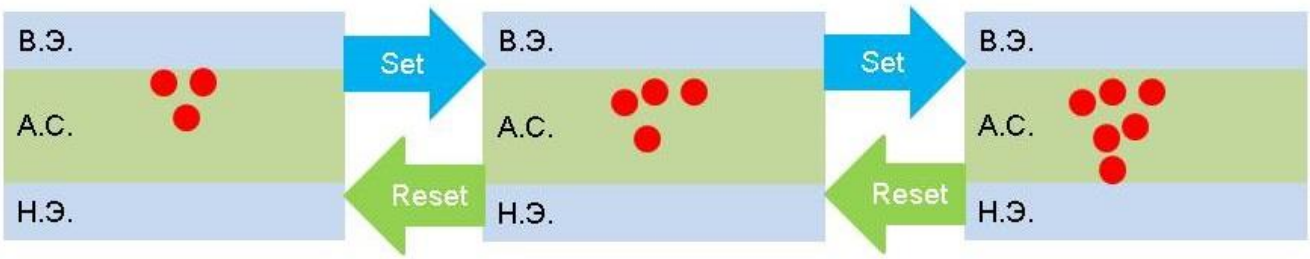


Рисунок 28. Процесс обратимого переключения электрической проводимости. В.Э. – верхний электрод, Н.Э. – нижний электрод, А.С. – активный слой, Set – переключение в низкорезистивное состояние, Reset – переключение в высокорезистивное состояние.

Существующие мемристоры могут быть разделены по механизму обратимого переключения электрической проводимости на биполярные и униполярные. К мемристорам с униполярным механизмом переключения относятся структуры в которых эффект обратимого переключения электрической проводимости зависит от абсолютного значения приложенного напряжения или абсолютного значения приложенного тока. К мемристорам с биполярным механизмом переключения относятся структуры в которых эффект обратимого переключения электрической проводимости зависит не только от значения приложенного напряжения или тока, но и от направления протекания тока. Вольт-амперные характеристики с отображением петель гистерезиса схематично представлены на рис. 29.

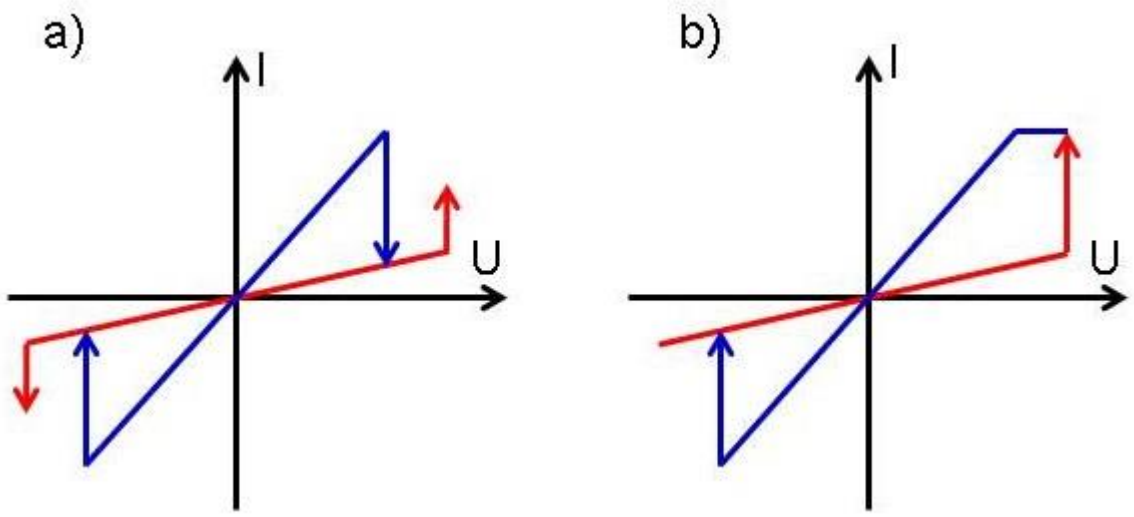


Рисунок 29. a) ВАХ униполярного мемристора b) ВАХ биполярного мемристора. Красная линия операция Set. Синяя линия операция Reset.

Применение униполярных мемристоров в качестве элемента коссбара существенно усложнит процесс переключения и настройки связи искусственного нейрона. Проблемы возникают в связи с большим разбросом параметров на элементах кроссбара, результатом которого могут являться: неполное (не до требуемого уровня проводимости) переключение элементов, частичное увеличение электропроводимости соседних элементов на кроссбаре, частичное уменьшение электропроводимости соседних элементов на кроссбаре. Применение в качестве элемента синапса на кроссбаре биполярного мемристора позволит избежать таких негативных эффектов как частичного уменьшения электропроводимости соседних элементов и неполного переключения элементов, что достигается за счет разной полярности операций переключения между состояниями проводимости.

Существует ряд подходов к описанию физических процессов механизма переключения электрической проводимости в тонких пленках. В качестве основного подхода к описанию применяется филаментарная модель проводимости и модель проводимости на основе барьера Шотки [104]. Под филаментом подразумевается локальная область активного слоя с низким сопротивлением, через которую протекает ток. Процесс образования филамента по данным ряда исследований может быть связан для униполярного механизма с процессами образования вакансий и нанокристаллитов в объеме активного слоя [105]. Процесс образования филамента в биполярных мемристорах на основе оксидов переходных металлов, как правило, связывают с окислительно-восстановительными реакциями и миграцией кислородных вакансий в объеме активного слоя [106]. Механизм обратимого переключения электрической проводимости на основе барьера Шотки исследователи связывают с изменением. В рамках механизма переключения проводимости на основе барьера Шотки один из интерфейсов электрода формирует омический контакт на стороне металла или оксида, на другой стороне формируется сам барьер, контактное сопротивление которого формируется распределением потенциала барьера. Исследования по

описанию изменения параметров барьера Шотки представлены в работах [107, 108].

Помимо механизмов обратимого переключения электрической проводимости, исследователи так же применяют разнообразные подходы к описанию физического процесса протекания тока (переноса зарядов между электродами). В качестве основных физических эффектов отвечающих за ток в структуре выступают следующие типы проводимости. Перенос заряда по ловушкам в объеме активного слоя [109]. Туннелирование электронов сквозь барьер с учетом температурного фактора [110]. Туннелирование электронов сквозь барьер [111]. Термоэлектронный механизм Шотки [112]. Электропроводность на основе эффекта Пула-Френкеля [113]. Механизм проводимости ограничения тока пространственного заряда [114]. Механизм Омической проводимости [115]. Помимо указываемых механизмов электропроводности в активном слое мемристоров, так же возможны и комбинации указанных механизмов проводимости в активном слое структуры.

Непосредственно переключение состояния проводимости в мемристивных структурах может осуществляться как полностью из высокопроводящего состояния в низкопроводящее состояние, так и частично. Данное свойство напрямую следует из протяженности процесса переключения во времени. Существующие экспериментальные методы [116] как правило, базируются на импульсах тока (напряжения) незначительно превышающих порог переключения мемристора (0.15В) и коротких по протяженности во времени (десятки наносекунд). Данный подход позволяет получать постепенное (ступенчатое) изменение проводимости активного слоя, что в свою очередь позволяет выделить промежуточные состояния проводимости из диапазона, задаваемого высокорезистивным и низкорезистивным состояниями.

Исходя из рассмотренных механизмов переключения проводимости и механизмов транспорта заряда, для ряда из них будет справедливо влияние температурных факторов на скорость переключения, а так же влияние температуры и геометрии ячейки на проводимость мемристивных структур [117].

2.3 Verilog-A описание мемристивных элементов

Первые подходы к описанию мемристивных компонентов базировались на модели линейного ионного дрейфа представленной в работе [102] и строились с применением известной компонентной базы [118] и были ориентированы преимущественно на биполярный механизм переключения.

$$V = R(w, I)$$

$$\frac{dw}{dt} = f(w, I)$$

Параметр w – демонстрирует ширину допированного кислородом региона мемристора, имеющего прямое линейное (в рамках модели линейного ионного дрейфа) влияние на проводимость мемристора. Модель линейного дрейфа на языке Verilog-A представлена в [119]. Более поздние модели и SPICE описания учитывали нелинейность ионного дрейфа [120] и позволяли производить боле точное моделирование вольт-амперных характеристик мемристоров и предусматривали применение нормировочной «функции окна» $f_w(x)$ (функции ограничивающей выход за допустимые пределы численного значения текущего внутреннего параметра мемристора определяющего проводимость или сопротивление [window function – англ.]) впервые предложенной в [121].

$$V(t) = R(x)I(t)$$

$$R(x) = R_{off} - x\Delta R, \Delta R = R_{off} - R_{on}$$

$$\frac{dx}{dt} = kI(t)f_w(x), k = \frac{\mu_V R_{on}}{D^2}, x = \frac{w}{D} \quad (5)$$

$$f_w(x) = 1 - \big(x - stp(-I)\big), stp = \begin{cases} 1, pro\ I \geq 0 \\ 0, pro\ I < 0 \end{cases}$$

Описание модели нелинейного дрейфа на языке Verilog-AMS продемонстрировано в [122] Впоследствии свои вариации аналогичных функций были предложены в других работах [123, 124]. Подход, базирующийся на применении более ресурсоемких вычислений для повышения точности модели

без применения «функции окна памяти» представлен в статье [125] и основан на туннельном эффекте. Менее ресурсоемкое, с точки зрения авторов, описание механизмов изменения проводимости основанное на токе, превышающем пороговое значение, представлено в статье [126]. Общее описание рассмотренных механизмов и моделей на языке Verilog-A (включая листинг программ) можно обнаружить в работе [127].

Помимо рассмотренных выше моделей, использующих для описания проводимости ток через структуру мемристора, существуют модели описывающие зависимость сопротивления от превышающего некоторый порог напряжения и аналогично моделям с током через структуру использующие функцию «окна памяти» [128], так же преимуществом данной работы является возможность моделировать униполярный механизм переключения. Пример Verilog-A описания униполярного мемристора 4x4 кроссбара на основе электродов из различных допированных областей кремния и оксида кремния в качестве активного слоя отражен в работе [129], указанные мемристоры в последующем были экспериментально использованы в трехмерной матрице мемристоров. Вариации моделей с пороговым напряжением существуют и для описания модели нелинейного ионного дрейфа [130]. Концепция описания мемристивных элементов для биполярного и униполярного механизмов управляемого напряжением переключения с возможностью настройки граничных условий «функции окна» выражена в работе [131]. Модель с управляющим внутренним состоянием проводимости мемристора пороговым напряжением, обобщающая основные мемристивные компоненты и различные подходы к их описанию, представлена в публикации [130].

Помимо моделирования механизма переключения мемристивных устройств необходимость в отражении зависимостей ВАХ от количества циклов переключения, разброса технологических параметров от устройства к устройству, разброса параметров при множественном считывании (Random telegraph noise – RTN) и фактора влияния температуры нашла отражение в модели переключения

управляемой электрическими порогами [132]. Более подробно аспекты зависимости ВАХ от текущего цикла переключения, параметров ячеек (толщина активного слоя, концентрация ионов и т.д.) и RTN рассмотрены в работе [133]. Примером разработки модели, исходя из математических особенностей описания петель гистерезиса, выступает работа [134], автором берутся за основу виды ВАХ зависимостей (гистерезисные петли) и вслед за этим производится выбор способов физического моделирования целого ряда различных устройств.

Отдельным направлением исследований является применение простых нелинейных пороговых модельных представлений реализованных на различных языках описания для моделирования поведения мемристорных компонентов в схемах смешанных и аналоговых сигналов. К таким схемам можно отнести: кроссбары мемристоров и ячеек памяти (1T-1R) [135, 136], с комплиментарной парой биполярных мемристоров [136, 137], программируемый генератор на основе биполярного мемристора [129], синапс искусственного нейрона [138] на основе биполярного мемристора. Биполярный пороговый механизм с управлением переключения по напряжению и «функцией окна» на основе многочлена десятого порядка для точного моделирования комплементарной пары мемристоров на кроссбаре представлен в [139]. Пример построения полного по Посту минимального базиса логики на основе униполярного мемристора отражен в работе [140]. Работа, посвященная моделированию датчика газа на основе кроссбара, представлена в публикации [141]. В качестве основных направлений развития компактных моделей мемристора могут быть выделены следующие подходы табл. 4. При этом стоит отдельно выделить тот факт, что пороговое процесса переключения электрической проводимости приводящееся в большинстве моделей способствует более сжатому описанию и менее ресурсоемкому процессу вычислений при моделировании. С другой стороны, применение множественных граничных условий для представления множественности состояний приводит увеличению описания элемента и ограничивает гибкость модели, приводя ее к фактически табличному виду.

Таблица 4. Модели мемристоров.

Наименование модели	Управляется током(**I**)/ Управляется напряжением(**U**)	Биполярный механизм (**Bi**)/ Униполярный механизм (**Un**)	Пороговые функции описания	Применение функций окна	Возможность задания нелинейности в модели	Учет разброса параметров	Учет температуры
Линейный дрейф [102]	I	Bi					
Линейный дрейф [119]	U	Bi	+	+			
Нелинейный дрейф [120]	I	Bi		+	+		
Нелинейный дрейф [129]	U	Bi	+	+	+		
TEAM [124]	I	Bi	+	+	+		
Туннелирование сквозь барьер [125]	I	Bi			+		
Memristor Device Model [127]	U	Bi+Un	+	+	+		
SPICE Verilog-A model [128]	U	Un	+		+		
VTEAM [131]	U	+ Bi	+	+	+		
SPICE model [130]	U	Bi+Un	+	+	+	+	+

Описание мемристора. Поскольку в качестве блока учета вклада сигнала предполагается использование биполярного мемристора, далее приводится описание мемристивного элемента с биполярным механизмом переключения. Из рассмотренного выше следует, что все подходы к Verilog-A описанию можно разделить на модельные представления с зависимостью проводимости от тока,

превышающего порог и модели с зависимостью проводимости от приложенного напряжения.

В данной работе будет использоваться механизм переключения проводимости с зависимостью от превышения порога по напряжению. Причинами указанного

выбора являются следующие аспекты физических процессов происходящих в активном слое биполярных мемристоров:

1. Приложение напряжения, превышающего порог Von или Voff, приводит к переключению в низкопроводящее и высоко проводящее состояние соответственно. Исходя из модели миграции кислородных вакансий, пороговые значения напряжения являются необходимой минимальной величиной напряженности электрического поля для начала миграции вакансий и начала процессов переключения.
2. Подача коротких во времени импульсов, порядка 10нс, и незначительно превышающих напряжение порога, не более 0.15В, не прекращает процесс переключения, но придает ему ступенчатый характер, из чего следует его зависимость во времени.
3. Подача импульсов напряжения превышающего пороговое значение сопротивления амплитудой в диапазоне от 0.15-0.5В приводит к более быстрому переключению мемристора, из чего следует зависимость от напряженности электрического поля.
4. Процесс переключения, стимулированный превышением порога электрического тока через активный слой структуры, не согласуется с экспериментальными данными.

Предлагаемая автором модель описания мемристора средствами САПР Cadence на языке высокого уровня Verilog-A основывается на модели линейного дрейфа кислородных вакансий с механизмом переключения по напряжению [129]. Отличие заключается в том, что изменение состояния проводимости мемристора зависит от приложенного напряжения, а не протекающего через структуру заряда, и учитывает параметры разброса при переключении между состояниями проводимости. В общем виде зависимости имеют вид:

$$I(t) = V(t)/R(V_c, x, t)$$

$$R(V_c, x, t) = R_s - x(V_c, t)$$

$$\frac{dx}{dt} = a * f_w(x) * \begin{cases} \int_0^{t_S} (V_c - V_{thS})\, dt, V_c > V_{thS} \\ \int_0^{t_{RS}} (V_c - V_{thRS})\, dt, V_c < V_{thRS} \\ 0, V_{thRS} < V_c < V_{thS} \end{cases}$$

$$f_w(x) = \begin{cases} 1, R_{on} + \Delta R_{on} < R(x,t) < R_{off} + \Delta R_{off} \\ 0, else \end{cases}$$

где, R_s – начальное состояние проводимости мемристора после электроформовки; a – подгоночный коэффициент; $V(t)$ – приложенное между контактными площадками напряжение, x – изменение сопротивления мемристора; $R(x,t)$ – текущее состояние сопротивления мемристора; $I(t)$ – выходной ток мемристора; V_c – текущее напряжение на контактных площадках мемристора; t_S – время превышения импульсом порога записи; t_{RS} – время превышения импульсом порога стирания; V_{thS} – текущее значение порога переключения в высокопроводящее состояние, задаваемое случайно по усеченному симметричному распределению Гаусса в границах $\pm\Delta V_{thS}$; V_{thRS} – текущее значение порога переключения в низкопроводящее состояние, задаваемое случайно, по усеченному симметричному распределению Гаусса в границах $\pm\Delta V_{thRS}$; R_{on} – текущее значение высокопроводящего состояния, задаваемое случайно по усеченному симметричному распределению Гаусса в границах $\pm\Delta R_{on}$; R_{off} – текущее значение низкопроводящего состояния, задаваемое случайно по усеченному симметричному распределению Гаусса в границах $\pm\Delta R_{off}$.

Модель позволяет учитывать количество циклов переключения мемристора, задаваемых напрямую параметром модели, аналогично подходу, применяемому в работе [130]. Отличие модели, применяемой в указанной работе, заключается в использовании не количества циклов напрямую, а применении переменного значения ресурса переключения мемристора определяемого по формуле:

$$Cyc = NumCyc * \left(R_{off} + \Delta R_{off} - R_{on} - \Delta R_{on}\right) * 2$$

где, Cyc – значение ресурса на текущей итерации; $NumCyc$ – количество циклов переключения. Изменение значений ресурса переключений описывается выражением:

$$Cyc(t+1) = Cyc(t) - x(V_c, t)$$

Указанные особенности описания мемристора позволяют добиться дисперсии выходного сигнала при изменении состояния проводимости мемристора в зависимости от порогов переключений в высокорезистивное и низкорезистивное состояния, а так же в учета разброса параметров низкорезистивного и высокорезистивного состояний при переключении между различными состояниями проводимости. Введение переменной ресурса переключения позволяет моделировать случаи отказа мемристоров в высокопроводящем, низкопроводяшщем и промежуточном состоянии проводимости, а также учитывать различное время отказа компонентов в зависимости от разброса параметров.

Модель переключения, учитывающая только временную зависимость, может быть описана следующими параметрами:

- Von – порог начала переключения мемристора в высокопроводящее состояние
- Voff – порог начала переключения мемристора в низкопроводящее состояние
- Ron – сопротивление мемристора в высокопроводящем состоянии
- Roff – сопротивление мемристора в низкокопроводящем состоянии
- dt – количество шагов до переключения мемристора в единицах времени.

В качестве переменных описываемого модуля применяются: t – верхний электрод; b – нижний электрод; x – внутреннее состояние мемристора, сопротивление в текущий момент времени; xdt – минимальное приращение за шаг моделировнаия.

В рамках модели на начальном шаге потребуется ввести начальное состояние проводимости и рассчитать элементарный шаг итерации, что в конструкциях языка Verilog-A может быть выражено следующим образом:

```
@ (initial_step) begin //инициация задания начальных условий моделирования
   x=Roff; //Задание начального состояния
   xdt=(Roff-Ron)/dt; //Расчет минимального изменения проводимости
end  //окончание задания начальных условий моделирования
```

Имплементация функции переключения может одновременно включать граничные условия корректной работы механизма переключения для ограничения выхода значений проводимости за границы допустимого диапазона. Вызов функции реализуется одновременно с передачей в нее параметров структуры и значений переменных и присвоения результата функции переменной внутреннего состояния:

```
x=dFx(Von, Voff, Ron, Roff, xdt, V(t,b), x); //Вычисление состояния
```

сопротивления.

Полное описание процесса вычислений и функции переключения имеет вид:

```
analog function real dFx; //Функция переключения
   input Vn, Vff, Rn, Rff, wdt, cv, w; //Входные данные функции
   real Vn, Vff, Rn, Rff, wdt, cv, w; //Локальные переменные функции
      if ( (cv > Vn) && (w > Rn) ) begin //Граничные условия операции SET
         dFx = w - wdt;
      end
    else if ( (cv < Vff) && (w < Rff) ) begin //Граничные условия операции RESET
         dFx = w + wdt;
      end
      else begin
         dFx = w; //Режим работы функции без переключений
      end
endfunction
analog begin //start of description
   @ (initial_step) begin
      x=Roff;
      xdt=(Roff-Ron)/dt;
```

end

x=dFx(Von, Voff, Ron, Roff, xdt, V(t,b), x);

I(t,b) <+ V(t,b) / x; //Вычисление выходного тока мемристора

end

Среда САПР Cadence позволяет для указанной модели регулировать шаг вычислений через задание времени шага. Задание времени моделирования допускает использовать масштабы от 1 секунды до 1 фемтосекунды. Моделирование ВАХ в режиме постоянного тока позволяет применять опцию петли гистерезиса для структуры. В результате моделирования на каждом шаге итерации выполняется вычисление по изменяемому от шага к шагу напряжению, а весь цикл моделирования состоит из изменения от минимального значения напряжения до максимального и обратно.

Указанный метод с абсолютным значением величины изменения напряжения в 0.01В позволяет получить ВАХ для модели, представленные на рис. 30. На рисунке мемристор изначально находится в высокорезистивном состоянии, порог напряжения для начала переключения в высокопроводящее состояние равен порогу переключения в низко проводящее состояние и составляет 0.5В. Низкопроводящее состояние мемристора имеет величину в 200 Ом, высокопроводящее состояние равно сопротивлению в 2200 Ом. Мемристор напрямую подключен к источнику напряжения, схема моделирования представлена на рис. 31.

При необходимости все параметры модели могут быть перенастроены с учетом конкретных физических реализаций. Например, могут быть заданы различные значения порогов переключения в высокорезистивное и низкорезистивное состояние, так же изменению могут быть подвержены значения параметров высокорезистивного и низкорезистивного состояний. Изменение крутизны ВАХ петли гистерезиса мемристора может быть достигнуто путем изменения параметров модели отвечающих за скорость изменения сопротивления мемристора.

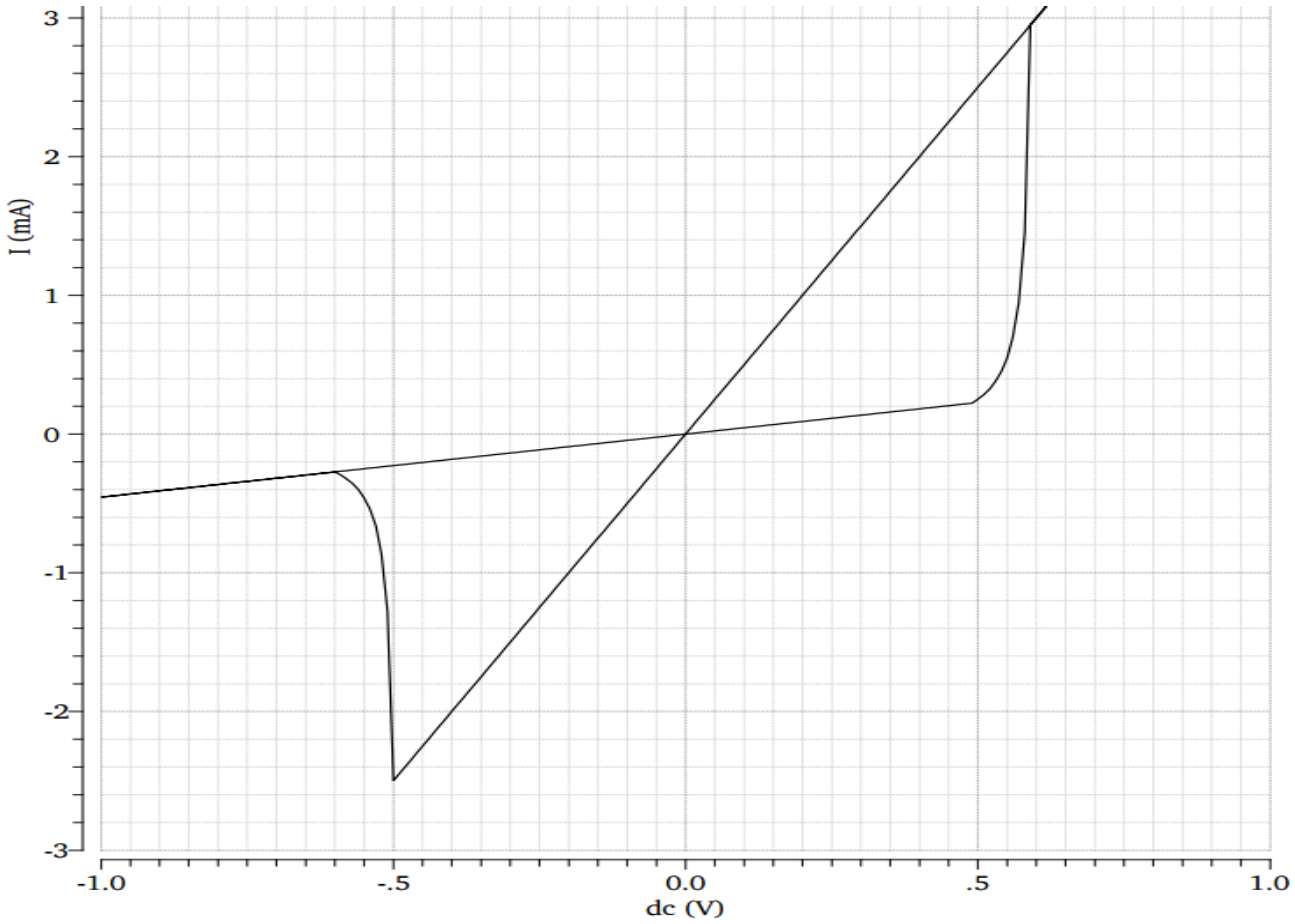


Рисунок 30. ВАХ мемристора базовой модели.

В случае изначального задания низкорезистивного состояния мемристора, на ВАХ структуры при проведении моделирования образуется дополнительный отрезок кривой, демонстрирующий переход от низкорезистивного состояния в высокорезистивное рис. 1 ПРИЛОЖЕНИЕ 2. Корректность модели подтверждается графиком изменения ВАХ при анализе постоянного тока от -1 В до 1 В рис. 2, 3 ПРИЛОЖЕНИЕ 2. Последовательное переключение мемристра из состояний с низкой проводимостью в состояния с высокой проводимостью мемристора показано на рис. 32. Изменение параметра dt=1 переводит схему в режим мгновенного переключения между состояниями. Работа в режиме мгновенного переключения показана на рис. 33, так же на ВАХ продемонстрирована корректная работа модели в диапазонах напряжений, не приводящих к изменению проводимости, соответствует уровням сигнала 0.4 В и позволяет говорить о корректной работе граничных условий переключения. Циклирование мемристора, в режиме мгновенного переключения,

прямоугольными импульсами в масштабе времени наносекунд показано на рис. 4 ПРИЛОЖЕНИЕ 2. Из диаграммы работы по шкале времени видно, что переключение проводимости происходит мгновенно.

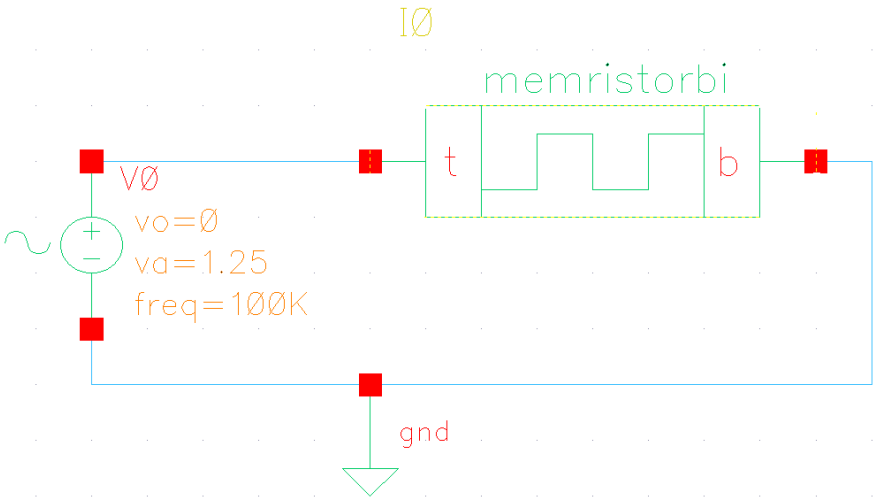


Рисунок 31. Схема моделирования мемристора. V0 – источник напряжения, I0 – биполярный мемристор, gnd – шина земли.

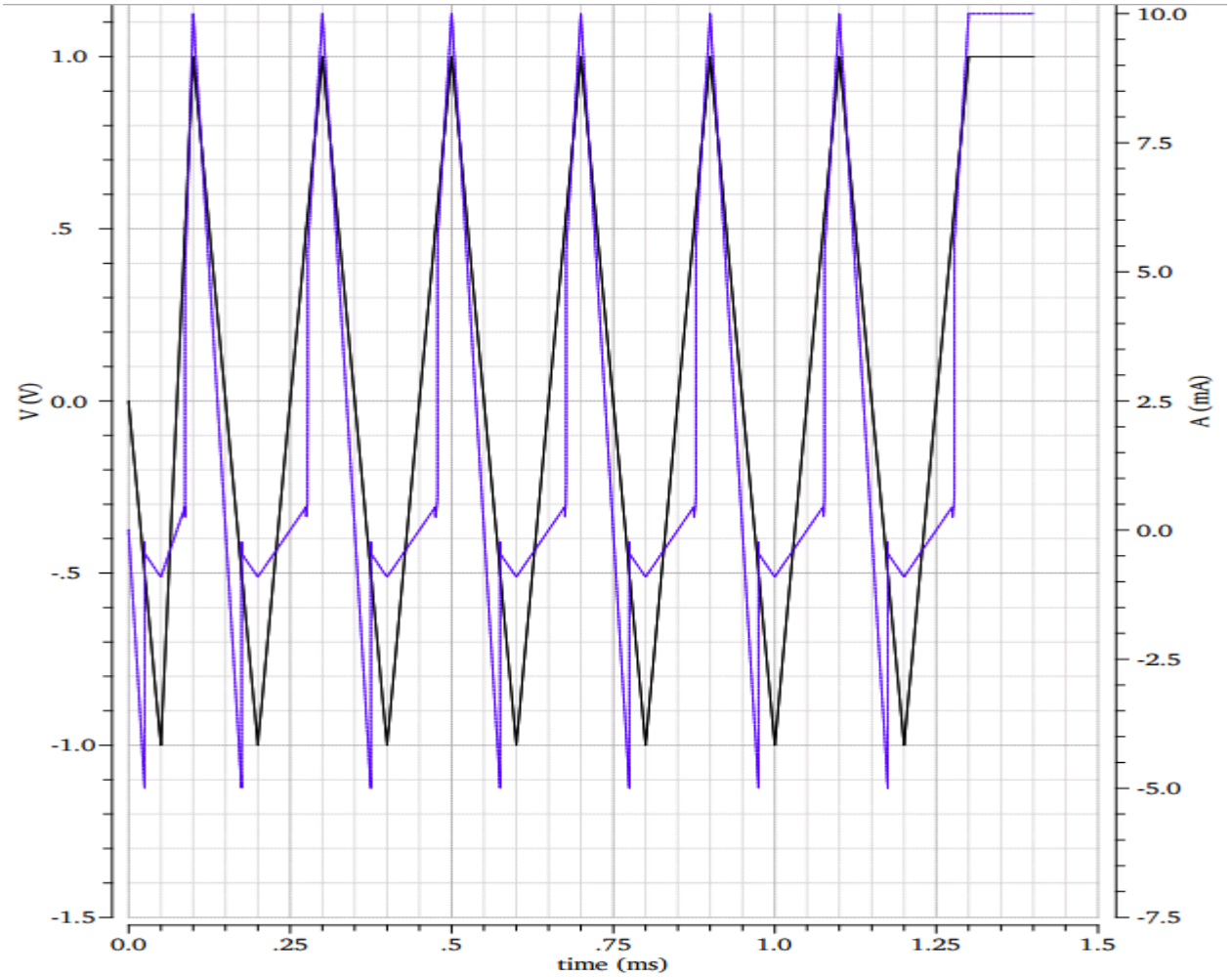


Рисунок 32. Циклирование мемристора. Уровень входного сигнала в вольтах представлен черной линией, выходной ток синей пунктирной линией.

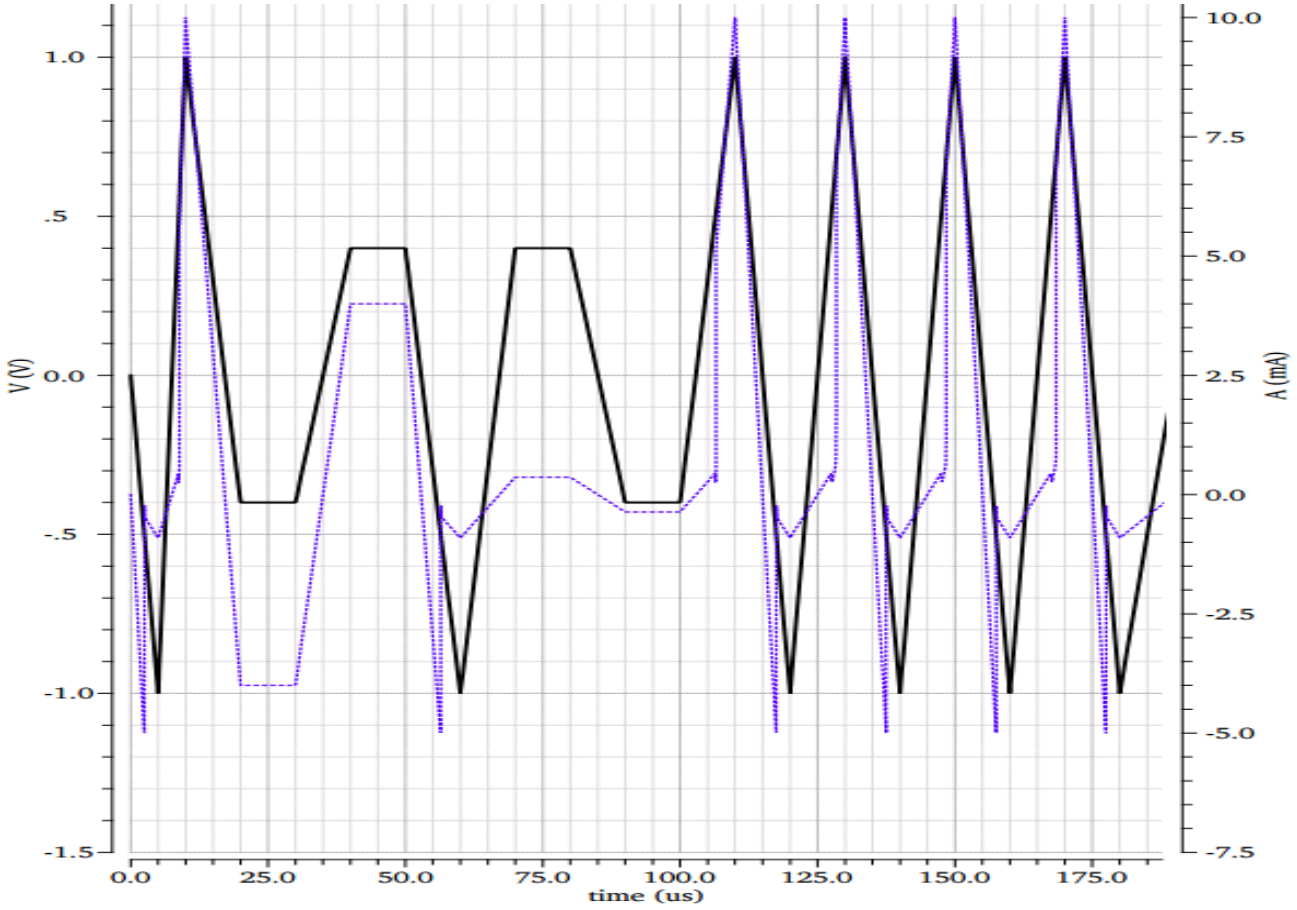


Рисунок 33. Циклирование мемристора, режим мгновенного переключения. Уровень входного сигнала в вольтах представлен черной линией, выходной ток синей пунктирной линией.

Для учета возможности отказа ячеек требуется введение дополнительной переменной учета ресурса переключений в модель и параметра модели отвечающего за общее количество циклов переключения. Численное значение переменной ресурса может быть вычислено путем удвоения разности высокопроводящего и низкопроводящего состояний умноженного на количество циклов, что очевидно.

*Cyc = (Ron-Roff) 8 NumCyc * 2;* //Вычисление ресурса переключений

Cyc – переменная ресурса, *NumCyc* – |параметр циклов переключения. Указанная операция используется на начальном этапе моделирования, а последующий учет изменения ресурса выполняется при каждой итерации функции переключения.

Cyc = Cyc - xdt; //Перерасчет ресурса переключений

Применение операций учета ресурса переключений требует наложения дополнительных условий на функцию переключения. Данные условия должны запрещать изменение состояния проводимости мемристора в случае $Cyc \leq 0$. Моделирование позволяет получить два состояния окончания ресурса переключения. Первое состояние соответствует мемристору не переключающейся из состояния высокой проводимости в результате исчерпания ресурса переключений рис. 34. Второе состояние соответствует мемристору не переключающемуся из состояния низкой проводимости рис. 35. Оба состояния неоднократно получались различными группами исследователей и имеют экспериментальные подтверждения.

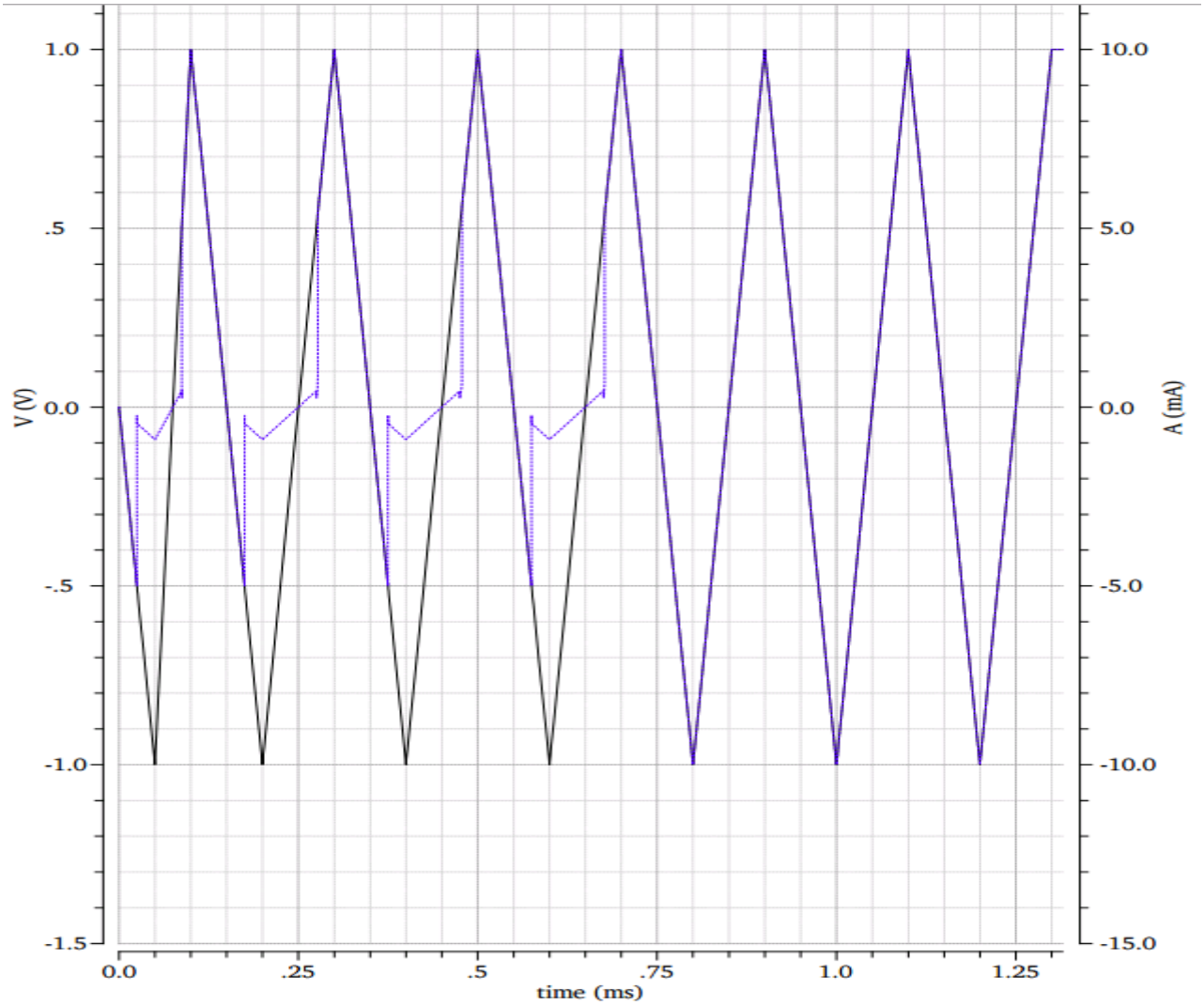


Рисунок 34. Циклирование мемристора. Исчерпание ресурса переключений в высоком состоянии проводимости. Уровень входного сигнала в вольтах представлен черной линией, выходной ток синей пунктирной линией.

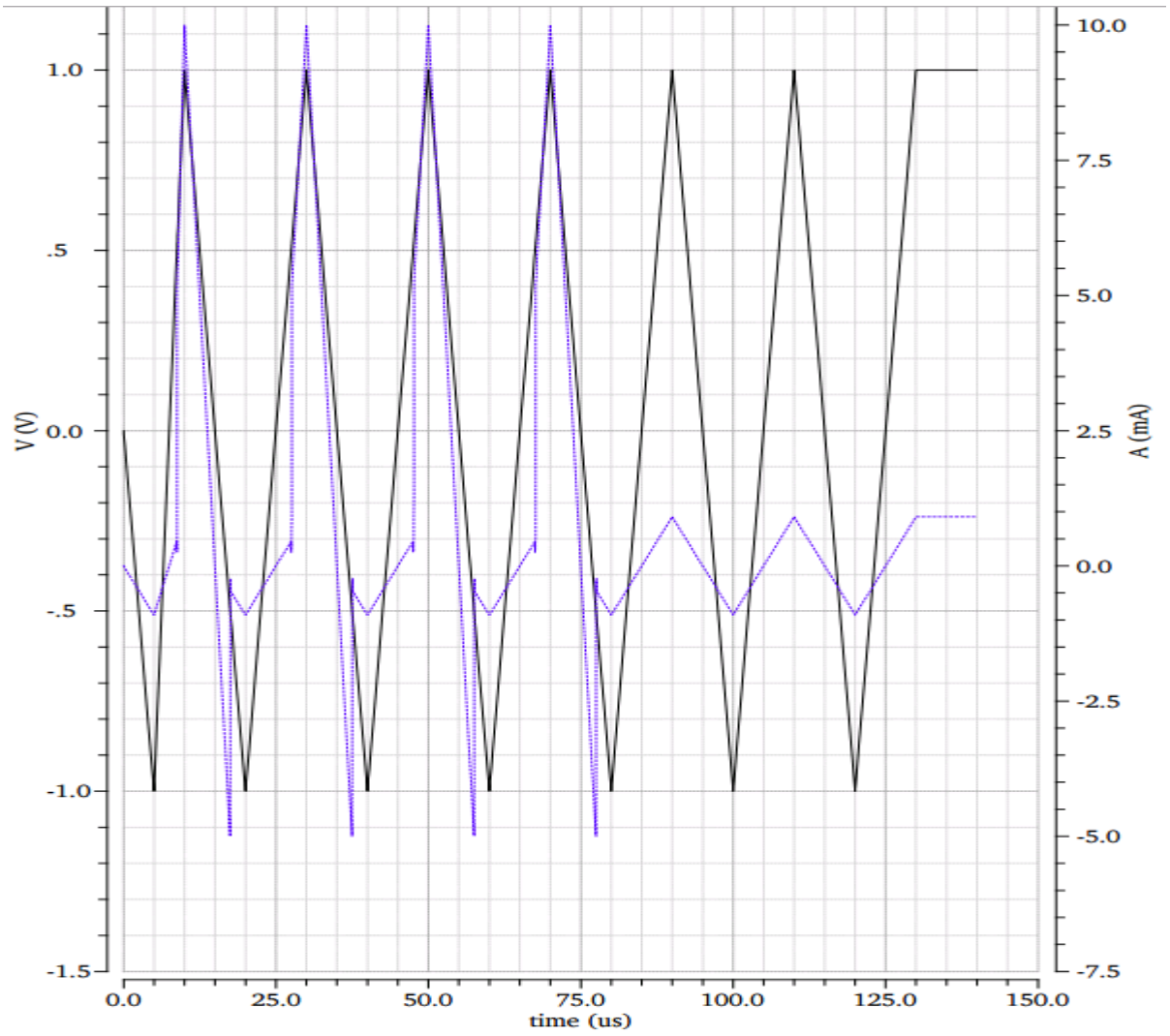


Рисунок 35. Циклирование мемристора. Исчерпание ресурса переключений в низком состоянии проводимости. Уровень входного сигнала в вольтах представлен черной линией, выходной ток синей пунктирной линией.

Для учета амплитуды сигнала необходима модификация функции переключения и введение дополнительных параметров. Введение следующих параметров: dtsc единичный вектор времени расчет; dtsn общее время операции SET в единицах dtsc; dtsff общее время операции RESET в единицах dtsc; Vgrw единичный вектор роста филамента; Vmlt единичный вектор растворения филамента – позволяет реализовать учет амплитуды и времени превышения сигналом абсолютного значения порога напряжения в виде интеграла от приложенного напряжения. Также требуется введение дополнительных переменных. Алгоритм вычисления представляет поточечный учет площади превышающей порог амплитуды, что в свою очередь позволяет реализовать множественность уровней состояний проводимости рис. 36.

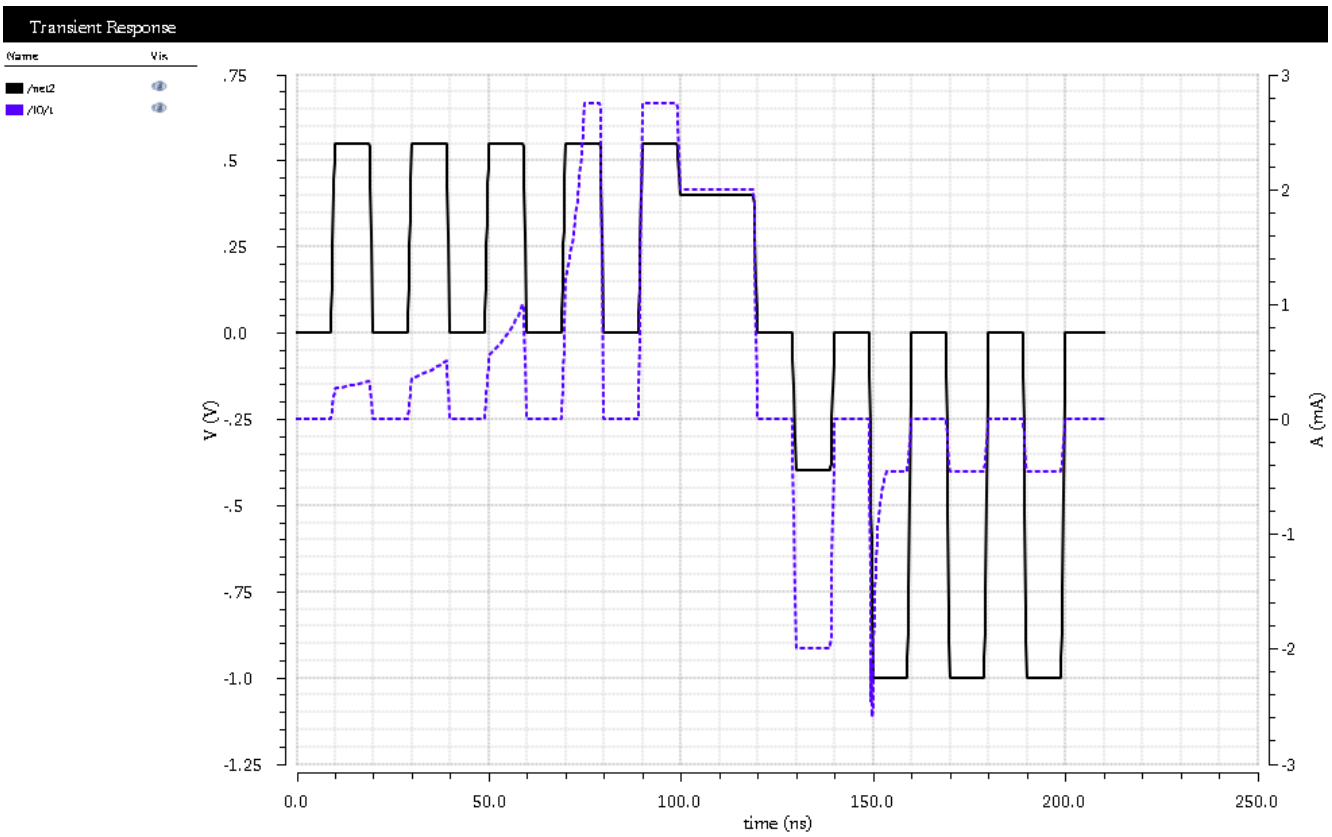


Рисунок 36. Циклирование мемристора. Множественность состояний получаемых путем задания коротких небольших по амплитуде импульсов. Уровень входного сигнала в вольтах представлен черной линией, выходной ток синей пунктирной линией.

Графики с результатами моделирования множественности с отказом при переключении и частичном переключении состояний мемристора получаемых серией коротких низких по амплитуде импульсов приведены в ПРИЛОЖЕНИИ 2 рис. 5 и 6 соответственно.

Разброс параметров от цикла переключения к циклу для мемристора может быть задан по закону нормального распределения. При этом следует учесть возможность разброса таких параметров как Von, Voff, Ron, и Roff. Для задания нормального распределения так же требуется введение дополнительной целочисленной переменной на основе, которой будет выбрана псевдослучайная последовательность и параметров среднеквадратичного отклонения. После чего указанные параметры будут использованы в библиотечной функции языка Verilog-A.

```
RRon = $rdist_normal(seed, Ron, DltRon );
RRoff = $rdist_normal(seed, Roff, DltRoff );
```

RVon = $rdist_normal(seed, Von, DltVon);

RVoff = $rdist_normal(seed, Voff, DltVoff);

Здесь переменные Dlt* представляют среднеквадратичное отклонение, а переменные, начинающиеся с R* текущее значение параметра для итерации. График разброса параметров представлен на рис. 37.

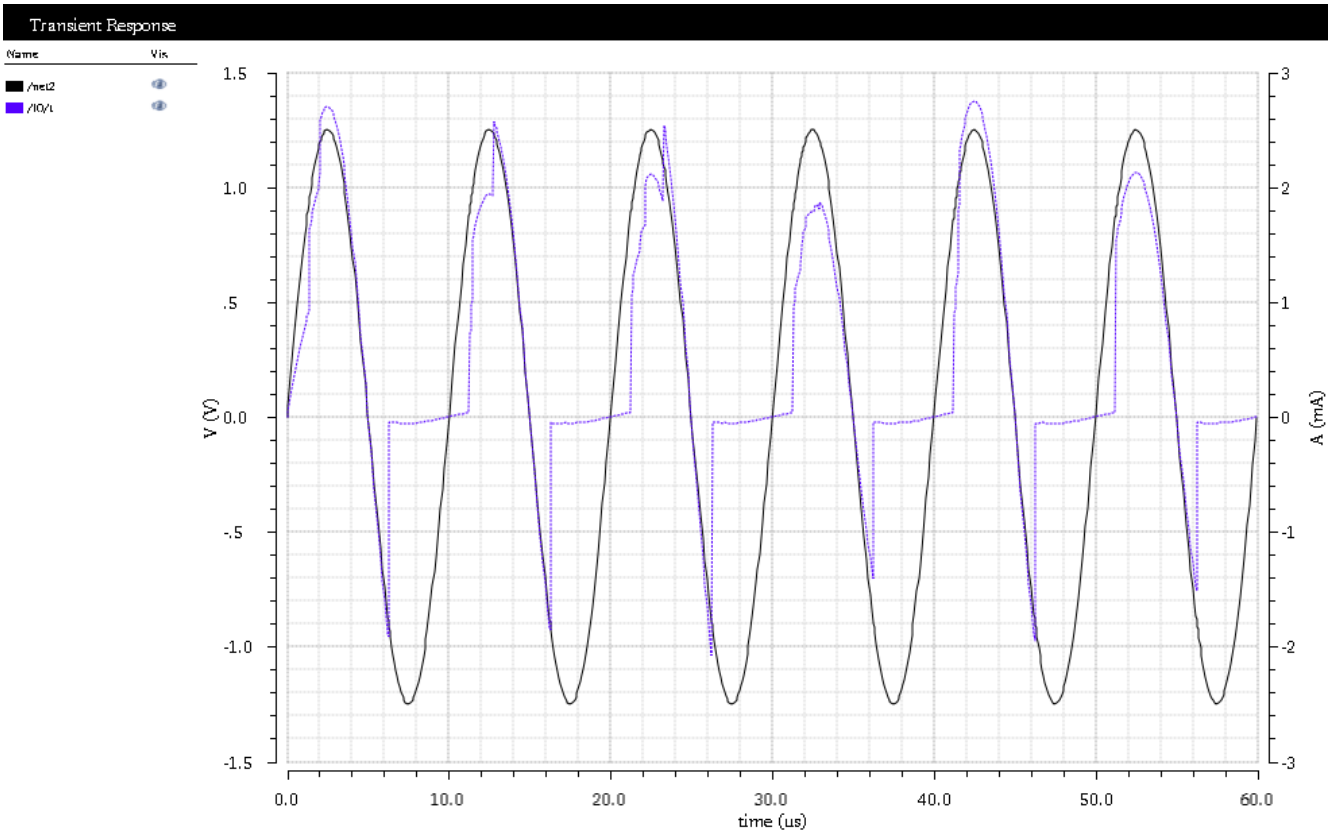


Рисунок 37. Циклирование мемристора. Множественность состояний и разброс параметров состояний высокой проводимости. Уровень входного сигнала в вольтах представлен черной линией, выходной ток синей пунктирной линией.

Из рисунка видно поэтапное переключение мемристора в зависимости от приложенного напряжения и времени, а так же разброс характеристик высокого состояния проводимости от цикла к циклу. График, отображающий разброс состояний низкой проводимости представлен на рис. 38. На рисунке представлен итеративный процесс переключения мемристора до исчерпания ресурса переключения и разброс от цикла к циклу состояний низкой проводимости мемристора. Как видно из сравнения графиков, разброс параметров в высокопроводящем состоянии будет вносить существенные искажения в обрабатываемый сигнал в сравнении с влиянием разброса в низкопроводящем состоянии. Дополнительные результаты моделирования циклов переключения и

разброса параметров представлены в ПРИЛОЖЕНИИ 2 рис. 7-9. Из графиков следует выполнение механизма ступенчатого изменения проводимости мемристоров в зависимости от времени и амплитуды с учетом разброса.

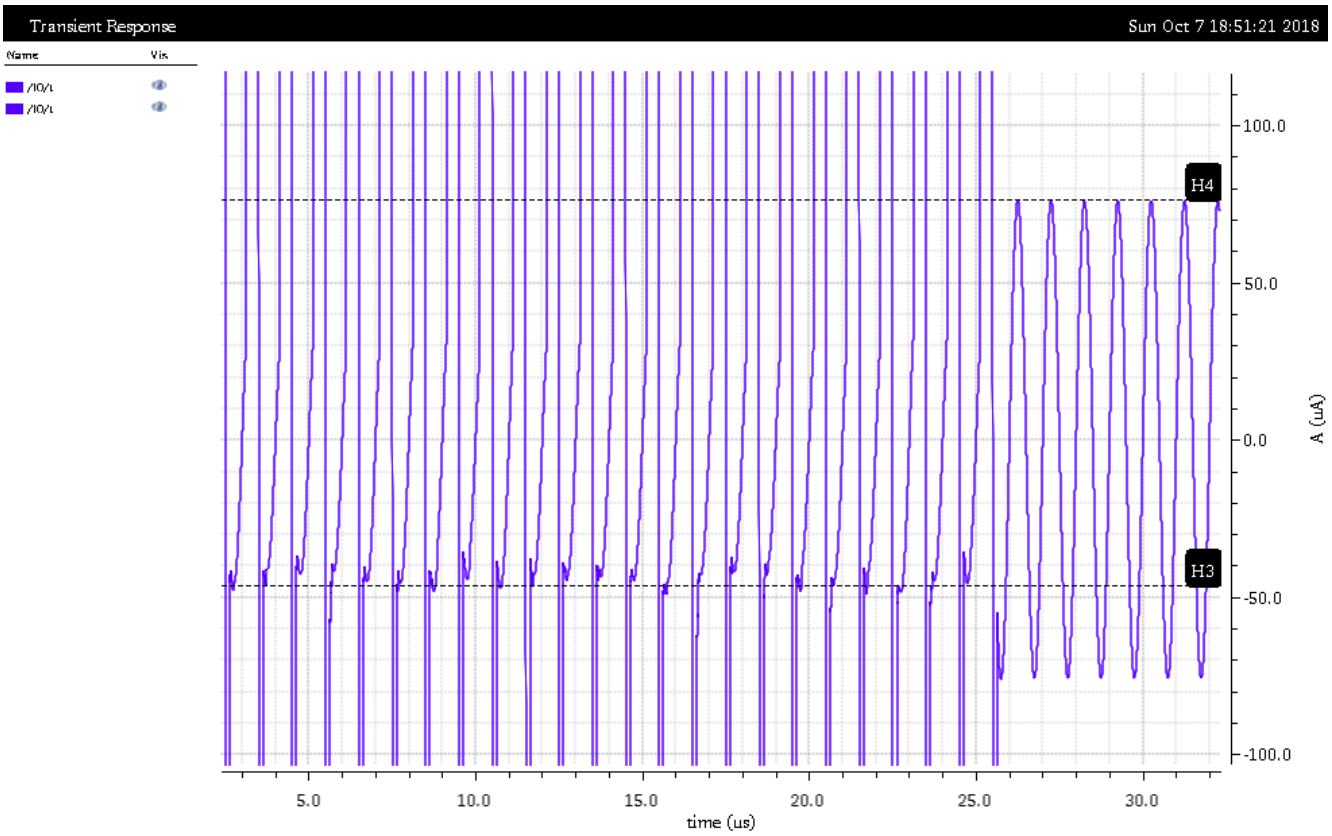


Рисунок 38. Циклирование мемристора. Множественность состояний и разброс параметров состояний низкой проводимости. Выходной ток обозначен синей пунктирной линией.

На рис. 39 представлены параметры переключения модели мемристора в низкорезистивное и высокорезистивное состояние соответственно. Как следует из временных диаграмм переключения, время переключения составляет 60 нс, что согласуется с радом экспериментальных данных представленных в литературе для механизма переключения на основе дрейфа кислородных вакансий.

Представленные диаграммы временных параметров переключения мемристора завершают рассмотрение результатов моделирования Verilog-A описания компонента. Следствием из результатов моделирования является приоритет задания множественности состояний путем подачи коротких импульсов с малой амплитудой по напряжению перед длительными уровнями с малой амплитудой, либо короткими уровнями с большой амплитудой

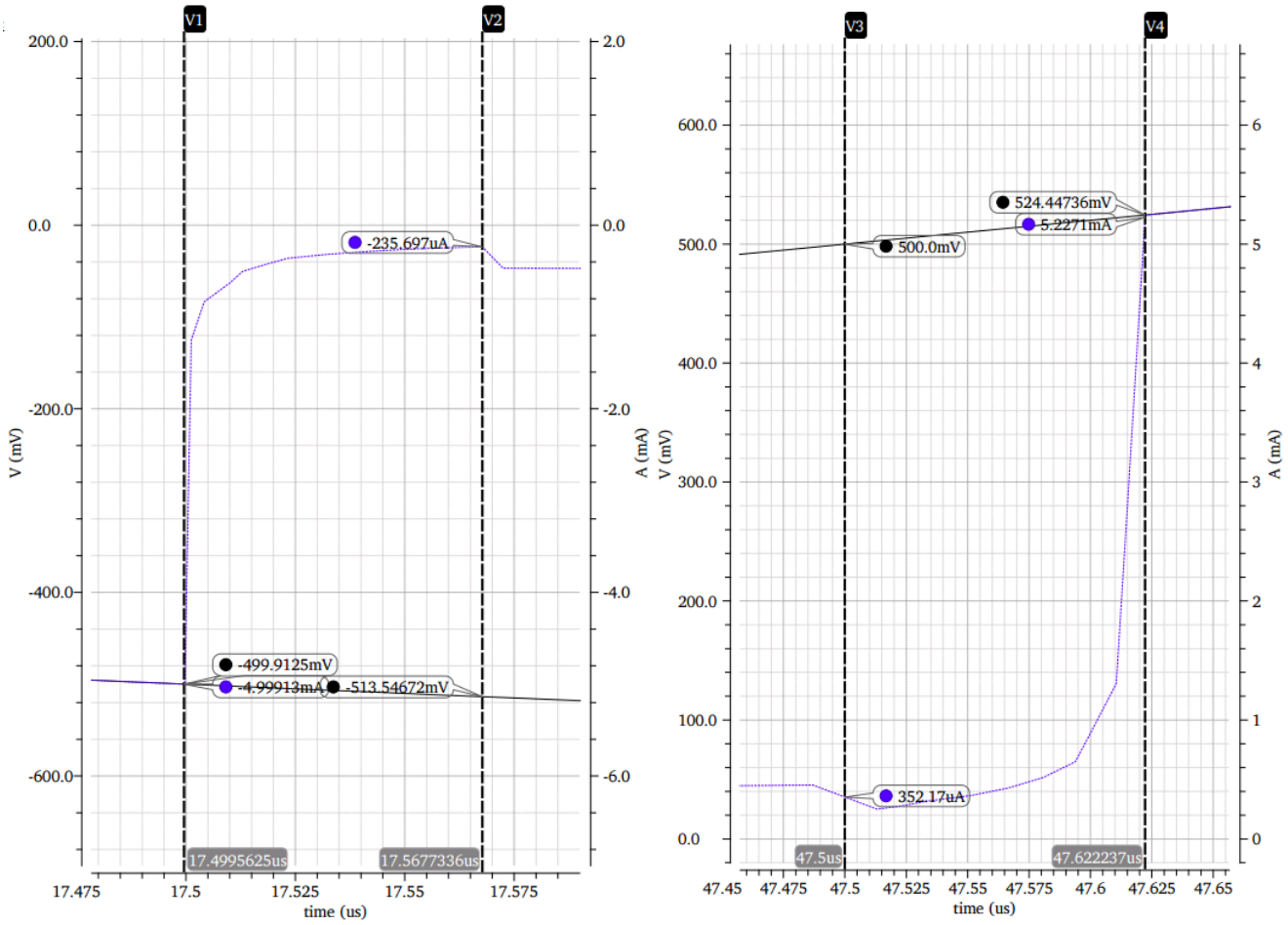


Рисунок 39. Слева график переключения мемристора в высокорезистивное состояние. Справа график переключения мемристора в низкорезистивное состояние. Выходной ток представлен синей пунктирной линией, входное напряжение черная сплошная линия.

Исходя из всего изложенного выше, можно заключить следующее, представленное описание биполярного мемристора позволяет производить подстройку высокоуровневой модели мемристора с учетом подстройки порогов переключения, времени переключения, амплитуды сигнала переключения, учесть фактор влияния температуры на проводимость структуры, учесть множественность состояний проводимости и разброс таких параметров как высокорезистивное и низкорезистивное состояния, а так же разброс параметров порогов переключения в зависимости от цикла к циклу.

Полное описание модели на языке Verilog-A представлено в ПРИЛОЖЕНИИ № 1.

2.4 Выводы по главе 2

Последние достижения в области компонентной базы, среди которых можно выделить такие как PCM RAM, SOT-элементы, FeRAM и мемристивные элементы, существенно расширили возможности для реализации нейроморфных систем аппаратными средствами. Текущие исследования в области архитектур привели к использованию концепции реализации нейроядер как базовых блоков имплементации нейрочипов.

Из всего массива доступных исследований проведенных разными независимыми группами исследователей теоретически обосновывающих применение мемристивных элементов в качестве блоков искусственного нейрона и экспериментальных данных предоставляющих бесспорные доказательства преимущества применения мемристоров в качестве элементов синапсов можно выделить два основных преимущества их применения:

1. Мемристивные структуры поддаются масштабированию вплоть до размеров реального биологического синапса, что позволяет в свою очередь говорить об экономии площади необходимой для их размещения в сравнении с реализацией средствами цифровой схемотехники.
2. Ключевым преимуществом мемристоров выступает множественность состояний проводимости, что при задании весового коэффициента позволит сэкономить площадь на кристалле в сравнении с тем же количеством состояний PCM, FeRAM, или SOT структур.

Предложенное описание мемристивного элемента средствами языка Verilog-A позволяет учесть следующие технические и физические особенности имплементации мемристоров:

1. Множественность состояний проводимости;
2. Разброс параметров в зависимости от цикла переключения;
3. Ограниченность количества циклов переключения;
4. Время переключения между состояниями проводимости;

5. Влияние амплитуды импульса на скорость переключения мемристора;

Представленное модельное описание позволяет использовать мемристор как библиотечный элемент при проектировании микросхем с нейроморфной структурой и может быть модифицировано в зависимости от конкретных параметров создаваемых структур, а так же, с учетом при необходимости, конкретных механизмов переключения и проводимости.

Прямым следствием из результатов моделирования является предпочтительность снижения разброса параметров высокопроводящего состояния и параметров разброса пороговых напряжений для изготавливаемых МДМ структур, так как данные разбросы будут напрямую влиять на корректную обработку информационных сигналов и проведение операций переключения. Немаловажным следствием из результатов моделирования является приоритет задания множественности состояний путем подачи коротких импульсов с малой амплитудой по напряжению перед длительными уровнями с малой амплитудой, либо короткими уровнями с большой амплитудой.

3 ГЛАВА. Техническая реализация модели КААН на базе мемристоров.

Предметом рассмотрения данной главы выступает конкретизация обобщённой модели КААН. Обобщенная схема конкретизируется до структурной схемы аппаратной реализации искусственного нейрона. Структурная схема включает определение основных блоков КААН. Далее производится уточнение структурной схемы с учетом применения при имплементации мемристивных компонентов. В заключение главы предоставляются выводы по результатам указанных уточнений.

В первой части главы производится построение структурной схемы. В процессе конкретизации обобщенная схема КААН конкретизируется до блоков искусственного нейрона. В процессе определения блоков формируются требования к их реализации с указанием ключевых целевых параметров. При построении блоков учитывается алгоритм работы разделе 1.3.

Вторая часть главы уточняет структуру некоторых блоков и их состав, а так же определяет отдельные параметры указанных блоков. В данном разделе учитывается специфика применения мемристивных компонентов. В процессе уточнения состава реализуемых блоков указываются блоки использующие мемристоры.

Глава завершается описанием полученных результатов в виде структурной схемы с уточнёнными блоками.

3.1 Особенности технического решения модели КААН с динамической функцией активации

Существующие технические решения имплементации функции активации аппаратными средствами искусственного нейрона, при необходимости задания функции активации произвольного типа и ускорения процесса вычислений используют LUT [142]. Указанный метод применяется не только для срабатывающих по совпадению нейронов, но и для интегрирующих, интегрирующих с утечками нейронов и связывающих нейронов [100].

Существующие аппаратные реализации имеют следующую схему работы. Взвешенные входные сигналы подаются на блок агрегации, где суммируются либо перемножаются и в качестве выходного сигнала блока агрегации выдается результат n-мерной операции сложения или умножения. Для срабатывающих по совпадению нейронов блок LUT фактически является блоком имплементирующим функцию активации. Выходной сигнал от блока агрегации поступает на LUT, где сравнивается с диапазонами значений. Каждому диапазону значений соответствует свой выходной сигнал функции активации. По результатам сравнения определяется диапазон текущего сигнала и соответствующее ему значение функции активации подается на выход нейрона.

Нейроны связывающего или интегрирующего типа имеют дополнительный блок сумматора с накоплением расположенный между блоком агрегации и блоком LUT. Отличие в работе алгоритма обработки сигналов заключается в учете временной составляющей приходящих на нейрон сигналов. Нейрон имеет три режима работы: режим активации, режим срабатывания и рефрактерный режим. Режим активации реализует учет сигналов от блока агрегации в течение заданного промежутка времени на сумматоре с накоплением. По окончании указанного режима результат последовательного сложения сигналов от блока агрегации передается на LUT. Режим срабатывания реализует определение текущего диапазона результата и выдачу сигнала активационной функции на выход нейрона с последующим его удержанием в течение периода срабатывания. Режим

рефрактерного состояния нейрона сопровождается отсутствием приема и обработки сигналов на входах нейрона и последующих блоках обработки информации в течение рефрактерного периода.

Использование в целях учета текущего состояния активации в интегрирующих и связывающем нейронах блока суммирования с накоплением позволяет учитывать только обобщенный уровень активности. Последовательность агрегированных сигналов обобщается до уровня общей активности за период активации без учета последовательности и амплитуды, агрегированных в момент времени входных сигналов. Предложенная модель КААН с динамической функцией активации не имеет указанного недостатка за счет применения двух блоков LUT и регистра M_k. Уточненная структурная схема аппаратной реализации представлена на рис. 40.

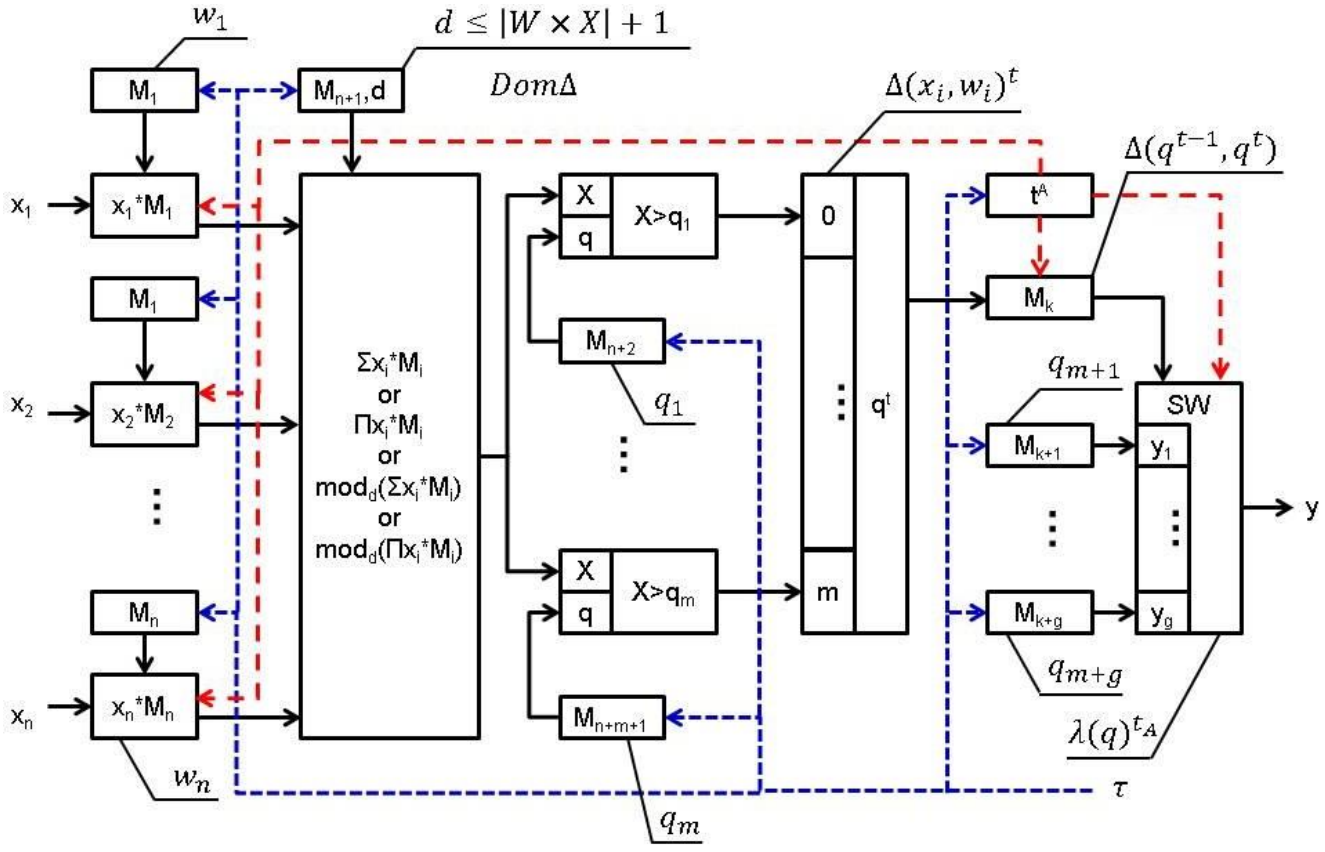


Рисунок 40. Уточненная структурная схема КААН с динамической функцией активации. M – регистры хранения данных, x*M – блок умножения на весовой коэффициент, Σ or Π or mod_d(Σ or Π) – блок агрегации входных сигналов, X>q – блоки сравнения с диапазоном значений LUT1, Δ(x,w) – блок генерации выходного сигнала LUT1, $\Delta(q^{t-1},q^t)$ – сдвиговый регистр, SW – блок генерации выходного сигнала LUT2, t^A – блок таймера

Уточнённая структурная схема включает обобщенное представление блока агрегации входного сигнала. С учетом возможных подходов к реализации блока агрегации, на структурной схеме блок агрегации представлен следующими вариантами построения блоков: сумматор взвешенных входных сигналов, блок перемножения взвешенных входных сигналов, программируемый блок взятия модуля от суммы или произведения взвешенных входных сигналов. На схеме отмечен предел принимаемых блоком выполнения операции взятия модуля значений определяемый как произведение множеств, инкрементированное на единицу. Пример аппаратного решения рассмотрен в работе [143]. Применение программируемого блока взятия модуля от суммы или произведения может быть использовано для периодических функций активации. Операция взятия модуля по программируемому значению фактически представляет собой определение периода функции активации соответствующего агрегированным сигналам входов. Дальнейшее сравнение с диапазонами значений и формирование сигнала на выходе LUT1 обеспечивает учет значения агрегированных сигналов в рамках периода.

Отдельно стоит отметить, что для задания почти периодических функций количество диапазонов LUT1 фактически должно соответствовать количеству всех возможных значений операции агрегирования входных сигналов. Данное условие в явном виде требует существенных аппаратных затрат на реализацию и влечет существенное увеличение энергопотребления, что приводит к неэффективности подходов при его имплементации в искусственном нейроне.

Управляющие сигналы представлены на структурной схеме КААН пунктирной линией. Сигналы от блока таймера представлены красной пунктирной линией. Информационные сигналы представлены сплошной черной линией. Направление распространения сигнала обозначено стрелками. По линиям передачи управляющих сигналов производится установка весовых коэффициентов, параметра операции взятия по модулю, в случае реализации данного блока в составе блока функции агрегации, диапазонов границ между элементами множества Q модели КААН, соответствующих текущим значениям

уровня активации на входах нейрона, значений функции активации LUT 2, и настройка параметров таймера отвечающих за сдвиговый регистр и время функционирования в каждом из режимов работы нейрона. Сигналы таймера реализуют контроль: блоков умножения на весовой коэффициент, не позволяя осуществлять операции в рефрактерном режиме; блока переключения выходного сигнала, контролируя время удержания выходного сигнала; сдвигового регистра, обеспечивая сброс по окончании режима активации, количество используемых на текущий момент ячеек регистра, за счет чего осуществляется перезадание функций активации в процессе вычислений.

Как отмечалось выше, применение двух блоков LUT позволяет учитывать амплитуду текущих уровней активации. Указанная опция реализуется за счет применения сдвигового регистра. Выходной сигнал LUT1 фактически является частью адреса подаваемого в сдвиговый регистр по сигналу таймера. Работа нейрона в режиме активации осуществляет постепенное формирование адреса ячейки памяти в LUT2, что позволяет учесть не только общий уровень активности на входах нейрона, но и амплитуду уровней агрегированных сигналов в каждый момент времени.

3.2 Описание модели КААН с применением мемристивных компонентов

Из обзора литературы известно, что применение мемристоров в качестве элементов синапса искусственного нейрона позволяет уменьшить площадь, занимаемую на кристалле путем непосредственного масштабирования элемента памяти и реализации множества уровне состояний проводимости. Сочетание данного подхода с реализацией на структуре кроссбара позволяет дополнительно снизить используемую на кристалле площадь, поскольку кроссбар с мемристивными элементами может рассматриваться как часть ANC реализующая связность между входами нейронов и самими нейронами по параметрическому шаблону «каждый с каждым» рис. 12.

С учетом применения мемристора с биполярным механизмом переключения и реализуемой моделью управления переключением по превышению порога напряжения, блок учета вклада входного сигнала (см. рис.8, рис. 40) конкретизируется до следующих блоков: блок контроля учета сигнала по команде таймера, представляется элементом И; блок ЦАП входного сигнала; мемристор. Общее предлагаемое количество синапсов искусственного нейрона 64. Предполагаемое количество дискретных состояний реализуемых на мемристоре 8. Входной сигнал представляется однобитным. В качестве блока агрегации применяется нижняя шина кроссбара с присоединенными к ней синапсами искусственного нейрона, осуществляющая сложение токов от мемристоров. Сигнал от блока агрегации поступает на усилитель и далее на три программируемых компаратора с разрядностью кода в 10 бит. Программирование компараторов предполагает использование блока мемристоров для задания требуемого сигнала сравнения. Сигнал от блока мемрситоров подается на преобразователь сигналов и далее после преобразования передается на инвертирующий вход. Программируемые компараторы генерируют двоичный позиционный код частичного адреса ячейки LUT2 в конце каждого цикла учета текущей активности на входах искусственного нейрона. Двоичный позиционный код поступает на шифратор, где преобразуется в выходной двухбитный сигнал.

Шифратор выступает в качестве ответной части блока LUT1, представляющей агрегированный уровень текущей активности на входах нейрона в виде части двоичного адреса ячейки памяти LUT2.

Двоичный двухбитный сигнал от LUT1 передается в сдвиговый двухбитный регистр. На следующем шаге вычислений процесс повторяется за исключение того аспекта, что перед передачей очередной части адреса ячейки производится сдвиг регистра. В процессе активации искусственного нейрона под управлением таймера осуществляется вычисление адреса ячеек, которые осуществляют хранение значений выходного сигнала. По окончании вычисления адреса осуществляется его передача от сдвигового регистра на блок формирования выходного сигнала LUT2. Блок формирования выходного сигнала LUT2 выдает значения ячейки адреса на выход нейрона и удерживает их до команды таймера. Уточненная структурная схема представлена на рис. 41.

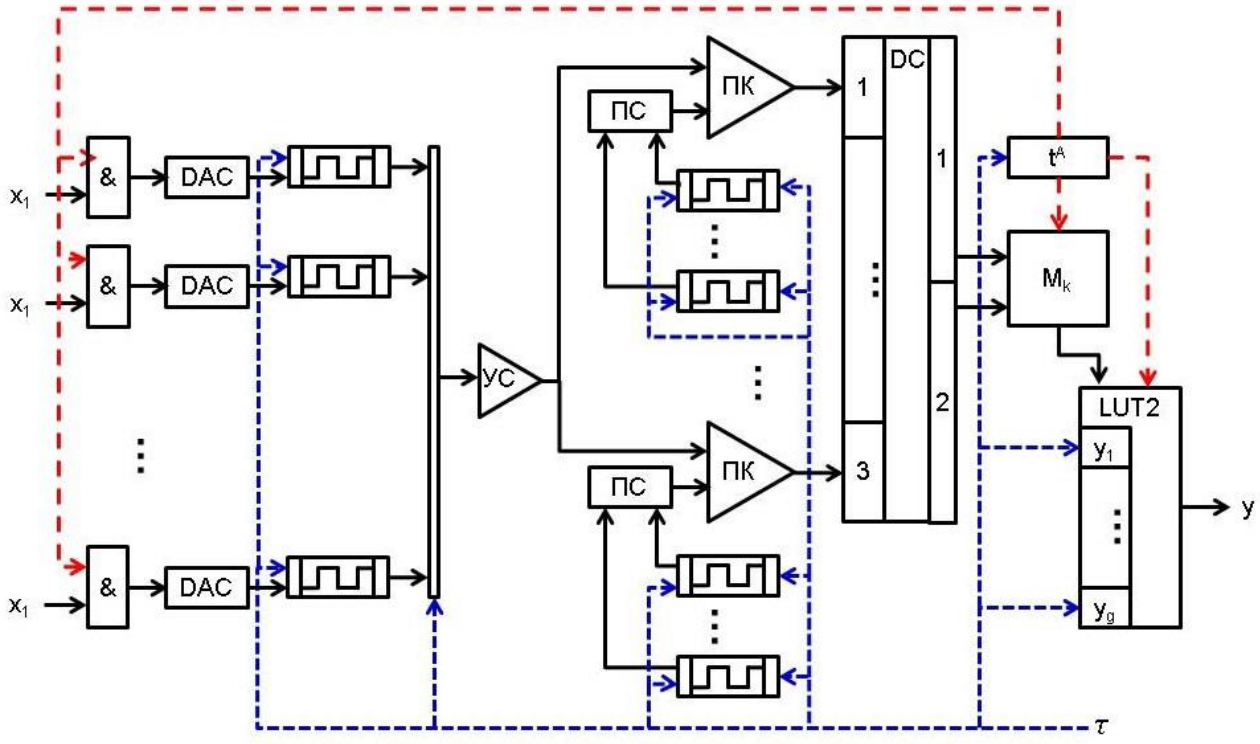


Рисунок 41. Уточненная структурная схема КААН с динамической функцией активации на базе мемристивных компонентов. DAC – цифроаналоговый преобразователь сигнала, УС – блок усиления сигнала, ПК – программируемый компаратор, DC – дешифратор; Mk – сдвиговый регистр, LUT2 – блок генерации выходного сигнала, t^A – блок таймера. Синяя пунктирная линия – линии команд. Красная пунктирная линия – линии сигналов таймера. Черная сплошная линия – линия передачи информационного сигнала

3.3 Выводы по главе 3.

В данной главе произведена конкретизация обобщенной схемы аппаратной реализации КААН с динамической функцией активации до структурной схемы аппаратной реализации с применением мемристоров. Предложенная схема относится к гибридной схемотехнике и учитывает такие физические особенности механизма переключения мемристоров как множественность уровней состояний проводимости элемента и управление механизмом переключения мемристоров приложенным напряжением.

В рамках главы показывается преимущество применения двух блоков LUT при реализации блока функции агрегации и блока функции активации перед техническим решением с одним блоком. Предлагаемый подход позволяет учесть не только обобщенный уровень возбуждения нейрона за период активации, но и амплитуду агрегированных сигналов в момент времени вычисления текущего уровня активации.

Схемотехническое решение включает применение мемристивных компонентов не только в блоках учета вклада сигнала, но и структуре блока LUT1, относящегося к функции активации искусственного нейрона. Немаловажным эффектом от применения блоков LUT является относительная универсальность задания функций активации, с возможностью переключения между функциями по сигналам от таймера. Реализация динамической функции активации позволяет производить переключение между функциями активации нейрона в процессе вычислений.

ЗАКЛЮЧЕНИЕ

Основным результатом диссертационной работы является решение актуальной научной и технической задачи, направленной на теоретическое исследование применения мемристивных компонентов при аппаратной реализации искусственного нейрона с динамической функцией активации, имеющей существенной значение для проектирования нейроморфных систем и блоков.

Основными выводами работы являются:

1. Произведено математическое описание искусственного нейрона с динамической функцией активации, переключаемой либо в процессе функционирования нейрона, либо в процессе обучения. В рамках модели осуществлено обобщение агрегационных функций, реализуемых операциями сложения и умножения. Модель позволяет применять метод неэквидистантного задания значений весовых коэффициентов синапсов, что в предельном случае увеличить область определения агрегационной функции искусственного нейрона до $I*N^{W}$, где N — количество синапсов, W — мощность множества значений весовых коэффициентов, I — мощность множества значений входных сигналов.

2. Установлено соответствие между физическими параметрами известных мемристивных компонентов (вольт-амперные характеристики процессов переключения, механизм переключения, и другие) и формальными параметрами описания на языке Verilog-A в среде САПР Cadence, позволяющее использовать мемристор как библиотечный элемент САПР. Предлагаемое описание позволяет произвести описание множественности состояний проводимости с учетом девиации напряжений порога переключения между циклами переключения, произвести учет параметров разброса высокорезистивного и низкорезистивного состояний между циклами переключения, а также учет параметров разброса количества циклов переключения.

3. Продемонстрировано преимущество реализации модели КААН с динамической функцией активации, включающей два блока LUT и сдвиговый

регистр, заключающееся в учете амплитуд агрегированного сигнала в каждый момент времени активации в дополнении к учету обобщенного уровня возбуждения нейрона.

4. Из результатов моделирования следует предпочтительность снижения разброса параметров высокопроводящего состояния и параметров разброса пороговых напряжений для изготавливаемых МДМ структур. Разброс параметров в высокопроводящем состоянии будет вносить большие искажения в обрабатываемый сигнал в сравнении с влиянием разброса в низкопроводящем состоянии. Разброс параметров пороговых напряжений при переключении от цикла к циклу напрямую влияет на точность получаемых промежуточных состояний проводимости мемристора.

5. Следствием из результатов моделирования является приоритет задания множественности состояний путем подачи коротких (не более 15 нс) импульсов с малой амплитудой по напряжению (не более 0.15 В выше порога переключения) перед длительными уровнями с малой амплитудой, либо короткими уровнями с большой амплитудой.

В ходе диссертационной работы достигнута её цель в виде выполнения теоретического исследования по всему комплексу поставленных задач, а именно: произведен синтез абстракции формального нейрона и конечного автомата, разработано описание высокоуровневой функциональной модели искусственного нейрона с динамической функцией активации, произведено модельное представление мемристора средствами САПР Cadence на языке высокого уровня Verilog-A. Также в ходе анализа модельных представлений мемристора получены практически значимые результаты, с учетом которых произведено описание нейрона с динамической функцией активации в виде структурной схемы с учетом физических особенностей мемристивных элементов.

Результаты диссертационной работы могут быть положены в основу дальнейших экспериментальных и теоретических исследований мемристивных элементов и нейроморфных архитектур на их основе.

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

Входной сигнал – элемент множества входных сигналов, являющийся абстрактным представлением данных поступающих на вход синапс искусственного нейрона. В контексте данной работы следует руководствоваться определением данных согласно ISO/IEC 2382:2015.

Синапс, коннекционный элемент, входящая информационная линия связи – элемент нейрона i отвечающий за преобразование сигнала с i-го входа искусственного нейрона в соответствии с применяемой моделью нейрона.

Весовой коэффициент (вес связи, вес синапса, внутренний параметр синапса) – элемент wi множества значений синапса W, предназначенный для обозначения вклада сигнала i в уровень возбуждения нейрона на текущем этапе определения его активности.

Функция агрегации – функция учета сигналов на выходах синапсов искусственных нейронов. Может являться n-мерной математической операцией (сложение, умножение, дизъюнкция, конъюнкция) логическим выражением, функцией нечеткого множества, сложной функцией представляющей комбинацию сложения и умножения от различных групп синапсов и т. д.

Функция активации – функция или композиция функций, отображения результата функции агрегации во множество выходных сигналов.

Аксон, выходная информационная линия связи – элемент нейрона, отвечающий за преобразование выходного сигнала в соответствующий i-му присоединенному нейрону уровень сигнала, примером использования аксонов в ИНС может служить outstar Grossberg`a.

Синхронный режим работы нейронов – режим функционирования сети, в котором каждый нейрон сети срабатывает одновременно с другими нейронами сети, т. е. функция активации нейронов вычисляется для всех нейронов в один и тот же момент времени работы сети.

Асинхронный режим работы нейронов – режим функционирования сети, в котором есть хотя бы один нейрон, который срабатывает не одновременно с другими нейронами сети, т. е. функция активации хотя бы одного нейрона в хотя бы один момент времени (на хотя бы одном такте) не вычисляется в отличие от функций активации других нейронов или же вычисляется в момент когда другие функции активации не вычисляются.

Искусственные нейроны, срабатывающие по совпадению (детектор совпадений) – класс нейронов, для которых вычисление функции активации не зависит от предыдущего состояния нейрона.

Динамические искусственные нейроны – класс нейронов, для которых возможен режим функционирования нейрона, в процессе которого функция активации зависит от предыдущего состояния нейрона.

Однонаправленная связь – см. синапс.

Двунаправленная связь – вид связи между двумя нейронами H1 и H2, имеющий общий весовой коэффициент для сигналов, передаваемых от нейрона H1 к нейрону H2 и от нейрона H2 к нейрону H1. В случае если весовые коэффициенты для передачи сигналов по разным направлениям разные, в рамках данной работы такая связь рассматривается как две однонаправленные связи.

Шаблон связности – схема организации передачи информации между нейронами, определяющая наличие либо отсутствие связи передачи информации от нейрона H1 к нейрону H2.

Случайный шаблон связности – шаблон связности, согласно которому существование связи между нейронами H1 и H2 определяется случайным образом.

Параметрический шаблон связности – шаблон связности, согласно которому существование связи между нейронами H1 и H2 определяется исследователем или алгоритмом в соответствии с наперед заданными критериями.

СПИСОК СОКРАЩЕНИЙ

ИНС – искусственная нейронная сет, искусственные нейронные сети

WTA – правило обучения нейрона «победитель забирает все»

RBF – радиально базисная функция

FFNN – искусственная нейронная сеть прямого распространения сигнала

RNN – искусственная нейронная сеть с рекуррентными связями

BNN – искусственная нейронная сеть с распространением сигнала в обоих направлениях

SOM – самоорганизующаяся карта Кохонена

BM – машина Больцмана

RBM – ограниченная машина Больцмана

HNN – иерархическая искусственная нейронная сеть

DBN – глубинная искусственная нейронная сеть доверия

CNN – сверточная искусственная нейронная сеть

DN – развертывающая искусственная нейронная сеть

КААН – конечный автомат абстрактного нейрона

LUT – блок таблицы соответствия выходных сигналов входным

PDP – параллельная распределенная обработка данных

GRID – виртуальный суперкомпьютер

CPU – центральный процессор

GPU – графический сопроцессор

ИС – интегральные схемы

FPGA – программируемые интегральные схемы

PCM – память на основе изменения фазового состояния активного слоя

SOT – память на основе вращения спина

MTJ – память на основе туннельного спинового тока

DNC – цифровое нейроядро

ANC – аналоговое неййроядро

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

1 **Г.С. Теплов, И.В. Матюшкин, Е.С. Горнев** Принципы повышения наработки до отказа схем управления на основе клеточно-автоматных и нейроподобных структур //Атомный проект, Информационно-аналитический журнал для специалистов в области атомного машиностроения –2015 – №. 22. – С. 48-49

2 **Матюшкин И. В., Теплов Г. С., Горнев Е. С.** Особенности реализации микросхем с клеточно-автоматной архитектурой на основе эффекта резистивного переключения //Электроника-2015 Тезисы. Международной научно-технической конференции, г. Зеленоград, 19-20 ноября 2015, стр. 51-52.

3 **Горнев Е.С., Матюшкин И.В., Теплов Г.С.** Анализ концепций неклассического компьютинга и парадигмы коннекционизма //Электронная техника. Серия 3: Микроэлектроника. – 2015. – №: 2 (158). – С. 45-66.

4 **Stempkovsky A.L., Gavrilov S.V., Matyushkin I.V., Teplov G.S.** On the issue of application of cellular automata and neural networks methods in VLSI design //Optical memory & Neural Networks. – 2016. – Том 2. – №. 2. – С. 72-78

5 **Горнев Е.С., Теплов Г.С.** Математическая модель конечного автомата абстрактного нейрона и сетей на его основе //Нано- и микросистемная техника. – 2018. – Т. 20. – №. 7. – С. 434-442.

СПИСОК ЦИТИРУЕМОЙ ЛИТЕРАТУРЫ

1. **McCulloch W.S. Pitts W.A.** A logical calculus of the ideas immanent in nervous activity //The bulletin of mathematical biophysics. – 1943. – Т. 5. – №4. – С. 115-133.
2. **Rosenblatt F.** The Perceptron: a probabilistic model for information storage and organisation in the brain //Psychological review. – 1958. – Т. 65. – №6. – С. 386.
3. **Widrow B., Hoff M.E.** Adaptive switching circuits. – STANFORD UNIV CA STANFORD ELECTRONICS LABS, 1960. – № TR-1553-1.
4. **Widrow B., Smith F.W.** Pattern-recognizing control systems //Computer and Information Sciences (COINS) Proceedings. – 1964.
5. **Clarke T.L.** Generalization of neural networks to the complex plane //1990 IJCNN International Joint Conference on Neural Networks. – IEEE, 1990. – С. 435-440.
6. **Grossberg S.** Some nonlinear networks capable of learning a spatial pattern of arbitary complexity //Proceedings of the National Academy of Science. – 1968. – Т. 59. – №. 2. – С. 368-372.
7. **Grossberg S.** Neural pattern discrimination //Jornal of Theoretical Biology. – 1970. – Т. 27. – №. 2. – С. 291-337.
8. **Kaski S., Kohonen T.** Winner-take-all networks for physiological models of competitive learning //Neural Networks. – 1994. – Т. 7. – №. 6. – С. 973-984.
9. **Specht D.F.** Probabilistic neural networks for classification, mapping, or associative memory //IEEE international conference on neural networks. – 1988. – Т. 1. – №. 24. – С. 525-535.
10. **Specht D.F.** A general regression neural networks //IEEE transactions on neural networks. – 1991. – Т. 2. – №. 6. – С. 568-576.

11. **Koch C., Poggio T.** Multiplying with synapses and neurons //Single Neuron Computation. – 1992. – С. 315-345.

12. **Mel B.W.** The sigma-pi model neuron: roles of the dendritic tree in associative learning //Soc Neurosci Abstr. – 1990. – Т. 16. – С. 205.4.

13. **Gurney K.** An introduction to neural networks. – CRC press., 2014. – С. 231-233.

14. **Chaturvedi D.K.** Soft computing: techniques and its applications in electrical engineering. – Springer, 2008. – Т. 103.

15. **Шибзухов З.М., Чередников Д. Ю.** О моделях нейронов агрегирующего типа //Машинное обучение и анализ данныхю – 2015. – Т. 1. – №. 12. – С. 1706-1716.

16. **Матюшкин И.В., Соловьев Р.А.** Модель адаптивного нейрона и его аппаратная реализация на плис //Электронная техника. Серия 3: Микроэлектроника. – 2017. – №. 3. – С. 53-61.

17. **Gupta M.M., Qi J.** On fuzzy neuron models //Neural Networks, 1991, IJCNN-91-Seattle International Joint Conference on. – IEEE, 1991. – Т. 2. – С. 431-436.

18. **Hopfield J.J.** Neural networks and physical systems with emergent collective computatuional abilities //Proceedings of the national academy of science. – 1982. – Т. 79. – №. 8. – С. 2554-2558.

19. **Abbott L.F.** Lapicque`s introdaction of the intagrate-and-fire model neuron (1907) //Brain research bulletin. – 1999. – Т. 50. – №. 5-6. – С. 303-304.

20. **Tuckwell H.C.** Introduction to theoretical neurobiology: volume 2, nonlinear and stochastic theories. – Cambridge University Press, 2005. – Т. 8.

21. **Vidybida A.K.** Inhibition as binding controller at the single neuron level //BioSystems. – 1998. – Т. 48. – №. 1-3. – С. 263-267.

22. **Parlos A.G., Chong K. T., Atiya A.F.** Application of the recurrent multilayer perceptron in modelling complex process dymamics //IEEE Transactions on Neural Networks. – 1994. – Т. 5. – №. 2. – С. 255-266.

23. **Kosko B.** Bidirectional associative memories //IEEE Transactions on Systems, man, and Cybernetics. – 1988. – Т. 18. – №. 1. – С. 49-60.

24. **Ruck D.W. et al.** The multilayer perceptron as an approximation to a Bayes optimal discriminant function //IEEE Transactions on Neural Networks. – 1990. – Т. 1. – №. 4. – С. 296-298.

25. **Kohonen T.** The self-organizing map //Proceedings of the IEEE. – 1990. – Т. 78. – №. 9. – С. 1464-1480.

26. **Kohonen T.** Self-organized formation of topologically correct feature maps //Biological cybernetics. – 1982. – Т. 43. – №. 1. – С. 59-69.

27. **Cheung K.F., Atlas L.E., Marks R.J.** Synchronous vs asynchronous behavior of Hopfield`s CAM neural net //Applied Optics. – 1987. – Т. 26. – №. 22. – С. 4808-4813.

28. **Ackley D.H., Hinton G.E., Sejnowski T.J.** A learning algorithm for Boltzmann machines //Cognitive science. – 1985. – Т. 9. – №. 1. – С. 147-169.

29. **Aarts E.H.L., Krost J.H.M.** Boltzmann machines for traveling salesman problems //European Journal of Operational Research. – 1989. – Т. 39. – №. 1. – С. 79-95.

30. **Prager R.W., Harrison T.D., Fallside F.** Boltzmann machines for speech recognition //Computer Speech & Language. – 1986. – Т. 1. – №. 1. – С. 3-27.

31. **Aarts E.H.L., Korst J.H.M.** Boltzmann machines and their applications //International Conference on Parallel Architectures and Languadges Europe. – Springer, Berlin, Heidelberg, 1987. – С. 34-50.

32. **Larochelle H., Bengio Y.** Classification using discriminative restricted Boltzmann machines //Proceedings of the 25th international conference on Machine learning. – ACM, 2008. – С. 536-543.

33. **Hjelm R.D. et. al.** Restricted Boltzmann machines for neuroimaging: an application in identifying intrinsic networks //NeuroImage. – 2014. – Т. 96. – С. 245-260.

34. **Orr M.J. et. al.** Introdaction to radial basis function networks. – 1996.

35. **Mavrovouniotis M.L., Chang S.** Hierarchical neural networks //Computers & chemical engineering. – 1992. – Т. 16. – №. 4. – С. 347-369.

36. **Fukushima K., Miyake S.** Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition //Competition and cooperation in neural nets. – Springer, Berlin, Heidelberg, 1982. – С. 267-285.

37. **Fukushima K.** Artificial vision by muti-layered neural networks: Neocognitron and its advanses //Neural networks. – 2013. – Т. 37. – С. 103-119.

38. **Weng J.J., Ahuja N., Huang T.S.** Cresceptron: a self-organizing neural network wich grows adaptively //Neural Networks, 1992. IJCNN., International Joint Conference on. – IEEE, 1992. – Т. 1. – С. 576-581.

39. **Weng J.J., Ahuja N., Huang T.S.** Learning recognition and segmentation using the cresceptron //International Journal of Computer Vision. – 1997. – Т. 25. – №. 2. – С. 109-143.

40. **Nossek J.A. et. al.** Cellular neural networks: Theory and circuit design //International Journal of Circuit Theory and Applications. – 1992. – Т. 20. – №. 5. – С. 533-553.

41. **Chua L.O., Yang L.** Cellular neural networks: Applications //IEEE Transactions on circuits and systems. – 1998. – T. 35. – №. 10. – C. 1273-1290.

42. **Chicco D., Sadowski P., Baldi P.** Deep autoencoder neural networks for gene ontology annotation predictions //Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatiocs. – ACM, 2014. – C. 533-540.

43. **Glorot X., Bordes A., Bengio Y.** Deep sparse rectifier neural networks //Proceedings of the fourteenth international conference on artificial intelligence and statistics. – 2011. – C. 315-323.

44. **Im D.J., et. al.** Denoising Criterion for Variational Auto-Encoding Framework //AAAI. – 2017. – C. 2059-2065.

45. **Le Roux N., Bengio Y.** Representational power of restricted Boltzmann machines and deep belief networks //Neural computation. – 2008. – T. 20. – №. 6. – C. 1631-1649.

46. **Lee H. et. al.** Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations //Proceedings of 26th annual international conference on machine learning. – ACM, 2009. – C. 609-616.

47. **Mohamed A., Dahl G., Hinton G.** Deep belief networks for phone recognition //Nips workshop on deep learning for speech recognition and related applications. – 2009. – T. 1. – №. 9. – C. 39.

48. **Zeiler M.D. et. al.** Deconvolutional networks. //2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco – 2010. – C. 2528-2535.

49. **Dhanajay S. Phatak** Fault Tolerant Artificial Neural Networks //Proc. Of the 5th Dual Use Technologies and Applications Conference, May 1995 – C. 1-7.

50. **Kelemen J., Sosrik P**. (eds.): Fault-Tolerant Structures: Towards Robust Self-Replication in a Probabilistic Environment //ECAL 2001, LNAI 2159 – 2001. – C. 90-99.

51. **Janke S., Whitehead M.** Partical Fault Tolerant 2D Cellular Automata //Proceedings of the European Conference on Artificial Life. – 2015. – C. 158-165.

52. **Dipti Deodhare, Vidyasagar M., Sathiya Keerthi S.** Synthesis of Fault-Tolerant Feedforward Neural Networks Using Minmax Optimization //IEEE Transactions on Neural Networks. – 1998. – Т. 9. – №. 5. – С. 891-900.

53. **Nugent A., Kenyon G., Porter R.** Unsupervised adaptation to improve fault tolerance of neural network classifiers //Evolvable Hardware, Proceedings DoD conference on. – 2004. – С. 146-149.

54. **Piuri V.** Analysis of fault tolerance in artificial neural networks //Journal of Parallel and Distributed Computing. – 2001. – С. 18-48.

55. **Elko B. Tchernev, Rory G. Mulvaney, Dhananjay S. Phatak** Investigating the Fault Tolerance of Neural Networks //Neural Computation 17. – 2006. – С. 1646-1664.

56. **Fontes P., Borralho R., Antunes A. et. al.** Fault tolerance simulation and evalution tool for artificial neural networks //Proceedings 8th Portuguese Control Conference. – 2008. – С. 1-6.

57. **Moore E.P.** GEDANKEN-EXPERIMENTS ON SEQUENTIAL MACHINES //Automata Studies, Annals of Mathematical Studies. — Princeton University Press — 1956. — Т. 34. — С. 129-153.

58. **Medler, David A.** A Brief History of Connectionism// Neural Computing Surveys. — 1998. —№ 1(2), – p.18-72.

59. **Савельев А.В., Янковская Е.А.** К истокам и будущему нейронаук. Юбилейный междисциплинарный симпозиум «150 лет «РЕФЛЕКСАМ ГОЛОВНОГО МОЗГА» ИМ Сеченова»

60. **D. E. Rumelhart et. al.** Parallel Distributed Processing. – Cambridge, MA: MIT press, 1987. – Т. 1.

61. **W. Bechtel and A. Abrahamsen.** Connectionism and the Mind: An Introduction to Parallel Processing in Networks. – Basil Blackwell, MA, 1991.

62. **фон Нейман Дж.** «Теория самовоспроизводящихся автоматов» //М.: Мир, 1971. С.382

63. **Ulam S.** Random Processes and Transformations //Proceedings of International Congress Mathematics. – 1952. – Т. 2. – С. 264-275.

64. **Винер Н., Розенблют А.** «Проведение импульсов в сердечной мышце. Математическая формулировка проблемы проведения импульсов в сети связанных возбудимых элементов, в частности в сердечной мышце» //Кибернетический сборник. Вып.3. – М.: Изд. иностр. лит., 1961. С.7– 56

65. **Scharf J. H. K. Zuse**, Rechnender Raum (Schriften zur Datenverarbeitung, Band 1). VIII+ 70 S. m. 74 Abb. Braunschweig 1969. Friedr. Vieweg & Sohn. Preis brosch. DM 16, 80 //ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik. – 1971. – Т. 51. – №. 8. – С. 649-650.

66. **Hecht-Nielsen R.** Kolmogorov's mapping neural network existence theorem. //In Proceedings of the First International Conference on Neural Networks. – 1987. – Т. III. – С. 11–14

67. **Е.Н. Бендерская, А.А. Толстов** Тенденции развития средств аппаратной поддержки нейровычислений //Научно технические ведомости СПбГПУ Информатика. Телекоммуникации. Управление. – 2013. – №. 3. (174). – С. 9-18.

68. **Morgado Dias F.M., Antunes A., Mota A.M.** Artificial neural networks: a review of commercial hardware //Engineering Applications of Artificial Intelligence. – 2004. – Т. 17. – №. 8. – С. 945–952.

69. **Misra J., Saha I.** Artificial neural networks in hardware: A survey of two decades of progress // Neurocomputing. – 2010. – Т. 74. – №. 1-3. – С. 239-255.

70. **Markram H.** The human brain project //Scientific American. – 2012. – Т. 306. – №. 6. – С. 50-55.

71. **Hsu J.** IBM's new brain [News] //IEEE spectrum. – 2014. – Т. 51. – №. 10. – С. 17-19.

72. **Woods D., Naughton T. J.** Optical computing: Photonic neural networks //Nature Physics. – 2012. – Т. 8. – №. 4. – С. 257.

73. **Younger A. S., Redd E.** Computing by Means of Physics-Based Optical Neural Networks // Proceedings Sixth Workshop on Developments in Computational Models: Causality, Computation, and Physics (DCM 2010), Electronic Proceedings in Theoretical Computer Science. – 2010 – Т. 26. С. 159–167.

74. **Larger L. et al.** Photonic information processing beyond Turing: an optoelectronic implementation of reservoir computing //Optics express. – 2012. – Т. 20. – №. 3. – С. 3241-3249.

75. **Khan F. N. et al.** Non-data-aided joint bit-rate and modulation format identification for next-generation heterogeneous optical networks //Optical Fiber Technology. – 2014. – Т. 20. – №. 2. – С. 68-74.

76. **Tait A. N. et al.** Demonstration of WDM weighted addition for principal component analysis //Optics Express. – 2015. – Т. 23. – №. 10. – С. 12758-12765.

77. **Dejonckheere A. et al.** All-optical reservoir computer based on saturation of absorption //Optics express. – 2014. – Т. 22. – №. 9. – С. 10868-10881.

78. **Merrikh-Bayat F., Bagheri-Shouraki S.** Efficient neuro-fuzzy system and its Memristor Crossbar-based Hardware Implementation //arXiv preprint arXiv:1103.1156. – 2011.

79. **Indiveri G. et al.** Integration of nanoscale memristor synapses in neuromorphic computing architectures //Nanotechnology. – 2013. – Т. 24. – №. 38. – С. 384010.

80. **Sharad M. et al.** Spin-based neuron model with domain-wall magnets as synapse //IEEE Transactions on Nanotechnology. – 2012. – Т. 11. – №. 4. – С. 843-853.

81. **Zhu L. Q. et al.** Artificial synapse network on inorganic proton conductor for neuromorphic systems //Nature communications. – 2014. – Т. 5. – С. 3158.

82. **Sharad M., Fan D., Roy K.** Spin-neurons: A possible path to energy-efficient neuromorphic computers //Journal of Applied Physics. – 2013. – Т. 114. – №. 23. – С. 234906.

83. **Sengupta A. et al.** Spin orbit torque based electronic neuron //Applied Physics Letters. – 2015. – Т. 106. – №. 14. – С. 143701.

84. **Biswas A. K., Atulasimha J., Bandyopadhyay S.** The straintronic spin-neuron //Nanotechnology. – 2015. – Т. 26. – №. 28. – С. 285201.

85. **Pershin Y. V., Di Ventra M.** Memcapacitive neural networks //Electronics Letters. – 2014. – Т. 50. – №. 3. – С. 141-143.

86. **Nordström T., Svensson B.** Using and designing massively parallel computers for artificial neural networks //Journal of parallel and distributed computing. – 1992. – Т. 14. – №. 3. – С. 260-285.

87. **Seiffert U.** Artificial neural networks on massively parallel computer hardware //Neurocomputing. – 2004. – Т. 57. – С. 135-150.

88. **Farabet C. et al.** Hardware accelerated convolutional neural networks for synthetic vision systems //Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on. – IEEE, 2010. – С. 257-260.

89. **Zhu J., Sutton P.** FPGA implementations of neural networks–a survey of a decade of progress //International Conference on Field Programmable Logic and Applications. – Springer, Berlin, Heidelberg, 2003. – С. 1062-1066.

90. **Lysaght P. et al.** Artificial neural network implementation on a fine-grained FPGA //International Workshop on Field Programmable Logic and Applications. – Springer, Berlin, Heidelberg, 1994. – С. 421-431.

91. **E. Won E.** A hardware implementation of artificial neural networks using field programmable gate arrays //Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. – 2007. – Т. 581. – №. 3. – С. 816-820.

92. **Dinu A., Cirstea M. N., Cirstea S. E.** Direct neural-network hardware-implementation algorithm //IEEE Transactions on Industrial Electronics. – 2010. – Т. 57. – №. 5. – С. 1845-1848.

93. **Atibi M. et al.** Parallel and Mixed Hardware Implementation of Artificial Neuron Network on the FPGA Platform //International Journal of Engineering & Technology. – 2014. – Т. 6. – №. 5. – С. 0975-4024.

94. **Schemmel J. et al.** A wafer-scale neuromorphic hardware system for large-scale neural modeling //Circuits and systems (ISCAS), proceedings of 2010 IEEE international symposium on. – IEEE, 2010. – С. 1947-1950.

95. **Dlugosz R. et al.** Realization of the conscience mechanism in CMOS implementation of winner-takes-all self-organizing neural networks //IEEE Transactions on Neural Networks. – 2010. – Т. 21. – №. 6. – С. 961-971.

96. **Tanaka G. et al.** Regularity and randomness in modular network structures for neural associative memories //Neural Networks (IJCNN), 2015 International Joint Conference on. – IEEE, 2015. – С. 1-7.

97. **Ji Y. et al.** A hardware implementation of a radial basis function neural network using stochastic logic //Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition. – EDA Consortium, 2015. – С. 880-883.

98. **Su Z. et al.** Adaptive integratable hardware realization of analog neural networks for nonlinear system //Industrial Informatics (INDIN), 2015 IEEE 13th International Conference on. – IEEE, 2015. – С. 521-526.

99. **Arthur J. V. et al.** Building block of a programmable neuromorphic substrate: A digital neurosynaptic core //IJCNN. – 2012. – С. 1-8.

100. **Merolla P. A. et al.** A million spiking-neuron integrated circuit with a scalable communication network and interface //Science. – 2014. – Т. 345. – №. 6197. – С. 668-673.

101. **Chua L.** Memristor-the missing circuit element //IEEE Transactions on circuit theory. – 1971. – Т. 18. – №. 5. – С. 507-519.

102. **Strukov D.B. et al.** The missing memristor found //nature. – 2008. – Т.453. – №. 7191. – С. 80.

103. **Di Ventra M., Pershin Y. V., Chua L. O.** Circuit elements with memory: memristors, memcapacitors, and meminductors //Proceedings of the IEEE. – 2009. – Т. 97. – №. 10. – С. 1717-1724.

104. **Hu S. G. et al.** Review of nanostructured resistive switching memristor and its applications //Nanoscience and Nanotechnology Letters. – 2014. – Т. 6. – №. 9. – С. 729-757.

105. **Захаров П.С., Итальянцев А.Г.** Модель эффекта переключения электрической проводимости в структурах резистивной памяти на основе нестехиометрического оксида кремния //Известия вузов. ЭЛЕКТРОНИКА. – 2016. – Т. 21. – No 4. – С. 309–315.

106. **Bocquet M. et al.** Compact modeling solutions for oxide-based resistive switching memories (OxRAM) //Journal of Low Power Electronics and Applications. – 2014. – Т. 4. – №. 1. – С. 1-14.

107. **Wang W. et al.** Memristive behavior of ZnO/Au film investigated by a TiN CAFM tip and its model based on the experiments //IEEE Transactions on Nanotechnology. – 2012. – Т. 11. – №. 6. – С. 1135-1139.

108. **Sun X. et al.** Coexistence of the bipolar and unipolar resistive switching behaviours in Au/SrTiO3/Pt cells //Journal of Physics D: Applied Physics. – 2011. – Т. 44. – №. 12. – С. 125404.

109. **Aziza H. et al.** Oxide based resistive RAM: ON/OFF resistance analysis versus circuit variability //Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT), 2014 IEEE International Symposium on. – IEEE, 2014. – С. 81-85.

110. **Kang J. F. et al.** Modeling and design optimization of ReRAM //Design Automation Conference (ASP-DAC), 2015 20th Asia and South Pacific. – IEEE, 2015. – С. 576-581.

111. **Ascoli A. et al.** The art of finding accurate memristor model solutions //IEEE Journal on Emerging and Selected Topics in Circuits and Systems. – 2015. – Т. 5. – №. 2. – С. 133-142.

112. **Yan P. et al.** Conducting mechanisms of forming-free TiW/Cu2O/Cu memristive devices //Applied Physics Letters. – 2015. – Т. 107. – №. 8. – С. 083501.

113. **Liu Y. et al.** Percolation mechanism through trapping/de-trapping process at defect states for resistive switching devices with structure of Ag/SixC1− x/p-Si //Journal of Applied Physics. – 2014. – Т. 116. – №. 6. – С. 064505.

114. **Kurnia F. et al.** Compliance current induced non-reversible transition from unipolar to bipolar resistive switching in a Cu/TaOx/Pt structure //Applied Physics Letters. – 2015. – Т. 107. – №. 7. – С. 073501.

115. **Zhong L. et al.** Nonpolar resistive switching in Cu/SiC/Au non-volatile resistive memory devices //Applied Physics Letters. – 2014. – Т. 104. – №. 9. – С. 093507.

116. **Kim W. et al.** Multistate memristive tantalum oxide devices for ternary arithmetic //Scientific reports. – 2016. – Т. 6. – С. 36652.

117. **Abunahla H., Mohammad B., Homouz D.** Effect of device, size, activation energy, temperature, and frequency on memristor switching time //Microelectronics (ICM), 2014 26th International Conference on. – IEEE, 2014. – С. 60-63.

118. **Benderli S. Wey T.A.** On SPICE macromodeling of TiO 2 memristors //Electronics letters. – 2009. – Т.45. – № 7. – С. 377-379.

119. **Emara A. A., Aboudina M. M., Fahmy H. A. H.** Corrected and accurate Verilog-A for linear dopant drift model of memristors //Circuits and Systems (MWSCAS), 2014 IEEE 57th International Midwest Symposium on. – IEEE, 2014. – С. 499-502.

120. **Biolek Z., Biolek D., Biolkova V.** SPICE Model of Memristor with Nonlinear Dopant Drift //Radioengineering. – 2009. – Т.18. – № 2.

121. **Joglekar Y.N., Wolf S.J.** The elusive memristor: propeties of basic electrical circuits //European Journal of Physics. – 2009/ – Т.30. – № 4. – С. 661.

122. **da Costa H.J. B., de Assis Brito Filho F., do Nascimento P.I.A.** Memristor behavioural modeling and simuations using Verilog-AMS //Circuits and Systems (LASCAS), 2012 IEEE Third Latin American Symposium on. – IEEE, 2012. – С. 1-4.

123. **Prodromakis T. et al.** A versatile memristor model with nonlinear dopant kinetics //IEEE transactions on electronic devices. – 2011. – Т.58. – №. 9. – С. 3099-3105.

124. **Kvatinsky S. et al.** TEAM: Threshold adaptive memristor model //IEEE Transactions on Circuits and Systems I: Regular Papers. – 2013. – Т. 60. – №. 1. – С. 211-221.

125. **Pickett M.D. et al.** Switching dynamics in titanium dioxide memristive devices //Jornal of Applied Physics. – 2009. – Т. 106. – №. 7. – С. 074508

126. **Kvatinsky S. et al.** Models of memristors for SPICE simulations //Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of. – IEEE, 2012. – С. 1-5.

127. **Yakopcic C. et al.** A memristor device model //IEEE electron device letters. – 2011. – Т. 32. – №. 10. – С. 1436-1438.

128. **Li C. et al.** Three-dimensional crossbar arrays of self-rectifying Si/SiO 2/Si memristors //Nature Communications. – 2017. – Т. 8. – С. 15666.

129. **Zeng G. et al.** Polynominal Metamodel integrated Verilog-AMS for memristor-based mixed-signal system design //Circuits and Systems (MWSCAS), 2013 IEEE 56th International Midwest Symposium on. – IEEE, 2013. – С. 916-919.

130. **Corinto F. Ascoli A.** A boundary condition-based approach to the modeling of memristor nanostructures //IEEE Transactiion on Circuits and Systems I: Regular Papers. – 2012. – Т. 59. – №. 11. – С. 2713-2726.

131. **Kvatinsky S. et al.** VTEAM: A general model for voltage-controlled memristors //IEEE Transactions on Circuits and Systems II: Express Briefs. – 2015. – Т. 62. – №. 8. – С. 786-790.

132. **Garcia-Redondo F. et al.** SPICE compact modeling of bipolar/unipolar memristor switching governed by electrical thresholds //IEEE Transactions on Circuits and System I: Regular Papers. – 2016. – Т. 63. – №. 8. – С. 1255-1264.

133. **Garcia-Redondo F., Lôpez-Vallejo M., Barrio C. L.** Advanced integration of variability and degradation in RRAM SPICE compact models //Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2017 14th International Conference on. – IEEE, 2017. – С. 1-4.

134. **Wang T.** Modelling multistability and hysteresis in ESD clamps, memristors and other devices //Custom Integrated Circuits Conference (CICC), 2017 IEEE. – IEEE, 2017. – С. 1-10.

135. **Lupo N. et al.** An Approximated Verilog-A Model for Memristive Devices //Circuits and Systems (ISCAS), 2018 IEEE International Symposium on. – IEEE, 2018. – С. 1-5.

136. **Yang Y. et al.** Verilog-A based effective complementary resistive switch model for simulations and analysis //IEEE Embedded Systems Letters. – 2014. – Т. 6. – №. 1. – С. 12-15.

137. **Wang X., Xu B., Chen L.** Efficient memristor model implementation for simulation and application //IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. – 2017. – Т. 36. – №. 7. – С. 1226-1230.

138. **Danilin S.N., Shchanikov S.A., Galushkin A.I.** The research of memristor-based neural network components operation accuracy in control and communication systems //Control and Communications (SIBCON), 2015 International Siberian Conference on. – IEEE, 2015. – С. 1-6.

139. **Li G. et al.** Multinominal based memristor modelling methodology for simulations and analysis //International Journal of Electronics Letters. – 2015. – Т. 3. – №. 1. – С. 1-12.

140. **Amrani E., Drori A., Kvatinsky S.** Logic design with unipolar memristors //Very Large Scale Integration (VLSI-SoC), 2016 IFIP/IEEE International Conference on. – IEEE, 2016. – С. 1-5.

141. **Adeyemo A. et. al.** Efficient sensing approaches for high-destiny memristor sensor array //Journal of Computational Electronics. – 2018. – С. 1-12.

142. **Piazza F., Uncini A., Zenobi M.** Neural networks with digital LUT activation functions //Neural Networks, 1993. IJCNN'93-Nagoya. Proceedings of 1993 International Joint Conference on. – IEEE, 1993. – Т. 2. – С. 1401-1404.

143. **Тельпухов Д. В. и др.** Методы построения прямых преобразователей модулярной логарифметики ориентированных на ЦОС //Проблемы разработки перспективных микро-и наноэлектронных систем (МЭС). – 2010. – №. 1. – С. 374-377.

ПРИЛОЖЕНИЕ № 1 Verilog-A описание биполярного мемристора

```
// VerilogA for memristors, memristorbi, veriloga

//GST -- great smart technology

//Behavioral model of memristor.
//Created in the JSC MERI by Department of Functional Electronics
//Author: research fellow Georgii Sergeevitch Teplov

`include "disciplines.vams"
`include "constants.vams"

module memristorbi (t, b); //t--top electrode, b--bottom electrode
    inout t, b;
    electrical t, b;
    parameter real Ron=1000.0; //Min value Ron
    parameter real Roff=25000.0; //Min value Roff
    parameter real Von=0.6; //Max value Vset threshold in integers millivolts
    parameter real Voff=-0.6; //Max value Vreset threshold in integers millivolts
    parameter real Rstart=1000; //Initial iternal state
    parameter real Vgrw=0.01; //Start point of filament grow, in Volts
    parameter real Vmlt=0.01; //Start point of filament melt, in Volts
    parameter real dvsn=100; //Time switching on in volt switching scale
    parameter real dvsff=100; //Time switching off in volt switching scale
    parameter real dtsc=0.000000001; //Time switching scale
    parameter real dtsn=100; //Time switching on in time switching scale
    parameter real dtsff=100; //Time switching off in time switching scale
    parameter integer NumCyc = 100; //Cycling parameter
    parameter real DltRon=100; //Absolute deviation of Ron in integers
    parameter real DltRoff=2500; //Absolute deviation of Ron in integers
    parameter real DltVon=0.1; //Absolute deviation of Von in integers millivolts
    parameter real DltVoff=0.1; //Absolute deviation of Voff in integers millivolts
    parameter integer pseed=31; //Random core;

    real x; //Internal state
    real xdtn; //Unit increment vector Set
    real xdtff; //Unit increment vector Reset
    real tx; //Time point one
    real txp1; //Time point two
    real Vx; //Voltage point one
    real Vxp1; //Voltage point two
    real Cyc; //Variable of cyclic parameter
    real RRon; //Variable of random resistance, Set
```

```
real RRoff; //Variable of random resistance, Reset
real RVon; //Variable of random resistance, Set
real RVoff; //Variable of random resistance, Reset
integer seed; //Random core variable;

   analog function real dFx; //function of Set-Reset
      input Vxp1, Vx, Vnff, tx, txp1, Vgm, x, xdt, Rn, Rff, dts;
      real Vxp1, Vx, Vnff, tx, txp1, Vgm, x, xdt, Rn, Rff, dts, y1, y2;
         begin //start of function
            y1 = x - (Vxp1-Vnff)/Vgm*xdt; //case first cross border
            y2 = x - (Vxp1+Vx-2*Vnff)/Vgm*0.5*(txp1-tx)/dts*xdt; //other cases
         if ((tx==0)&&(Rn<y1)) begin //branches first crossing border
            dFx = y1; //Resistance change for first step
         end
         else begin if ((tx!=0)&&(Rn<y2)) begin //branches other crossing border
            dFx = y2; //Resistance change for other steps
         end else begin //case last cross border
            dFx = Rn; //Resistance change for final step
         end
         end
         end //end of function
   endfunction

   analog function real dFrx; //function of Set-Reset
      input Vxp1, Vx, Vnff, tx, txp1, Vgm, x, xdt, Rn, Rff, dts;
      real Vxp1, Vx, Vnff, tx, txp1, Vgm, x, xdt, Rn, Rff, dts, y1, y2;
         begin //start of function
            y1 = x - (Vxp1-Vnff)/Vgm*xdt; //case first cross border
            y2 = x - (Vxp1+Vx-2*Vnff)/Vgm*0.5*(txp1-tx)/dts*xdt; //other cases
         if ((tx==0)&&(y1<Rff)) begin //branches first crossing border
            dFrx = y1; //Resistance change for first step
         end
         else begin if ((tx!=0)&&(y2<Rff)) begin //branches other crossing border
            dFrx = y2; //Resistance change for other steps
         end else begin //case last cross border
            dFrx = Rff; //Resistance change for final step
         end
         end
         end //end of function
   endfunction

   analog function real dFC; //function control of switching
      input Vxp1, Vx, Vnff, tx, txp1, Vgm, x, xdt, Rn, Rff, dts, Cyc;
      real Vxp1, Vx, Vnff, tx, txp1, Vgm, x, xdt, Rn, Rff, dts, Cyc, y1, y2;
         begin //start of function
```

```
            y1 = x - (Vxp1-Vnff)/Vgm*xdt; //case first cross border
            y2 = x - (Vxp1+Vx-2*Vnff)/Vgm*0.5*(txp1-tx)/dts*xdt; //other cases
         if ((tx==0)&&(Rn<y1)) begin //branches first crossing border
            dFC = Cyc-x+y1; //Cycling parameter change for first step
         end
         else if ((tx!=0)&&(Rn<y2)) begin //branches other crossing border
            dFC = Cyc-x+y2; //Cycling parameter change for other step
         end else begin //case last cross border
            dFC = Cyc-x+Rn; //Cycling parameter for final step
         end
         end //end of function
   endfunction

   analog function real dFrC; //function control of switching
      input Vxp1, Vx, Vnff, tx, txp1, Vgm, x, xdt, Rn, Rff, dts, Cyc;
      real Vxp1, Vx, Vnff, tx, txp1, Vgm, x, xdt, Rn, Rff, dts, Cyc, y1, y2;
         begin //start of function
            y1 = x - (Vxp1-Vnff)/Vgm*xdt; //case first cross border
            y2 = x - (Vxp1+Vx-2*Vnff)/Vgm*0.5*(txp1-tx)/dts*xdt; //other cases
         if ((tx==0)&&(y1<Rff)) begin //branches first crossing border
            dFrC = Cyc-y1+x; //Cycling parameter change for first step
         end
         else if ((tx!=0)&&(y2<Rff)) begin //branches other crossing border
            dFrC = Cyc-y2+x; //Cycling parameter change for other step
         end else begin //case last cross border
            dFrC = Cyc-Rff+x; //Cycling parameter for final step
         end
         end //end of function
   endfunction

analog begin //start of description
   @ (initial_step)
      begin
         seed = pseed;
         Cyc = NumCyc * (Roff-Ron) * 2;
         x = Rstart;
         xdtn = (Roff-Ron) / dtsn / dvsn;
         xdtff = (Roff-Ron) / dtsff / dvsff;
         tx = 0;
         txp1 = 0;
         Vx = 0;
         Vxp1=0;
         RRon = $rdist_normal(seed, Ron, DltRon );
         RRoff = $rdist_normal(seed, Roff, DltRoff );
         RVon = $rdist_normal(seed, Von, DltVon );
```

```
            RVoff = $rdist_normal(seed, Voff, DltVoff );
        end

    if ((Cyc>0) && (V(t,b)>RVon) && (x>RRon)) begin //Vset process
        txp1=$abstime; //Normalizing time, initializing time point two
        Vxp1=V(t,b); //Initializing the point two of voltage
        x=dFx(Vxp1, Vx, RVon, tx, txp1, Vgrw, x, xdtn, RRon, RRoff, dtsc);
//Calculating resistance
        Cyc = dFC(Vxp1, Vx, RVon, tx, txp1, Vgrw, x, xdtn, RRon, RRoff, dtsc, Cyc);
//Calculating cycling resource
        tx=txp1; //Initializing time point one
        Vx=Vxp1; //Initializing voltage point one
    end
    else if ((Cyc>0) && (V(t,b)<RVoff) && (x<RRoff)) begin //Vreset process
        txp1=$abstime; //Normalizing time, initializing time point two
        Vxp1=V(t,b);  //Initializing the point two of voltage
        x=dFrx(Vxp1, Vx, RVoff, tx, txp1, Vmlt, x, xdtff, RRon, RRoff, dtsc);
//Calculating resistance
        Cyc = dFrC(Vxp1, Vx, RVoff, tx, txp1, Vmlt, x, xdtff, RRon, RRoff, dtsc, Cyc);
//Calculating cycling resource
        tx=txp1; //Initializing time point one
        Vx=Vxp1; //Initializing voltage point one
    end else begin
        tx=0; //Reset of point
        txp1=0; //Reset of point
        Vx=0; //Reset of point
        Vxp1=0; //Reset of point
        RRon = $rdist_normal(seed, Ron, DltRon );
        RRoff = $rdist_normal(seed, Roff, DltRoff );
        RVon = $rdist_normal(seed, Von, DltVon );
        RVoff = $rdist_normal(seed, Voff, DltVoff );
    end
    I(t,b) <+ V(t,b) / x; //Calculating current
    end //end of description

endmodule
```

ПРИЛОЖЕНИЕ № 2. Графики результатов моделирования.

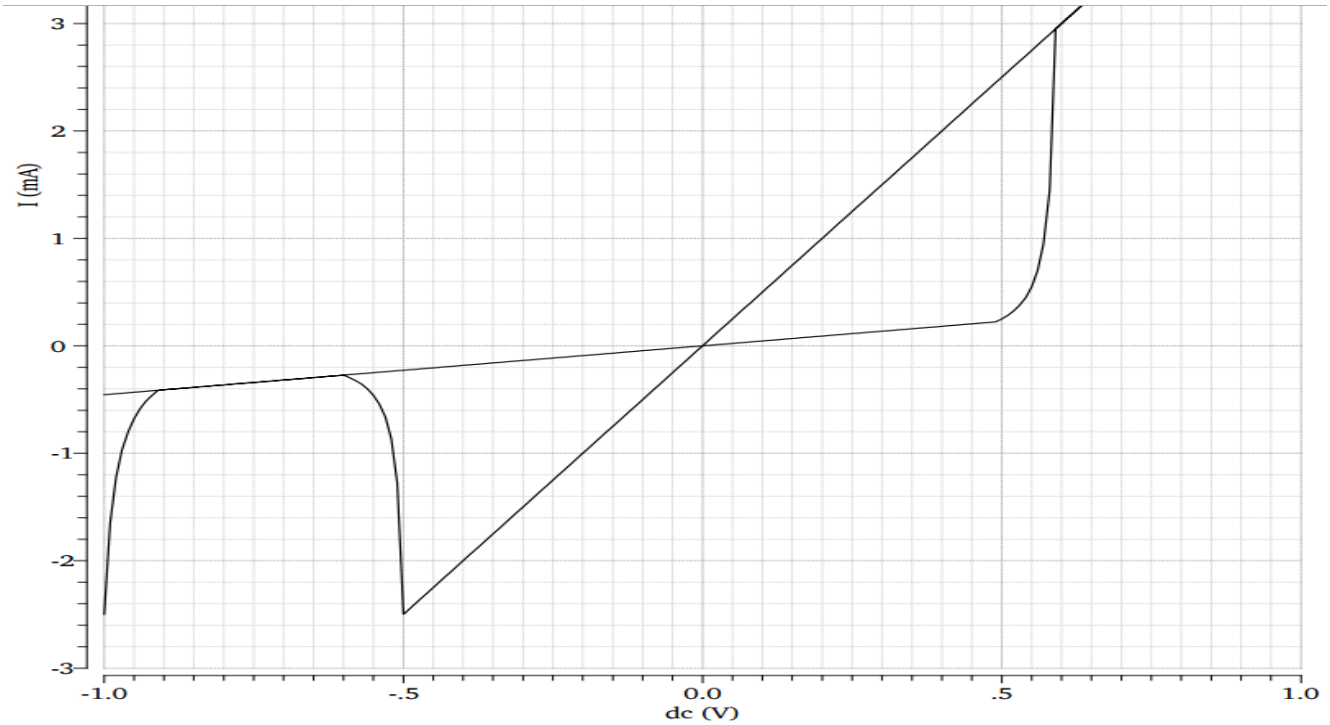


Рисунок 1. ВАХ мемристора базовой модели. Начальное состояние низкорезистивное. Слева на ВАХ виден переход из начального состояния в высокорезистивное. Von= Voff 0.5В, Ron=200 Ом,Roff=2200 Ом.

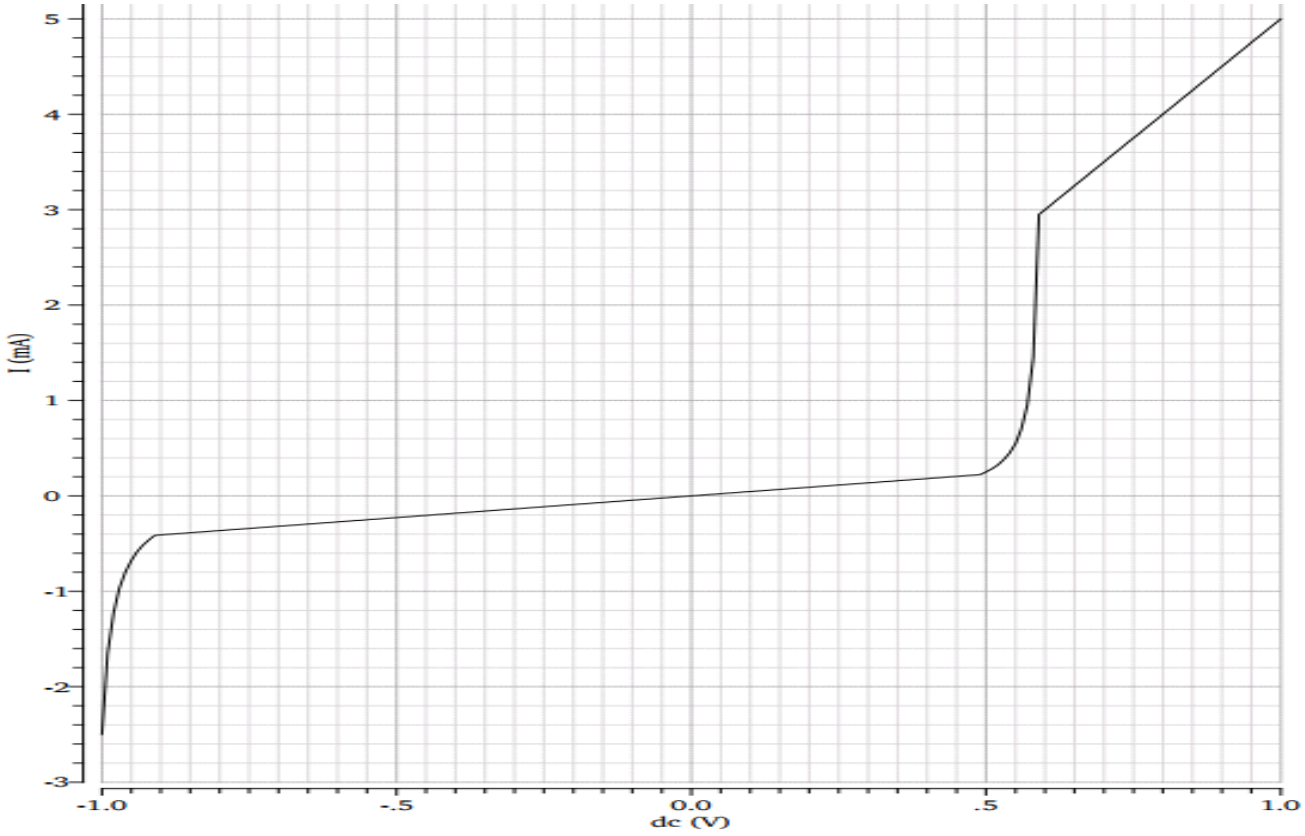


Рисунок 2. ВАХ мемристора базовой модели. Начальное состояние низкорезистивное. DC анализ от -1В до 1 В

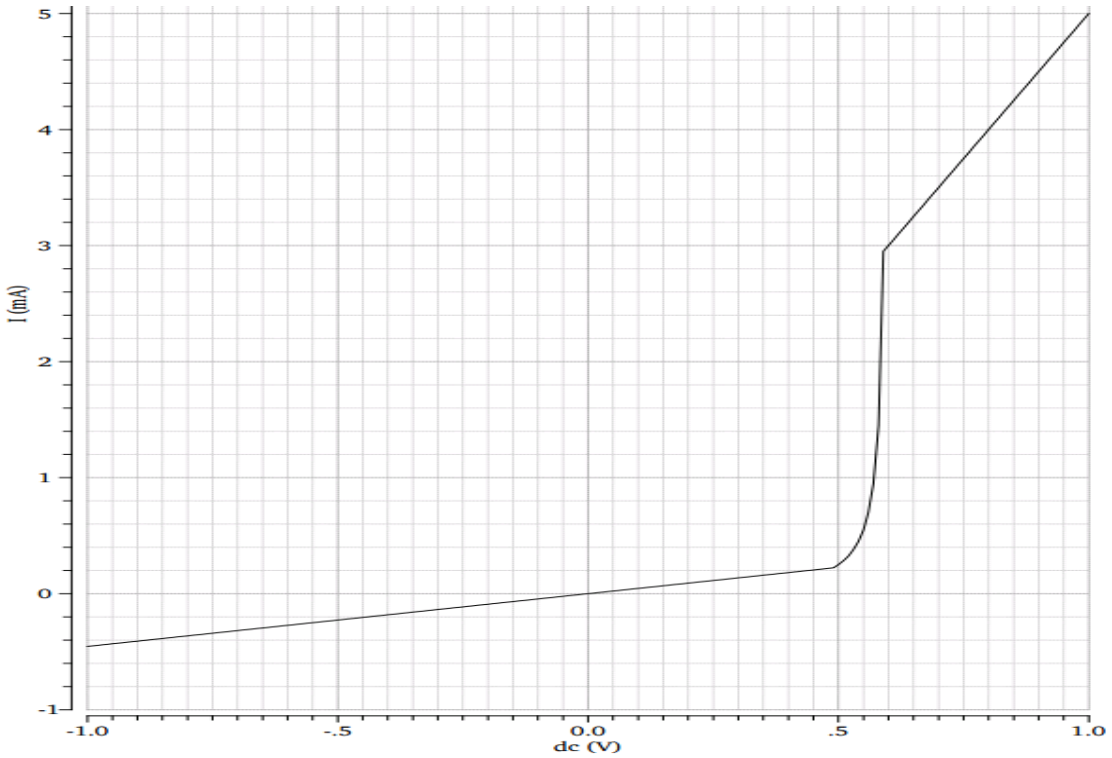


Рисунок 3. ВАХ мемристора базовой модели. Начальное состояние высокорезистивное. DC анализ от -1В до 1 В

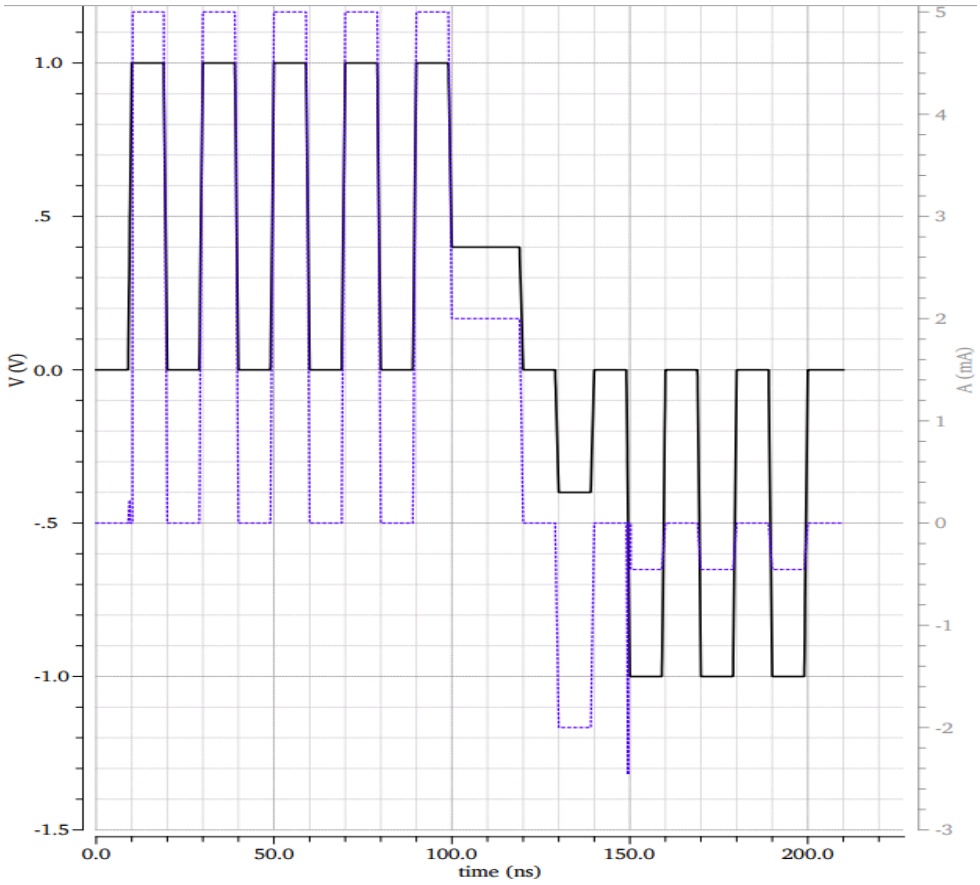


Рисунок 4. Циклирование мемристора, в режиме мгновенного переключения, прямоугольными импульсами. Уровень входного сигнала в вольтах представлен черной линией, выходной ток синей пунктирной линией.

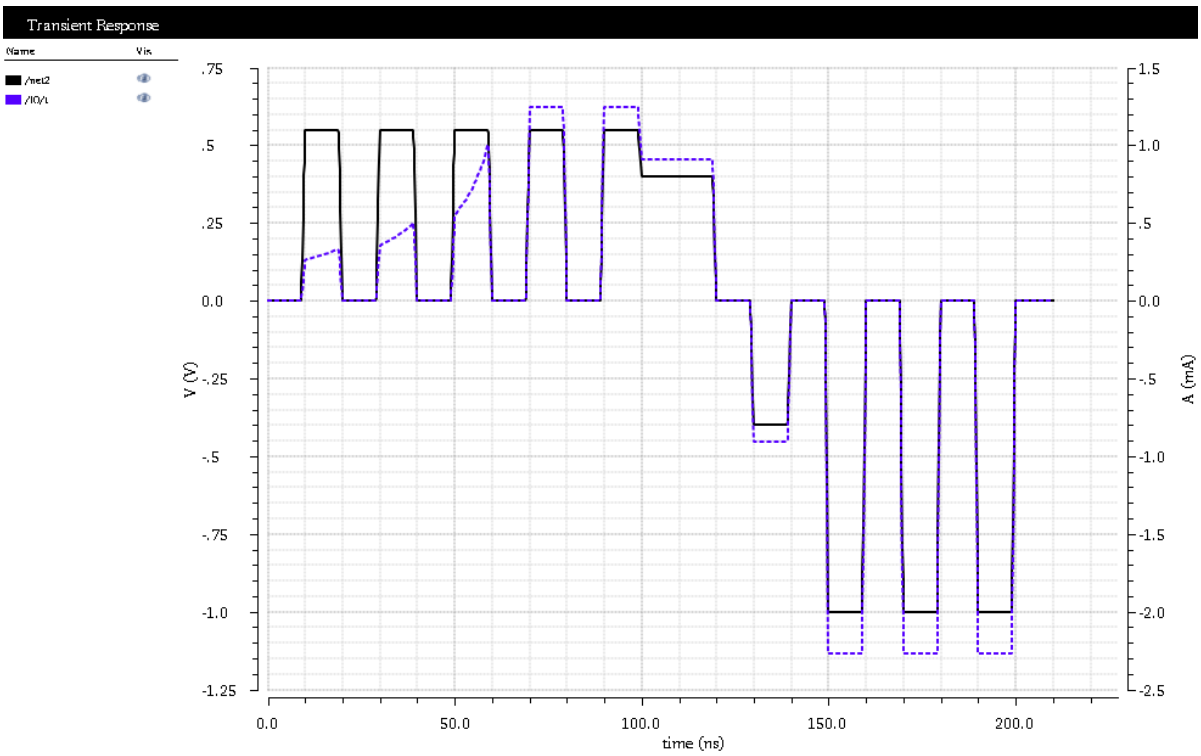


Рисунок 5.Циклирование мемристора. Множественность состояний получаемых путем задания коротких небольших по амплитуде импульсов. Ресурс переключений 0.5 цикла. Уровень входного сигнала в вольтах представлен черной линией, выходной ток синей пунктирной линией.

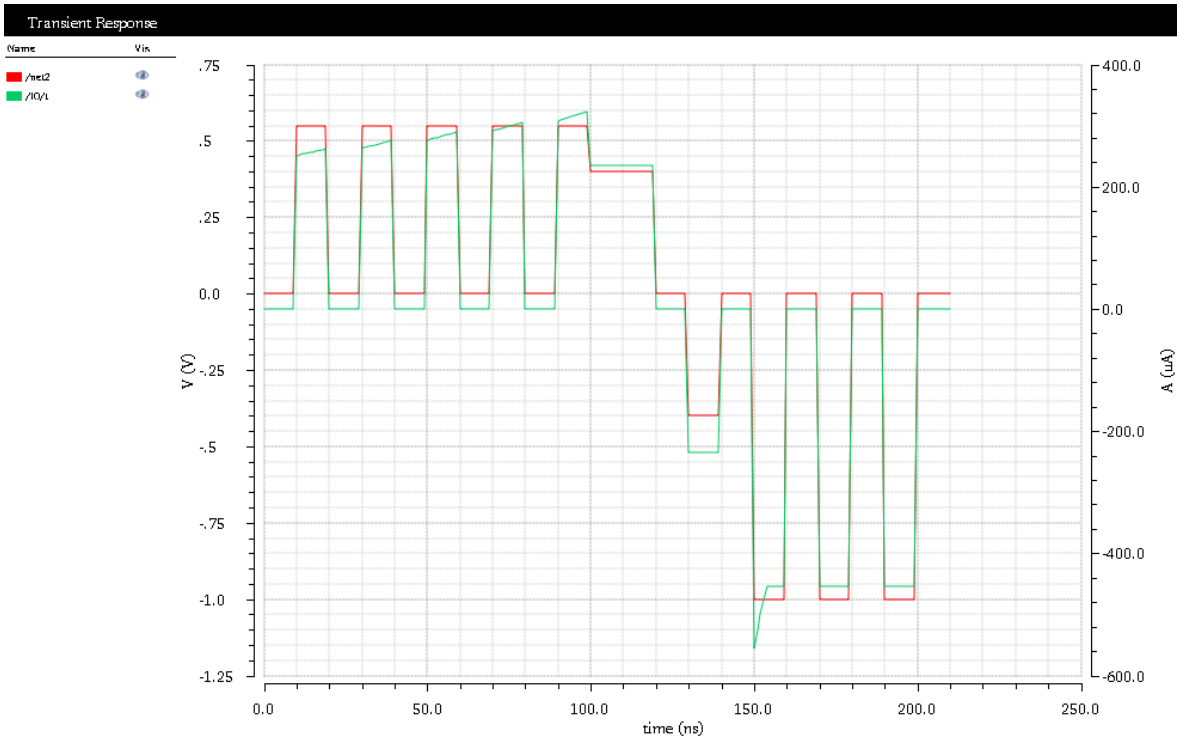


Рисунок 6.Циклирование мемристора. Множественность состояний получаемых путем задания коротких небольших по амплитуде импульсов. Частичное переключение. Уровень входного сигнала в вольтах представлен красной линией, выходной ток другой линией.

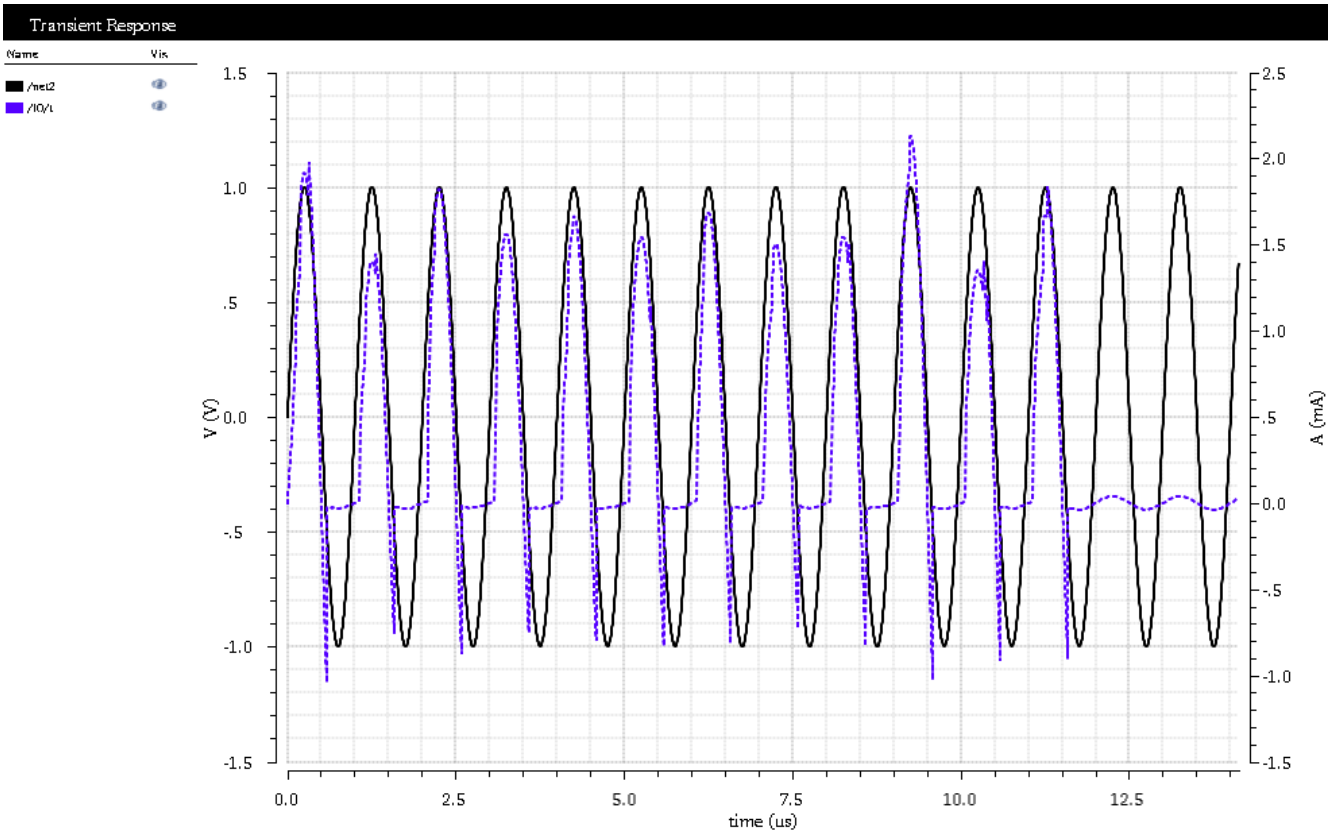


Рисунок 7.Циклирование мемристора. Множественность состояний исчерпание ресурса переключений в состоянии низкой проводимости. Уровень входного сигнала в вольтах представлен черной линией, выходной ток другой линией.

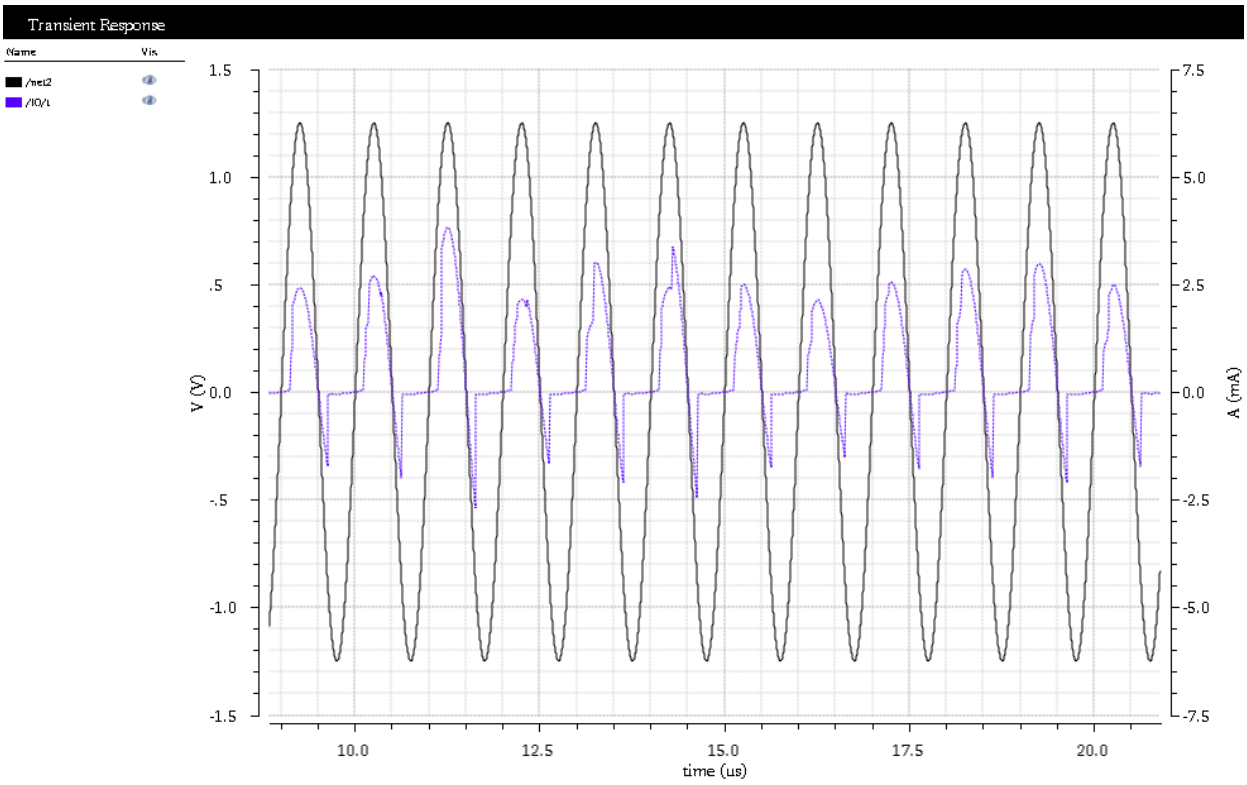


Рисунок 8.Циклирование мемристора. Множественность состояний разброс параметров состояний высокой проводимости. Уровень входного сигнала в вольтах представлен черной линией, выходной ток другой линией.

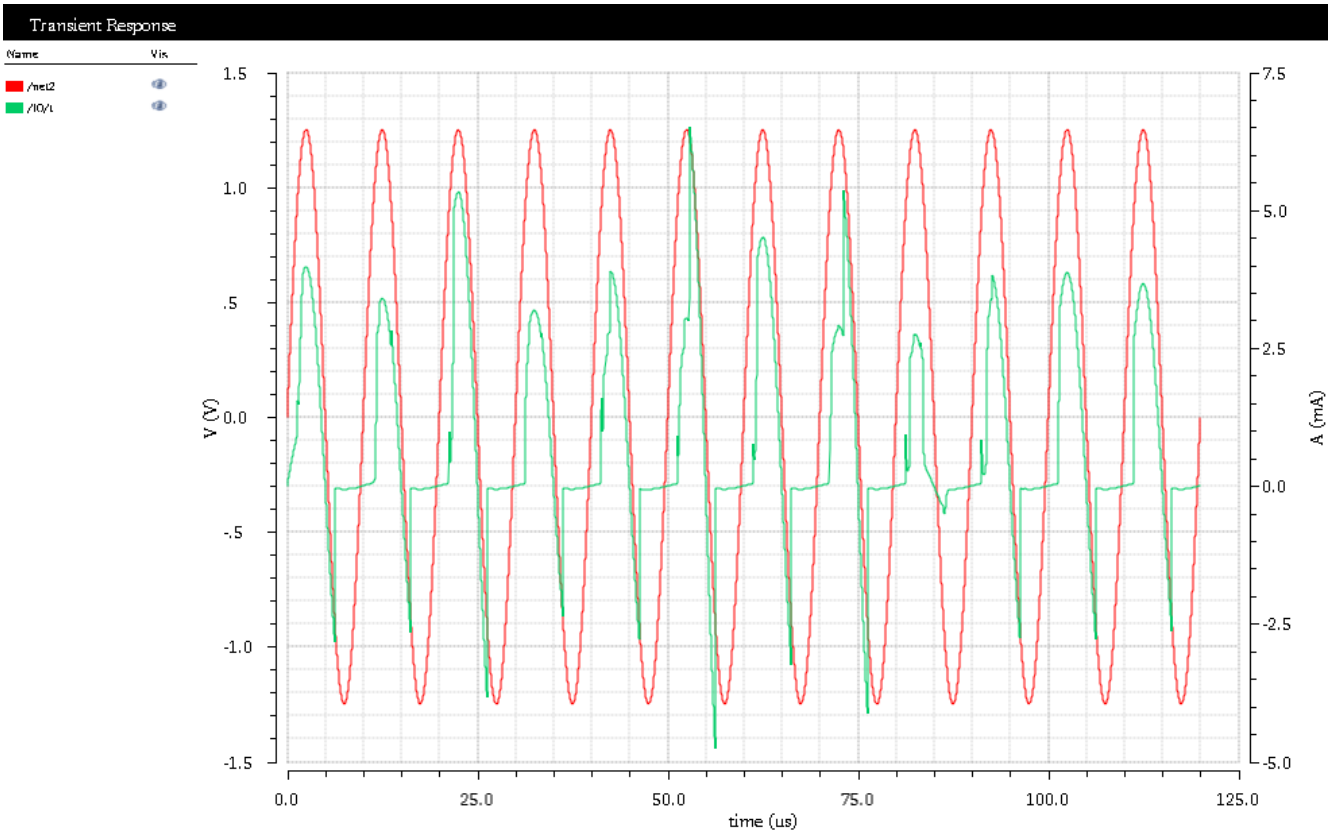


Рисунок 9.Циклирование мемристора. Множественность состояний разброс параметров состояний высокой проводимости. Уровень входного сигнала в вольтах представлен красной линией, выходной ток другой линией.