УДК 004.383.8.032.26

Современное состояние исследований в области вычислительных систем на базе мемристивных устройств с распределенной архитектурой

Никишов Д.А., Данилин С.Н., Щаников С.А.

В статье представлен аналитический обзор научно-технических публикаций о состоянии исследований в области инженерного проектирования, производства и применения вычислительных систем с распределенной архитектурой нового поколения на базе мемристивных устройств. Показаны достижения наиболее известных научных коллективов, и актуальные проблемы, замедляющие темп развития данной области науки и техники.

Ключевые слова: мемристоры, искусственные нейронные сети, распределенные системы, масштабирование.

Введение

Масштабные работы мировых лидеров в области разработки и производства вычислительной техники с программной эмуляцией искусственных нейронных сетей (ИНС) на базе CPU (Central Processing Unit), GPU (Graphics Processing Unit), TPU (Tensor Processing Unit), FPGA (Field-Programmable Gate Array) достигли значительных успехов. Однако, по основным техническим характеристикам, прежде всего по энергоэффективности, этот способ реализации ИНС приблизился к своему теоретическому пределу. По указанной причине активно развиваются и средства непосредственной аппаратной реализации принципов функционирования ИНС, как на цифровой (специализированные нейроморфные процессоры), так и аналоговой элементной базе. Наиболее перспективной элементной базой в настоящее время являются наноразмерные мемристоры. Работы в данной области достигли уровня распределенных систем. Состояния исследований в данном направлении представлено ниже.

«Атомарным» компонентом любой распределенной масштабируемой системы является микропроцессор. Именно микропроцессор выполняет те вычислительные задачи, которые на него возлагает устройство управления всей распределенной системой. Мемристоры позволяют создавать специализированные микропроцессоры для выполнения векторно-матричного умножения, являясь таким же «атомарным» компонентом для специализированных вычислительных систем, работающих на основе принципов ИНС - нейрокомпьютеров и супернейрокомпьютеров [1, 2]. ИНС очень требовательны к вычислительной мощности средств их реализации. Для того, чтобы обрабатывать информацию быстрее, чем компьютеры на базе CPU, необходимо применять графические процессоры с несколькими ядрами или тензорные процессоры – GPU и TPU. По сравнению с CPU, они позволяют более эффективно распараллеливать вычисления, что значительно ускоряет работу глубоких ИНС. TPU – это процессоры, в основе которых лежат аппаратные ускорители матричного умножения, называемые тензорами [3]. Названная операция наиболее часто повторяется в ИНС (умножения вектора входных данных нейрона на матрицу весовых коэффициентов синапсов). TPU процессоры имеют близкие к GPU показатели производительности [4]. Реконфигурируемые FPGA являются хорошей альтернативой рассмотренным процессорам, для ускорения моделей DNN (Deep Neural Network) [5]. Они позволяют обеспечить параллелизм выполнения операций за счет системы вентилей. FPGA предназначены для обработки нерегулярного параллелизма и нестандартных типов информации, что хорошо вписывается в современные тенденции развития CNN (Convolutional Neural Network) [6]. Аппаратная реализация ИНС в настоящее время является очень перспективным направлением, так как она позволяет с наибольшей точностью реализовать работу биологического аналога. Кроме того, эти системы обладают низким потреблением

энергии и высокой скоростью обработки информации

Как было сказано выше, при аппаратной реализации ИНС в качестве базовых компонентов наилучшим образом подходят мемристоры – новые элементы, которые естественным образом реализуют функции синапсов нейронов ИНС различного назначения. Мемристор – энергонезависимый переменный резистор. Он может изменять свое сопротивление и сохранять его продолжительное время без затрат энергии. Программирование мемристора осуществляется напряжением или током. Запрограммированная информация сохраняется до тех пор, пока не будет применено следующее входное напряжение или ток [7].

Создание и применение в вычислительных системах нового поколения мемристивных устройств с распределённой архитектурой позволит значительно ускорить обработку информации и вывести выполнение функций искусственного интеллекта техническими средствами на значительно более высокий уровень [8, 9]

Архитектуры распределенных и масштабируемых вычислительных систем

Распределенной является вычислительная система, компоненты которой расположены на разных компьютерах, объединенных в сеть, которые общаются и координируют свои действия, передавая друг другу необходимые сообщения для достижения общей цели. Для нормальной работы распределенных систем необходимо с достаточной точностью решать три сложные задачи: поддержание параллелизма компонентов, преодоление отсутствия глобальных часов и управление независимым отказом компонентов. Когда компонент системы выходит из строя, вся система не должна выходить из строя [10].

Компьютерная программа, которая выполняется в распределенной системе, называется распределенной программой, а распределенное программирование – это процесс написания таких программ. Существует множество различных типов реализации механизма передачи сообщений, включая чистый HTTP (HyperText Transfer Protocol), RPC-подобные (Remote Procedure Call) соединители и очереди сообщений [11].

Распределенные вычисления относятся к использованию распределенных систем для решения вычислительных задач. В распределенных вычислениях проблема делится на множество задач, каждая из которых решается одним или несколькими компьютерами, которые взаимодействуют друг с другом посредством передачи сообщений. Помимо понятия «распределённые вычисления» существует понятие «параллельные вычисления». Эти понятия схожи, одна и та же система может характеризоваться как «параллельная» так и «распределённая». Параллельные вычисления можно рассматривать как особую форму распределенных вычислений, а распределенные вычисления можно рассматривать как слабо связанную форму параллельных вычислений. Однако разница в них есть. Так, при параллельных вычислениях, все процессоры могут иметь доступ к общей памяти для обмена информацией между процессорами. В распределенных вычислениях каждый процессор имеет свою собственную частную память (распределенную память). Обмен информацией происходит путем передачи сообщений между процессорами.

Основные характеристики распределенных систем

А. Масштабируемость – способность системы, процесса или сети расти и справляться с возрастающим спросом на обработку информации. Любая распределенная система, которая может постоянно развиваться, чтобы поддерживать растущий объем работы, является масштабируемой.

Масштабирование системы может потребоваться по многим причинам, таким как увеличение объема информации или увеличение объема работы, например, количества транзакций. Для пользователей желательно осуществление масштабирования без потери

производительности системы, процесса или сети.

Как правило, производительность системы, разработанной как масштабируемая, снижается с увеличением размера системы из-за затрат времени на управление и на прохождение сигналов между компьютерами по каналам связи. В более общем случае некоторые вычислительные задачи не могут быть распределены, либо из-за присущей им атомарной природы, либо из-за какого-то технического ограничения в конструкции системы. В какой-то момент такие задачи ограничат рост скорости вычислений, достигаемое за счет их распределения по компонентам. Масштабируемая архитектура системы, процесса или сети позволяет избежать такой ситуации за счет равномерно распределяемой нагрузки на все участвующие в работе компоненты [12].

Существует два основных вида масштабирования – горизонтальное и вертикальное. Горизонтальное масштабирование производится путем добавления большего количества серверов в пул ресурсов. Вертикальное масштабирование производится путем добавления большей мощности к существующему серверу. На практике более эффективным является первое из них [12].

По определению, приведенному в стандарте [13], надежность – это свойство объекта сохранять во времени способность выполнять требуемые функции в заданных режимах и условиях применения, технического обслуживания, хранения и транспортирования. Основные стандартные показатели надежности: средняя наработка до отказа $T_{ср}$ и (или) вероятность безотказной работы $P(t)$ [14].

В ряде зарубежных публикаций в качестве характеристики надежности распределенных вычислительных систем применяют ее отказоустойчивость: «распределенная система считается надежной, если она продолжает предоставлять свои услуги даже при отказе одного или нескольких программных, или аппаратных компонентов» [15]. Надежность представляет собой одну из основных характеристик любой распределенной вычислительной системы [16].

B. Доступность – это время, в течение которого система остается работоспособной для выполнения требуемой функции в определенный период. Это мера процента времени, в течение которого система, услуга или машина остается работоспособной в нормальных условиях [9] (по стандарту РФ [13] аналог доступности - гамма процентная наработка между отказами).

Надежность – это доступность в течение долгого времени с учетом всего спектра возможных реальных условий, которые могут возникнуть [9].

C. Надежность в сравнении с доступностью. Если система надежна, то она доступна. Однако, если система доступна, она не обязательно надежна. Другими словами, высокая надежность способствует высокой доступности, но можно достичь высокой доступности даже с ненадежным объектом, минимизировав время ремонта (повысив ремонтопригодность [13]) и обеспечив постоянную доступность запасных частей, когда они необходимы [10].

D. Эффективность. Двумя стандартными показателями эффективности распределенной информационной системы являются время отклика (или латентность), обозначающее задержку получения на выходе первого информационного элемента, и пропускная способность, обозначающая количество информационных элементов, выданных системой за единицу времени. Эти два показателя соответствуют следующим удельным затратам:

1. Количество сообщений, отправленных узлами системы, независимо от размера сообщения.

2. Размер сообщений, представляющих объем обмена информации.

Сложность операций, поддерживаемых распределенными структурами (например, поиск определенного ключа в распределенном индексе), может быть охарактеризована как функция одной из рассмотренных единиц

затрат. По мнению ряда исследователей, анализ распределенной структуры с точки зрения «количества сообщений» является слишком упрощенным. Он игнорирует влияние многих существенных факторов: топологию сети, нагрузку на сеть и ее вариации, возможную неоднородность программных и аппаратных компонентов, участвующих в обработке и маршрутизации информации и т.д. Однако до настоящего времени не разработана адекватная системная модель затрат, которая бы учитывала все факторы производительности. Поэтому на практике применяются достаточно грубые оценки поведения распределенных систем [17].

Е. Удобство обслуживания или управляемость. Еще одним важным показателем при проектировании распределенной системы является простота ее эксплуатации и обслуживания. Удобство обслуживания или управляемость - это простота и скорость, с которой систему можно отремонтировать или обслужить; если время на исправление отказавшей системы увеличивается, то доступность снижается. Для управляемости необходимо учитывать простоту диагностики и понимания проблем при их возникновении, простоту внесения обновлений или изменений, а также простоту эксплуатации системы (т.е. работает ли она регулярно без сбоев и исключений). Аналог управляемости по стандарту РФ [13] является ремонтопригодность. Свойство объекта, заключающееся в его приспособленности к поддержанию и восстановлению состояния, в котором объект способен выполнять требуемые функции, путем технического обслуживания и ремонта.

Раннее обнаружение неисправностей может уменьшить или избежать простоя системы. Например, некоторые корпоративные системы могут автоматически вызывать сервисный центр (без вмешательства человека), когда в системе возникает системный сбой.

Так как распределённая система оперирует несколькими процессорами, которые в свою очередь включают в себя несколько ядер, а задачи могут разбиваться на отдельные потоки, необходимо уточнить термин многопроцессорность [18].

Многопроцессорность – это использование двух или более центральных процессоров (ЦП) в одной компьютерной системе [19, 20]. Этот термин также относится к способности системы поддерживать более одного процессора или возможности распределять задачи между ними. На уровне операционной системы многопроцессорность иногда используется для обозначения выполнения нескольких параллельных процессов в системе, при этом каждый процесс выполняется на отдельном процессоре или ядре, в отличие от одного процесса в любой момент времени. При использовании этого определения многопроцессорность иногда противопоставляется многозадачности, которая может использовать только один процессор, но переключать его на временные отрезки между задачами (т.е. система с разделением времени). Многопроцессорностью является реальное параллельное выполнение нескольких информационных процессов с использованием более чем одного процессора. Многопроцессорность не обязательно означает, что один процесс или задача использует более одного процессора одновременно; для обозначения такого сценария обычно используется термин параллельная обработка [21].

В таксономии Флинна мультипроцессоры, как определено выше, являются MIMD-машинами (multiple instruction stream / multiple data stream) [22, 23]. Поскольку термин «мультипроцессор» обычно относится к тесно связанным системам, в которых все процессоры совместно используют память, мультипроцессоры не являются всем классом MIMD-машин, который также содержит мультикомпьютерные системы с передачей сообщений [21]. Все категории Флинна представлены в таблице 1.

Многопроцессорные системы классифицируются в зависимости от того, как обрабатывается доступ к памяти процессора и являются ли системные процессоры одного типа или разных.

В слабосвязанных многопроцессорных системах каждый процессор имеет собственную локальную память, каналы ввода/вывода (I/O) и операционную систему. Процессоры обмениваются данными через высокоскоростную коммуникационную сеть, посылая сообщения с помощью техники, известной как «передача сообщений». Слабосвязанные многопроцессорные системы также известны как системы с распределенной памятью, поскольку процессоры не разделяют физическую память и имеют индивидуальные каналы ввода/вывода [25].

Таблица 1 - Категории Флинна [24]

SISD	(single instruction stream / single data stream) - одиночный поток команд и одиночный поток данных. К этому классу относятся, прежде всего, классические последовательные машины, или иначе, машины фон-неймановского типа, например, PDP-11 или VAX 11/780. В таких машинах есть только один поток команд, все команды обрабатываются последовательно друг за другом и каждая команда инициирует одну операцию с одним потоком данных. Не имеет значения тот факт, что для увеличения скорости обработки команд и скорости выполнения арифметических операций может применяться конвейерная обработка - как машина CDC 6600 со скалярными функциональными устройствами, так и CDC 7600 с конвейерными попадают в этот класс.
SIMD	(single instruction stream / multiple data stream) - одиночный поток команд и множественный поток данных. В архитектурах подобного рода сохраняется один поток команд, включающий, в отличие от предыдущего класса, векторные команды. Это позволяет выполнять одну арифметическую операцию сразу над многими данными - элементами вектора. Способ выполнения векторных операций не оговаривается, поэтому обработка элементов вектора может производится либо процессорной матрицей, как в ILLIAC IV, либо с помощью конвейера, как, например, в машине CRAY-1.
MISD	(multiple instruction stream / single data stream) - множественный поток команд и одиночный поток данных. Определение подразумевает наличие в архитектуре многих процессоров, обрабатывающих один и тот же поток данных. Однако ни Флинн, ни другие специалисты в области архитектуры компьютеров до сих пор не смогли представить убедительный пример реально существующей вычислительной системы, построенной на данном принципе. Ряд исследователей относят конвейерные машины к данному классу, однако это не нашло окончательного признания в научном сообществе. Будем считать, что пока данный класс пуст.
MIMD	(multiple instruction stream / multiple data stream) - множественный поток команд и множественный поток данных. Этот класс предполагает, что в вычислительной системе есть несколько устройств обработки команд, объединенных в единый комплекс и работающих каждое со своим потоком команд и данных

Симметричные многопроцессорные системы - системы, работающие под управлением одной ОС (операционной системы) с двумя или более однородными процессорами и с централизованной общей оперативной памятью [26].

Симметричная многопроцессорная система (SMP) - это система с пулом однородных процессоров, работающих под управлением одной ОС с централизованной общей основной памятью. Каждый процессор, выполняя различные программы и работая с различными наборами данных, имеет возможность совместно использовать общие ресурсы (память, устройство ввода/вывода, систему прерываний и так далее), которые соединены системной шиной, кросс-шиной, или их смесью, или шиной адреса и кросс-шиной данных [27].

Каждый процессор имеет свою собственную кэш-память, которая действует как мост между процессором и основной памятью. Функция кэш-памяти заключается в том, чтобы облегчить доступ к данным основной памяти, тем самым снижая нагрузку на системную шину [28].

Известно, что SMP-система имеет ограниченную масштабируемость. Для преодоления этого ограничения обычно используется архитектура под названием «cc-NUMA» (cache coherency-non-uniform memory access). Основной характеристикой системы cc-NUMA является наличие общей глобальной памяти, распределенной между всеми узлами, хотя

эффективный «доступ» процессора к памяти удаленной компонентной подсистемы, или «узла», медленнее по сравнению с доступом к локальной памяти, поэтому доступ к памяти является «неравномерным» [29].

Система cc-NUMA представляет собой кластер SMP-систем - каждая из которых называется «узел», который может иметь один процессор, многоядерный процессор или их смесь, одной или другой архитектуры - соединенных через высокоскоростную «сеть подключения», которая может быть «каналом», который может быть одинарным или двойным обратным кольцом, или несколькими кольцами, соединениями «точка-точка», или их смесью , шинное соединение, «перекрестная шина», «сегментированная шина», «ячеистый маршрутизатор» и т. д [30].

Разница во времени доступа между локальной и удаленной памятью может быть также на порядок больше, в зависимости от типа используемой сети подключения (быстрее при сегментированной шине, перекрестной шине и соединении «точка-точка»; медленнее при последовательном кольцевом соединении) [31].

Суперкомпьютер – это компьютер с высоким уровнем производительности по сравнению с компьютером общего назначения. Производительность суперкомпьютера принято измерять в операциях с плавающей запятой в секунду (FLOPS), а не в миллионах инструкций в секунду (MIPS). С 2017 года существуют суперкомпьютеры, производительность которых превышает 10^{17} FLOPS (сто квадриллионов FLOPS, 100 петаFLOPS или 100 PFLOPS) [32].

Для сравнения, производительность настольного компьютера составляет от сотен гигафлопс до десятков терафлопс [33, 34].

С ноября 2017 года все 500 самых быстрых суперкомпьютеров в мире работают на операционных системах на базе Linux [35]. В США, Европейском союзе, Тайване, Японии и Китае проводятся дополнительные исследования для создания более быстрых, мощных и технологически совершенных экзафлопсных суперкомпьютеров [36].

Суперкомпьютеры играют важную роль в области вычислительной науки и используются для решения широкого спектра задач с интенсивными вычислениями в различных областях, включая квантовую механику, прогнозирование погоды, исследование климата, разведку нефти и газа, молекулярное моделирование (вычисление структур и свойств химических соединений, биологических макромолекул, полимеров и кристаллов) и физическое моделирование (например, моделирование ранних моментов Вселенной, аэродинамики самолетов и космических кораблей, детонации ядерного оружия и ядерного синтеза). Они сыграли важную роль в области криптоанализа [37].

Суперкомпьютеры появились в 1960-х годах, и в течение нескольких десятилетий самые быстрые из них были созданы Сеймуром Крэем в компаниях Control Data Corporation (CDC), Cray Research и последующих компаниях, носящих его имя или монограмму. Первые такие машины представляли собой обычные конструкции, которые работали быстрее, чем их современники общего назначения. В течение десятилетия добавлялось все большее количество параллельных вычислителей, причем типичными были от одного до четырех процессоров. В 1970-х годах стали преобладать векторные процессоры, работающие с большими массивами данных. Ярким примером является успешный Cray-1 1976 года. Векторные компьютеры оставались доминирующей конструкцией до 1990-х годов. С тех пор и до сегодняшнего дня массово-параллельные суперкомпьютеры с десятками тысяч готовых процессоров стали нормой [38, 39].

США долгое время были лидером в области суперкомпьютеров, сначала благодаря почти непрерывному доминированию компании Cray, а затем благодаря целому ряду технологических компаний. Япония добилась значительных успехов в этой области в 1980-х и 90-х годах, а Китай становится все более активным в этой области. По состоянию на май 2022 года самым быстрым суперкомпьютером в списке суперкомпьютеров TOP500 является американский Frontier с результатом

1,102 ExaFlop/s в бенчмарке LINPACK, за ним следует Fugaku. [40] У США пять из 10 лучших; у Китая - два; у Японии, Финляндии и Франции - по одному. [41] В июне 2018 года все суперкомпьютеры, входящие в список TOP500, преодолели отметку в 1 exaFLOPS. [42]

Состояние исследований мемристивных вычислительных систем за рубежом

За рубежом наибольших успехов в разработке многоядерных вычислителей на базе мемристивных устройств добились ученые из Университета Стенфорда (США) и Университета Цинхуа (Китай), разработавшие и продемонстрировавшие архитектуру NeurRAM – 48-ми ядерного устройства с распределением вычислений на чипах, основанных на резистивной памяти с произвольным доступом (RRAM). Далее рассмотрим публикации членов данного коллектива, в которых изложены основные достижения и решенные задачи на пути к созданию чипа NeurRAM.

В статье [43] изложены последние достижения в области бинарной металлоксидной резистивной переключаемой памяти с произвольным доступом. Обсуждается физический механизм, свойства материалов и электрические характеристики различных бинарных металлоксидных RRAM с акцентом на использование RRAM для энергонезависимой памяти. Приводится обзор последних разработок крупномасштабных массивов RRAM. Обсуждаются такие вопросы, как однородность, долговечность, сохранение, многоразрядная работа и тенденции масштабирования.

В статье [44] дан обзор последних достижений в области памяти с фазовыми изменениями (PCM). Рассмотрены электрические и тепловые свойства материалов с фазовыми изменениями с акцентом на масштабируемость материалов и их влияние на дизайн устройств. Описываются инновации в структуре устройства, селектор ячеек памяти и стратегии для достижения многоразрядной работы и трехмерных многослойных массивов памяти высокой плотности. Масштабирующие свойства PCM иллюстрируются последними экспериментальными результатами с использованием специальных тестовых структур устройства и нового синтеза материала. Обсуждаются факторы, влияющие на надежность PCM.

В статье [45] представлено текущее состояние понимания факторов, ограничивающих дальнейшее масштабирование технологии Si комплементарных металлоксид-полупроводниковых транзисторов (КМОП), и проведен анализ того, как связанные с применением факторы влияют на определение этих пределов. Физические истоки этих пределов лежат в основном в туннельных токах, просачивающихся через различные барьеры в МОП полевом транзисторе (МОП-транзисторе), когда он становится очень маленьким, и в термически генерируемых подпороговых токах. Обсуждается зависимость этих утечек от геометрии и структуры МОП-транзистора, а также критерии проектирования для минимизации короткоканальных эффектов и другие вопросы, связанные с масштабированием.

В работе [46] показано, что нейроморфные вычисления – это развивающаяся область, цель которой – расширить возможности информационных технологий за пределы цифровой логики. В качестве строительного блока для нейроморфных вычислительных систем необходимо компактное наноразмерное устройство, эмулирующее биологические синапсы.

Авторы работы [47] дали анализ достижений в области синаптической электроники. Приведены основы биологической синаптической пластичности и обучения. Обсуждены свойства материалов и электрические характеристики переключения различных синаптических устройств, с акцентом на использование синаптических устройств для нейроморфных вычислений. Показаны метрики производительности, желательные для крупномасштабных реализаций синаптических устройств. Представлен обзор последних работ по целевым вычислительным приложениям с использованием синаптических устройств.

В статье [48] рассмотрены новые устройства энергонезависимой памяти, которые хранят информацию, используя физические механизмы, отличные от тех, которые имеют место в современных запоминающих устройствах, и могут обеспечить существенное повышение производительности вычислений и энергоэффективности.

В работе [49] говорится о том, что современные аппаратные платформы потребляют огромное количество энергии для когнитивного обучения из-за перемещения информации между процессором и внечиповой памятью. Технологии устройств на базе ИНС, с использованием аналоговой весовой памяти позволяют выполнять когнитивные задачи более эффективно. Авторы представляют аналоговую энергонезависимую резистивную память (электронный синапс) с использованием материалов, дружественных к современному производству микросхем. Устройство демонстрирует двунаправленную непрерывную модуляцию веса. Экспериментально продемонстрирована классификация лиц в черно белых тонах с использованием интегрированного массива из 1024 ячеек с параллельным онлайн-обучением. Энергопотребление аналоговых синапсов на каждой итерации на 1 000 × (20 × 20) меньше по сравнению с реализацией на процессоре Intel Xeon Phi с внечиповой памятью (с гипотетической цифровой резистивной памятью с произвольным доступом на кристалле). Точность на тестовых наборах близка к результату при использовании центрального процессора. Эти экспериментальные результаты подтверждают осуществимость аналогового синаптического массива и прокладывают путь к созданию энергоэффективной и крупномасштабной нейроморфной системы.

В работе [50] показано, что аппаратная реализация импульсных нейронов может быть чрезвычайно полезной для большого числа приложений, начиная от высокоскоростного моделирования крупномасштабных нейронных систем и заканчивая системами поведения в реальном времени и двунаправленными интерфейсами «мозг-машина». Конкретные схемотехнические решения, используемые для реализации кремниевых нейронов, зависят от требований приложения. В этой статье описываются наиболее распространенные строительные блоки и методы, используемые для реализации этих схем, и представляем обзор широкого спектра нейроморфных кремниевых нейронов, которые реализуют различные вычислительные модели, в частности биофизически реалистичных и основанных на проводимости моделей Ходжкина-Хаксли. Авторы сравнивают различные методологии проектирования, используемые для каждой описанной конструкции кремниевого нейрона, и демонстрируют их особенности с помощью экспериментальных результатов, измеренных на широком спектре изготовленных микросхем VLSI (Very Large-Scale Integration).

В статье [51] показано, что ограниченные машины Больцмана (RBM) и глубокие сети продемонстрировали свою эффективность в различных приложениях, таких как уменьшение размерности, обучение признаков и классификация. Их реализация на нейроморфных аппаратных платформах, эмулирующих крупномасштабные сети спайковых нейронов, может иметь значительные преимущества с точки зрения масштабируемости, рассеиваемой мощности и взаимодействия с окружающей средой в реальном времени. Однако традиционная архитектура RBM и широко используемый алгоритм обучения, известный как Contrastive Divergence (CD), основаны на дискретных обновлениях и точной арифметике, которые не могут быть напрямую отображены на динамическую нейронную подложку. Авторы представили событийно-управляемую вариацию CD для обучения RBM, построенного с нейронами Integrate & Fire (I&F), которая ограничена свойствами существующих и перспективных нейроморфных аппаратных платформ. Стратегия основана на нейронной выборке, что позволяет синтезировать нейронную спайковую сеть, которая делает выборку из целевого распределения Больцмана. Рекуррентная активность сети заменяет дискретные шаги алгоритма

CD, в то время как Spike Time Dependent Plasticity (STDP) осуществляет обновление веса в режиме онлайн, асинхронно. Авторы демонстрируют подход, обучая RBM, состоящую из нейронов с утечкой I&F (leakage integrate and fire) с синапсами STDP, для обучения генеративной модели набора данных рукописных цифр MNIST, и тестируя ее в задачах распознавания, генерации и интеграции подсказок.

В работе [52] авторы изложили перспективы развития нейроморфной инженерии. Нейроморфная инженерия (NE) охватывает широкий спектр подходов к обработке информации, которые основаны на процессах в нейробиологических системах. Мозг эволюционировал в течение миллиардов лет, чтобы решать сложные инженерные задачи с помощью эффективных, параллельных, маломощных вычислений. Целью НЭ является разработка систем, способных выполнять вычисления, подобные мозгу. В последнее время появилось множество крупномасштабных нейроморфных проектов. Эта междисциплинарная область вошла в список 10 лучших технологических прорывов 2014 года по версии MIT Technology Review и в список 10 лучших развивающихся технологий 2015 года по версии Всемирного экономического форума. НЭ имеет двусторонние цели: во-первых, научная цель – понять вычислительные свойства биологических нейронных систем с помощью моделей, реализованных в интегральных схемах (ИС); во-вторых, инженерная цель – использовать известные свойства биологических систем для проектирования и реализации эффективных устройств для инженерных приложений. Создание аппаратных нейроэмуляторов может быть чрезвычайно полезным для моделирования крупномасштабных нейронных моделей для объяснения того, как в мозге возникает разумное поведение. Основные преимущества нейроморфных эмуляторов заключаются в том, что они обладают высокой энергоэффективностью, параллельностью и распределенностью, а также требуют небольшой площади кремния. Таким образом, по сравнению с обычными процессорами, нейроморфные эмуляторы полезны во многих инженерных приложениях, например, для переноса алгоритмов глубокого обучения для различных задач распознавания.

Работа [53] дает представление о том, как авторы добились полной аппаратной реализации нейронной сети на базе мемристоров. Авторы сообщают об изготовлении не дорогих, высокопроизводительных и однородных мемристорных перекрестных массивов для реализации CNN, которые объединяют восемь 2048-ячеечных мемристорных массивов для повышения эффективности параллельных вычислений. Кроме того, авторы предлагают эффективный метод гибридного обучения для адаптации к несовершенству устройства и повышения общей производительности системы. Была создана пятислойная CNN на основе мемристоров для распознавания изображений MNIST10 и авторы добились высокой точности, превышающей 96 процентов. Помимо параллельного свертывания с использованием различных ядер с общими входами, была продемонстрирована репликация нескольких идентичных ядер в массивах мемристоров для параллельной обработки различных входов. Нейроморфная система CNN на основе мемристоров имеет энергоэффективность более чем на два порядка выше, чем у современных графических процессоров, и демонстрирует масштабируемость на более крупные сети, такие как RNN (residual neural networks). Ожидается, что достигнутые результаты позволят создать высоконадежное аппаратное решение на основе мемристоров без архитектуры фон-Неймана для глубоких нейронных сетей и граничных вычислений.

В мемристорной ячейке используется стек материалов TiN/TaOx/HfOx/TiN и демонстрируется способность непрерывной настройки проводимости как в режиме потенцирования (SET), так и в режиме депрессии (RESET). Материалы и процесс изготовления совместимы с обычным КМОП (комплементарный металл-оксид-полупроводник), так что массивы мемристоров можно удобно создавать в конце линии на кремниевом заводе, чтобы уменьшить вариации процесса и достижения высокой воспроизводимости. Изготовленные

массивы мемристоров демонстрируют однородное аналоговое поведение переключения при идентичных условиях программирования. Была построена аппаратная система с использованием специализированной печатной платы и программируемой вентильной матрицы. Как видно из схемы системы, она состоит в основном из восьми мемристорных вычислительных элементов (ВЭ). Каждый ВЭ имеет свой собственный интегрированный 2 048 ячеек массив мемристоров. Каждый мемристор подключен к стоковой клемме транзистора. Каждый мемристорный массив имеет сборку из 128 × 16 ячеек. Имеется 128 параллельных линий слов и 128 исходных линий по горизонтали и 16 битовых линий по вертикали. Показано распределение 1024 мемристоров в 32 различных состояниях проводимости состояния, где все кривые разделены без какого-либо перекрытия. Идентичные последовательности импульсов SET и RESET с длительностью импульса 50 нс были использованы в 24 операции программирования замкнутого цикла для достижения определенного состояния проводимости.

В работе [54] авторы предлагают аппаратными средствами обучать множества моделей ИНС, используя NeuRRAM – микросхему CIM (calculate in memory) на основе памяти с произвольным доступом. Эта микросхема обеспечивает гибкость в реконфигурировании ядер CIM, более высокую энергоэффективность, чем все современные чипы на основе RRAM и точность, сопоставимую с программными реализациями моделей ИНС, а также высокую инвариантность в области распараллеливания информации на чипе.

Чип NeuRRAM состоит из 48 ядер CIM, которые могут выполнять вычисления параллельно. Ядро может быть выборочно отключено через блокировку питания, когда оно не используется активно, в то время как вес модели сохраняется энергонезависимыми устройствами RRAM. Центральным элементом каждого ядра является TNSA, состоящий из 256 × 256 ячеек RRAM и 256 нейронных цепей CMOS, которые реализуют аналого-цифровые преобразователи и функции активации. Дополнительные периферийные схемы по краю обеспечивают управление логическими выводами и управляют программированием RRAM.

Архитектура TNSA (transposable neurosynaptic array) предназначена для обеспечения гибкого управления направлениями потоков информации, что имеет решающее значение для использования различных архитектур моделей с различными шаблонами потоков данных. Например, в CNN, которые обычно применяются для задач, связанных с машинным зрением, информация проходит в одном направлении через слои для создания представлений данных на разных уровнях абстракции. В LSTM (long short-term memory), которые используются для обработки аудио сигналов, информация периодически проходит через один и тот же уровень в виде нескольких временных блоков. В вероятностных графических моделях, таких как ограниченная машина Больцмана, вероятностная выборка выполняется между уровнями до тех пор, пока сеть не сойдется к высоковероятностному состоянию. Помимо вывода, обратное распространение ошибки во время обучения градиентному спуску нескольких моделей ИНС требует изменения направления потока информации через сеть.

Традиционные архитектуры RRAM-CI ограничены выполнением MVM (Matrix-Vector-Multiply) в одном направлении из-за жесткого подключения строк и столбцов матрицы RRAM к выделенным схемам на периферии для управления входными данными и измерения выходных данных. В некоторых исследованиях реализуются реконфигурируемые направления потока информации путем добавления дополнительного оборудования, что влечет за собой значительные потери энергии, задержки и площади.

Ранние исследования в области вычислений в памяти (CIM), резистивной памяти с произвольным доступом (RRAM) были сосредоточены на демонстрации функций искусственного интеллекта (ИИ). В устройствах RRAM применялось внешнее программное и

аппаратное обеспечение для реализации аналогово-цифрового преобразования и реализации функций активации нейронов. В этих исследованиях были предложены различные методы для уменьшения влияния разброса параметров аналогового оборудования на точность выходного результата.

Чип NeuRRAM состоит из 48 ядер CIM, которые могут выполнять вычисления параллельно. Ядро может быть выборочно отключено через блокировку питания, когда оно не используется активно, в то время как вес модели сохраняется энергонезависимыми устройствами RRAM. Центральным элементом каждого ядра является TNSA, состоящий из 256 × 256 ячеек RRAM и 256 нейронных цепей CMOS, которые реализуют аналого-цифровые преобразователи (ADC) и функции активации. Дополнительные периферийные схемы по краю обеспечивают управление логическими выводами и управляют программированием RRAM. Чтобы максимизировать пропускную способность вывода ИИ на 48 ядрах CIM, реализуется широкий выбор стратегий сопоставления весов, которые позволяют использовать как параллелизм моделей, так и параллелизм информации через многоядерные параллельные MVM. Используя CNN в качестве примера, чтобы максимизировать параллелизм информации, дублируются веса наиболее ресурсоемких слоев (ранние сверточные слои) на несколько ядер для параллельного вывода по нескольким данным. Для максимизации параллелизма модели, сопоставляются разные сверточные слои с разными ядрами и выполняется параллельный вывод конвейерным способом.

Одновременно происходит деление слоёв, весовые размеры которых превышают размер RRAM array, на несколько сегментов и назначаются нескольким ядрам для параллельного выполнения. Буферы промежуточной информации и частичные суммирующие накопители реализуются программируемой вентильной матрицей (FPGA), интегрированной на той же плате, что и микросхема NeuRRAM.

Представленная архитектура чипа, а также аппаратно-алгоритмическая оптимизация позволяет достичь в системе точности выполнения функций ИИ, сопоставимой с их программной реализацией на универсальных компьютерах. Чип NeuRRAM одновременно повышает эффективность, гибкость и точность функционирования нового класса вычислительных систем по сравнению с существующим аппаратным обеспечением RRAM-CIM. Полученные результаты достигнуты за счет инноваций во всей иерархии конструкции, от архитектуры TNSA, обеспечивающей реконфигурируемое направление потока информации, до энергосберегающей схемы нейронов, работающей в режиме напряжения, и ряда методов совместной оптимизации алгоритма и оборудования. Эти методы могут быть более широко применены к другим технологиям энергонезависимой резистивной памяти, таким как память с фазовым переходом [55,56,57,58,59], магниторезистивное ОЗУ [60] и сегнетоэлектрические полевые транзисторы [61].

Состояние исследований мемристивных вычислительных систем в России

В настоящее время в России также существует значительный научный задел, который позволит в ближайшее время перейти к созданию многоядерного нейрочипа на базе мемристивных устройств. В частности, учеными из Университета Лобачевского под руководством к.ф.-м.н. Михайлова А.Н. разработаны авторские технологии производства неорганических мемристивных устройств на базе оксида циркония. Данные мемристивные устройства совместимы с КМОП процессом и интегрируются с периферийной КМОП электроникой на уровне верхних слоев металлизации. Основной метод осаждения тонких пленок, применяемый авторами – метод магнетронного распыления. Коллективом также разработаны конструктивные варианты объединения мемристивных устройств в массивы с архитектурой кросс-поинт и кросс-бар с числом мемристоров до 256. На базе данных массивов разработаны архитектуры многослойных персептронов для решения задач в области создания нейроинтефейсов [62-67].

Разработанные устройства и подходы легли в основу концепции нейрогибридного мемристивного чипа, основанного на комбинации живых нейронных сетей, культивируемых в микрофлюидной / микроэлектродной системе и КМОП интегрированных массивов металлоксидных мемристивных устройств для обработки декодированной информации и организации обратной стимуляции биологической культуры в качестве части двунаправленного нейроинтерфейса [62].

Коллектив исследователей из НИЦ «Курчатовский институт» во главе с к.ф.-м.н. Деминым В.А. достигли значительных успехов в области аппаратной реализации различных архитектур нейронных сетей, особенно спайковых [68,69,70]. Они одни из первых в мире продемонстрировали работу многослойных песептронов на базе мемристоров, спайковых нейронных сетей, в которых обучение происходит на основе локальных правил, а также сетей, работающих по принципам обучения с подкреплением. В своих работах данный коллектив применяет как органические мемристивные устройства (на основе полианилина и поли-пара-ксилилена), так и неорганические нанокомпозитные.

В ЮФУ исследованием материалов и наноструктур, обладающих эффектом резистивного переключения занимается группа исследователей во главе с к.т.н. Смирновым В.А. в том числе для создания сверхбыстродействующих элементов энергонезависимой памяти, а также нейроморфных систем. Особое внимание авторы уделяют моделированию элекро-физических характеристик мемристивных устройств и природы резистивного переключения [71-73].

В ЛЭТИ коллектив ученых под руководством д.ф.-м.н. Андреевой Н.В. решает задачи формирования физико-технологических основ элементной базы нового поколения на основе мемристивных нанослоевых композиций с многоуровневым переключением сопротивления, ориентированной на бионические принципы функционирования аналоговых нейроморфных электронных систем. [74]. Ими продемонстрированы спайковые нейронные сети на базе мемристоров, применяемые для распознавания аудио информации. Другая группа исследователей из ЛЭТИ, которую возглавляет к.т.н. Бутусов Д.Н. [75-77], разрабатывает системы автоматизированного проектирования, моделирования, исследования цепей и устройств с мемристивными элементами.

Галушкиным А.И. [1,2] рассмотрены проблемы создания суперкомпьютеров на базе мемристивных устройств. Автор отмечает характерные ошибки при их создании, обосновывает перспективные направления и этапы работ.

Коллектив исследователей из ТюмГУ под руководством д.ф.-м.н. Удовиченко С.Ю. занимается моделированием и проведением экспериментов с мемристорами, а также симуляцией процессов в нейропроцессорах. Под его руководством проводятся работы по созданию матрицы нейропроцессора, реализованного на основе комбинированного мемристорно-диодного кроссбара – нового компонента наноэлектроники. Предложена концепция биоморфного нейропроцессора, реализующего аппаратную спайковую нейронную сеть, для задач обработки информации, способного имитировать кору головного мозга или ее часть [78, 79].

Группа исследователей, которую возглавляет д.т.н. Кулик С.Д., проводит многопрофильные работы в области теории и практики современных и перспективных ИНС для решения сложных задач ИИ. В работах широко применяются основные положения системного анализа [80-83].

Научный коллектив под руководством Данилина С.Н., решил актуальную научную проблему – разработал методологию автоматизированного инженерного проектирования электронных систем искусственного интеллекта (гражданского, промышленного, военного назначения), работающих по принципам искусственных нейронных сетей на базе новых перспективных видов электронных компонентов — мемристивных устройств (НСМ).

Методология включает в себя полный комплекс компонентов, предусмотренных международными стандартами: общие подходы, технологии, методы, алгоритмы, программное обеспечение, программно-аппаратные средства их реализации. Впервые методология проектирования содержит положения системной инженерии и имитационного моделирования. Показано, что для адекватного проектирования, моделирования, исследования на всех стадиях жизненного цикла НСМ следует рассматривать как единые физическо-информационные системы, реализованные аппаратно-программными обучаемыми, мультирежимными средствами. Впервые для обеспечения наибольшей защищенности программно-технических средств реализации НСМ, выполняющих функции искусственного интеллекта, выполнен комплекс мероприятий, рекомендованных правовыми актами РФ [84-95].

Заключение

Рассмотренные в статье материалы позволяют сделать вывод о том, что как в России, так и за рубежом активно ведутся работы в области проектирования, исследования и производства мемристивных нейроморфных устройств и систем различного назначения. В области систем с распределенными структурами возможно достижение наиболее весомых результатов в решении проблемы создания суперкомпьютеров сверхвысокой производительности и энергоэффективности.

Авторами сделан акцент на материалы об архитектурах распределенных мемристивных вычислительных систем, которые рассмотрены как единые физическо-информационные объекты, реализованные аппаратно-программными распределенными обучаемыми средствами. Вышеназванные составляющие ИНС оказывают совместное, в общем случае зависимое влияние на все их параметры и характеристики.

Для создания рассматриваемых систем применяются знания более чем из 20 областей науки и техники. По этой причине все передовые научно-производственные коллективы ведут работы только в отдельных направлениях. Эта особенность отчетливо показана в различных обзорных статьях и материалах профильных НТК

В данных обстоятельствах авторы делают вывод, что для обеспечения высоких темпов создания мемристивных устройств и систем различного назначения необходимо тесное сотрудничество всех профильных научных коллективов.

Литература

1. Галушкин А. И. Новые технологии микроэлектроники и разработки перспективных нейрокомпьютеров //Информационные технологии. – 2016. – Т. 22. – №. 7. – С. 550-555.

2. Галушкин А.И. Пантюхин Д.В. СуперЭВМ и мемристоры // Информационные технологии. 2016. №4. Т.22. С. 304-312

3. Dakkak A. et al. Accelerating reduction and scan using tensor core units //Proceedings of the ACM International Conference on Supercomputing. – 2019. – С. 46-57.

4. Wang Y. et al. Benchmarking the performance and energy efficiency of AI accelerators for AI training //2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID). – IEEE, 2020. – С. 744-751.

5. Sefat M. S. et al. Accelerating hotspots in deep neural networks on a CAPI-based FPGA //2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). – IEEE, 2019. – С. 248-256.

6. Abdelouahab K. et al. Accelerating CNN inference on FPGAs: A survey //arXiv preprint arXiv:1806.01683. – 2018.

7. Handbook of Memristor Networks / L. Chua, G. Sirakoulis, A. Adamatzky. Springer, 2019. 1357 p.

8. Нейроморфные вычисления и их перспективы для искусственного интеллекта // Экспресс-информация по зарубежной электронной технике. 2020. № 6 (6705). С 17-20.

9. Повышение способности искусственного интеллекта к самостоятельному обучению при помощи ReRAM// Экспресс-информация по зарубежной электронной технике. 2020. № 20 (6719). С 30-34

10. Tanenbaum, Andrew S.; Steen, Maarten van (2002). Distributed systems: principles and paradigms. Upper Saddle River, NJ: Pearson Prentice Hall. ISBN 0-13-088893-1.

11. Distributed Programs". Texts in Computer Science. London: Springer London. 2010. pp. 373–406. doi:10.1007/978-1-84882-745-5_11. ISBN 978-1-84882-744-8. ISSN 1868-0941

12. Key Characteristics of Distributed Systems – URL: https://medium.com/geekculture/key-characteristics-of-distributed-systems-e5b873bcdfae

13. ГОСТ 27.002-2015 Надежность в технике. Термины и определения. Официальное издание. Москва: Стандартинформ, 2016. 28 с.

14. ГОСТ Р 51901.14 – 2007 (МЭК 61078:2006). Менеджмент риска. 8.2 Структурная схема надежности и булевы методы. IEC 61078:2006 Analysis techniques for dependability — Reliability block diagram and boolean methods (MOD) M.: Стандартинформ 2008. 28 с.

15. Галушкин А.И. Нейронные сети: основы теории / А.И. Галушкин. Москва: Горячая линия-Телеком, 2013. 496 с.

16. How to Measure Availability in Distributed Systems – URL: https://towardsdatascience.com/availability-in-distributed-systems-adb43df78b9a

17. How to judge a Distributed System based on its Scalability, Reliability, Availability, Efficiency, and Manageability. – URL: https://medium.com/geekculture/key-characteristics-of-distributed-systems-e5b873bcdfae

18. Distributed Systems – URL:http://www.csc.villanova.ed/~schragge/ CSC8530/Intro.html

19. Raj Rajagopal (1999). Introduction to Microsoft Windows NT Cluster Server: Programming and Administration. CRC Press. p. 4. ISBN 978-1-4200-7548-9.

20. Mike Ebbers; John Kettner; Wayne O'Brien; Bill Ogden (2012). Introduction to the New Mainframe: z/OS Basics. IBM. p. 96. ISBN 978-0-7384-3534-3.

21. Deborah Morley; Charles Parker (13 February 2012). Understanding Computers: Today and Tomorrow, Comprehensive. Cengage Learning. p. 183. ISBN 978-1-133-19024-0.

22. Ran Giladi (2008). Network Processors: Architecture, Programming, and Implementation. Morgan Kaufmann. p. 293. ISBN 978-0-08-091959-1.

23. Sajjan G. Shiva (20 September 2005). Advanced Computer Architectures. CRC Press. p. 221. ISBN 978-0-8493-3758-1.

24. Flynn, Michael J. (September 1972). "Some Computer Organizations and Their Effectiveness". IEEE Transactions on Computers. C-21 (9): 948–960. doi:10.1109/TC.1972.5009071. S2CID 18573685.

25. Loosely Coupled Multiprocessor System – URL: https://www.techopedia.com/definition/30839/loosely-coupled-multiprocessor-system

26. SMP (symmetric multiprocessing) – URL: https://www.techtarget.com/searchdatacenter/definition/SMP

27. Symmetric Multiprocessing – URL:https://www.tutorialspoint.com /Symmetric-Multiprocessing

28. What is SMP (Symmetric Multi-Processing)? – URL: https://www.geeksforgeeks.org/what-is-smp-symmetric-multi-processing/

29. SourceForge - URL:http://lse.sourceforge.net/numa/faq/system_descriptions.html

30. Bull HN F. Zulian – A. Zulian patent – Computer system with a bus having a segmented structure – URL:http://www.freepatentsonline.com/6314484

31. NUMA Architecture – URL: http://www.dba-oracle.com/real_application_clusters_rac_grid/numa.html

32. The List: June 2018. Top 500. – URL:https://www.top500.org/lists/top.

33. AMD Playstation 5 GPU Specs. TechPowerUp. – URL: https://www.techpowerup.com/gpu-specs/playstation-5-gpu.c3480

34. NVIDIA GeForce GT 730 Specs. TechPowerUp. – URL: https://www.techpowerup.com/gpu-specs/geforce-gt-730. c1988

35. Operating system Family / Linux. TOP500.org. – URL: https://www.top500.org/statistics/details/osfam/1/

36. Anderson, Mark (21 June 2017). Global Race Toward Exascale Will Drive Supercomputing, AI to Masses. Spectrum.IEEE.org. – URL: https://spectrum.ieee.org/global-race-toward-exascale-will-drive-supercomputing-ai-to-masses

37. Lemke, Tim (8 May 2013). NSA Breaks Ground on Massive Computing Center. – URL: https://patch.com/maryland/odenton/nsa-breaks-ground-on-massive-computing-center

38. Hoffman, Allan R.; et al. (1990). Supercomputers: directions in technology and applications. National Academies. pp. 35–47. ISBN 978-0-309-04088-4.

39. Hill, Mark Donald; Jouppi, Norman Paul; Sohi, Gurindar (1999). Readings in computer architecture. pp. 40–49. ISBN 978-1-55860-539-8.

40. Paul Alcorn (30 May 2022). AMD-Powered Frontier Supercomputer Breaks the Exascale Barrier, Now Fastest in the World. Tom's Hardware. – URL: https://www.tomshardware.com/news/amd-powered-frontier-supercomputer-breaks-the-exascale-barrier-now-fastest-in-the-world

41. Japan Captures TOP500 Crown with Arm-Powered Supercomputer - TOP500 website. - URL: www.top500.org.

42. Performance Development. – URL: www.top500.org.

43. Wong H. S. P. et al. Metal–oxide RRAM //Proceedings of the IEEE. – 2012. – Т. 100. – №. 6. – С. 1951-1970.

44. Wong H. S. P. et al. Phase change memory //Proceedings of the IEEE. – 2010. – Т. 98. – №. 12. – С. 2201-2227.

45. Frank D. J. et al. Device scaling limits of Si MOSFETs and their application dependencies //Proceedings of the IEEE. – 2001. – Т. 89. – №. 3. – С. 259-288.

46. Kuzum D. et al. Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing //Nano letters. – 2012. – Т. 12. – №. 5. – С. 2179-2186.

47. Kuzum D., Yu S., Wong H. S. P. Synaptic electronics: materials, devices and applications //Nanotechnology. – 2013. – Т. 24. – №. 38. – С. 382001.

48. Wong H. S. P., Salahuddin S. Memory leads the way to better computing //Nature nanotechnology. – 2015. – Т. 10. – №. 3. – С. 191-194.

49. Yao P. et al. Face classification using electronic synapses //Nature communications. – 2017. – Т. 8. – №. 1. – С. 1-8.

50. Indiveri G. et al. Neuromorphic silicon neuron circuits //Frontiers in neuroscience. – 2011. – Т. 5. – С. 73.

51. Neftci E. et al. Event-driven contrastive divergence for spiking neuromorphic systems //Frontiers in neuroscience. – 2014. – Т. 7. – С. 272.

52. Thakur C. S. et al. Large-scale neuromorphic spiking array processors: A quest to mimic the brain //Frontiers in neuroscience. – 2018. – Т. 12. – С. 891.

53. Yao P. et al. Fully hardware-implemented memristor convolutional neural network //Nature. – 2020. – Т. 577. – №. 7792. – С. 641-646.

54. Wan W. et al. A compute-in-memory chip based on resistive random-access memory //Nature. – 2022. – Т. 608. – №. 7923. – С. 504-512.

55. Khaddam-Aljameh, R. et al. HERMES core-A 14nm CMOS and PCM-based in-memory compute core using an array of 300ps/LSB linearized CCO-based ADCs and local digital processing. In IEEE Symposium on VLSI Circuits, Digest of Technical Papers JFS2-5 (IEEE, 2021).

56. Narayanan, P. et al. Fully on-chip MAC at 14 nm enabled by accurate row-wise programming of PCM-based weights and parallel vector-transport in duration-format. IEEE Trans. Electron Devices 68, 6629–6636 (2021).

57. Joshi, V. et al. Accurate deep neural network inference using computational phase-change memory. Nat. Commun. 11, 2473 (2020).

58. Eryilmaz, S. B. et al. Experimental demonstration of array-level learning with phase change synaptic devices. In International Electron Devices Meeting (IEDM), Technical Digest 25.5.1–25.5.4 (IEEE, 2013).

59. Burr, G. W. et al. Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element. IEEE Trans. Electron Devices 62, 3498–3507 (2015).

60. Jung, S. et al. A crossbar array of magnetoresistive memory devices for in-memory computing. Nature 601, 211–216 (2022).

61. Jerry, M. et al. Ferroelectric FET analog synapse for acceleration of deep neural network training. In International Electron Devices Meeting (IEDM), Technical Digest 6.2.1–6.2.4 (IEEE, 2018).

62. Makarov V.A., Lobov S.A., Shchanikov S.A., Mikhaylov A. & Kazantsev V. B. Towards reflective spiking neural networks exploiting memristive devices Frontiers in Computational Neuroscience. 2022. V. 16. № 859874.

63. Mikhaylov A., Pimashkin A., Pigareva Y. et al. // Front. Neurosci. 2020. V. 14. P. 358.

64. Surazhevsky I.A., Demin V.A., Ilyasov A.I., Emelyanov A.V., Nikiruy K.E., Rylkov V.V., Shchanikov S.A., Bordanov I.A., Gerasimova S.A., Guseinov D.V., Malekhonova N.V., Pavlov D.A., Belov A.I., Mikhaylov A.N., Kazantsev V.B., Valenti D., Spagnolo B. & Kovalchuk M. V. Noise-assisted persistence and recovery of memory state in a memristive spiking neuromorphic network. Chaos, solitons & fractals. V. 146. № 110890.

65. Shchanikov S.A., Zuev A.D., Bordanov I.A., Danilin S.N., Lukoyanov V., Korolev D.S., Belov A.I., Pigareva Y., Gladkov A., Pimashkin A., Mikhaylov A.N., Kazantsev V.B. & Serb A. Designing a bidirectional, adaptive neural interface incorporating machine learning capabilities and memristor-enhanced hardware. Chaos, solitons & fractals. 2021. V. 142. № 110504

66. Mikhaylov A.N., Pimashkin A., Pigareva Y., Gerasimova S.A., Gryaznov E., Shchanikov S.A., Zuev A.D., Talanov M., Lavrov I., Demin V.A., Erokhin V., Lobov S.A., Mukhina I., Kazantsev V.B., Wu H. & Spagnolo B. Neurohybrid memristive CMOS-integrated systems for biosensors and neuroprosthetics. Frontiers in neuroscience. 2020. V. 14. № 358.

67. Щаников С.А., Зуев А.Д., Борданов И.А. Данилин С.Н., Лукоянов В., Королев Д., Белов А., Пигарева Я., Гладков А., Пимашкин А., Михайлов А., Казанцев В. Искусственная нейронная сеть на основе мемристивных устройств для двунаправленного адаптивного нейроинтерфейса . Электроника: Наука, технология, бизнес. 2020. № 9(200). С. 86-95

68. Демин В. А. и др. Нейроморфные элементы и системы как основа для физической реализации технологий искусственного интеллекта //Кристаллография. – 2016. – Т. 61. – №. 6. – С. 958-968.

69. Demin V.A., Surazhevsky I.A., Emelyanov A.V. et al. // J. Comput. Electron. 2020. V. 19. P. 565.

70. Демин В.А. Surazhevsky I.A., Demin V.A., Ilyasov A.I., Emelyanov A.V., Nikiruy K.E., Rylkov V.V., Shchanikov S.A., Bordanov I.A., Gerasimova S.A., Guseinov D.V., Malekhonova N.V., Pavlov D.A., Belov A.I., Mikhaylov A.N., Kazantsev V.B., Valenti D., Spagnolo B. & Kovalchuk M. V. Noise-assisted persistence and recovery of memory state in a memristive spiking neuromorphic network. Chaos, solitons & fractals. V. 146. № 110890.

71. Tominov R. et al. Multilevel resistive switching of ZnO memristive crossbar prototype for artificial

neural systems //2022 Fourth International Conference Neurotechnologies and Neurointerfaces (CNN). – IEEE, 2022. – С. 198-201.

72. Il'ina M. V. et al. Memristors based on strained multi-walled carbon nanotubes //Diamond and Related Materials. – 2022. – Т. 123. – С. 108858.

73. Tominov R. V. et al. Synthesis and memristor effect of a forming-free zno nanocrystalline films //Nanomaterials. – 2020. – Т. 10. – №. 5. – С. 1007.

74. Андреева Н. В. Физико-технологические основы мемристивных нанослоевых композиций для аналоговых нейроморфных электронных систем. Диссертация на соискание учёной степени доктора физико-математических наук. СПб, ЛЭТИ, 2022,- 303 с.

75. Островский В.Ю. Автоматизация исследовательского проектирования цепей с мемристивными элементами Диссертация на соискание учёной степени кандидата технических наук. СПб, ЛЭТИ, 2022,- 183.с.

76. Ostrovskii, V., Fedoseev, P., Bobrova, Y., Butusov, D. Structural and parametric identification of Knowm memristors //Nanomaterials. – 2022. – Т. 12. – №. 1. – С. 63.

77. Ostrovskii V.Y., Zubarev A.V., Rybin V.G., Karimov T.I., Control of switching the dynamical modes of a memristor based chaotic circuit //IV International Conference on Control in Technical Systems (CTS). – IEEE, 2021. – С. 49-52.

78. Pisarev A. D. et al. A biomorphic neuroprocessor based on a composite memristor-diode crossbar //Microelectronics Journal. – 2020. – Т. 102. – С. 104827.

79. Бобылев А.Н., Бусыгин А.Н., Губин А.А., Писарев А.Д., Удовиченко С.Ю. Изготовление и тестирование аппаратной импульсной нейросети с мемристорными синапсами для биоморфного нейропроцессора // Российские нанотехнологии. 2021. Т.16. № 6. С. 793-798.

80. Кулик С.Д., Штанько А.Н. Элементы системного анализа, фактографические системы и сверточные нейронные сети: Монография. – М.: НИЯУ МИФИ, 2022. – 208с.

81. Кулик С.Д., Штанько А.Н. Элементы системного анализа программ, использующих сверточные нейронные сети. Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки. 2022. № 6. С. 98-105.

82. Кулик С.Д., Ткаченко К.И. Оценка эффективности технических систем с использованием нейронных сетей. Нейрокомпьютеры: разработка, применение. 2009. № 9. С. 47-60.

83. S.D. Kulik, “Neural network model of artificial intelligence for handwriting recognition,” Journal of Theoretical and Applied Information Technology, Vol.73, No.2, 2015, pp. 202-211.

84. Щаников С. А. Методология программно-аппаратного моделирования нейроморфных вычислительных систем на базе мемристивных устройств. Российские нанотехнологии. 2021. Т. 16. № 6. С. 816-824.

85. Данилин С.Н., Щаников С.А. Исследование точности функционирования нейросетевых компонентов РТС на основе мемристоров. Радиотехнические и телекоммуникационные системы. 2015. № 1(17). С. 39-48.

86. Данилин, С.Н., Щаников С.А., Сакулин А.Е. Определение функциональных допусков искусственных нейронных сетей на основе наномемристоров. Вестник Рязанского государственного радиотехнического университета. 2017. № 61. С. 25-31

87. Галушкин, А.И., Данилин С.Н., Щаников С.А. Нейросетевой контроль точности функционирования технических средств на основе мемристоров. Радиотехнические и телекоммуникационные системы. 2016. № 2(22). С. 44-51.

88. Данилин С.Н. Зуев А.Д. Особенности обеспечения отказоустойчивости нейронных сетей на базе мемристоров на схемотехническом структурно-функциональном уровне // Радиотехнические и телекоммуникационные системы. 2019. № 4 (36). С. 32-43

89. Борданов И.А., Щаников С.А. Перспектива применения имитационного моделирования при проектировании искусственных нейронных сетей на базе мемристоров // Алгоритмы, методы и системы обработки данных. 2019. № 2(40). С. 13-19.

90. Danilin S.N., Shchanikov S.A., Zuev A.D., Bordanov I.A., Sakulin A.E. The Research of Fault Tolerance of Memristor-Based Artificial Neural Networks // 2019 12th International Conference on Developments in eSystems Engineering (DeSE). 2019. PP. 539-544.

91. Данилин С.Н., Щаников С.А., Борданов И.А. Системная инженерия в области мемристивных нейронных сетей // Телекоммуникации. 2021. № 6. С. 27-39

92. Борданов И.А., Данилин С.Н., Щаников С.А. Программный комплекс «МемриСим» для имитационного моделирования искусственных нейронный сетей на базе мемристоров // Информационные системы и технологии - 2021 Сборник материалов XXVII Международной научно-технической конференции. 2021. С. 79-87.

93. Данилин С.Н., Щаников С.А., Зуев А.Д. Борданов И.А., Сакулин А.Е. Проектирование искусственных нейронных сетей на основе мемристоров с заданной отказоустойчивостью .Радиотехнические и телекоммуникационные системы. 2019. № 2(34). С. 41-50.

94. Данилин С.Н., Щаников С.А., Борданов И.А., Зуев А.Д. Разработка методов определения и обеспечения надежности искусственных нейронных сетей на базе мемристивных устройств // Телекоммуникации. 2021. № 4. С. 20-26.

95. Shchanikov S.A. Methodology fo Hrdware-in-the-Loop Simualation of Memristive Neuromorphic

Systems // Nanobiotechnology Reports. 2021, Vol.16,
No. 6, pp. 782-789.

Работа выполнена при поддержке стипендии Президента РФ СП-5411.2021.5.

Поступила 24 октября 2022 г.

The article presents an analytical review of scientific and technical publications on the state of research in the field of engineering design, production and application of computing systems with a new generation distributed architecture based on memristive devices. The achievements of the most famous scientific teams are shown, and current problems that slow down the pace of development of this field of science and technology.

Key words: memristors, artificial neural networks, distributed systems, scaling.

Никишов Даниил Андреевич – стажер-исследователь лаборатории разработки систем искусственного интеллекта Муромского института (филиала) ФГОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых».

E-mail: daniilnikisov74@gmail.com.

Данилин Сергей Николаевич – кандидат технических наук, ведущий научный сотрудник лаборатории разработки систем искусственного интеллекта Муромского института (филиала) ФГОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых».

E-mail: dsn-55@mail.ru.

Щаников Сергей Андреевич – кандидат технических наук, ведущий научный сотрудник лаборатории разработки систем искусственного интеллекта Муромского института (филиала) ФГОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых».

E-mail: seach@inbox.ru.

Адрес: 602264, Муром, ул. Орловская, д. 23.