ФИЗМАТ, 2023, том 1, № 1, с. 42–64

УДК 621.38-022.532

НА ПУТИ К РЕАЛИЗАЦИИ ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ ВЫЧИСЛЕНИЙ В ПАМЯТИ НА ОСНОВЕ МЕМРИСТОРНОЙ ЭЛЕКТРОННОЙ КОМПОНЕНТНОЙ БАЗЫ

© 2023 г. А. Н. Михайлов[1,*], Е. Г. Грязнов[1], В. И. Лукоянов[1], М. Н. Коряжкина[1], И. А. Борданов[2], С. А. Щаников[1,3], О. А. Тельминов[4], М. В. Иванченко[1], В. Б. Казанцев[1,3]

[1]Нижегородский государственный университет им. Н.И. Лобачевского, Нижний Новгород, Россия

[2]Муромский институт Владимирского государственного университета им. А.Г. и Н.Г. Столетовых, Муром, Россия

[3]Московский физико-технический институт, Москва, Россия

[4]Научно-исследовательский институт молекулярной электроники, Зеленоград, Россия

**e-mail: mian@nifti.unn.ru*

Поступила в редакцию 15.09.2023 г.
После доработки 14.11.2023 г.
Принята к публикации 14.11.2023 г.

Статья посвящена анализу современного состояния и перспектив развития высокопроизводительных вычислений на основе принципов хранения и обработки информации в биологических нейронных сетях, которые обеспечены возможностями новой электронной компонентной базы (ЭКБ), представленной мемристорами (нелинейными резисторами с памятью или элементами резистивной памяти с произвольным доступом (англ.: Resistive Random Access Memory (RRAM)). Мемристоры могут быть реализованы на базе различных материалов и наноструктур, совместимых со стандартным технологическим процессом микроэлектроники, и позволяют производить “вычисления в памяти”. Естественным образом такие вычисления реализуются в нейроморфных системах, использующих для выполнения векторно-матричного умножения архитектуру “кроссбар”, в которой мемристоры на пересечениях проводящих шин выступают в качестве синаптических весов – пластичных соединений между искусственными нейронами в полносвязной архитектуре нейронной сети. В статье рассмотрены общие подходы к разработке и созданию новой ЭКБ на основе технологии RRAM, совместимой с комплементарными структурами металл-оксид-полупроводник, разработке искусственных нейронных сетей и нейропроцессора, использующих мемристорные матрицы кроссбар как вычислительные ядра и масштабируемые многоядерные архитектуры для реализации как формальных, так и импульсных нейросетевых алгоритмов. Описаны технические решения, обеспечивающие аппаратную реализацию мемристорных кроссбаров достаточно большой размерности, а также решения, компенсирующие некоторые недостатки или принципиальные ограничения, присущие современным мемристорам на стадии взросления технологии. Проведен анализ производительности и энергоэффективности для опубликованных прототипов таких нейроморфных систем и сделан вывод о существенном (на порядки величины) выигрыше с точки зрения этих параметров по сравнению с вычислительными системами на основе традиционной элементной базы (в том числе нейроморфными). Технологическое освоение новой элементной базы и создание мемристорных нейроморфных вычислительных систем обеспечит не только своевременную диверсификацию аппаратного обеспечения для непрерывного развития и массового внедрения технологий искусственного интеллекта, но и позволит поставить задачи совершенно нового уровня по созданию гибридного интеллекта на основе симбиоза искусственных и биологических нейронных сетей. Среди этих задач первоочередными являются задачи по созданию мозгоподобных самообучающихся спайковых нейросетей и адаптивных нейроинтерфейсов на основе мемристоров, которые также обсуждаются в работе.

DOI: 10.56304/S2949609823010021, **EDN:** HTSRZQ

ВВЕДЕНИЕ

Четвертая промышленная революция, на пороге которой стоит человечество, предъявляет совершенно новые требования к аппаратному обеспечению технологий искусственного интеллекта (ИИ), которое должно приближаться по своим возможностям к возможностям человеческого мозга (интеллекта естественного). Кроме требований по компактности и энергоэффективности, к новым аппаратным средствам ИИ предъявляются требования по совместимости с существующей кремниевой технологией микроэлектроники и совместимости с живыми системами. Удовлетворение этих требований обеспечит массовое производство аппаратных систем ИИ и реализацию новых гибридных форм ИИ, причем второе требование подразумевает, что новые электронные системы ИИ должны не только по форме (как сейчас), но и по функциональности воспроизводить свойства элементов нервной системы и мозга.

На удовлетворение этих требований направлены текущие парадигмальные изменения в электронной технике, связанные с переходом от традиционной архитектуры фон Неймана (в которой запоминание и обработка информации разделены в пространстве) к аналоговым вычислениям в памяти и массовому параллелизму в обработке информации подобному тому, как это имеет место в мозге. В основе новой “пост-цифровой” парадигмы лежит мозгоподобная электронная компонентная база (ЭКБ), представленная мемристорами (аналоговыми резистивными элементами с памятью) и мемристорными устройствами, которые имитируют функции элементов живой нервной системы (нейронов и синапсов). Многообразие возможных вычислительных архитектур обеспечивается универсальным характером мемристорного эффекта, поскольку он может быть реализован как в классических, так и в квантовых системах, как в различных искусственных материалах и структурах (неорганических, органических, молекулярных и т.д.), так и в живых системах.

Результаты всестороннего исследования и разнообразных применений мемристорных устройств стали предметом многочисленных публикаций за последние годы (см., например, [1–7], в том числе дорожные карты, обзоры и перспективы 2020–2023 годов в журналах уровня Springer Nature и Willey Advanced), которые свидетельствуют о важности и актуальности данного направления на мировом уровне, а также о необходимости реализации генерального плана (комплексных и междисциплинарных проектов) в области биоинспирированных систем, нацеленного на технологическое освоение новой элементной базы и создание прототипов информационно-вычислительных систем нового поколения.

В данной работе проведен анализ современного состояния и перспектив развития высокопроизводительных вычислений на основе мемристоров. Проведено сравнение достигнутых параметров нейроморфных вычислительных систем на новой и традиционной ЭКБ. Рассмотрены общие подходы к разработке технологии резистивной памяти с произвольным доступом (англ.: Resistive Random Access Memory (RRAM)), совместимой с комплементарными структурами металл-оксид-полупроводник (КМОП). Такая технология необходима для создания элементов и функциональных блоков мемристорного нейропроцессора, а также применения новых вычислительных систем в технологиях искусственного и гибридного интеллекта.

Структура статьи: раздел 1 посвящен обсуждению актуальности и перспектив исследования и развития мемристоров и нейроморфных и нейрогибридных систем на их основе; в разделе 2 обсуждается многоуровневый и междисциплинарный подход к разработке нейроморфных систем на основе КМОП-совместимых мемристорных устройств; в разделе 3 рассмотрены различные варианты масштабирования КМОП-интегрированных мемристорных кроссбаров для увеличения скорости передачи сигналов в искусственных нейронных сетях; раздел 4 содержит сравнение нейроморфных вычислительных систем на основе традиционной и новой компонентной баз. Заключение подводит итоги исследования.

1. МЕМРИСТОР, НЕЙРОМОРФНЫЕ И НЕЙРОГИБРИДНЫЕ СИСТЕМЫ НА ОСНОВЕ МЕМРИСТОРОВ

В течение последних пяти десятилетий мировая микроэлектроника развивалась в соответствии с законом Мура, который предсказывает экспоненциальное увеличение числа транзисторов на чипе, соответствующее увеличение скорости вычислений и снижение энергопотребления для каждого нового поколения технологий. В настоящее время эта тенденция приблизилась к физическому пределу – дальнейшее увеличение числа транзисторов уже не приводит ни к увеличению тактовой частоты, ни к снижению энергопотребления. “Узким местом” является процесс обмена данными между центральным процессором и рабочей памятью вне кристалла, что

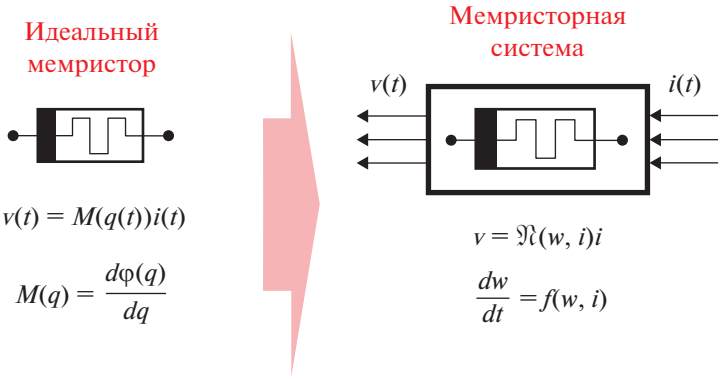

Рис. 1. Исходное и обобщенное определения мемристора [8, 13].

делает цифровые процессоры на базе традиционной архитектуры фон Неймана крайне неэффективными с точки зрения энергопотребления и временных задержек. Между тем, объем цифровых данных, требующих обработки, продолжает лавинообразно увеличиваться. Каждые два года создается больше данных, чем было создано за всю предыдущую историю. Неструктурированные данные составляют уже более 80% от общего объема ежедневно генерируемых данных. Таким образом, потребности растут быстрее, чем возможности современных компьютеров. Требуется разработка прорывных технологических решений, которые бы сняли проблему “узкого места” архитектуры фон Неймана. Исследования, ведущиеся сегодня в мировых научных центрах, выявили два основных направления решения этой проблемы – совмещение вычислений и памяти в единых функциональных блоках и переход от традиционных фон-Неймановских архитектур к нейроморфным, воспроизводящим принципы хранения и обработки информации в нервной системе и мозге.

Новая парадигма в электронике, с развитием которой связывается прорыв в аппаратной реализации нейроморфных информационно-вычислительных систем, основана на использовании мемристора. Мемристор (англ.: memristor = memory + resistor) был теоретически описан Леоном Чуа в 1971 году как недостающий пассивный элемент электрических схем, который связывает изменение магнитного потока $\varphi(t)$ и электрического заряда $q(t)$ [8] (рис. 1). Легко показать, что этот элемент эквивалентен нелинейному резистору, который меняет свое сопротивление $M(q(t))$ в зависимости от предыстории протекания через него электрического заряда. Это определение идеального мемристора до сих пор вызывает сомнения и споры ученых [9–11] и стимулирует поиск материалов и сред, в которых реализуется физическая связь между магнитными и электрическим свойствами [12]. Однако в 1976 году Л. Чуа и С. Канг предложили обобщенное определение мемристора и мемристорных динамических систем [13], которые описываются портовым уравнением, эквивалентным закону Ома, и набором уравнений состояния, описывающих динамику параметров внутреннего состояния системы (w). Это определение является универсальным и описывает изменение сопротивления (эффект памяти) на основе различных явлений в неорганических и органических наноматериалах (миграция ионов, восстановительно-окислительные реакции, фазовые превращения, спиновые и сегнетоэлектрические явления) [1], а также в фотонных [14] и сверхпроводниковых [15, 16] схемах. Среди них необходимо выделить наноструктуры типа “металл-оксид-металл” (МОМ), которые идеально подходят для создания компактных (с нанометровым размером) и энергоэффективных (фемтоджоули на переключение) устройств памяти RRAM, интегрируемых в стандартный технологический КМОП-процесс. Такие устройства могут не только хранить логическое значение, задаваемое проводимостью, но и позволяют менять его в том же физическом месте, реализуя новые “не фон-Неймановские” парадигмы вычислений в памяти. Кроме того, простая структура мемристора обеспечивает создание сверхплотных и в перспективе трехмерных массивов “кроссбар”, которые естественным образом (на основе законов Ома и Кирхгофа и в аналоговой форме) реализуют операции векторно-матричного умножения (ВМУ), лежащие в основе инференса в традиционных искусственных нейронных сетях с глубоким обучением и новых алгоритмов обучения спайковых нейронных сетей [17].

Развитие технологий ИИ опирается на развитие нейроморфных вычислительных систем в соответствии с известным прогнозом в рамках международной дорожной карты технологий: “The

Future of AI is Neuromorphic". Мозгоподобная ЭКБ, представленная мемристорами и мемристорными системами, обеспечит своевременную диверсификацию аппаратного обеспечения, которое в основном накладывает фундаментальные ограничения на каждом цикле развития ИИ, и позволит избежать очередной "зимы" ИИ. Альтернативные нейроморфные технологии на новой ЭКБ только вступают в стадию зрелости, конкурируя с доминирующими сейчас цифровыми технологиями высокопроизводительных вычислений на основе чипов центрального процессорного устройства (ЦПУ), графических ускорителей (англ.: Graphic Processing Unit (GPU)), тензорных ускорителей (англ.: Tensor Processing Unit (TPU)) и т.д. Детальный анализ и сопоставление достигнутых характеристик нейроморфных вычислительных систем на основе мемристоров и традиционной ЭКБ были ранее представлены в литературе [4, 5], однако каждый год пополняются новыми прототипами и рекордами (см., например, [18–21]), которые детально обсуждаются в разделе 4. Согласно дорожной карте мозгоинспирированных вычислительных чипов [4], создание мемристорных нейропроцессоров общего пользования ожидается уже в течение ближайших 5–10 лет. Продемонстрированные на данный момент прототипы мемристорных вычислительных систем уже сейчас составляют конкуренцию известным нейроморфным процессорам на основе традиционных цифровых элементов и специализированных архитектур (англ.: Application Specific Integrated Circuit (ASIC)) [5].

При всех успехах в развитии технологий ИИ и впечатляющем прогрессе в разработке специализированных вычислительных систем, реализующих нейросетевые алгоритмы, все больше внимания уделяется перспективам существенно более глубокой адаптации нейроморфных принципов, чем достигнуто на настоящий момент [22]. Кроме того, что они не только по форме, но и по функциональности приближаются к принципам работы мозга, нейроморфные системы (в их узком понимании), реализуемые на базе мемристорных систем, обладают существенным потенциалом для достижения нового уровня когнитивных возможностей, в первую очередь – за счет возможности эффективной обработки в реальном времени электрической активности биологических нейронных систем в составе так называемых био- или нейрогибридных систем [23–25]. В то же время, первые известные из литературы примеры, в которых мемристорные устройства и массивы были использованы для обработки биоэлектрической активности, либо только фиксируют сам факт коммуникации электронных и биологических систем через единичные мемристорные устройства [26], либо делают это в отрыве от самих живых систем (например, в недавних работах [27–29] мемристорные чипы обрабатывают эмулированные последовательности прямоугольных "спайков" или сигналы нейрональной активности, взятые из общедоступных баз данных).

Существенный прогресс в создании мемристорных нейрогибридных систем был достигнут в работе [30], которая демонстрирует первый в мире двунаправленный адаптивный нейроинтерфейс на основе передовых решений в области мемристорной электроники и нейроинженерии (рис. 2).

Со стороны живой системы впервые использована культура клеток нейронов гиппокампа на мультиэлектродной матрице с функциональными связями между группами нейронов, пространственно упорядоченными с помощью микрофлюидного чипа. Мемристорная сеть впервые используется не только для решения задачи нелинейной классификации пространственно-временного отклика клеточной культуры на электрические стимулы, но и для контроля ее функционального состояния. А именно, выходные сигналы мемристорной сети соответствуют разным стимулам и используются для адаптивного управления стимуляцией, что позволяет восстанавливать нарушенные функциональные связи в нейрональной культуре.

Большой интерес вызывают перспективы использования таких нейрогибридных технологий для задач нейрореабилитации, восстановления или реорганизации биологических нейрональных функций после развития патологического состояния [31]. Крайне привлекательной как с точки зрения удобной экспериментальной модели, так и с точки зрения использования в реальной нейрогибридной технологии является перспектива создания клеточных культур, в высокой степени воспроизводящих архитектурные особенности мозга [32].

Таким образом, сочетание высокой энергоэффективности и уникальной масштабируемости мемристорных систем позволяет сделать решающий шаг от нейроморфных вычислительных систем к нейрогибридным системам на основе прямого (физиологического) и безопасного взаимодействия искусственных электронных систем и живых нейрональных систем [33]. Благодаря этому, мемристорные нейроморфные системы займут достойное место и в медицинских технологиях ИИ: обеспечат не только эффективное решение традиционных задач ИИ, связанных с обработкой и анализом биомедицинских данных, но и создание компактных и энергоэффектив-

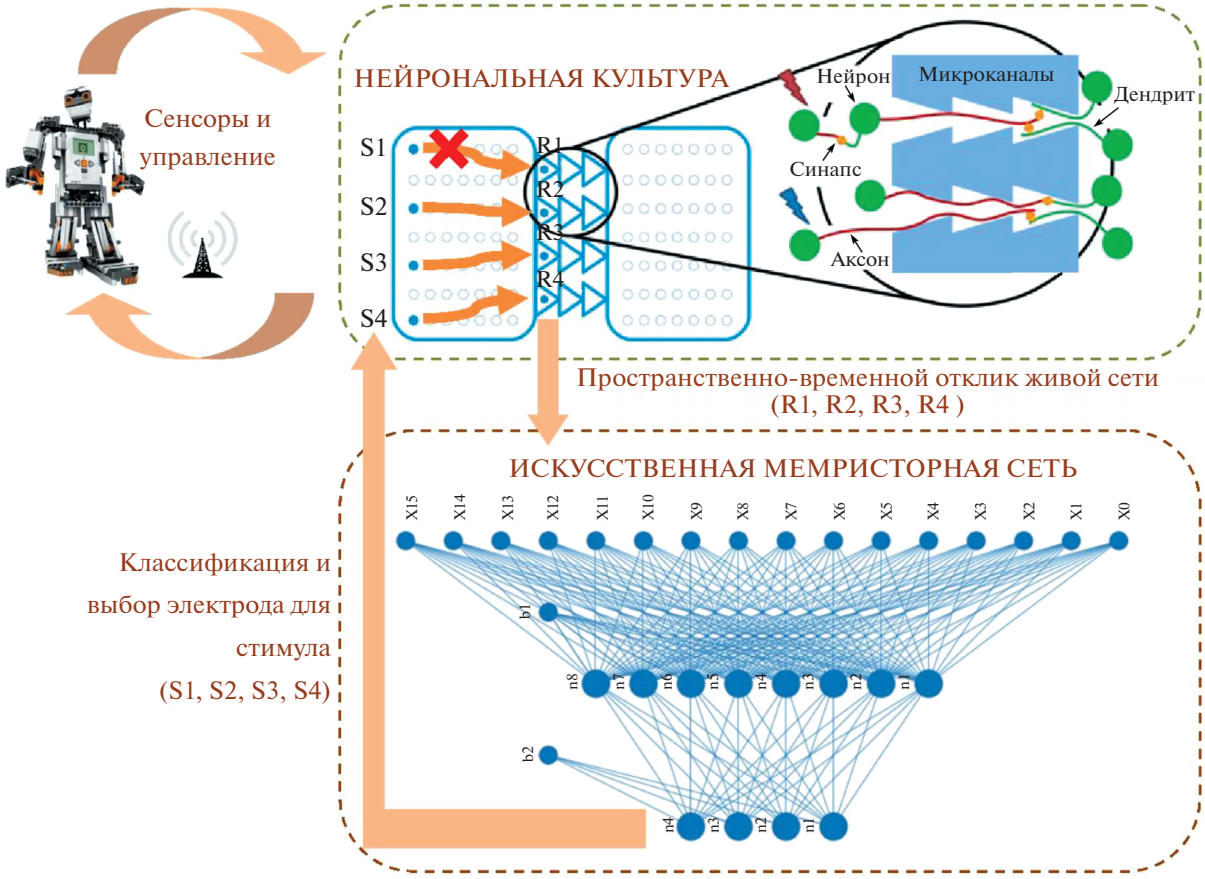

Рис. 2. Двунаправленный адаптивный нейроинтерфейс между упорядоченной нейрональной культурой и искусственной нейронной сетью на основе мемристоров [30].

ных адаптивных систем для замещения/восстановления утраченных или улучшения существующих функций мозга и нервной системы (нейропротезирования и инструментальной корректировки/поддержки/усиления когнитивных способностей человека).

2. ОБЩИЙ ПОДХОД К СОЗДАНИЮ НЕЙРОМОРФНЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ НА ОСНОВЕ МЕМРИСТОРОВ

Согласно недавним перспективам [7, 34], исследования и разработки в области нейроморфных и мозгоинспирированных вычислительных систем отличаются комплексным (многоуровневым) и междисциплинарным характером. Первая характеристика подразумевает, что новые функциональные изделия рождаются за счет совместной оптимизации решений на уровнях материалов, устройств и систем. Междисциплинарный характер не только требует объединения различных научных сообществ (хотя это уже само по себе большой вызов), но и реализации скоординированного плана, финансирования и поддержки (по сути – генерального плана, который мы видели в области цифровых или квантовых технологий, например). Рассмотрим в данном разделе, как этот комбинированный подход реализуется в случае разработки нейроморфных и нейрогибридных систем [33] на основе КМОП-совместимых устройств МОМ с резистивным переключением (рис. 3).

На уровне материалов формируются и изучаются наноструктуры МОМ, которые проявляют резистивное переключение (один из классических механизмов мемристорного эффекта). Причем для понимания закономерностей мемристорного эффекта и управления его параметрами недостаточно детального изучения физико-химических явлений на нано- или микроуровне. Например, совокупность разных транспортных явлений (фононов, электронов, ионов) в разных временных масштабах делает даже один мемристор сложной нелинейной системой с богатым динамическим откликом. Для того, чтобы двигаться дальше на пути к нейроморфным и нейро-

Рис. 3. Иллюстрация комплексного (многоуровневого) и междисциплинарного подходов к созданию нейроморфных и нейрогибридных систем на основе мемристоров.

гибридным системам, те же разработанные конструктивные варианты мемристорных структур формируются в интегральном исполнении – в виде устройств и чипов, входящих в состав различных функциональных схем на уровне систем. Экспериментальная работа всегда ведется параллельно с многомасштабным моделированием: от моделей физических явлений на микро-, мезо- и макроуровне до компактных моделей устройств и схемотехнических моделей, необходимых для автоматизированного проектирования электронных схем. В основе такого подхода лежит сквозная технология мемристорных устройств, совместимая с традиционной кремниевой технологией и обеспечивающая создание элементной базы новых мозгоподобных информационно-вычислительных систем с широким спектром применений, среди которых можно выделить традиционные и импульсные нейросетевые архитектуры, нейроинтерфейсы.

Междисциплинарный характер проекта также проиллюстрирован на рис. 3. Физика и технология мемристорных наноструктур является одной из ключевых областей, которая на основе традиционных и новых подходов в области микроэлектроники создает технологическую платформу для аппаратной реализации нейроморфных систем на основе мемристоров. Для интерпретации, описания и предсказания мемристорного эффекта следует использовать существенный задел научных школ в областях статистической радиофизики и нелинейной динамики, а на основе последних достижений в областях нейробиологии и нейротехнологий можно сделать следующий шаг к симбиозу искусственных электронных и живых биологических систем.

Для достижения поставленной цели должны быть решены взаимоувязанные задачи, включающие исследование новых материалов и устройств, разработку сквозной технологии новой элементной базы, разработку и аппаратную реализацию нейросетевых архитектур.

Применительно к мемристорам решение первой задачи осложняется тем, что совокупность разных транспортных явлений (фононов, электронов, ионов) в разных временных масштабах делает мемристорное устройство сложной нелинейной системой с богатым динамическим откликом и требует взаимоувязанных исследований на микро- и макроскопическом уровнях с привлечением задела из областей физики и химии твердотельных наноструктур, нелинейной динамики и статистической физики. Развитие данных междисциплинарных исследований позволяет найти новые применения обнаруженным явлениям и реализовать новые способы улучшения характеристик электронных устройств на основе мемристорных материалов. Решение данной задачи, по сути, означает решение основных фундаментальных проблем, связанных с необходимостью корректного описания мемристорного эффекта в различных структурах и ма-

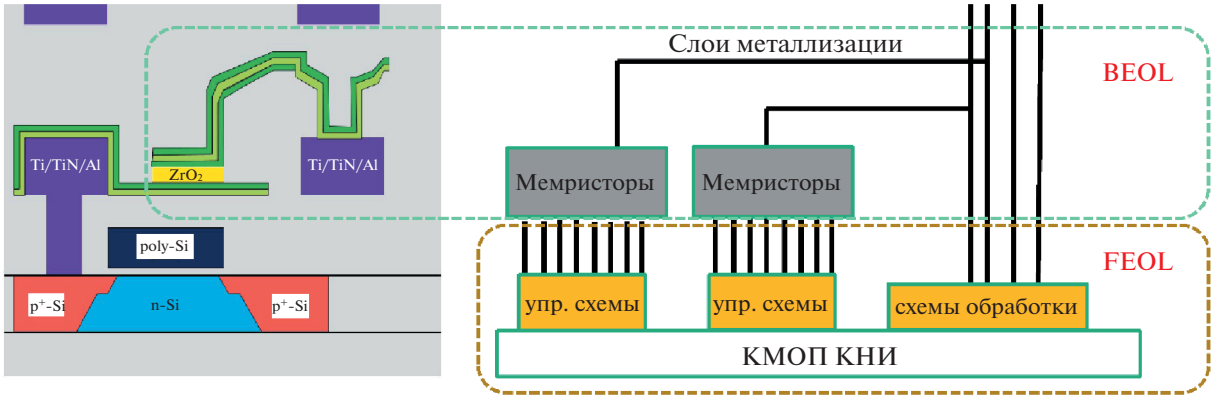

Рис. 4. Иллюстрация BEOL-интеграции мемристорных наноструктур со схемами КМОП.

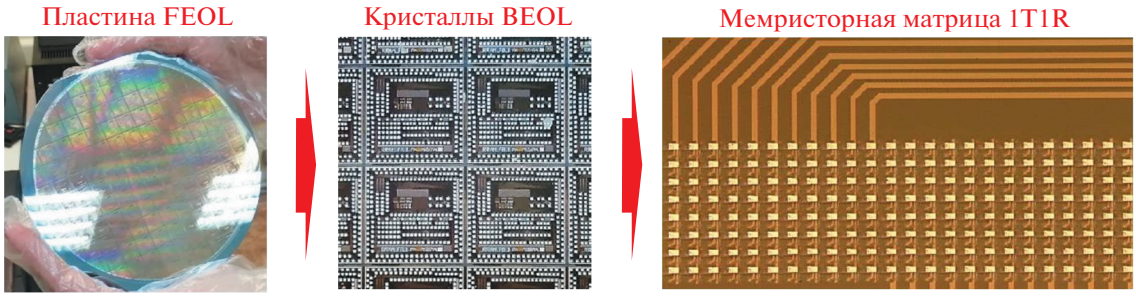

Рис. 5. Изображения пластины FEOL, ее фрагмента после завершения процесса BEOL и готового массива мемристоров 1T1R.

териалах, в том числе при внешних воздействиях, и сопровождающих проектирование и создание информационно-вычислительных систем ИИ на новой элементной базе.

Разработка сквозной технологии на основе устройств с резистивным переключением подразумевает разработку научно-технологических решений по созданию элементов и ячеек энергонезависимой резистивной памяти RRAM на основе мемристорных наноструктур с высоким выходом годных и высокими параметрами выносливости к многократному переключению и удержания резистивного состояния. Важнейшей характеристикой мемристорных устройств с точки зрения нейроморфных применений является их способность к многоуровневому хранению информации, и в этой области сегодня наблюдается существенный прогресс [35]. Основным решением при разработке технологии RRAM является изготовление функциональных блоков RRAM на основе интеграции мемристорных структур, изготавливаемых в лабораторных условиях в слоях металлизации (англ.: back-end-of-line (BEOL)), и приборного слоя КМОП (англ.: front-end-of-line (FEOL)), изготавливаемого в промышленных условиях (рис. 4). Примеры изображений для пластины FEOL, ее фрагмента после завершения процесса BEOL и готового массива мемристоров кроссбар 1T1R (один мемристор – один транзистор) приведены на рис. 5. В случае успешной реализации сквозная технология создания мемристорных микросхем обеспечит технологическую платформу для широкого спектра продуктов от микросхем спецстойкой памяти до нейрочипов, нейроинтерфейсов и нейропротезов для медицинских применений.

Результатом исследований и разработок в рамках этой задачи является проектирование и изготовление тестовых кристаллов с функциональными блоками энергонезависимой резистивной памяти (ячеек памяти и матриц RRAM), необходимых для демонстрации возможностей новых запоминающих устройств и базовых принципов нейроморфных вычислений (операций ВМУ).

Ключевой задачей в рамках данного научно-технического направления является разработка нейроморфного процессора с массивом синаптических весов на основе мемристоров в архитектуре кроссбар (наиболее популярный вариант управляемого кроссбара – RRAM 1T1M) с цифро-аналоговыми нейронами типа интегрирующего элемента с утечкой (англ.: leaky integrate and fire

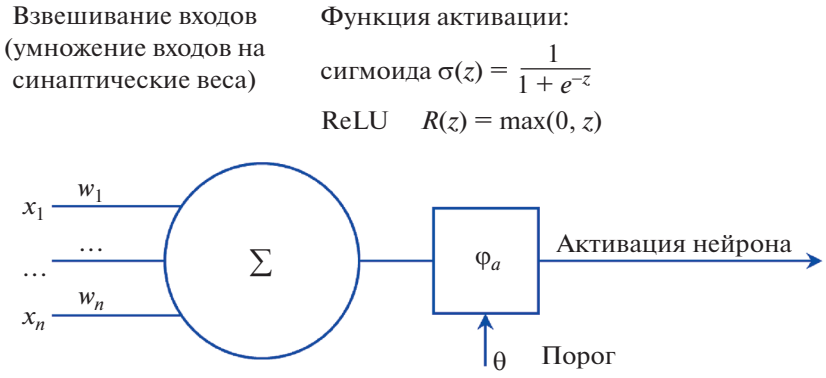

Рис. 6. Модель формального нейрона – взвешенная сумма входов подается на вход принципиально нелинейной функции активации, в качестве которой используется как сигмоида, так и более простая ReLU (Rectified linear unit – усеченное линейное преобразование).

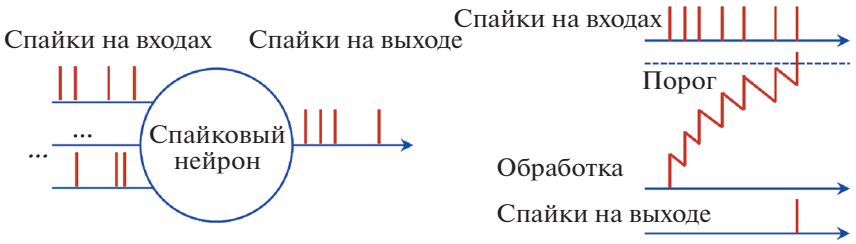

Рис. 7. Спайковый нейрон принимает на входы последовательности спайков и при определенных условиях формирует на своем выходе спайк; например, в модели LIF каждый спайк вносит вклад в статус – амплитуду нейрона, которая с течением времени затухает; если в определенном временном окне вклад внесет достаточное количество спайков, то амплитуда нейрона превысит порог и нейрон сгенерирует выходной спайк. Электрическая модель такого нейрона может быть реализована на ОУ с интегрирующим RC-контуром в плече инвертирующего входа и компаратором.

(LIF)) и другими настраиваемыми параметрами, с возможностями контроля и перезаписи произвольных мемристорных элементов, обучения с учителем и без учителя, в том числе на основе локальных правил, работы в режимах логического вывода алгоритмов как на основе формальных нейронных сетей, так и спайковых нейронных сетей с пространственно-временным кодированием многомерных паттернов решаемой задачи. Такой нейропроцессор в перспективе должен позволять решать различные задачи из области ИИ: распознавание визуальных образов, обработка текстов, речи, разных типов больших данных, предсказание временных рядов данных, сенсомоторный контроль мобильных объектов, оптимизационный контроль потоков данных в реальном времени и др.

Рассмотрим подробнее общий подход к построению искусственной нейронной сети, который базируется на модели нейрона. Различают формальную (рис. 6) и спайковую или импульсную (рис. 7) модели нейрона [36]. Главное отличие заключается в способе представления обрабатываемых сигналов: в формальном нейроне такие сигналы имеют непрерывную форму, в то время как в спайковом – импульсную. С одной стороны, аппаратная реализация спайковых нейронов имеет преимущество в несколько порядков по критерию энергоэффективности, с другой – резкие фронты импульсного сигнала затрудняют дифференцирование и, как следствие – применение зарекомендовавшего себя метода обратного распространения ошибки при обучении нейросети. Последнее обстоятельство приводит к необходимости разработки новых алгоритмов обучения спайковых нейронных сетей, основанных на биоподобных локальных правилах пластичности [37].

Формальная модель нейрона широко используется в различных типах современных сопроцессоров – процессорах цифровой обработки сигналов (англ.: Digital Signal Processing (DSP)), GPU, множестве нейроускорителей и TPU (“Google TPU” (фирма Google), “IVA TPU” (фирма “Ива Технолоджи”), “NM6408” (фирма НТЦ “Модуль”), “RoboDeus” (фирма НТЦ “Элвис”) и

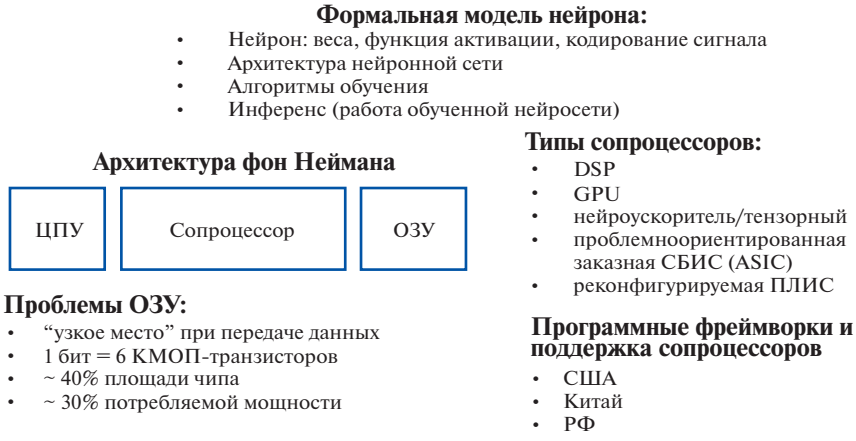

Рис. 8. Программно-аппаратная экосистема для реализации нейросетей на формальной модели нейрона: создан значительный задел, лучшие практики которого могут быть использованы для быстрой разработки и апробации перспективных нейроморфных систем.

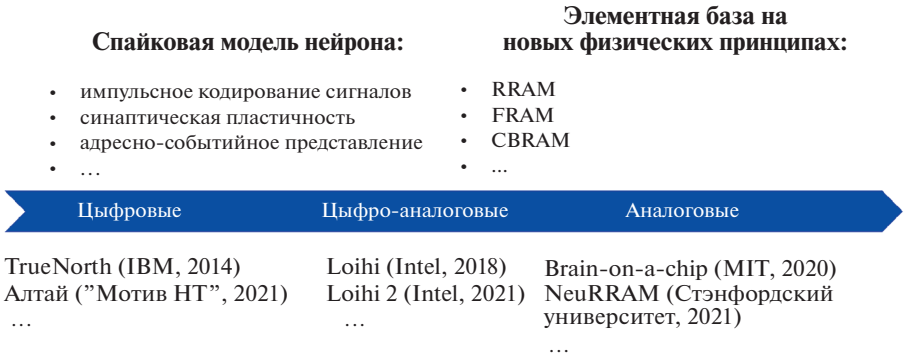

Рис. 9. Современные нейроморфные системы, основанные на спайковой модели нейрона.

многие другие), проблемно-ориентированных заказных сверхбольших интегральных схемах (т.е. ASIC), реконфигурируемых программируемых логических интегральных схемах (ПЛИС) (англ.: Field of Programming Gate Array (FPGA)). Разработаны и широко применяются фреймворки – программные среды для разработки нейросетей, их обучения и инференса с помощью вышеприведенных процессоров. Таким образом, к настоящему моменту наработаны полноценные аппаратно-программные комплексы для обработки нейросетей на формальной модели нейрона (рис. 8). Дальнейшее развитие формальной модели продолжается в направлении улучшения алгоритмов обработки и сокращения проектных норм процессоров, выполненных по кремниевой технологии [38].

Такой задел является ориентиром и может быть частично использован для аппаратно-программного обеспечения новых вычислителей на нейроморфных архитектурах и элементной базе на новых физических принципах. В настоящий момент нейроморфная модель основана на спайковой модели, но со временем нейрофизиологи откроют и обоснуют сверхновую, более правдоподобную модель работы нейрона. К настоящему моменту на спайковой модели разработаны цифровые (фирмами IBM, “Мотив НТ”), цифро-аналоговые (фирмой Intel) и аналоговые (MIT и др.) нейропроцессоры. Разумеется, существует множество других разработок, не указанных на рис. 9. Аналоговая реализация нейронов основана на применении операционных усилителей (ОУ), с помощью которых можно осуществить ряд математических операций с применением токов и напряжений в электрической цепи.

Основная операция, реализуемая при вычислении нейросетей – векторно-матричное и матричное умножение. Как было отмечено выше, ВМУ естественным образом и в аналоговой форме реализуется в мемристорном кроссбаре, представляющем собой набор параллельных шин в

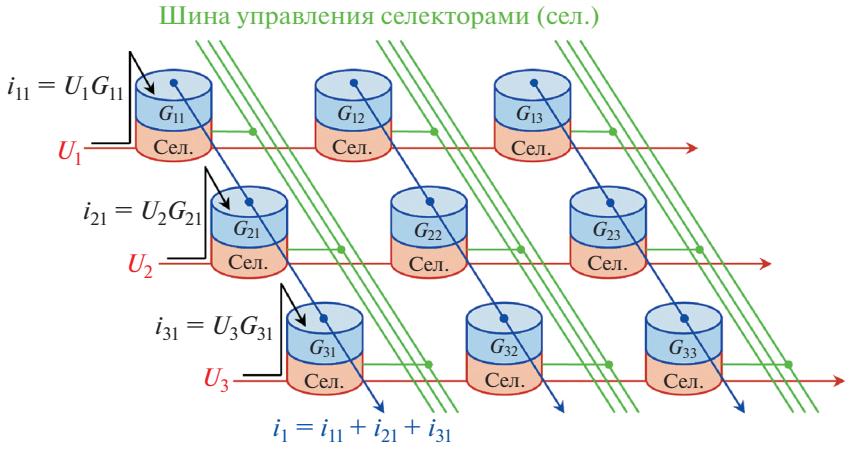

Рис. 10. Мемристорный кроссбар в режиме ВМУ (инференса): входные напряжения умножаются на проводимости G соответствующих мемристоров в определенном столбце, полученные токи суммируются в столбце. Селекторы обеспечивают подключение мемристоров к линиям кроссбара. В некоторых случаях селекторы обеспечивают реверсное подключение: входы меняются с выходами (на синие линии подаются напряжения, с красных линий снимаются токи) для осуществления процесса обучения.

Рис. 11. Уровни логических нуля и единицы при размахе напряжения питания 5 В, обеспечивающие высокую помехоустойчивость обрабатываемых сигналов. Диапазоны сигналов V_O (выход источника сигнала) для логических нуля и единицы уже, чем аналогичные диапазоны V_I (вход приемника сигнала), что также компенсирует возможные пульсации напряжения при передаче в линиях связи между логическими элементами, источником и приемником.

одной плоскости и второй набор параллельных шин, перпендикулярно ориентированных в другой параллельной плоскости. В узлах кроссбара размещаются мемристоры с программируемыми (самонастраиваемыми на основе локальных правил) значениями проводимости в паре с селекторами – элементами, обеспечивающими корректную адресацию при доступе к мемристорам (рис. 10) [5].

С одной стороны, аналоговое представление и обработка информации в отсутствие тактирования, характерного для современных процессоров и сопроцессоров на архитектуре фон Неймана, обеспечивает максимальное быстродействие и отсутствие конвейерной задержки при получении результата. С другой стороны, цифровое представление информации в виде логических нулей и единиц обеспечивает высокий уровень помехозащищенности в связи с тем, что весь размах напряжения питания разбивается на 3 зоны (рис. 11), средняя из которых не используется и минимизирует количество возможных ошибок. Использование же именно аналоговой, непрерывной шкалы амплитуды обрабатываемых сигналов автоматически накладывает ограничения на размерность кроссбара.

Мемристорные кроссбары являются основой аппаратного аналогового выполнения математических операций, присущих различным архитектурам нейроморфных устройств. В частности,

они позволяют выполнять ВМУ, занимающее наибольшую часть времени обработки данных в нейроморфных системах (инференса), параллельно для нескольких нейронов за один такт работы процессора с очень низким (пикоджоули) энергопотреблением. Однако потенциал высокого быстродействия и низкого энергопотребления не раскрывается автоматически – необходимо наиболее оптимальным образом организовать вычисления в системах на базе мемристорных кроссбаров. По аналогии с фон Неймановской архитектурой задача коммутирования сигналов и управления вычислительным устройством при её некорректном решении может стать "узким местом" нейроморфных систем.

Особенностью нейроморфных систем является то, что они подобно биологическим сетям нейронов на уровне обработки информации содержат большое число связанных друг с другом узлов, выполняющих принципиально одинаковые операции. Для реальных практических применений количество узлов (нейронов) может измеряться тысячами, а количество связей (синапсов) – миллионами. Отдельные мемристорные кроссбары, имея конкретное число мемристорных устройств, определяемое топологией кристалла и существующими технологическими ограничениями, физически реализуют лишь часть связей между нейронами разных слоев, при этом несколько кроссбаров могут относиться к одним и тем же нейронам. В данных условиях разрабатываемые архитектуры обязательно должны быть масштабируемыми.

Масштабируемость нейроморфных систем на базе мемристоров логически должна быть реализована как на уровне архитектуры нейроморфной модели – "по горизонтали" (для обеспечения необходимого числа слоев нейронов) и "по вертикали" (для обеспечения необходимого числа входов нейронов), так и на уровне распараллеливания потоков обработки данных несколькими одинаковыми по архитектуре нейромофрными моделями. При этом физически такое масштабирование также имеет несколько уровней – увеличение количества кроссбаров в одном нейропроцессоре, увеличение количества нейропроцессоров и объединение их в кластер и так далее. Базовым требованием для каждого уровня масштабирования является сохранение высокого уровня распараллеливания коммутирования сигналов для их одновременной подачи на эквивалентный кроссбар (единичный кроссбар или несколько кроссбаров, объединенных "по горизонтали" и "по вертикали") и управления ключами и селекторами.

3. ПОДХОДЫ К МАСШТАБИРОВАНИЮ НЕЙРОМОРФНЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ НА ОСНОВЕ МЕМРИСТОРОВ

Рассмотрим различные варианты масштабирования активных мемристорных кроссбаров в интегральном исполнении для повышения скорости передачи сигналов в мемристорных нейронных сетях со статическим и импульсным кодированием [39].

В классической архитектуре фон Неймана для хранения данных и вычислений используются отдельные устройства – оперативное запоминающее устройство (ОЗУ) и ЦПУ соответственно. Принципиально медленная динамическая память (англ.: Dynamic Random Access Memory (DRAM)) ограничивает скорость чтения/записи информации как исходных, так и результирующих данных вычислительного процесса. В связи с этим при вычислениях в памяти (англ.: In-Memory Computing (IMC)) отдельный чип снабжается собственной памятью и вычислительной схемой, которыми управляет чип ЦПУ – рис. 12 [5].

Вычислительный процесс организован следующим образом. Процессорное устройство при необходимости обновляет весовые коэффициенты в мемристорном кроссбаре, загружает входную матрицу во встроенную память (англ.: embedded DRAM (eDRAM)) в чипе для вычислений в памяти, выдает команду начала вычислений. Данные из eDRAM перемещаются во входной контур и преобразуются в напряжения, необходимые для работы мемристорного кроссбара. Каждый столбец мемристорного кроссбара обеспечивает суммирование произведений входного напряжения на проводимость мемристора в виде тока, выполняя аналоговую реализацию умножения с накоплением в памяти. В выходном контуре результаты преобразуются в выходную, результирующую матрицу и сохраняются в eDRAM для дальнейшего использования процессорным устройством в вычислительном процессе.

Входные и выходные контуры, обслуживающие работу мемристорного кроссбара, реализуются с помощью цифровых схем с применением аналого-цифровых и цифро-аналоговых преобразователей (АЦП и ЦАП соответственно), выполненных по КМОП-технологии – рис. 13 [5].

Наиболее простые двоичные нейронные сети требуют относительно небольшого процента КМОП-обрабатывающих схем в общей с учетом мемристорного кроссбара аппаратной реализации. Современный уровень развития проектирования и изготовления мемристоров отражает на-

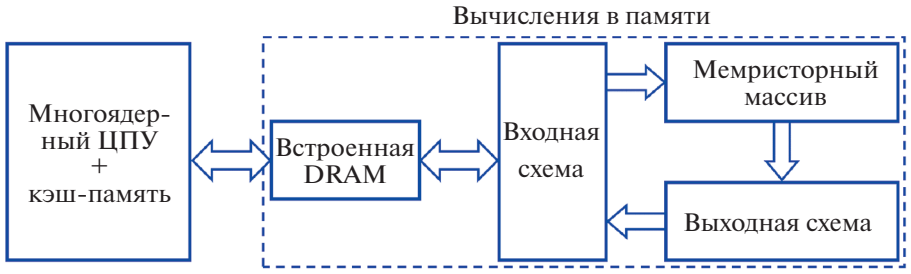

Рис. 12. Отдельный чип вычислений в памяти на основе мемристорного кроссбара, снабженный встроенным динамическим ОЗУ; входной и выходной контуры обеспечивают преобразование двоичных сигналов в напряжение и обратное преобразование результирующих токов в напряжение. Управление осуществляется многоядерным процессорным устройством с кэш-памятью, реализованными на другом чипе.

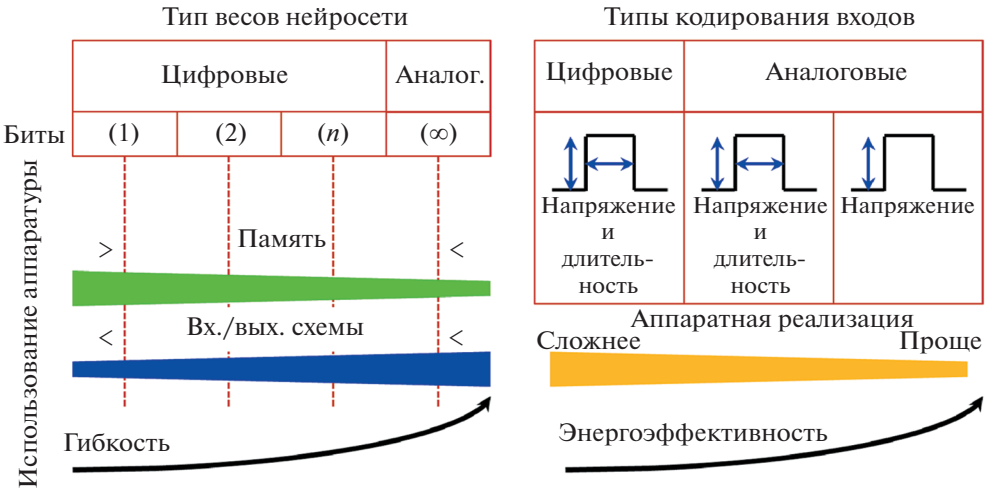

Рис. 13. Гибкость и энергоэффективность максимальны при использовании аналоговой обработки; аппаратная реализация постобработки входного сигнала (напряжения) также упрощается при переходе от цифрового представления к аналоговому.

личие таких приборов именно с двумя уровнями хранения информации. Бинарные сети весьма энергоэффективны, но способны решать достаточно простые задачи, например, несложную предобработку и обработку звуков и речи.

Переход к трем и более разрядам с одной стороны открывает возможность применения более сложных нейросетей, но и влечет за собой увеличение доли КМОП-схем в общей аппаратной реализации. Технологии таких многоуровневых мемристоров в данный момент активно разрабатываются [35].

Применение весов неограниченной (аналоговой) точности требует применения, с одной стороны, соответствующих мемристоров, которые широко не доступны в настоящее время, и, с другой стороны, значительного объема высокоточной КМОП элементной базы для обеспечения цифро-аналоговой и аналого-цифровой поддержки работы мемристорного кроссбара. Неоспоримым преимуществом нейронных сетей при такой реализации является высокая степень точности, достигаемая при их работе ввиду отсутствия необходимости сокращения разрядности весовых коэффициентов при конвертации модели в аппаратную реализацию.

Каждый столбец наиболее простой в реализации схемы с аналоговым кодированием амплитуды входного сигнала и мемристором без селектора-транзистора (0T1R) в узле кроссбара обслуживается ОУ с резистором в цепи обратной связи (рис. 13, аналоговое кодирование напряжения). Добавление длительности во входной сигнал требует функции интегрирования ОУ (рис. 13, аналоговое кодирование напряжения и длительности). Для обработки сигналов в полноценном мемристорном кроссбаре с элементами 1T1R и оцифрованных квантованной амплитуды и дискретизированной длительности входного сигнала потребуется наиболее сложная электрическая

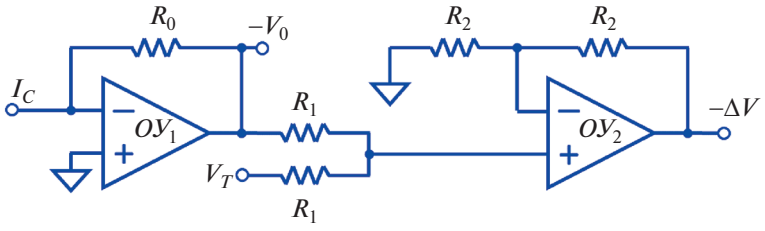

Рис. 14. Пример реализации функции активации на ОУ для полносвязного слоя нейронной сети при поступлении суммарного тока столбца I_c: V_0 – активированный выход, V_T – целевое значение для V_0, ΔV – ошибка рассогласования между V_0 и V_T.

схема с применением компаратора и счетчика (рис. 13, цифровое кодирование напряжения и длительности).

Функция активации также реализуется на схемах с применением ОУ (рис. 14). Схема, содержащая 2 ОУ и набор резисторов, обслуживает один столбец мемристорного кроссбара [40].

При хорошо проработанных вариантах электрических схем обвязки мемристорных кроссбаров ограничивающим фактором по увеличению их размерности является наличие паразитных путей в таких кроссбарах. Проблема заключается в том, что ток кроме заданного пути распространения строка-столбец также протекает и через смежные нежелательные пути. В работе [41] проведен анализ данной проблемы: вычислены отношения размаха напряжений в кроссбаре к размаху напряжений в одном мемристоре в зависимости от хранимых значений в кроссбаре и заземления строк и столбцов. Наличие паразитных путей существенно зависит от хранимых в мемристорном кроссбаре значений. Проведено сравнение зависимости параметра Δ', равного отношению разности напряжения питания и напряжения нуля (земли), для случая всего кроссбара к аналогичной разности для одного мемристора. В идеальном случае результат равен единице, в иных – менее единицы. Результаты моделирования показывают, что заметное падение анализируемого параметра наблюдается даже при относительно небольших размерностях матрицы 16×16 и 64×64.

Для решения проблемы, связанной с наличием паразитных путей, рассматривается ряд способов. Первый способ называется многостадийным чтением и включает в себя пять шагов: измерение тока целевой ячейки, установка целевой ячейки в состояние с высоким сопротивлением (англ.: High Resistance State (HRS)) и измерение для нее тока, аналогичная операция для состояния с низким сопротивлением (англ.: Low Resistance State (LRS)), сравнение измеренных токов, возврат ячейки в исходное состояние. Второй способ – развертывание архитектуры по принципу разделения столбцов для каждого мемристора. Третий и четвертый способы – применение диода и транзистора в качестве селектора (1D1R и 1T1R). Пятый способ – применение комплементарных мемристоров, обеспечивающих постоянное сопротивление $R_{LRS} + R_{HRS}$, существенно снижающих паразитные токи. Хотя ячейка 1T1R занимает большую площадь, а для управления затвором транзистора требуется дополнительный проводник, такой способ наиболее распространен.

В различных схемах кроссбаров используют как дублирование элементов для достижения заданной функциональности и повышения производительности, так и наоборот, мультиплексирование элементной базы для повторного ее использования для выполнения различных функций с разделением во времени.

Так, состояния HRS и LRS в бинарных RRAM положительны, поэтому для кодирования знакопеременных весов используют операцию “XNOR” и удваивают количество строк мемристорных кроссбаров (рис. 15) [30]. Для обработки сигналов столбцов мемристорной матрицы используются прецизионные инструментальные усилители в режиме напряжения, которые разделяются для обработки сигнала одного из восьми столбцов (битовой линии) с помощью мультиплексора. В первом случае налицо дублирование аппаратуры, во втором – экономия транзисторного бюджета за счет увеличения времени обработки сигналов. Применяются различные схемы адаптивной компенсации больших или малых значений токов в опрашиваемых мемристорах – например, схема управления напряжением подтяжки [42].

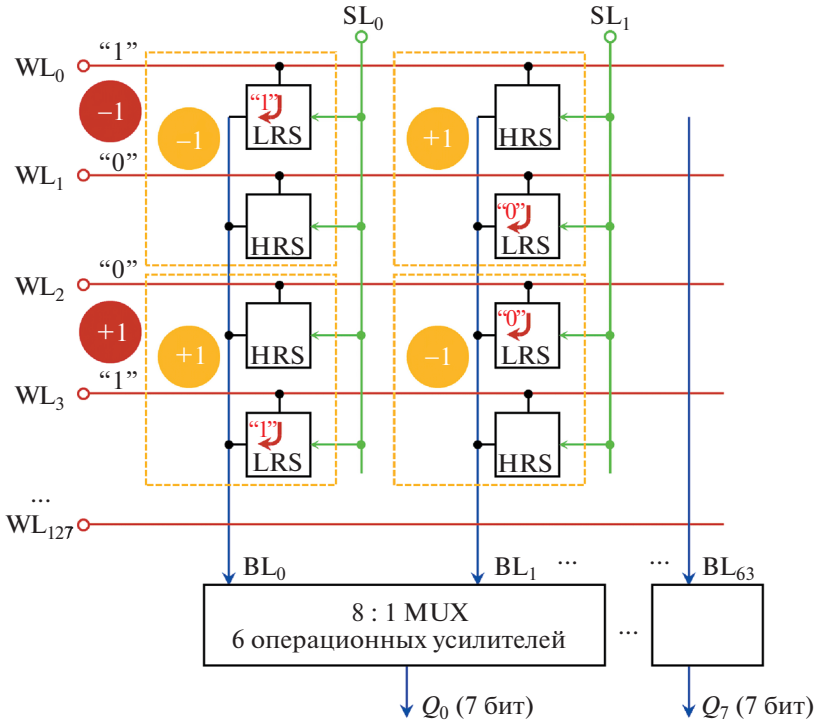

Рис. 15. Пример реализации мемристорной матрицы эффективной размерности 64 × 64: линии источника (англ.: Source Line (SL)), битов (англ.: Bit Line (BL)) и слов (англ.: Word Line (WL)); значение входного сигнала "–1" кодируется парой 1 и 0, значение "+1" – парой 0 и 1; вес "–1" кодируется парой LRS и HRS, вес "+1" – парой HRS и LRS. Битовые линии по 8 штук собираются в блок-обработки. На мультиплексоре выбирается очередная битовая линия, производится оцифровка (128 уровней) ее значения с помощью инструментальных усилителей.

Активное развитие схем на мемристорных кроссбарах, с одной стороны, подтверждается ростом размерности кроссбаров, с другой стороны – предложениями по переносу нейросетей на аппаратную реализацию таких процессоров.

Например, в [18] предложен процессор NeuRRAM с многоуровневой организацией процессорных устройств. На верхнем уровне реализуемая аппаратно нейросеть отображается на такой процессор, состоящий из 48 ядер, организованных в виде матрицы из 8 строк на 6 столбцов (рис. 16). Каждое ядро состоит из матрицы корлетов (англ.: corelet) размерностью 16 × 16. Эта матрица называется двунаправленный транспонируемый нейросинаптический массив (англ.: bidirectional Transposable NeuroSynaptic Array (TNSA)). Это означает, что исходные сигналы благодаря обслуживающим КМОП-схемам могут быть поданы как на строки, так и на столбцы. На этапе ввода матрично-векторного умножения драйверы преобразуют входы регистров (REG) и входы PRN в аналоговые напряжения и передают их на TNSA; на этапе вывода матрично-векторного умножения драйверы передают цифровые выходы с нейронов обратно на регистры через REG. Кроме того, в КМОП-схемах реализованы различные функции активации, включая стохастические.

На нижнем уровне корлет состоит из матрицы мемристоров размерностью 16 × 16 и одного КМОП-нейрона. Нейрон подключается к одной из 16 битовых линий и к одной из 16 линий выбора источника, проходящих через корлет. Он отвечает за интеграцию входов от всех 256 RRAM, подключенных к одной BL или SL: 16 RRAM в текущем корлете и 240 RRAM в других корлетах вдоль той же строки/столбца. Благодаря развитой системе маршрутизации каждое ядро способно выполнять прямое, обратное и рекуррентное матрично-векторное умножение всех 256 строк.

Рассмотренные выше мемристорные кроссбары с КМОП-схемами управления реализуются в виде монолитных микросхем с топологическими нормами 90 и 130 нм. Как было отмечено выше, КМОП-схемы управления размещаются в слое FEOL, а мемристорный кроссбар – между слоями металлизации в слое BEOL или поверх (рис. 4). Однако существует и другое, сравнительно новое направление реализации сложных приборов, в том числе и для случаев, когда их части из-

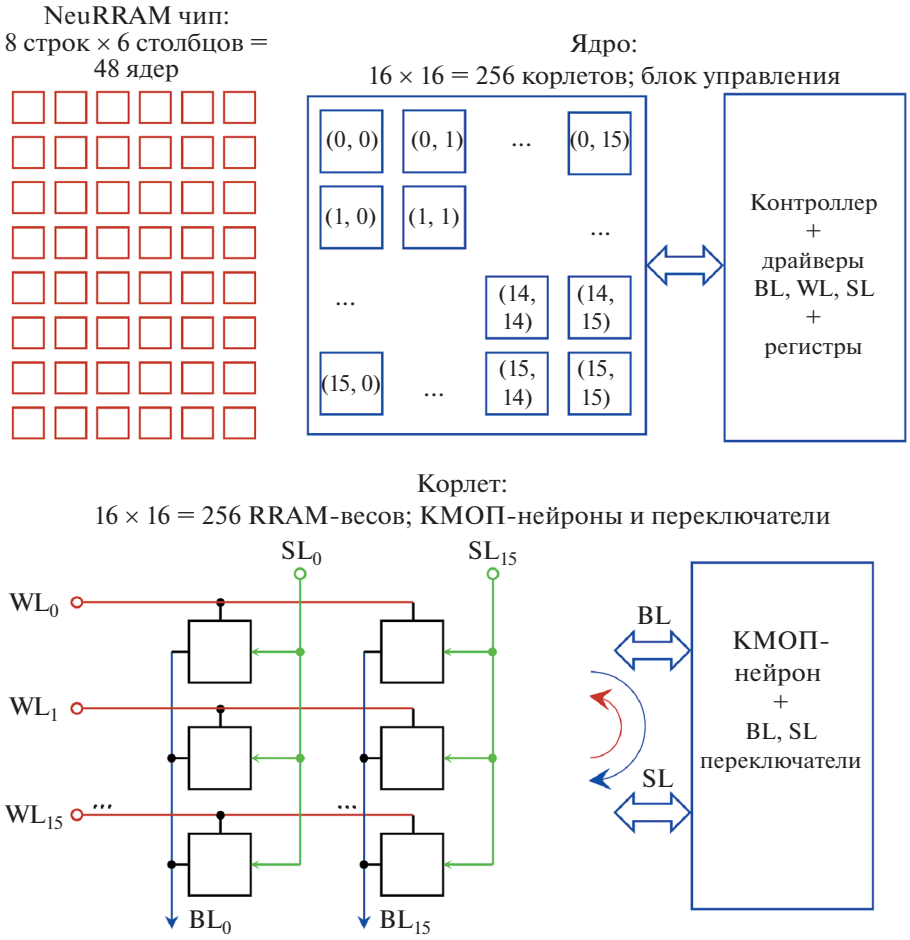

Рис. 16. Проект NeuRRAM [18]. Для работы нейросеть отображается на 48 ядер чипа одним из 6 способов: (1) 1 слой в 1 ядро, (2) дублирование в несколько ядер для увеличения пропускной способности, (3) несколько слоев в одно ядро, (4) переупорядочивание в одно ядро для увеличения степени использования, (5) и (6) распараллеливание на несколько ядер. Каждое ядро состоит из 256 корлетов, каждое из которых содержит 16 × 16 RRAM-весов и КМОП-нейрон. Одноразрядные BL- и SL- переключатели в составе корлета способны изменять направление обрабатываемого КМОП-нейроном сигнала от BL к SL или обратно.

готавливаются по разным и возможно несовместимым технологиям. В примере выше слой оксида в мемристоре может быть разрушен высокой для него температурой при формировании вышележащих слоев по КМОП-технологии – превышением температурного бюджета [43].

Идея разбиения большого по площади чипа на совокупность отдельных мини-чипов – чиплетов (англ.: chiplet) с их последующим размещением и соединением “сторона-к-стороне” на плоскости подложки-интерпозера (2.5D-интеграция) или в виде стека (этажерки, 3D-интеграция) с соединением вертикальными проводниками (англ.: Through-Silicon Via (TSV)) берет начало в 2015 г. [44].

Каждый чиплет обычно является модулем системы, выполненный по несовместимой с другими технологиями, либо реализующий сложно-функциональный блок. Паскаль Вивет из Европейского центра исследований в области микроэлектроники (англ.: Laboratory of Electronics and Information Technologies (LETI)) считает, что “Экосистемы на базе чиплетов будут быстро внедряться в высокопроизводительные вычисления и различные другие сегменты рынка, например, встраиваемые высокопроизводительные вычисления...” [44]. Центр LETI представил технологию активного интерпозера для чиплетов, с помощью которой собирается конструкция из 6 чиплетов в общей сложности с 96 ядрами (рис. 17).

В технологии чиплетов пока недостаточно хорошо решены вопросы по сборке, тестированию и выходу годных чиплетов, а также поддержке систем автоматизированного проектирования. Однако ведутся обширные работы по унификации технологий межкристалльных интерфейсов –

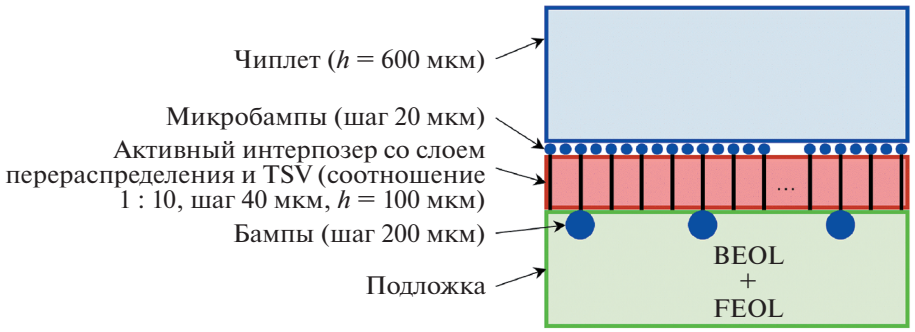

Рис. 17. Активный интерпозер, представленный центром LETI [44] для объединения 96 ядер на 6 чиплетах; активный интерпозер с помощью RDL (англ.: redistribution layer – слой перераспределения) позволяет совместить интерпозер с шагом бампов (контактных шариков) 200 мкм (внизу) с шагом микробампов 20 мкм (вверху на чиплете).

Advanced Interface Bus (AIB) фирмой Intel, CEI-112G-XSR форумом Optical Internetworking Forum, BoW (Bunch of Wires) и OpenHBI (High Bandwidth Interface) в проекте Open Domain-Specific Architecture.

Серьезность технологии чиплетов подтверждается участием таких известных фирм как Boeing, Cadence, Synopsys, Intel, Micron и других в проекте "Общие стратегии интеграции гетерогенных чипов и повторного использования сложно-функциональных блоков" (англ.: Common Heterogeneous Integration and IP Reuse Strategies) с 2017 г., а также GE, Intel, Keysight, Xilinx и других в проекте "Программа передовой упаковки гетерогенных чипов" (The State of The Art (SOTA) Heterogeneous Integrated Packaging) с 2019 г. для формирования стандартов интерфейса между чиплетами и обеспечения сборки систем из сложно-функциональных блоков. Оба проекта реализуются Управлением перспективных исследовательских проектов Министерства обороны США (англ.: Defense Advanced Research Projects Agency (DARPA)).

Реальным примером использования такой технологии для вычислителей в памяти на мемристорных кроссбарах является разработка основанного на чиплетах масштабируемого ускорителя вычислений в памяти для глубоких нейросетей (англ.: "Chiplet-based Scalable In-Memory Acceleration with Mesh for Deep Neural Networks" (SIAM)) [45], (рис. 18). На начальном этапе осуществляется переход от нейросети к архитектуре, при котором учитываются: режим чипа IMC, частота сети в корпусе (англ.: Network on Package (NoP)), размеры и количество чиплетов, отображение IMC, количество тайлов на чиплет, размер кроссбара, тип ячеек памяти, топологические нормы, размер аккумулятора; программа в составе системы автоматизированного проектирования (САПР) для разбиения и отображения: внутреннее планирование чиплета, размещение чиплетов; программы в составе САПР для NoP и ОЗУ; отображение на IMC-тайлы, внешнее планирование чиплета, трассировка и размещение; программу в составе САПР для электрической цепи и сети на кристалле; получение разбиения чипа как на рис. 18 слева.

4. СРАВНЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ НА ОСНОВЕ ТРАДИЦИОННОЙ И НОВОЙ ЭЛЕМЕНТНОЙ БАЗЫ

Рассмотрим более подробно результаты сравнения известных графических и нейроморфных процессоров на основе традиционных цифровых элементов с прототипами нейроморфных процессоров на базе мемристорных устройств. Для сравнения будем использовать два абсолютных критерия – количество ячеек и пиковая производительность (гигаопераций в секунду (ГОПС)), и два относительных – производительность на площадь чипа (ГОПС/мм^2) и производительность на ватт потребляемой энергии – энергоэффективность (ГОПС/Вт). Данные критерии рассчитаны для инференса нейронных сетей, при котором базовой операцией является ВМУ. Результаты сравнения показаны на рис. 19.

Для данного сравнения были выбраны специализированные нейроморфные процессоры Altai [46] и Tianjic [47], оптимизированные под работу спайковых нейронных сетей и самый производительный графический процессор от NVIDIA – Tesla V100 [48], являющийся по отношению к предыдущим более универсальным вычислителем, так как позволяет решать широкий круг задач из области обработки данных. Значения показателей брались из открытых источников (ссылки

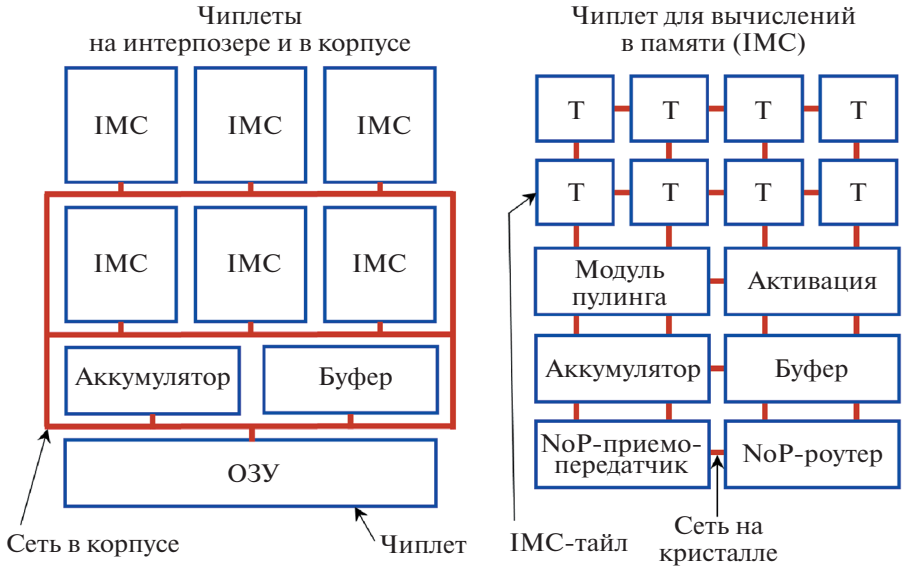

Рис. 18. Проект SIAM [45]: реализующие функции вычислителя чиплеты, глобальные аккумулятор и буфер, память DRAM размещены и соединены на интерпозере в составе корпуса (Package) и соединены NoP (слева). Каждый чиплет IMC состоит из IMC-тайлов, модулей расчета, связи и маршрутизации (справа). Каждый тайл состоит из нескольких процессорных элементов (PE), мультиплексора, АЦП, инструментального ОУ, устройства сдвига и сложения, буфера; каждый PE содержит мемристорный кроссбар (на рисунке не показано).

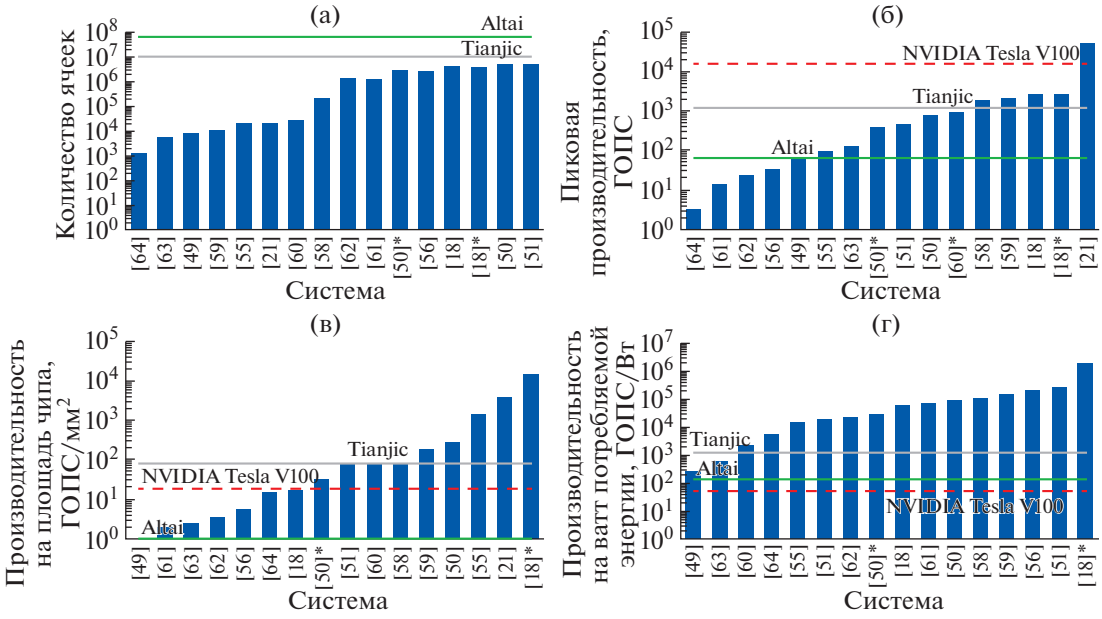

Рис. 19. Результаты сравнения вычислительных систем на основе мемристорных устройств (столбцы диаграмм) с нейроморфными (горизонтальные сплошные линии) и графическим (горизонтальная пунктирная линия) процессорами на основе традиционных цифровых элементов по следующим критериям: количество ячеек (а), пиковая производительность (б), производительность на площадь чипа (в) и производительность на ватт потребляемой энергии (г).

в подписи горизонтальной оси на рис. 19) так, как это было указано авторами. Все прототипы вычислителей на базе мемристорных устройств, выбранные для сравнения, выполнены по КМОП-совместимой технологии и имеют приборный слой с селекторами-транзисторами (за исключением [49]) и другой необходимой для работы электроникой.

Как видно из рис. 19а, вычислительные системы на основе мемристорных устройств имеют значительно меньшее количество ячеек, чем существующие процессоры. Однако это не является недостатком и объясняется тем, что в настоящее время представленные в обзоре системы пока еще прототипы, создающиеся в результате исследований и разработок. Тем не менее, даже такие относительно небольшие вычислители, имеющие до 4 млн ячеек, демонстрируют достаточно высокую производительность, в абсолютных величинах обгоняя процессоры Altai и Tianjic с 67 и 10 млн синапсов соответственно (см. рис. 19б).

Наиболее наглядно преимущества вычислительных систем на основе мемристорных устройств проявляются при сравнении по относительным критериям. Высокий потенциал по миниатюризации мемристорных устройств (до единиц нм) и ячеек RRAM (достаточно 1–2 транзисторов) позволяет эффективнее использовать площадь чипа, как это видно из рис. 19в. Например, “RAND-chip” (Resistive Analog Neuro Device [50]), выполненный по 40 нм технологии, имеет площадь 2.71 мм^2 при плотности размещения в 1.48 М синапсов на мм^2 с драйверами, контроллерами и мультиплексорами и при этом обеспечивает в 3 раза большую относительную производительность, чем процессор Tianjic, и в 12.6 раза большую производительность, чем NVIDIA Tesla V100. В свою очередь энергоэффективность вычислительных систем на основе мемристорных устройств на 2–3 порядка величины лучше, чем у существующих процессоров (см. рис. 19г). Например, чип “nvCIM macro” [51], выполненный по 22 нм технологии, демонстрирует 12–150-кратное преимущество по энергопотреблению перед Tianjic и 300–3700-кратное перед NVIDIA Tesla V100.

С взрослением технологии создания нейропроцессоров на основе мемристорных устройств увеличится и число ячеек, а это означает, что при большей плотности вычислений пиковая производительность превзойдет параметры производительности нейроморфных процессоров на основе традиционных цифровых элементов и специализированных архитектур, представленных на рис. 19б. Конечно, этот рост не сможет быть бесконечно большим, и на потенциально высокие производительность и энергоэффективность будут больше влиять конструктивные решения на уровне архитектуры процессоров и вычислительных систем, особенно растущие накладные расходы на маршрутизацию и ввод / вывод данных в цифровой форме (см. также раздел 3). Например, если говорить о потоковой обработке сигналов разной природы, производительность будет ограничиваться характеристиками сенсоров и интерфейсов передачи информации, поэтому для таких задач в настоящее время разрабатываются устройства “вычисления в сенсорах” с непосредственной передачей информации в аналоговом виде к вычислительному устройству на базе мемристоров [17, 52].

В данное сравнение не был включен ряд других ярких примеров прототипов вычислительных систем на основе мемристорных устройств, так как в публикациях авторы не всегда указывают значения критериев, использованных на рис. 19. Помимо них существуют более узкоспециализированные критерии, позволяющие оценивать производительность и энергоэффективность применительно к особенностям архитектуры вычислителя или решаемой им частной задачи. Это такие критерии, как количество гига или тера синаптических операций в секунду [47, 53] и количество гига- или тераопераций в расчете на 1 Мб RRAM [18], тера аналоговых вычислений в памяти в расчете на ватт потребляемой энергии [54], количество обработанных фреймов на ватт [21], произведение энергии на задержку [18]. Кроме этого, некоторые авторы используют программные симуляторы (например, “XPEsim” [55]) для оценки показателей производительности и энергоэффективности чипов на основе RRAM из-за высокой стоимости прототипирования. В будущем, интересным критерием сравнения систем вычислений в памяти будет цена 1 k/M/G байта памяти.

Ускорители нейроморфных вычислений (стандартные цифровые микросхемы ASIC на основе КМОП, системные решения и микросхемы на основе мемристоров), представленные на рис. 19, сравнивались по производительности и энергоэффективности с учетом того, что они подтвердили высокую (сопоставимую с программной эмуляцией) точность инференса моделей нейронных сетей при решении конкретных задач распознавания образов, классификации, сегментации и т.д. В табл. 1 занесены числовые значения характеристик вычислительных систем на основе мемристорных устройств с указанием задачи, архитектуры модели нейронной сети и достигнутых значений метрики точности.

Из табл. 1 видно, что рассматриваемые вычислители справляются на высоком уровне с общепринятыми тестовыми задачами классификации изображений из датасетов MNIST с точностью от 90.8 [50] до 99% [18], CIFAR-10 – от 85.7 [18] до 95.19 [51], CIFAR-100 – 65.71% [56], с 84.7% вероятностью распознают голосовые команды Google [18] и успешно решают другие задачи, при

Таблица 1. Численные характеристики вычислительных систем на основе мемристорных устройств

Ссылка	КМОП-технология	Тип мемристорной ячейки	Количество ячеек	Размер кроссбара	Пиковая производительность, ГОПС	Производительность на ватт потребляемой энергии, ГОПС/Вт	Производительность на площадь чипа, ГОПС/мм^2	Продемонстрированная точность в приложениях
[18]*	7 нм	1T1R	3M	16 × 16	2135	1360000	12800	99% MNIST, 85.7% CIFAR-10, 84.7% Google speech
[51]	22 нм	1T1R	4M	1024 × 512	394	194000	65.7	92.01–95.19% CIFAR-10
[50]	40 нм	1T1R	4M	n/a	660	66500	240	90.8% MNIST (MLP)
[18]	130 нм	1T1R	3M	16 × 16	2135	43000	13.4	see row [18]*
[58]	130 нм	2T2R	159k	n/a	1500	78400	71	94.4% MNIST (MLP)
[21]	130 нм	1T1R	16k	128 × 16	41900	14900	3100	40.21 dB PSNR and 22.38 dB SNR for MRI and CT images
[59]	2 мкм	1T1R	8k	128 × 64	1640	119700	150	n/a
[56]	22 нм	1T1R	2M	512 × 512	29	146000	4.8	90.88% CIFAR-10 (ResNet-20), 65.71% CIFAR-100 (ResNet-20)
[50]*	180 нм	1T1R	2M	n/a	330	21000	26	see row [50]
[60]	130 нм	1T1R	18k	256 × 16	780	1650	690	n/a
[55]	130 нм	1T1R	16k	128 × 16	81	11000	1160	96.92% MNIST (CNN)
[61]	55 нм	1T1R	1M	512 × 256	12	53170	1.6	88.52% CIFAR-10 (CNN)
[62]	65 нм	1T1R	1M	512 × 256	19	16950	3	98.8% MNIST (LeNet DNN)
[63]	150 нм	1T1R	4k	32 × 32	101	462	2	n/a
[64]	130 нм	2T2R	1k	32 × 32	2.7	4200	13	98.4% MNIST (MLP), 87% CIFAR-10 (CNN)
[49]	180 нм	0T1R	6k	54 × 108	57	187.6	0.9	94.6% breast cancer screening dataset

этом реализуя известные архитектуры нейронных сетей, такие как MLP (Multilayer Perceptron), DNN (Deep Neural Network), CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory) и модели ResNet-20, ResNet-50, VGG16.

Стоит отметить, что среди рассматриваемых прототипов наиболее универсальным с точки зрения возможности запуска разных архитектур нейронных сетей является чип NeuRRAM [18]. Как видно из рис. 19 и табл. 1, NeuRRAM уже сейчас при проектных нормах 130 нм имеет в 33–800 раз лучшую энергоэффективность, чем процессоры Tianjic, Altai и NVIDIA Tesla V100, и обеспечивает высокую относительную производительность в сравнении с ними. При этом выигрыш на много порядков величины по указанным и другим параметрам ожидается при масштабировании технологии до проектных норм 7 нм от норм на уровне 90–130 нм, которые используются сейчас при создании прототипов многоядерных процессоров на основе мемристорных устройств в структуре МОМ.

Таким образом, вычисления в памяти в настоящее время это единственный путь к повышению производительности и снижению энергопотребления вычислительных систем ИИ, поскольку с функциональной точки зрения – это наиболее биоподобный принцип обработки информации, а с точки зрения архитектуры он позволяет значительно сократить расстояние передачи данных и требуемый объем памяти (параметры модели хранятся в вычислителе постоянно), а также затрачиваемую на выполнение ВМУ энергию. Для вычислений в памяти могут быть использованы разные виды памяти [57]: SRAM, DRAM, Flash, однако наиболее подходящей является RRAM, поскольку другие виды памяти имеют недостатки (такие как низкая масштабируемость, дороговизна и энергозависимость для SRAM, плохая технологическая совместимость с КМОП-процессом для процессоров и необходимость регенерации десятки раз в секунду для DRAM, сложности в реализации записи по произвольному адресу для Flash и т.д.) и накладывают существенные ограничения при создании нейроморфных чипов.

5. ЗАКЛЮЧЕНИЕ

Мемристоры – это очень простые устройства и в то же время очень умные и сложные нелинейные системы, обещающие широкий спектр приложений от микросхем памяти и нейроморфных вычислительных систем в памяти до адаптивных нейроинтерфейсов. Реализация нейроморфных вычислительных систем на основе новой элементной базы требует проведения скоординированных и междисциплинарных исследований/разработок разного уровня. В основе соответствующего научно-технического направления лежит сквозная технология мемристорных устройств и схем, обеспечивающая создание элементной базы новых мозгоподобных информационно-вычислительных систем с широким спектром применений. Продемонстрированные на данный момент перспективы связаны с монолитной интеграцией мемристорных устройств со схемами КМОП, а также совместной оптимизацией материалов, устройств и архитектур, необходимыми для создания демонстрационных прототипов информационно-вычислительных систем на основе мемристоров.

Различные варианты масштабирования активных мемристорных кроссбаров в интегральном исполнении обеспечивают повышение скорости передачи сигналов в мемристорных нейронных сетях со статическим и импульсным кодированием. Анализ схемотехнических решений на основе КМОП-элементной базы, обеспечивающих эффективную работу мемристорного кроссбара при обучении и инференсе, показывает рост эффективной размерности кроссбаров в опубликованных зарубежных проектах последних лет. Альтернативное монолитному интегральному исполнению решение также представлено в статье на разных примерах реализации по технологии чиплетов.

Сравнение нейроморфных вычислительных систем на основе традиционной и новой элементной базы показало, что уже сейчас существующие прототипы значительно (на порядки величины) опережают известные вычислительные системы на основе традиционной элементной базы по производительности и энергоэффективности без снижения точности векторно-матричного умножения и инференса искусственных нейронных сетей.

Исследование выполнено в рамках научной программы Национального центра физики и математики, направление № 9 “Искусственный интеллект и большие данные в технических, промышленных, природных и социальных системах”.

СПИСОК ЛИТЕРАТУРЫ

1. *Wang Z., Wu H., Burr G.W., et al.* // Nat. Rev. Mater. 2020. V. 5. P. 173–195. https://doi.org/10.1038/s41578-019-0159-3
2. *Christensen D.V., Dittmann R., Linares-Barranco B., et al.* // Neuromorph. Comput. Eng. 2022. V. 2. № 2. Art. № 022501. https://doi.org/10.1088/2634-4386/ac4a83
3. *Xia Q., Yang J.J.* // Nat. Mater. 2019. V. 18. P. 309–323. https://doi.org/10.1038/s41563-019-0291-x
4. *Zhang W., Gao B., Tang J., et al.* // Nat. Electron. 2020. V. 3. P. 371–382. https://doi.org/10.1038/s41928-020-0435-7
5. *Amirsoleimani A., Alibart F., Yon V., et al.* // Adv. Intell. Syst. 2020. V. 2. № 11. Art. № 2000115. https://doi.org/10.1002/aisy.202000115
6. *Ham D., Park H., Hwang S., Kim K.* // Nat. Electron. 2021. V. 4. P. 635–644. https://doi.org/10.1038/s41928-021-00646-1
7. *Huang Y., Kiani F., Ye F., Xia Q.* // Appl. Phys. Lett. 2023. V. 122. № 11. Art. № 110501. https://doi.org/10.1063/5.0133044
8. *Chua L.* // IEEE Trans. Circ. Theor. 1971. V. 18. № 5. P. 507–519. https://doi.org/10.1109/TCT.1971.1083337
9. *Vongehr S., Meng X.* // Sci. Rep. 2015. V. 5. Art. № 11657. https://doi.org/10.1038/srep11657
10. *Demin V.A., Erokhin V.V.* // Int. J. Unconv. Comput. 2016. V. 12. P. 433–438.
11. *Kim J., Pershin Y. V., Yin M., et al.* // Adv. Electron. Mater. 2020. V. 6. № 7. Art. № 2000010. https://doi.org/10.1002/aelm.202000010
12. *Shen J., Shang D., Chai Y., et al.* // Phys. Rev. Appl. 2016. V. 6. № 6. Art. № 064028. https://doi.org/10.1103/PhysRevApplied.6.064028
13. *Chua L.O., Kang S.M.* // Proc. IEEE. 1976. V. 64. № 2. P. 209–223. https://doi.org/10.1109/PROC.1976.10092
14. *Spagnolo M., Morris J., Piacentini S., et al.* // Nat. Photon. 2022. V. 16. P. 318–323. https://doi.org/10.1038/s41566-022-00973-5
15. *Pfeiffer P., Egusquiza I.L., Di Ventra M., et al.* // Sci. Rep. 2016. V. 6. Art. № 29507. https://doi.org/10.1038/srep29507
16. *Schegolev A.E., Klenov N.V., Soloviev I.I., et al.* // Nanotechnol. Russia. 2021. V. 16. P. 811–820. https://doi.org/10.1134/S2635167621060227
17. *Makarov V.A., Lobov S.A., Shchanikov S., et al.* // Front. Comput. Neurosci. 2022. V. 16. Art. № 859874. https://doi.org/10.3389/fncom.2022.859874
18. *Wan W., Kubendran R., Schaefer C., et al.* // Nature. 2022. V. 608. P. 504–512. https://doi.org/10.1038/s41586-022-04992-8
19. *Bianchi S., Muñoz-Martin I., Covi E., et al.* // Nat. Commun. 2023. V. 14. Art. № 1565. https://doi.org/10.1038/s41467-023-37097-5
20. *Wang S., Li Y., Wang D., et al.* // Nat. Mach. Intell. 2023. V. 5. P. 104–113. https://doi.org/10.1038/s42256-023-00609-5
21. *Zhao H., Liu Z., Tang J., et al.* // Nat. Commun. 2023. V. 14. Art. № 2276. https://doi.org/10.1038/s41467-023-38021-7
22. *Pulvermüller F., Tomasello R., Henningsen-Schomers M.R., Wennekers T.* // Nat. Rev. Neurosci. 2021. V. 22. P. 488–502. https://doi.org/10.1038/s41583-021-00473-5
23. *Chiolerio A., Chiappalone M., Ariano P., Bocchini S.* // Front. Neurosci. 2017. V. 11. Art. № 70. https://doi.org/10.3389/fnins.2017.00070
24. *Roy K., Jaiswal A., Panda P.* // Nature. 2019. V. 575. P. 607–617. https://doi.org/10.1038/s41586-019-1677-2
25. *George R., Chiappalone M., Giugliano M., et al.* // iScience. 2020. V. 23. № 10. Art. № 101589. https://doi.org/10.1016/j.isci.2020.101589
26. *Serb A., Corna A., George R., et al.* // Sci. Rep. 2020. V. 10. Art. № 2590. https://doi.org/10.1038/s41598-020-58831-9
27. *Zhu X., Wang Q., Lu W.D.* // Nat. Commun. 2020. V. 11. Art. № 2439. https://doi.org/10.1038/s41467-020-16261-1
28. *Liu Z., Tang J., Gao B., et al.* // Nat. Commun. 2020. V. 11. Art. № 4234. https://doi.org/10.1038/s41467-020-18105-4

29. *Liu Z., Tang J., Gao B., et al.* // Sci. Adv. 2020. V. 6. № 41. Art. № eabc4797. https://doi.org/10.1126/sciadv.abc4797
30. *Shchanikov S., Zuev A., Bordanov I., et al.* // Chaos, Solitons & Fractals. 2021. V. 142. Art. № 110504. https://doi.org/10.1016/j.chaos.2020.110504
31. *Guggisberg A.G., Koch P.J., Hummel F.C., Buetefisch C.M.* // Clin. Neurophysiol. 2019. V. 130. № 7. P. 1098–1124. https://doi.org/10.1016/j.clinph.2019.04.004
32. *Chiaradia I., Lancaster M.A.* // Nat. Neurosci. 2020. V. 23. P. 1496–1508. https://doi.org/10.1038/s41593-020-00730-3
33. *Mikhaylov A., Pimashkin A., Pigareva Y., et al.* // Front. Neurosci. 2020. V. 14. Art. no. 358. https://doi.org/10.3389/fnins.2020.00358
34. *Mehonic A., Kenyon A.J.* // Nature. 2022. V. 604. P. 255–260. https://doi.org/10.1038/s41586-021-04362-w
35. *Rao M., Tang H., Wu J., et al.* // Nature. 2023. V. 615. P. 823–829. https://doi.org/10.1038/s41586-023-05759-5
36. *Miranda E., Suñé J.* // Materials. 2020. V. 13. Art. no. 938. https://doi.org/10.3390/ma13040938
37. *Demin V.A., Nekhaev D.V., Surazhevsky I.A., et al.* // Neural Netw. 2021. V. 134. P. 64–75. https://doi.org/10.1016/j.neunet.2020.11.005
38. *Красников Г.Я.* // Наноиндустрия. 2020. Т. 13. № S5–1 (102). С. 13–19.
39. *Telminov O., Gornev E.* // 2022 6th Scientific School Dynamics of Complex Networks and their Applications (DCNA). 14–16 September 2022, Kaliningrad, Russian Federation. P. 278–281. https://doi.org/10.1109/DCNA56428.2022.9923302
40. *Liu X., Zeng Z.* // Complex Intell. Syst. 2021. V. 8. P. 787–802. https://doi.org/10.1007/s40747-021-00282-4
41. *Zidan M.A., Fahmy H.A.H., Hussain M.M., Salama K.N.* // Microelectronics J. 2013. V. 44. № 2. P. 176–183. https://doi.org/10.1016/j.mejo.2012.10.001
42. *Yin S., Sun X., Yu S., Seo J.-S.* // IEEE Trans. Electron Devices. 2020. V. 67. № 10. P. 4185–4192. https://doi.org/10.1109/ted.2020.3015178
43. *Zhuk M., Zarubin S., Karateev I., et al.* // Front. Neurosci. 2020. V. 14. Art. № 94. https://doi.org/10.3389/fnins.2020.00094
44. https://semiengineering.com/chiplet-momentum-rising/
45. *Krishnan G., Mandal S.K., Pannala M., et al.* // ACM Trans. Embed. Comput. Syst. 2021. V. 20. № 5s. Art. № 68. https://doi.org/10.1145/3476999
46. https://motivnt.ru/neurochip-altai/
47. *Pei J., Deng L., Song S., et al.* // Nature. 2019. V. 572. P. 106–111. https://doi.org/10.1038/s41586-019-1424-8
48. https://www.nvidia.cn/content/dam/en-zz/Solutions/Data-Center/tesla-product-literature/volta-architecture-whitepaper.pdf
49. *Cai F., Correll J.M., Lee S.H., et al.* // Nat. Electron. 2019. V. 2. P. 290–299. https://doi.org/10.1038/s41928-019-0270-x
50. *Mochida R., Kouno K., Hayata Y., et al.* // 2018 IEEE Symposium on VLSI Technology. 18–22 June 2018, Honolulu, HI, USA. P. 175–176. https://doi.org/10.1109/VLSIT.2018.8510676
51. *Hung J.M., Xue C.X., Kao H.Y., et al.* // Nat. Electron. 2021. V. 4. P. 921–930. https://doi.org/10.1038/s41928-021-00676-9
52. *Vasileiadis N., Ntinas V., Sirakoulis G.C., Dimitrakis P.* // Materials. 2021. V. 14. № 18. Art. № 5223. https://doi.org/10.3390/ma14185223
53. *Akopyan F., Sawada J., Cassidy A., et al.* // IEEE T. Comput. Aid. D. 2015. V. 34. № 10. P. 1537–1557. https://doi.org/10.1109/TCAD.2015.2474396
54. *Cai F., Yen S., Uppala A., et al.* // Adv. Intell. Syst. 2022. V. 4. № 8. Art. № 2200014. https://doi.org/10.1002/aisy.202200014
55. *Yao P., Wu H., Gao B., et al.* // Nature. 2020. V. 577. P. 641–646. https://doi.org/10.1038/s41586-020-1942-4
56. *Xue C.X., Chiu Y.C., Liu T.W., et al.* // Nat. Electron. 2021. V. 4. P. 81–90. https://doi.org/10.1038/s41928-020-00505-5
57. *Sebastian A., Le Gallo M., Khaddam-Aljameh R., Eleftheriou E.* // Nat. Nanotechnol. 2020. V. 15. P. 529–544. https://doi.org/10.1038/s41565-020-0655-z

58. *Liu Q., Gao B., Yao P., et al.* // 2020 IEEE International Solid- State Circuits Conference – (ISSCC). 16–20 February 2020, San Francisco, CA, USA. P. 500–502. https://doi.org/10.1109/ISSCC19947.2020.9062953

59. *Li C., Hu M., Li Y., et al.* // Nat. Electron. 2018. V. 1. P. 52–59. https://doi.org/10.1038/s41928-017-0002-z

60. *Wu T.F., Le B.Q., Radway R., et al.* // 2019 IEEE International Solid- State Circuits Conference – (ISSCC). 17–21 February 2019, San Francisco, CA, USA. P. 226–228. https://doi.org/10.1109/ISSCC.2019.8662402

61. *Xue C.-X., Chen W.-H., Liu J.-S., et al.* // 2019 IEEE International Solid- State Circuits Conference – (ISSCC). 17–21 February 2019, San Francisco, CA, USA. P. 388–390. https://doi.org/10.1109/ISSCC.2019.8662395

62. *Chen W.H., Dou C., Li K.X., et al.* // Nat. Electron. 2019. V. 2. P. 420–428. https://doi.org/10.1038/s41928-019-0288-0

63. *Su F., Chen W.-H., Xia L., et al.* // 2017 Symposium on VLSI Technology. 5–8 June 2017, Kyoto, Japan. P. T260–T261. https://doi.org/10.23919/VLSIT.2017.7998149

64. *Bocquet M., Hirztlin T., Klein J.-O., et al.* // 2018 IEEE International Electron Devices Meeting (IEDM). 1–5 December 2018, San Francisco, CA, USA. P. 20.6.1–20.6.4. https://doi.org/10.1109/IEDM.2018.8614639

ON THE WAY TO IMPLEMENTATION OF HIGH-PERFORMANCE IN-MEMORY COMPUTING BASED ON MEMRISTIVE ELECTRONIC COMPONENT BASE

A. N. Mikhaylov[1,#], E. G. Gryaznov[1], V. I. Lukoyanov[1], M. N. Koryazhkina[1], I. A. Bordanov[2], S. A. Shchanikov[1,3], O. A. Telminov[4], M. V. Ivanchenko[1], and V. B. Kazantsev[1,3]

[1]*Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia*

[2]*Murom Institute of Vladimir State University named after Alexander and Nikolay Stoletovs, Murom, Russia*

[3]*Moscow Institute of Physics and Technology, Moscow, Russian Federation*

[4] *Molecular Electronics Research Institute, Zelenograd, Russia*

[#]*e-mail: mian@nifti.unn.ru*

The article is devoted to the analysis of the current state and perspectives for the development of high-performance computing based on the principles of information storage and processing in biological neural networks, which are provided with the capabilities of a new electronic component base (ECB), represented by memristors (nonlinear resistors with memory or Resistive Random Access Memory (RRAM) elements). Memristors can be implemented on the basis of various materials and nanostructures that are compatible with the standard technological process of microelectronics and allow performing "in-memory computing". Naturally, such computing are implemented in neuromorphic systems that use the crossbar architecture to perform vector-matrix multiplication, in which memristors at the intersections of conductive buses act as synaptic weights – plastic connections between artificial neurons in a fully connected neural network architecture. The article considers general approaches to the development and creation of a new ECB based on RRAM technology compatible with complementary metal-oxide-semiconductor structures, the development of artificial neural networks and a neuroprocessor using memristor crossbar arrays as computational cores and scalable multi-core architectures for implementing both formal and spiking neural network algorithms. Technical solutions are described that provide hardware implementation of memristor crossbars of a sufficiently large dimension, as well as solutions that compensate for some of the deficiencies or fundamental limitations inherent in modern memristors at the stage of technology maturation. The analysis of performance and energy efficiency for the reported prototypes of such neuromorphic systems was carried out and a conclusion was made about a significant (by orders of magnitude) gain in terms of these parameters compared to computing systems based on traditional element base (including neuromorphic ones). The technological development of a new element base and the creation of memristor-based neuromorphic computing systems will not only ensure the timely diversification of hardware for the continuous development and mass implementation of artificial intelligence technologies, but will also allow us to set tasks for a completely new level of creating hybrid intelligence based on the symbiosis of artificial and biological neural networks. Among these tasks are the primary ones of creating brain-like self-learning spiking neural networks and adaptive neurointerfaces based on memristors, which are also discussed in the paper.