

ВАРИАНТЫ АРХИТЕКТУР КМОП- И МЕМРИСТОРНЫХ ПРОЦЕССОРОВ

Тельминов О. А.¹

¹АО «НИИМЭ», г. Москва, Российская Федерация

Рассматривается направление искусственного интеллекта, связанное с работой глубоких нейронных сетей. Различают классическую формальную модель нейрона, которая реализуется на процессорах с архитектурой фон Неймана (x86, ARM, RISC-V). Увеличение производительности может достигаться параллелизацией вычислений на многоядерных и многопроцессорных системах. Описаны варианты зарубежных и отечественных нейропроцессоров — специализированных процессоров, ускоряющих вычисления классических нейросетей. Такие процессоры изготавливаются по комплементарной металл-оксид-полупроводниковой (КМОП) технологии, основанной на работе взаимодополняющих пар транзисторов разных типов проводимости. Благодаря отлаженным промышленным процессам производства транзисторы обладают высокими надежностью и предсказуемостью параметров. Однако, снижение топологических норм приводит к необходимости поиска новых объемных конструкций таких транзисторов, учету квантовых эффектов, росту сложности и стоимости производства, увеличению тепловыделения при кратном росте плотности транзисторов. В качестве альтернативы КМОП технологии для простых с точки зрения вычислений процессоров рассматриваются возможности печатной электроники.

Поиск идей для более производительных и энергоэффективных процессоров осуществляется в области достижений в исследованиях мозга нейробиологами. Такие решения называются нейроморфными или подобными работе биологических нейронных сетей. К нейроморфным процессорам с точки зрения аппаратной реализации предъявляются требования по архитектуре, отличной от архитектуры фон Неймана, по кодированию сигнала, отличному от используемого в формальной модели, по замене элементной базы с КМОП на использующую новые физические принципы. Рассматриваются варианты архитектур для обработки спайковых сигналов как с элементной базой как на основе КМОП технологии, так и на комбинированной с мемристорными кроссбрами.

Ключевые слова – архитектуры процессоров, нейроморфные процессоры.

Благодарности: 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».

Архитектуры КМОП- и мемристорных нейроморфных процессоров

Тельминов О. А.¹

¹АО «НИИМЭ», г. Москва, Российская Федерация

Введение

Искусственный интеллект (ИИ) сначала реализовывался в виде баз знаний и экспертных систем, затем активно применялись методы машинного обучения — различные виды [1] регрессии, методы k средних и главных компонент (рис. 1). Искусственные нейронные сети состоят из упорядоченных в слои нейроны, которые с помощью связей объединяются в сеть. Различают входной, выходной и, при наличии, скрытые слои между ними. С 2012 г. наибольший интерес и эффективность демонстрируют именно глубокие нейронные сети, имеющие несколько скрытых слоев; машинное обучение в этом случае называется глубоким обучением.

Искусственный нейрон является простейшим вычислителем, а их нахождение в нейронной сети реализует параллельные вычисления [2, 3]. Именно распараллеливание действий на уровне архитектуры нейросети приводит к более быстрому получению результата. Поскольку реализация нейросети выполняется на процессоре, то для сохранения эффективности вычислений его архитектура в идеале должна полностью совпадать с архитектурой нейросети. Однако, размеры практически значимых глубоких сетей кратно превышают возможности по их полноразмерному размещению в современных процессорах, поэтому такие сети вычисляются по фрагментам с сохранением и последующим использованием промежуточных результатов.

На активное развитие математического, программного и аппаратного обеспечения для глубокого обучения обратило внимание правительство Канады и в марте 2017 г. выпустило первую в мире собственную стратегию развития ИИ. В нашей стране такая инициатива была

реализована в октябре 2019 г., в современной редакции отечественной стратегии развития ИИ [4] к технологиям ИИ относятся компьютерное зрение, обработка естественного языка, распознавание и синтез речи, интеллектуальная поддержка принятия решений, перспективные методы ИИ.

Одним из примеров мировых событий явилось открытое письмо от 22 марта 2023 г. с предложением о временной приостановке «гигантских ИИ-экспериментов» [5]. В частности, в нем декларируются требования к ИИ по точности, безопасности, интерпретируемости, прозрачности, надежности, согласованности, доверия и лояльности. В настоящее время наблюдается рост ограничений на использование передовых аппаратно-программных решений в ряде стран, что косвенно стимулирует собственные разработки в таких странах.

Ниже будут рассмотрены популярные архитектуры процессоров, применяемые в глубоком обучении. Хотя структура дальнейшего изложения подразумевает деление по виду моделей нейрона — формальной и спайковой, можно отметить применение устаревших архитектур в новейших разработках для спайковых нейросетей. Эти модели различаются видом представления сигнала и его математической обработкой.

1 Формальные нейросети и архитектуры процессоров

1.1 Формальные нейросети

Головной мозг человека по разным оценкам имеет мощность около 10-20 Вт, в пересчете на 1 см^2 — 10 мВт, тактовая частота работы составляет порядка 10 Гц.

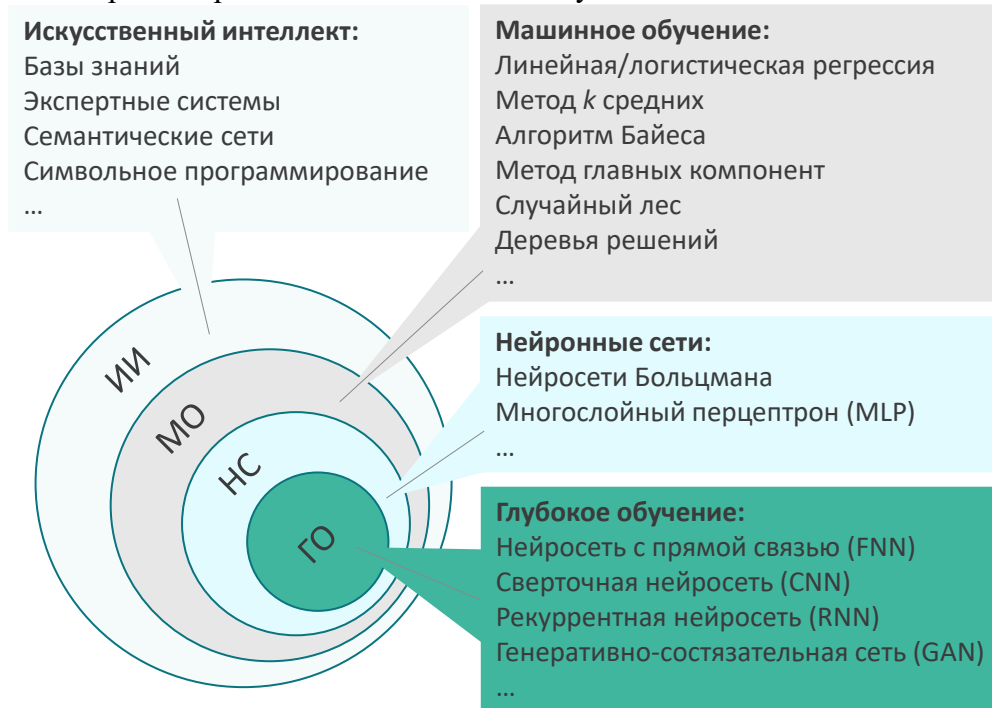


Рис. 1. Этапы развития ИИ: наиболее результативным и масштабируемым являются глубокие нейросети, в связи с чем их аппаратная реализация имеет наибольший интерес

К таким характеристикам стремятся разработчики современных процессоров, однако разрыв все еще велик. Мозг содержит порядка 86 млрд. нейронов и 100 трлн. связей между ними — синапсов [6].

Нейробиологи в ходе проведения исследований мозга открывают новые законы его функционирования. Специалисты в области математики и вычислительной техники обращаются к таким результатам и на основе модели искусственного нейрона разрабатывают математическое, программное и аппаратное обеспечение для работы нейросетей.

В настоящее время самой распространенной моделью нейрона является формальная модель, характеризующаяся непрерывным представлением сигнала. Математическая функция обработки сводится к взвешиванию входных сигналов путем их умножения на синаптические веса, последующему суммированию и прохождению через нелинейную функцию активации. Нелинейность может быть описана различными законами [7], однако чаще всего используется функция ReLU — нулевая в отрицательной полуплоскости и равная аргументу в остальной области. Настройка весовых коэффициентов

нейросети производится с помощью хорошо отработанного метода обратного распространения ошибки.

1.2. Классическая архитектура фон Неймана для процессоров

Наиболее популярной архитектурой построения вычислительных систем с разделенными процессором и оперативной памятью остается архитектура фон Неймана (1940-е гг.). Процессор состоит из устройства управления, арифметико-логического устройства и внутренней регистровой памяти. Реализуется процессор на транзисторах по КМОП-технологии, при этом на каждый бит памяти затрачивается 6 транзисторов, что делает такую память весьма затратным по площади на кристалле и по стоимости устройством — обычно процессор содержит всего порядка 32 регистров. Поэтому для хранения команд и данных используют более дешевую память — оперативное запоминающее устройство (ОЗУ), в котором значения битов хранятся в конденсаторах; размер современных ОЗУ достигает десятков ГБ. Обратной стороной медали

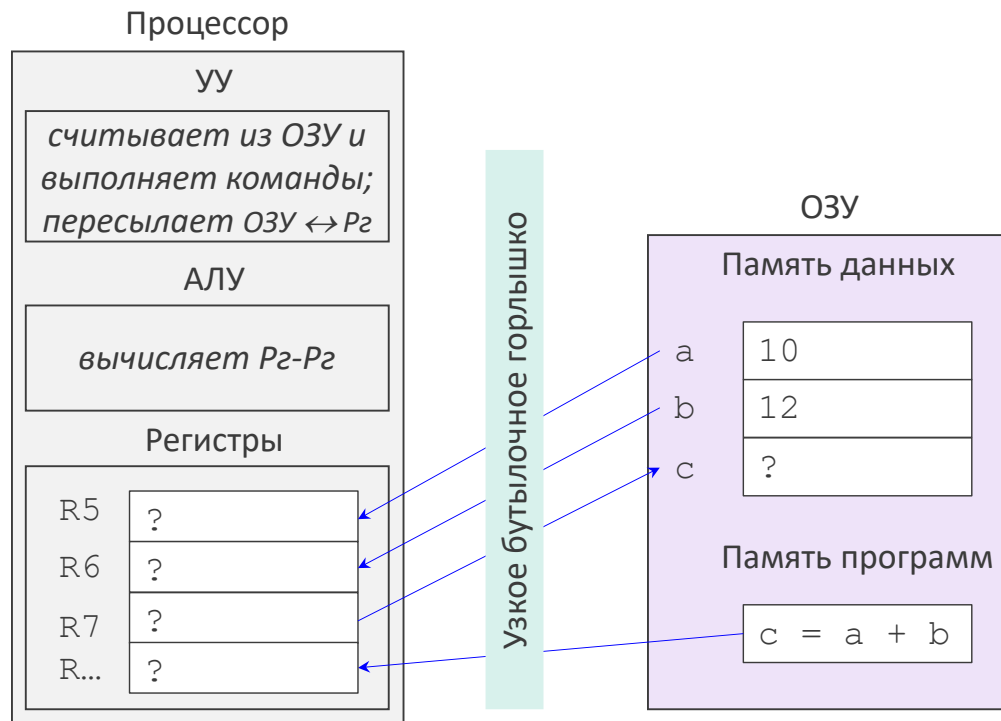


Рис. 2. Упрощенный пример архитектуры фон Неймана: процессор взаимодействует с ОЗУ, извлекая очередные команды и выполняя операции по чтению и записи данных. Однако распознавание команды и операции над данными выполняются только в регистровой памяти процессора. Принципиальная разница в скорости работы ОЗУ и процессора является бутылочным горлышком такой архитектуры.

является на порядок более медленная скорость работы ОЗУ и необходимость перезаписывать состояние всех конденсаторов 30 раз в секунду. Процессоры в зависимости от назначения тактируются частотами от десятков МГц до примерно 5 ГГц, что ограничено физическими особенностями транзисторов.

Рассмотрим упрощенный пример выполнения арифметической операции. Процессор с помощью УУ считывает очередную команду из ОЗУ в регистр и дешифрирует ее, раскладывая выполнения на простые шаги. АЛУ оперирует только данными в регистрах, поэтому сначала значения операндов копируются из ОЗУ в регистры, затем УУ дает команду АЛУ выполнить сложение и разместить сумму в регистре. Далее результат копируется в ОЗУ.

Интерфейс процессора и ОЗУ является бутылочным горлышком, поскольку высокое быстродействие процессора резко ограничивается медленной внешней памятью. Его компенсируют путем размещения промежуточной транзисторной 2-х или 3-х уровневой кэш-памяти ограниченного

объема, в которой хранится информация, которая с высокой вероятностью будет использована процессором в ближайшее время. Также физически разделяют потоки команд и данных.

АЛУ процессора выполняет наиболее популярные арифметико-логические операции. В различных приложениях требуется быстрое выполнение специфичных операций, прямая реализация на типовых процессорах будет занимать неприемлемо большое время. К таким операциям относятся, например, алгоритмы шифрования, операция умножения с накоплением и ряд других. Тогда в систему вводят дополнительные устройства — так называемый сопроцессор для выполнения специфичных операций и контроллер прямого доступа к памяти (кПДП, рис. 3). Процессор, встретив специфичную операцию в программе, делегирует ее выполнение сопроцессору, который с помощью собственных ресурсов без привлечения центрального процессора выполняет необходимые действия по чтению-записи ОЗУ с помощью кПДП.



Рис. 3. Центральный процессор (CPU) обращается к ОЗУ для выполнения операций чтения и записи, управляя сигналами на шинах управления, адреса и данных. С помощью контроллера прямого доступа к памяти (кПДП) сопроцессор или внешнее устройство может выполнять аналогичные операции в обход CPU для ускорения работы.

1.3 Три ключевые архитектуры, многоядерные процессоры и многопроцессорные системы

Для процессора первичной является архитектура набора команд (ISA), которая и определяет его возможности. Изначально выделялись архитектуры CISC с набором из большого количества сложных и медленно выполняемых команд и RISC — малым набором простых команд, большинство из которых выполняется за один такт работы процессора. В настоящее время граница между ними стирается.

Компания Intel первой в 1971 г. разработала и начала производить процессоры с архитектурой команд x86. За это время архитектура многократно модифицировалась и уже насчитывает около 1 300 команд. Компания ARM Limited владеет архитектурой ARM (проект начат в 1983 г.), которая обеспечивает низкое энергопотребление и применяется во встраиваемых системах. Широкое использование объясняется также и наличием значительного количества надежного программного обеспечения. Новой и динамично развивающейся архитектурой является открытая и модульная архитектура RISC-V (разработка с 2010 г.), позволяющая с помощью модулей расширения конфигурировать ISA для кроссплатформенных решений.

Наличие развитой теории цифровой обработки сигналов привело к разработке и выпуску специализированных цифровых сигнальных процессоров (DSP), например, серии TMS320 компании Texas Instruments с 1983 г. Основой математического аппарата является быстрое преобразование Фурье, ключевой операцией

которого является умножение с накоплением (MAC). Изображения являются частным случаем сигналов, но они выделились в отдельное направление благодаря развитию компьютерной 3D-графики. Так, в 1999 г. компания NVIDIA выпускает графический процессорный модуль (GPU), а в 2007 г. — GPU общего назначения (GPGPU), включающего программно-аппаратную архитектуру параллельных вычислений (CUDA).

При этом в процессорах параллельность вычислений может обеспечиваться как небольшим количеством (единицы и десятки) ядер, так и значительным (тысячи) ядер. В зарубежной терминологии они называются multicore и manycore соответственно. Тем не менее возможности распараллеливания ограничены. В 1967 г. Джин Амдал открыл закон, согласно которому возможность масштабирования вычислений за счет параллелизма ограничена, если в задаче присутствует синхронизация. Например, при возможности распараллеливания 95 % алгоритма и бесконечном количестве ядер достигаемая кратность ускорения стремится лишь к 20. Производительность многоядерной системы (рис. 4) ограничивается наиболее узким местом — доступом к общей памяти ядер.

Во многих приложениях используются пары процессоров с архитектурой ARM различных серий, например, Cortex-A72 и Cortex-A53. Первый обладает высокой производительностью второй — высокой энергоэффективностью. Оба построены по архитектуре ARMv8-A и содержат по 4 ядра и кэш

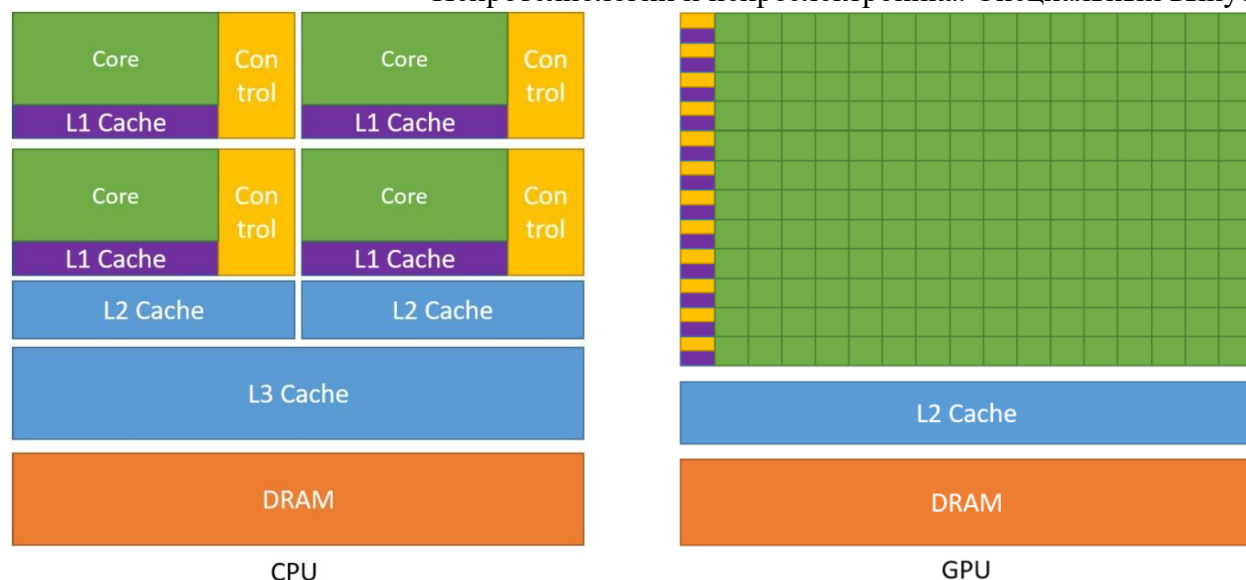


Рис. 4. Архитектуры многоядерных центральных процессоров (слева) и графических процессоров (справа): показано взаимное расположение вычислительных ядер (Core) с УУ (Control), кэш-памятью трех уровней (Cache L1...L3) и ОЗУ (DRAM) [8]

1-го и 2-го уровня, таймеры, средства отладки и доступ к высокоскоростной шине. Их парное применение предусмотрено архитектурой big.LITTLE, предназначенной для балансировки производительности и энергоэффективности процессоров в составе SoC: например, A72 в качестве big — для обработки видео, A53 в качестве LITTLE — для фоновых задач и обмена данными. Используют 3 стратегии распределения задач между 8-ю ядрами такой пары процессоров: переключение кластеров, каждый из которых содержит по 4 мощных или слабых ядра; переключение между мощным и слабым ядром внутри каждой из 4-х пар; гетерогенный мультипроцессинг — автономный запуск на каждом из 8-и ядер.

Так, фирма Rockchip (основана в 2001 г.) в своем нейропроцессоре RK3588 (2021 г.) [9] использует пару Cortex-A76 и A55, которая управляется тремя микроконтроллерами с простейшей архитектурой Cortex-M0. Наряду с различными сопроцессорами в состав входит и GPU с архитектурой Mali (ARM).

Следует отметить, что при аппаратной реализации нейросетей сигналы квантуются — разрядность может достигать 64, 32, 16, 8, 4 и даже 2 бит. Представление чисел с плавающей точкой регламентировано стандартом IEEE 754 (1985 г.) и чаще всего применяется с двойной (64 разряда) и одинарной (32 разряда) точностью. Число

представляется в виде знака, мантиссы и показателя степени (экспоненты). Некоторые компании вводят собственные форматы с сокращением разрядности мантиссы и экспоненты, например, NVIDIA — TensorFloat (TF32), Google Brain — Brain Floating Point (BF16) [10].

Исторически третья из рассматриваемых архитектур после x86 и ARM является RISC-V (Калифорнийский университет в Беркли, 2010 г.), развитием которой с 2015 г. занимается RISC-V Foundation. Эта архитектура процессоров является открытой и безвозмездной, а также модульной — в базовый набор входит 47 команд. К ним можно добавить команды из следующих модулей: M — целочисленное умножение и деление, F и D — вычисления с плавающей запятой (float, double), A — атомарные операции с памятью, C — сжатые команды для IoT (интернета вещей), E — команды для встраиваемых систем. Эта архитектура еще недостаточно распространена на рынке как x86 и ARM, развитие системного программного обеспечения, компиляторов и конвертеров находится в начальной стадии. Применение отмечается в областях IoT, встраиваемых систем, высокопроизводительных вычислений, автоэлектроники, телекома, а также в составе



Рис. 5. Отечественные процессоры для нейросетей (верхний ряд): 1879BM8Я или NM6408 (НТЦ «Модуль»), RoboDeus и СКИФ (НПЦ «ЭЛВИС»), IVA TPU или H1 («ХайТэк»); решения на основе NM6408 и RK3588 (нижний ряд): модули NM Card mini и NM Quad, блоки NM Vision и NM Pilot

специализированных процессоров для нейросетей.

Кроме использования многоядерных процессоров, применяют и многопроцессорные системы. Миниатюризация систем из дискретных компонентов на печатных платах привела к разработке заказных интегральных схем (ASIC) в виде систем-на-кристалле (SoC). Развитие телекоммуникаций послужило объединению SoC в сети-на-кристалле (NoC), а рост в области высокопроизводительных вычислений — к реализации многопроцессорных систем на кристалле (MPSoC) [11].

1.4. Отечественные нейропроцессоры

В России две компании ведут эволюционную разработку нейропроцессоров, имея опыт в несколько десятилетий по разработке процессоров цифровой обработки сигналов. Так, НТЦ «Модуль» в нейропроцессоре 1879BM8Я (2019 г.) использует 4 нейропроцессинговых кластера, в состав каждого из которых входят 4 модуля с фирменной архитектурой NeuroMatrix версии 4, и один процессор Cortex-A5.

Cortex-A5.

На основе нейропроцессора 1879BM8Я (другое название — NM6408) выпускается модуль NM Card mini для установки в материнскую плату компьютера (рис. 5, по материалам разработчиков). В состав модуля NM Quad входят 4 таких процессора, позволяя распараллеливать выполнение задачи или параллельно запускать несколько задач. В блоках NM Vision для видеоаналитики и NM Pilot для робототехники используется один нейропроцессор 1879BM8Я (NM6408) и рассмотренный выше RK3588.

В качестве перспективных разработок (начальный анонс [15]) на 2025 — 2027 гг. компанией планируется разработка и выпуск нейропроцессоров по технологии 5 нм, отличающихся по производительности и потребляемой мощности: 100 Вт для серверных решений, 25 Вт для бортовых решений и 2,5 Вт для носимой электроники. Все три решения основаны на архитектуре NeuroMatrix версии 5.

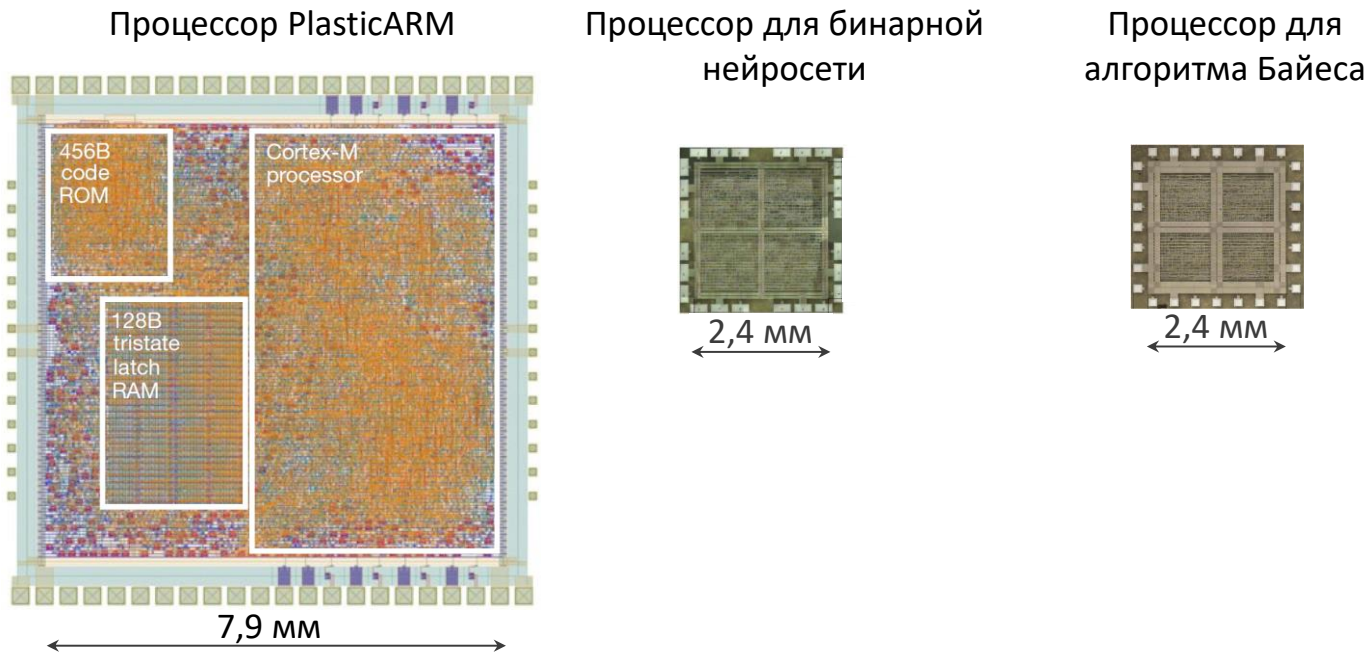


Рис. 6. Топологии процессоров, изготовленные по печатной технологии [12, 13, 14]

К относительно недавним разработкам второй отечественной компании с длительной историей успешных разработок в области цифровой обработки сигналов — НПЦ «ЭЛВИС», — относятся 50-ядерная микросхема RoboDeus (2020 г.) и нейропроцессор СКИФ (2022 г.). Третья, относительно молодая компания «ХайТэк», основанная в 2013 г., разработала и выпускает нейропроцессор Н1 (предыдущее название — IVA TRU).

1.5 Одна из альтернатив КМОП технологии для изготовления процессоров

Указанные выше процессоры изготавливаются по классической КМОП-технологии [16]. Однако с 2019 г. компании ARM Ltd., PragmatIC Semiconductor Ltd. и Манчестерский университет разработали и изготовили образцы 3-х простейших микросхем с помощью печатной технологии: микроконтроллер с архитектурой Cortex-M0, процессор для бинарной нейросети и процессор для реализации одномерного байесовского классификатора (машинное обучение без нейросетей) [12, 13, 14, 17, 18, 19]. Чип PlasticARM имеет 32-разрядную архитектуру, ОЗУ 128 Б, ПЗУ 456 Б, площадь кристалла составила 59,2 мм², связь осуществляется по 28-ми контактам. В состав входят 39 157 транзисторов и 17 183

резистора. При напряжении 3 В и частоте 29 кГц потребление составляет 21 мВт. Чип для бинарной нейросети имеет площадь 5,86 мм², содержит 2 084 транзисторов и 1 048 резисторов, потребляет 1,1 мВт при 3 В и 22 кГц. Чип для байесовского классификатора имеет подобные параметры, но потребляет 7,2 мВт при 4,5 В и 104 кГц.

Процесс разработки заказных процессоров на печатной электронике в целом имеет следующую последовательность. На вход подаются модели и данные, выполняется обучение и инференс (запуск обученной нейросети), с применением параметрической библиотеки моделей процессоров для машинного обучения генерируется и валидируется проект заказного процессора. Далее с применением стандартных библиотечных элементов FlexIC подготавливается топология для передачи на печатное производство с реализацией на пластине либо на пластиковой основе.

Технология FlexIC интегрирована с линейкой изготовления кремниевых пластин на подложках диаметром 300 мм: гибкая подложка крепится к стеклу и проходит необходимые технологические операции нанесения материала, формирования и травления. Входные данные передаются в формате GDSII, набор правил проектирования (PDK)

содержит библиотеку примитивов, проверку правил, технологические данные, файлы правил, параметризованные ячейки. Возможно изготовление транзисторов n-типа с каналом min 600 нм, резисторов с поверхностным сопротивлением 200 кОм/квадрат, конденсаторов High-k, ММСАР с емкостью 2,7 фФ/мкм². Максимальное количество слоев для металлизации составляет 4 слоя, для материалов — 13 слоев, общая толщина до 30 мкм.

В России крупный научный проект Минобрнауки России по теме «Аддитивное формирование активных элементов микроэлектроники методами печати» в период 2024-2026 гг. выполняется консорциумом из МФТИ, ИСПМ РАН, ИПТМ РАН, АО «НИИМЭ».

1.6. Фреймворки для формальных нейросетей

Анализируя ландшафт программного обеспечения для разработки, обучения, запуска и отладки нейросетей, следует выделить так называемые фреймворки. К ним относятся программные платформы и библиотеки, предназначенные для разработки, моделирования и обучения нейросетей. Для запуска таких нейросетей на конкретных моделях нейропроцессоров разработчик фреймворка добавляет их поддержку. При доступности известных фреймворков (США), например, PyTorch, TensorFlow, Caffe2, Deeplearning4j и Apache MXNet, китайские специалисты весьма активно вкладывают ресурсы в разработку собственных фреймворков (PaddlePaddle, MindSpore) для обеспечения доверенности и технологической безопасности от возможных программных закладок и конкурентного анализа своих нейросетевых решений, размещенных на зарубежных серверах. В России разработку одного из отечественных фреймворков «Платформа-ГНС» [20] с 2018 г. ведет ФАУ «ГосНИИАС», обеспечивая поддержку отечественных нейропроцессоров.

2 Спайковые нейросети и архитектуры процессоров для их обработки

2.1 Нейроморфные процессоры и спайковые нейросети

Формальная модель нейрона, хотя и используется в

работе подавляющего большинства современных решений, приводит к их высокому энергопотреблению. Причина кроется в необходимости постоянно поддерживать уровни напряжения на синаптических связях. Более экономным вариантом может служить посылка короткого импульса при необходимости передать сообщение от нейрона к нейрону.

Как было показано выше, при реализации классической архитектуры имеется проблема с физической реализацией памяти: либо быстрая на транзисторах, но имеющая большую площадь на кремниевом чипе и высокую стоимость, либо дешевая и на порядок более медленная на конденсаторах. Оба вида памяти являются энергозависимыми и теряют информацию при отключении питания, поэтому для долговременного хранения используют твердотельную, магнитную, Flash и другие виды памяти. Энергетические и временные затраты связаны с пересылкой информации между такими видами памяти. Из этого можно сформулировать два требования к разработчикам. Во-первых, в идеале вычисления должны выполняться непосредственно в ячейках памяти, где хранятся операнды. Во-вторых, требуется память на новых физических принципах, которая обеспечит высокое быстродействие, долговременное хранение информации и возможность масштабирования при доступной стоимости. При этом стремятся сократить разрыв между памятью и вычислительными блоками, применяя подходы «вычисления рядом с памятью» и «вычисления в памяти» [21].

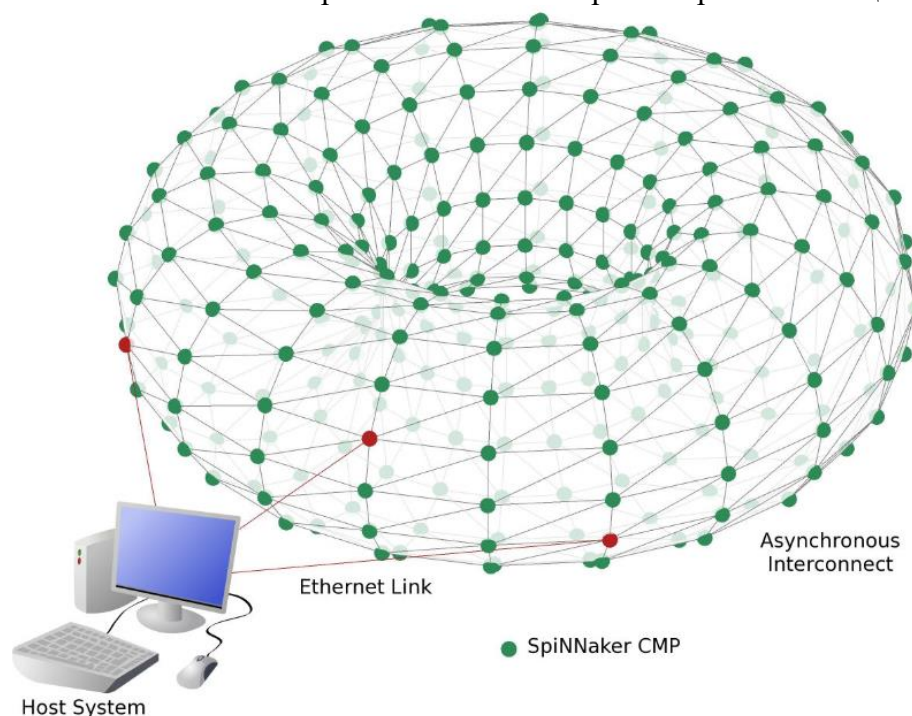


Рис. 7. Соединение узлов «чип-мультипроцессоров» (SpiNNaker CMP) в виде тора, связь между которыми обеспечивается асинхронно через Ethernet-соединения[22].

За новым направлением закрепилось название «нейроморфные» процессоры, которые в некоторой степени эмулируют структуру и функции мозга. В аппаратной части они отличаются от реализующих формальную модель нейрона хотя бы первыми двумя пунктами из трех: (1) архитектура, отличная от архитектуры фон Неймана, (2) новый способ кодирования (представления) сигналов, (3) элементная база на новых физических принципах. Нейроморфный подход, являющийся основной чертой таких процессоров, обеспечивает применение биологически правдоподобных моделей в организации вычислительного процесса для нейросетей.

Наиболее популярной из вновь предложенных нейробиологами моделей нейрона является спайковая модель. Показано, что в мозге генерируются и распространяются импульсные (спайковые, spike) сигналы, причем только в те моменты, когда передается информация от нейрона к нейрону. Такая особенность относится к пункту о кодировании сигналов.

Разработаны различные модели нейронов: порогового интегратора с утечкой (LIF), Ижикевича [23], Ходжкина-Хаксли [24] и многие другие.

Переход к спайкам затрудняет возможность дифференцирования сигналов и применения метода обратного распространения ошибки для обучения, что привело к разработке адаптированных подходов, например, обратного распространения на основе событий E-prop (Event-based Propagation), обратного распространения во времени BPTT (Backpropagation Through Time), методов суррогатного градиента (Surrogate gradient). Один из применяемых методов — спайковая пластичность, зависящая от времени (STDP) [25, 26]. Синаптическая связь между пресинаптическим и постсинаптическим нейронами усиливается, если спайк на первом послужил причиной возникновения спайка на втором. В противном случае синаптический вес снижается.

2.2 Примеры нейроморфных процессоров

Одним из знаковых нейроморфных процессоров является SpiNNaker (Манчестерский университет), который реализует массово-параллельную многоядерную систему из 10^9 нейронов и 10^{12} синапсов с потреблением 90 кВт [22]. В один многопроцессорный чип входят 18 ARM ядер и синхронная DRAM память. В систему входят 57 тысяч таких узлов, соединенных между собой в виде

Нейротехнологии и нейроэлектроника. Специальный выпуск. 2025

тора и общающихся между собой через пакеты размером по 5 или 9 Б.

Как и в многих других нейроморфных процессорах, кодирование и передача данных выполняется по распространенному коммуникационному протоколу Address Event Representation (адресно-событийное представление), позволяющему передавать по шине данные с указанием их источника (адрес) и времени их появления (событие). Этот протокол в различных вариациях находит применение и при передаче сенсорных данных [27].

Компания IBM в 2014 г. выпускает нейроморфный процессор TrueNorth емкостью 1 млн. нейронов и 256 млн. синапсов [28]. В основе лежат 7 принципов: архитектура, управляемая событиями; электронная компонентная база с низким энергопотреблением (КМОП); массовый параллелизм; операции в реальном времени с обновлением в 1 мс; масштабируемая архитектура; устойчивость к ошибкам, особенно для схем памяти; применение неспециализированного САПР для разработки.

Вдохновленные таким решением, разработчики компании «Мотив НТ» разработали и выпустили нейроморфный чип «Алтай» (2020 г.), построенный на архитектуре с глобальной асинхронной и локально синхронной связями (GALS). Такой подход позволяет снизить динамическое энергопотребление КМОП-схемы, поскольку оно пропорционально тактовой частоте. Этот же подход используется и в TrueNorth. При необходимости обмена частота ядер синхронизируется в момент передачи данных в виде спайка, которая осуществляется через соседние ядра к целевому. Частота появления спайков составляет 1 кГц, в связи с чем одно ядро моделирует работу 512 нейронов. Все спайки разбиты на группы по времени прибытия, за квант времени принят тик. Спайки передаются через соседние ядра к целевым. Отправку межъядерных спайков, передачу транзитных спайков и прием спайков осуществляет маршрутизатор. Более детально архитектура представлена в [29].

Компания Intel разрабатывает и выпускает нейроморфный процессор Loihi (2017 г.) и Loihi 2 (2021 г.) [30, 31, 32]. Последняя разработка содержит матрицу из 8x16 нейроморфных ядер, каждое из

которых эмулирует работу 8 912 нейронов, весь чип — 1 048 576 нейронов. Каждый нейрон содержит блок для обучения с программируемой пользователем функцией. Для управления используются 6 процессоров Lakemont с архитектурой x86.

В [33] представлена информация по ряду нейроморфных процессоров. Если в рассмотренных выше процессорах в качестве архитектур управляющих ядер использовались в подавляющем большинстве ARM, то компания Intel — собственную архитектуру x86. Архитектура RISC-V постепенно набирает обороты: именно ее компания Innantera использовала в своем новом энергоэффективном нейроморфном процессоре T1 (2024 г.) для обработки сенсорных данных с малой задержкой.

Выполнение всех трех отличительных свойств нейроморфного процессора присуще приборам с применением электронной компонентной базы на новых физических принципах. К ним можно отнести мемристоры [34] и фотонные интегральные схемы.

2.3 Нейроморфные процессоры на мемристорных кроссбарах

Мемристоры впервые описаны Леоном Чуа (1971 г.) и изготовлены компанией Hewlett Packard (2008 г.). Их основным преимуществом является изменение проводимости в зависимости от протекшего через них заряда. К потребительским характеристикам можно отнести стабильность резистивного состояния, выносливость к циклическим переключениям, пластичность резистивного состояния, напряжение и время переключения и напряжение чтения. Активные работы по такой тематике ведутся многими коллективами во всем мире, однако минимально допустимые значения одновременно по всем потребительским характеристикам пока не достигнуты. Это препятствует выходу на массовое производство таких приборов. К наиболее важным характеристикам мемристоров относятся:

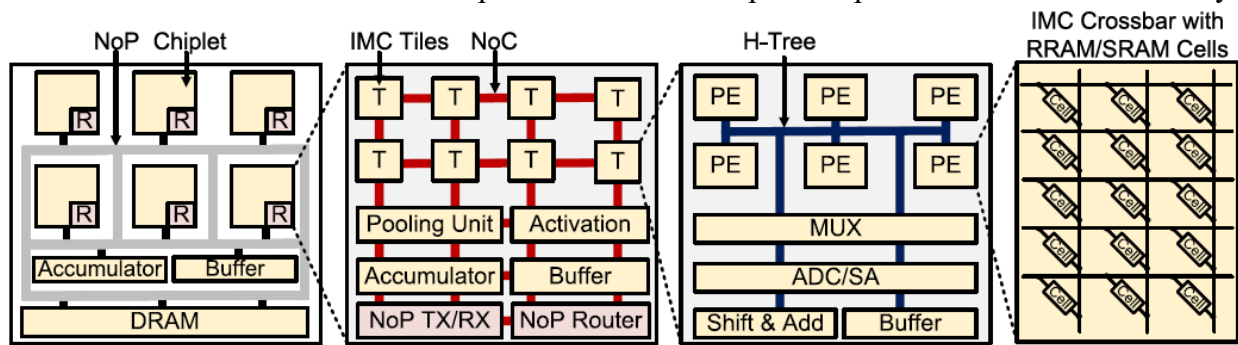


Рис. 8. Архитектура процессора SIAM (справа налево): (1) вычисления в памяти (IMC) реализуются на мемристормом кроссбаре с использованием ячеек резистивной памяти и статического ОЗУ, которые образуют процессорный элемент (PE); (2) результаты работы шести PE, объединенных в дерево вида «Н», через мультиплексор MUX передаются на АЦП (ADC) и прецизионный операционный усилитель (SA), сдвигаются со сложением (Shift & Add) и буферизируются (Buffer); (3) тайлы (T), модули пулинга (Pooling) и активации (Activation), аккумулятор (Accumulator), буфер (Buffer), передатчик (TX/RX) и маршрутизатор (Router) объединены в сеть-на-кристалле и реализованы в виде чиплета; (4) 6 таких чиплетов, аккумулятор, буфер и динамическое ОЗУ объединяются в сеть-в-корпусе и размещаются на интерпозере как готовый нейроморфный процессор [35].

стабильность резистивного состояния (retention), выносливость по отношению к циклическим переключениям (endurance), пластичность резистивных состояний, напряжение и время переключения, напряжение чтения.

Если в архитектуре фон Неймана проблемой является разнесенные процессорные модули и память, то в архитектурах типа «вычисления рядом с памятью» (NMC) часть процессорных модулей располагается в непосредственной близости к ОЗУ и энергонезависимой памяти. Наилучший эффект достигается в архитектурах типа «вычисления в памяти», когда и расчеты, и хранение информации выполняется на едином приборе.

Кроссбар является структурой из проводников, часть из которых размещена параллельно между собой в одной плоскости, оставшаяся часть — также параллельно между собой в другой параллельной плоскости и перпендикулярно проводникам из первой. На пересечении размещаются элементы, каждый из которых адресуется по выбранным горизонтальному (строка) и вертикальному (столбец) проводникам. В такой элемент входят мемристор и селектор (транзистор или диод).

С помощью такого активного кроссбара возможно «вычисление в памяти» векторно-матричного умножения, как основной операции в нейросетях. После программирования проводимостей в матрице мемристоров на горизонтальные проводники

подаются напряжения, соответствующие входному вектору. Попадая на столбец мемристоров, напряжение умножается на проводимость соответствующего мемристора и в виде тока суммируется с аналогичными токами других мемристоров. Полученные токи по столбцам преобразуются в напряжения на прецизионных инструментальных операционных усилителях. Преобразование из цифры во входное напряжение кроссбара и обратно на выходе кроссбара осуществляется на цифро-аналоговых (ЦАП) и аналого-цифровых (АЦП) преобразователях. Управление таким мемристормым кроссбаром осуществляется с помощью специализированной КМОП-схемы. Интеграция мемристормой матрицы и КМОП-схемы может выполняться по монолитной, чиплет-, и 3D-технологиям сборки.

В качестве примера можно привести основанный на чиплетах масштабируемый ускоритель вычислений в памяти для глубоких нейросетей проект SIAM (Chiplet-based Scalable In-Memory Acceleration with Mesh for Deep Neural Networks) [35]. Более детальную информацию по сопоставлению развития КМОП- и мемристормых процессоров и их применению можно найти в [36].

Работы в этой области ведутся в том числе и в рамках научной программы Национального центра физики и математики (г. Саров) консорциумом из организаций ННГУ им. Н.И. Лобачевского,

Нейротехнологии и нейроэлектроника. Специальный выпуск. 2025
АО «НИИМЭ» и ряда других.

2.4. Фреймворки для спайковых нейросетей

Для нейроморфных процессоров используются специализированные фреймворки, не совместимые с фреймворками для нейропроцессоров. Например, для Loihi/Loihi 2 применяется открытый фреймворк Intel Lava. Для создания фреймворка отечественного нейроморфного процессора «Алтай» компания «Мотив НТ» с 2022 г. привлекла разработчиков из АО «Лаборатория Касперского». В итоге в 2024 г. выпущен Kaspersky Neuromorphic Platform [37], который предназначен как для разработки спайковых нейросетей, так и для преобразования формальных нейросетей в импульсные.

Исходя из особенностей нейроморфных вычислений, можно выделить три основных специализации таких фреймворков. Для крупномасштабных исследований биологически правдоподобных нейросетей, включая изучение их динамики, могут применяться NEST (Neural Simulation Tool) и Brian2. Фреймворки Intel Lava и SpiNNaker удобны для аппаратной реализации и отладки. Гибридные подходы с классическими фреймворками реализуются в BindsNET, SpikingJelly, PySNN. Среди перспективных отечественных фреймворков можно выделить AgNI-X [38].

ЗАКЛЮЧЕНИЕ

Целью данного обзора не является исчерпывающее перечисление всех существующих нейро- и нейроморфных процессоров и соответствующих фреймворков, как, например, в [39, 40, 41]. Акцент сделан на вариантах архитектур нескольких таких процессоров, рассмотрение которых демонстрирует ключевые тенденции в рассматриваемой области. В их числе: нейро- и нейроморфные процессоры; стремление к высокой энергоэффективности и реализации «вычислений в памяти»; поиск и применение новых биологически правдоподобных моделей работы для реализации пластичности и обучения; совмещение в кристалле и корпусе гетерогенных процессоров для классических и нейроморфных алгоритмов; тенденция разработки и выпуска линейек нейропроцессоров с разделением по потребляемой

мощности; использование ядер x86, ARM и RISC-V в современных нейроморфных процессорах, также как и в классических нейропроцессорах.

Благодарности: 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».

Ссылки

1. Li, S.; Deng, Y.-Q.; Zhu, Z.-L.; Hua, H.-L.; Tao, Z.-Z. A Comprehensive Review on Radiomics and Deep Learning for Nasopharyngeal Carcinoma Imaging. *Diagnostics* 2021, 11, 1523, doi:10.3390/diagnostics11091523.
2. von Neumann, J. *The Computer and the Brain*; Yale University Press: New Haven, CT, USA, 1958.
3. Rosenblatt, F. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*; Spartan Books: Washington, DC, USA, 1962.
4. Указ Президента РФ от 10.10.2019 N 490 (ред. от 15.02.2024) «О развитии искусственного интеллекта в Российской Федерации» (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 года»).
5. Pause Giant AI Experiments: An Open Letter. *Future of Life Institute*. URL: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/> (дата обращения 20.01.2025).
6. Herculano-Houzel, S. The Human Brain in Numbers: A Linearly Scaled-up Primate Brain. *Front. Hum. Neurosci.* 2009, 3, doi:10.3389/neuro.09.031.2009.
7. Shivappriya, S.N.; Priyadarsini, M.J.P.; Stateczny, A.; Puttamadappa, C.; Parameshachari, B.D. Cascade Object Detection and Remote Sensing Object Detection Method Based on Trainable Activation Function. *Remote Sensing* 2021, 13, 200, doi:10.3390/rs13020200.
8. CUDA C и графический процессор. URL: https://blog.csdn.net/qq_29788741/article/details/125411010 (дата обращения 20.01.2025).

9. Сайт компании RockChip. URL: www.rockchips.com (дата обращения 20.01.2025).
10. Сравнение точности при обучении и инференсе. URL: <https://frankdenneman.nl/2022/07/26/training-vs-inference-numerical-precision/> (дата обращения 20.01.2025).
11. Lionel Torres, et al. An Introduction to Multi-Core System on Chip - Trends and Challenges. 2011. URL: <https://hal-lirmm.ccsd.cnrs.fr/lirmm-00574947/document> (дата обращения 20.01.2025).
12. Ozer, E.; Kufel, J.; Biggs, J.; Myers, J.; Reynolds, C.; Brown, G.; Rana, A.; Sou, A.; Ramsdale, C.; White, S. Binary Neural Network as a Flexible Integrated Circuit for Odour Classification. In Proceedings of the 2020 IEEE International Conference on Flexible and Printable Sensors and Systems (FLEPS); IEEE: Manchester, United Kingdom, August 16 2020; pp. 1–4.
13. Ozer, E.; Kufel, J.; Myers, J.; Biggs, J.; Brown, G.; Rana, A.; Sou, A.; Ramsdale, C.; White, S. A Hardwired Machine Learning Processing Engine Fabricated with Submicron Metal-Oxide Thin-Film Transistors on a Flexible Substrate. *Nat Electron* **2020**, 3, 419–425, doi:10.1038/s41928-020-0437-5.
14. Biggs, J.; Myers, J.; Kufel, J.; Ozer, E.; Craske, S.; Sou, A.; Ramsdale, C.; Williamson, K.; Price, R.; White, S. A Natively Flexible 32-Bit Arm Microprocessor. *Nature* **2021**, 595, 532–536, doi:10.1038/s41586-021-03625-w.
15. НТЦ «Модуль» создаёт 3 нейропроцессора с топологией 5 нм. URL: <https://russianelectronics.ru/2022-02-25-modul-5-nm> (дата обращения 20.01.2025).
16. Красников Г.Я., Горнев Е.С., Матюшкин И.В. Общая теория технологий и микроэлектроника. М.: ТЕХНОСФЕРА, 2020.
17. Biggs, J.; Myers, J.; Kufel, J.; Ozer, E.; Craske, S.; Sou, A.; Ramsdale, C.; Williamson, K.; Price, R.; White, S. A Natively Flexible 32-Bit Arm Microprocessor. *Nature* **2021**, 595, 532–536, doi:10.1038/s41586-021-03625-w.
18. Ozer, E.; Kufel, J.; Biggs, J.; Brown, G.; Myers, J.; Rana, A.; Sou, A.; Ramsdale, C. Bespoke Machine Learning Processor Development Framework on Flexible Substrates. In Proceedings of the 2019 IEEE International Conference on Flexible and Printable Sensors and Systems (FLEPS); IEEE: Glasgow, United Kingdom, July 2019; pp. 1–3.
19. Гид по разработке на основе технологии FlexIC. URL: <https://www.pragmaticsemi.com/app/uploads/2023/07/Pragmatic-Pocket-Guide-to-Designing-for-FlexICs.pdf> (дата обращения 20.01.2025).
20. Программное обеспечение «Унифицированная программная платформа для разработки конечно ориентированных программных комплексов автоматического распознавания объектов на основе нейросетевых подходов» – «Платформа». URL: <https://www.gosniias.ru/platform.html> (дата обращения 20.01.2025).
21. Amirsoleimani, A.; Alibart, F.; Yon, V.; Xu, J.; Pazhouhandeh, M.R.; Ecoffey, S.; Beilliard, Y.; Genov, R.; Drouin, D. In-Memory Vector-Matrix Multiplication in Monolithic Complementary Metal–Oxide–Semiconductor-Memristor Integrated Circuits: Design Choices, Challenges, and Perspectives. *Advanced Intelligent Systems* **2020**, 2, 2000115, doi:10.1002/aisy.202000115.
22. Furber, S.B.; Lester, D.R.; Plana, L.A.; Garside, J.D.; Painkras, E.; Temple, S.; Brown, A.D. Overview of the SpiNNaker System Architecture. *IEEE Trans. Comput.* **2013**, 62, 2454–2467, doi:10.1109/TC.2012.142.
23. Izhikevich, E.M. Simple Model of Spiking Neurons. *IEEE Trans. Neural Netw.* **2003**, 14, 1569–1572, doi:10.1109/TNN.2003.820440.
24. Hodgkin, A.L.; Huxley, A.F. A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve. *The Journal of Physiology* **1952**, 117, 500–544, doi:10.1113/jphysiol.1952.sp004764.
25. Первые шаги в импульсных нейронных сетях. URL: <https://habr.com/ru/articles/746762/> (дата обращения 20.01.2025).

26. Andrade-Talavera, Y.; Fisahn, A.; Rodríguez-Moreno, A. Timing to Be Precise? An Overview of Spike Timing-Dependent Plasticity, Brain Rhythmicity, and Glial Cells Interplay within Neuronal Circuits. *Mol Psychiatry* **2023**, *28*, 2177–2188, doi:10.1038/s41380-023-02027-w.
27. Purohit, P.; Manohar, R. Field-Programmable Encoding for Address-Event Representation. *Front. Neurosci.* **2022**, *16*, doi:10.3389/fnins.2022.1018166.
28. Merolla, P.A.; Arthur, J.V.; Alvarez-Icaza, R.; Cassidy, A.S.; Sawada, J.; Akopyan, F.; Jackson, B.L.; Imam, N.; Guo, C.; Nakamura, Y.; et al. A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface. *Science* **2014**, *345*, 668–673, doi:10.1126/science.1254642.
29. Нейроморфный процессор "Алтай" для энергоэффективных вычислений / Н. В. Гришанов, А. В. Зверев, Д. Е. Ипатов и др. // Наноиндустрия. – 2020. – № S96-2. – С. 531-538
30. N.L, M.; V.N, T.; D.P, V.; N.D, N.; L.H, N.; T.N, H. Bio-Realistic Neural Network in Speech Timing Learning Mechanism. In Proceedings of the 2018 4th International Conference on Green Technology and Sustainable Development (GTSD); IEEE: Ho Chi Minh City, November 2018; pp. 663–667.
31. Процессоры, ядра и потоки. Топология систем. URL: <https://habr.com/ru/companies/intel/articles/583006/> (дата обращения 20.01.2025).
32. Davies, M.; Srinivasa, N.; Lin, T.-H.; Chinya, G.; Cao, Y.; Choday, S.H.; Dimou, G.; Joshi, P.; Imam, N.; Jain, S.; et al. Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. *IEEE Micro* **2018**, *38*, 82–99, doi:10.1109/MM.2018.112130359.
33. Гид по нейроморфному аппаратному обеспечению. URL: <https://open-neuromorphic.org/neuromorphic-computing/hardware/> (дата обращения 20.01.2025).
34. На пути к реализации высокопроизводительных вычислений в памяти на основе мемристорной электронной компонентной базы / А. Михайлов и др. // Физмат. – 2023. – Т. 1, № 1. – С. 42-64. – DOI 10.56304/S2949609823010021. – EDN HTSRZQ.
35. Krishnan, G.; Mandal, S.K.; Pannala, M.; Chakrabarti, C.; Seo, J.-S.; Ogras, U.Y.; Cao, Y. SIAM: Chiplet-Based Scalable In-Memory Acceleration with Mesh for Deep Neural Networks. *ACM Trans. Embed. Comput. Syst.* **2021**, *20*, 1–24, doi:10.1145/3476999.
36. Huang, Y.; Kiani, F.; Ye, F.; Xia, Q. From Memristive Devices to Neuromorphic Systems. *Applied Physics Letters* **2023**, *122*, 110501, doi:10.1063/5.0133044.
37. Kaspersky Neuromorphic Platform (нейроморфная платформа Касперского). URL: <https://support.kaspersky.ru/neuromorphic-platform/1.0/240225> (дата обращения 20.01.2025).
38. Нейроморфный фреймворк ArNI-X. URL: <https://arni-x.pro/> (дата обращения 20.01.2025).
39. Neuromorphic Computing Based on CMOS-Integrated Memristive Arrays: Current State and Perspectives. *JSFI* **2023**, *10*, doi:10.14529/jsfi230206.
40. Ivanov, D.; Chezhegov, A.; Kiselev, M.; Grunin, A.; Larionov, D. Neuromorphic Artificial Intelligence Systems. *Front. Neurosci.* **2022**, *16*, 959626, doi:10.3389/fnins.2022.959626.
41. Киселев М. В. Импульсные нейронные сети. Представление информации, обучение, память. Palmarium academic publishing, Рига, 2020.