



**СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ/SYSTEM ANALYSIS,
MANAGEMENT AND PROCESSING OF INFORMATION**

DOI: <https://doi.org/10.60797/IRJ.2026.165.70> EDN: RWSQLD**ОРГАНИЗАЦИЯ ПАМЯТИ ВЫЧИСЛИТЕЛЬНЫХ И ИСКУССТВЕННЫХ КОГНИТИВНЫХ СИСТЕМ:
РАЗДЕЛЫ, ВИДЫ И ОБРАБОТКА ДАННЫХ**

Научная статья

Грибков А.А.^{1,*}, Зеленский А.А.²¹ ORCID : 0000-0002-9734-105X;² ORCID : 0000-0002-3464-538X;^{1,2} Технологиченский центр, Москва, Российская Федерация

* Корреспондирующий автор (andarmo[at]yandex.ru)

Аннотация

Статья посвящена рассмотрению проблематики организации памяти в системах различной природы и назначения. В исследовании предлагается новая функциональная классификация памяти: по длительности хранения данных, по степени формализации и по степени осознанности данных. Память также классифицируется по организации хранения данных и доступу к ним на: адресуемую, стековую, ассоциативную и семантическую. Приоритетным контекстом, в рамках которого необходимо рассматривать память, является реализация процессов управления объектами, в которых память играет центральную роль. Качество управления напрямую зависит от эффективности использования памяти, в свою очередь определяемой выбранной архитектурой вычислительной или когнитивной системы. Наилучший результат достигается в случае архитектуры обработки данных, обеспечивающей многопоточность обработки, память-ориентированность и разделение общей памяти на локальные, привязанные к функциональным подсистемам.

Ключевые слова: память, хранение данных, адресация, ассоциативная, семантическая, синтаксическая, обработка данных, вычисление, когнитивная система.

**THE ORGANISATION OF MEMORY IN COMPUTATIONAL AND ARTIFICIAL COGNITIVE SYSTEMS:
CATEGORIES, TYPES AND DATA PROCESSING**

Research article

Gribkov A.A.^{1,*}, Zelenskii A.A.²¹ ORCID : 0000-0002-9734-105X;² ORCID : 0000-0002-3464-538X;^{1,2} Technological Center, Moscow, Russian Federation

* Corresponding author (andarmo[at]yandex.ru)

Abstract

The article examines the issue of memory organisation in systems of various types and purposes. The study suggests a new functional classification of memory: based on the duration of data storage, the degree of formalisation, and the degree of data awareness. Memory is also classified according to the organisation of data storage and access to it as: addressed, stack-based, associative and semantic. The primary context within which memory must be considered is the implementation of object management processes, in which memory plays a central role. The quality of control depends directly on the efficiency of memory usage, which in turn is determined by the chosen architecture of the computational or cognitive system. The best results are achieved with a data processing architecture that supports multithreading, is memory-oriented, and divides shared memory into local memory units tied to functional subsystems.

Keywords: memory, data storage, addressing, associative, semantic, syntactic, data processing, computation, cognitive system.

Введение

Ключевым фактором, определяющим эффективность вычислений, когнитивной деятельности и решения прочих задач, связанных с обращением информации, является организация памяти. Под организацией памяти мы будем понимать ее составляющие и характер их взаимодействия, механизмы хранения и доступа к данным в памяти, индексации и поиска, а также реализующие их процессы.

Несмотря на активное развитие информационных технологий, в том числе построенных на их основе систем искусственного интеллекта, интеллектуальных систем управления и других когнитивных систем, основная часть знаний, которыми мы в настоящее время располагаем, — это знания о памяти естественных (биологических) систем, в основном человека. Это означает, что одной из первоочередных задач развития представлений об организации памяти является их универсализация, т.е. формализация посредством терминов и понятий, в равной степени применимых как к естественным, так и искусственным системам, наделенным памятью. Также необходимо учитывать возможность наличия таких качественных отличий искусственных систем от естественных, которые потребует введения дополнительных понятий или корректирования понятий, используемых в настоящее время для естественных систем, но неприменимых в полной мере для искусственных.

В рамках данной статьи авторами планируется рассмотрение следующих вопросов: «Каковы функциональные подсистемы памяти?», «Согласно каким критериям осуществляется их определение?», «Какая подсистема памяти ответственна за хранение и использование осознанных знаний?», «Как реализуются и соотносятся между собой понимание и осмысление знаний?», «Какие существуют виды памяти по организации хранения данных и используемым алгоритмам доступа?», «В чем заключается ключевая проблема сложных вычислительных и когнитивных систем и как она связана с положением памяти в архитектуре обработки данных?», «Что необходимо сделать (в том числе с памятью) для совершенствования архитектуры обработки данных?».

Разделы и виды памяти

2.1. Разделы памяти

В настоящее время комплексное описание системы памяти складывается из представлений о формирующих ее функциональных подсистемах — разделах памяти. По мнению авторов, можно выделить три основных критерия, согласно которым определяются функциональные подсистемы памяти: длительность хранения, формализация и осознанность данных.

По длительности хранения выделяют долговременную (для обычных вычислительных систем — постоянную) память, кратковременную память (для обычных вычислительных систем — регистры процессора и кэш-память первого уровня), а также рабочую (для обычных вычислительных систем — оперативную) память [1].

Долговременная память искусственных нейронных сетей, сходных по принципам построения с нейронными сетями естественных (биологических) когнитивных систем, реализуется в виде ее параметров (весов и смещений), изменяющихся в процессе обучения. Сравнительно недавно распространение получили нейронные сети с памятью (memory-augmented neural networks, MANN) [2], использующие специальные механизмы (подобные оперативной памяти компьютера) для хранения и извлечения информации из постоянной внешней памяти. Кратковременная память искусственных нейронных сетей формируется на базе рекуррентных нейронных сетей (recurrent neural networks, RNN) [3], рабочая память — в виде долговременной краткосрочной памяти (long short-term memory, LSTM) [4] или управляемых рекуррентных блоков (gated recurrent units, GRU) [5].

Длительность хранения данных в долговременной памяти когнитивных и вычислительных систем ограничена продолжительностью их жизненного цикла.

Длительность хранения данных в кратковременной памяти естественных когнитивных систем составляет несколько секунд (у человека — 25–30 секунд [6]), в рекурсивных нейронных сетях — от долей секунды до нескольких секунд (в зависимости от временного шага сети). Регистры процессора хранят данные и инструкции, которые обрабатываются в текущий момент, длительность хранения данных — от нескольких пикосекунд до нескольких наносекунд. Кэш-память первого уровня (L1 Кэш) — тип памяти, находящейся непосредственно на кристалле процессора, работающий на сопоставимой с ним скорости. Время доступа к L1 Кэшу составляет несколько наносекунд (в несколько раз больше, чем к регистрам процессора).

Длительность хранения данных в рабочей памяти жестко не ограничена и зависит от активности ее использования и подпитки энергией. В человеческой памяти в условиях постоянной активации данные рабочей памяти могут сохраняться до нескольких десятков минут, без активации — несколько минут. В искусственных нейронных сетях (память LSTM и GRU) — от десятков до сотен секунд. В обычных компьютерных системах данные в оперативной памяти сохраняются до тех пор, пока память получает электропитание или данные не вычищаются для освобождения места для более актуальных.

Рабочая память принципиально отличается от долговременной и кратковременной. Это отличие заключается в том, что рабочая память представляет собой не запись или состояние системы, ответственной за фиксацию памяти (как кратковременная и долговременная память), а процесс, протекание которого сопровождается изменением состояния системы — носителя памяти, в том числе рождением и трансформацией соответствующих этим изменениям информационных объектов. Данные в рабочей памяти сохраняются в виде указанных изменений и информационных объектов.

По степени формализации записи данных в памяти можно выделить информационную и неинформационную память [7]. Отличительными особенностями неинформационной памяти являются: во-первых, автономность (полная или частичная) от основной информационной памяти, существующей в виде виртуальной информационной системы, оперирующей формализованными информационными объектами; во-вторых, отсутствие трансляции сохраняемых данных в упорядоченную и систематизированную информационную запись. Неинформационная память содержит данные в неорганизованном и необработанном виде. В естественных (биологических) когнитивных системах к неинформационной относятся: сенсорная память [8], вегетативная память [9], структурная (например, мышечная) память, клеточная память [10, С. 225–259]. Сенсорная память является кратковременной (длительность хранения не более нескольких секунд), прочие указанные виды неинформационной памяти — долговременные.

По степени осознанности записанных в память данных обычно выделяют имплицитную и эксплицитную память [11]. Под имплицитной (от лат. *implicatus* — свёрнутый, закрытый) понимают бессознательный тип долговременной памяти, который позволяет использовать информацию из прошлого опыта без осознанного вспоминания, под эксплицитной (от лат. *explicitus* — развёрнутый, раскрытый) или декларативной (от лат. *declaratio* — заявление, объявление) — тип долговременной памяти, в которой имеющийся опыт или информация актуализируется произвольно и осознано.

Эксплицитная память в контексте естественных когнитивных систем в настоящее время представляется как совокупность эпизодической и семантической памяти. Эпизодическая память содержит все воспоминания о событиях, которые произошли непосредственно с носителем памяти (когнитивной системой) или о событиях, которые происходили вокруг. Семантическая память содержит знания о мире (факты, идеи, смыслы, понятия), которые могут

быть сформулированы. По степени осознанности знаний в семантической памяти можно выделить понятия (понятные) знания, которые могут быть объяснены посредством известных терминов и обобщенных понятий в рамках существующей парадигмы, и осмысленные знания, которые интегрированы в систему знаний о мире.

В искусственных когнитивных и вычислительных системах абсолютно преобладающая часть информации не является результатом личного опыта (и, соответственно, записями в эпизодической памяти), и также не может быть отнесена к понятным или осмысленным знаниям, а представляет собой формальную запись произвольной информации по заданным правилам. Память, содержащую информацию в таком виде, следует назвать синтаксической памятью. Запись в синтаксической памяти может быть запрошена, прочитана и (при необходимости и возможности) преобразована в понятные или осмысленные знания. Синтаксическая память, очевидно не является частью эпизодической или семантической, однако она входит в состав эксплицитной (декларативной) памяти, следовательно, последняя (в общем случае) состоит не из двух составляющих (эпизодической и семантической), а из трех (эпизодической, семантической и синтаксической).

Отсутствие у естественных когнитивных систем такой синтаксической памяти — интересный факт, заслуживающий отдельного рассмотрения. Предварительное объяснение данного феномена заключается в субъектности естественных (биологических) когнитивных систем, для которых (так сложилось эволюционно) не существует просто данных или информации — все данные поступают через органы чувств (сенсоры), т.е. являются личным опытом. В результате такие данные аккумулируются в эпизодической памяти, из которой уже в дальнейшем они извлекаются для обобщения и записи в семантическую память. Согласно такой интерпретации указанные три составляющие эксплицитной памяти можно назвать семантической, лично-синтаксической (эпизодической) и обезличенно-синтаксической.

Записываемые в семантическую память понятия и осмысленные знания, с одной стороны, являются двумя этапами на пути осознания знаний [7], но, с другой стороны, соответствуют противоположным подходам к формированию системы знаний. Понимание — определяющий процесс в построении иерархической системы знаний, в которой каждый (эволюционно) последующий уровень знаний формируется посредством обобщенных понятий и аксиом, инкапсулирующих знания предшествующих уровней. Принимая во внимание недерминированность большей части имеющихся знаний, данный подход к познанию представляется безальтернативным. Осмысление — процесс, служащий обеспечению целостности системы знаний, требующей их онтологизации [12], т.е. соотнесения с бытием. Понимание эту функцию выполнять не может, поскольку осуществляется на формальном уровне понятий, онтологичность которых не верифицируется. В идеале для осмысления знания необходимо осознание внутренних механизмов объектов и процессов, описанию которых служит это знание, однако на практике это невозможно. Одним из доступных утилитарных подходов к осмыслению знаний является их интерпретация в рамках общей теории систем посредством универсальных паттернов и форм, выявляемых из наблюдений объектов и процессов в различных предметных областях (в первую очередь относящихся к простейшим объектам и процессам, доступным для детерминированного представления), а также исходя из априорных метафизических знаний [13, С. 63–127, С. 207–248].

2.2. Виды памяти

Производительность обработки данных в значительной степени зависит от затрат времени на поиск необходимой для этой обработки информации в памяти. При решении сложных задач, в частности вычислительных, быстро увеличивается объем используемых данных и растут затраты времени на их адресацию и извлечение.

По организации хранения данных и используемым алгоритмам доступа к ним цифровая память делится на следующие основные виды: адресуемая память — наиболее распространенный вид памяти, в которой адресация осуществляется по местоположению данных; стековая или магазинная память (англ. stack memory), в которой данные упорядочены по последовательности записи в виде списка, доступ к которому осуществляется по принципу LIFO (англ. last in — first out, «последним пришёл — первым вышел»); ассоциативная или контекстно-адресуемая память (англ. associative memory, content-addressable memory, CAM) [14], в которой адресация блока связанных данных (файла) осуществляется посредством ассоциирования (группирования) исходя из его синтаксического содержания или ассоциативного признака; семантическая память (англ. semantic storage) [15], в которой адресация блока связанных данных осуществляется по его семантическому содержанию, т.е. согласно заданной структуре понятийных признаков или смысловых паттернов.

В искусственных нейронных сетях, не использующих дополнительных модулей памяти, хранение и доступ к данным соответствуют соединению ассоциативной и семантической памяти. При запросе данных указывается не конкретный адрес, а часть искомой информации (ассоциативный признак, частичный или полный паттерн), на основе которой находится и активируется связанный с ней полный или наиболее близкий аналог (по элементам, структуре или отношениям), хранящийся в синаптических связях (весах).

При использовании искусственных нейронных сетей с дополнительными модулями памяти, обращение к данным, хранящимся в этих модулях, может носить адресный характер. В частности, нейронные сети с памятью (MANN) предполагают использование позиционной и контентной адресации. Для этого используется специальный контроллер или механизм внимания (attention mechanism), регулирующий адресные обращения к внешней памяти: напрямую к записи в заданной позиции, либо к заранее неизвестному объекту определенного содержания через посредника в виде контроллера, выполняющего функцию библиотекаря, заранее выполняющего контекстную индексацию записей во внешней памяти.

Формирование аналогичной (по организации хранения данных и используемым алгоритмам доступа к ним) классификации памяти естественных когнитивных систем в полной мере не может быть реализовано: знания нейрофизиологии пока остаются для этого недостаточными. Можно предположить, что наиболее близким является

вариант соединения ассоциативной и синоптической памяти, присущий искусственным когнитивным системам, не использующим дополнительных (цифровых) модулей памяти.

Одним из ключевых различий перечисленных видов памяти является степень и характер используемой в них индексации данных. Напомним, что индексация осуществляется посредством анализа и классификация содержимого данных (по определенным критериям) и присвоения каждой единице хранения данных (например, файлу в памяти вычислительной машины) индивидуального индекса, который в дальнейшем позволяет существенно ускорить поиск данных, сверяя поисковый запрос с данными в индексах. Для адресуемой памяти индексация не является обязательной, однако широко используется, существенно повышая скорость поиска данных. В стековой памяти данные напрямую не индексируются, возможна лишь косвенная индексация через вычисление смещения ячейки данных относительно указателя стека. Для ассоциативной и семантической памяти индексация обязательна — посредством нее реализуется специфическая организация указанных видов памяти. Для ассоциативной памяти индексация задается исходя из синтаксического представления индексируемых данных, для семантической памяти — исходя из семантического представления индексируемых данных. В обоих случаях память проверяет наличие ячеек с заданным содержимым, и если таковые в памяти имеются, то возвращает их адреса и/или извлекает полное содержимое, а также (в качестве дополнительной опции) выдает дополнительные ассоциированные с ними данные.

Вычисление (обработка данных) в памяти

Согласно информационной концепции сознания память нужна для реализации процесса мышления и любых других процессов обращения информации, при которых последовательность изменений информационных объектов в сознании формируется на основе их предшествующих состояний. Наиболее жесткие требования предъявляются к рабочей (оперативной) памяти: именно в ней выполняются вычислительные и логические операции. Кратковременная и долговременная память влияют на производительность (в том числе, на быстродействие [16] — производительность в жестком реальном времени) косвенно, через воздействие на рабочую память.

Можно констатировать, что рабочая память — ключевая составляющая управления вычислительных и когнитивных систем, обеспечивающая их преемственность и изменения во времени. Повышение качества управления (увеличение точности и числа параметров, посредством которых задается объект управления) требует одновременного удовлетворения двух противоположных требований. С одной стороны, чем выше дискретность управления (т.е. частота управляющих воздействий) и больше управляемых параметров, тем сложнее и медленнее система управления. С другой стороны, чем более сложной является модель объекта управления (т.е., в рамках теоретико-множественного подхода к оценке сложности, чем выше точность и количество характеризующих его параметров), тем выше должна быть дискретность управления. Исследования авторов на примере систем управления технологическим оборудованием показали, что имеет место обратно пропорциональная зависимость между сложностью объекта управления и длительностью цикла его управления (интервалом времени между управляющими воздействиями) [17].

Каким образом можно удовлетворить указанные противоположные требования? Общий ответ — за счет совершенствования архитектуры обработки данных (в частности, их поиска и вычислений), потенциально позволяющей решить следующие задачи: во-первых, уменьшить объем обрабатываемого потока данных; во-вторых, увеличить скорость передачи данных; в-третьих, устранить очереди при одновременном обращении к одной памяти нескольких функциональных или структурных подсистем.

Уменьшить объем обрабатываемого потока данных можно за счет параллельного выполнения операций их обработки. Это путь реализуется в человеческом мышлении (параллельность в обработке задач [18], двухуровневое мышление [19], многозадачность [20]), а также в вычислительных системах, в том числе системах реального времени [21]. Наряду с дроблением исходного обрабатываемого потока данных (что уже частично решает проблему), также имеется возможность использования для разных потоков (отличающегося функционального назначения) альтернативных (по скорости и возможностям) инструментов обработки. Например, в системах управления технологическим оборудованием возможно использование цифровых, аппаратных и аналоговых компонентов. В настоящее время практическая реализация параллельной обработки информации в вычислительных системах основана на применении сопроцессоров (векторных, матричных, тензорных и др.) или ускорителей (графических, SoC и др.), которым переходит (перехватывается сопроцессором или передается ускорителю командой центрального процессора) часть вычислительной работы.

Наряду с использованием параллельности в обработке информации, задача уменьшения объема обрабатываемого потока также может решаться посредством функционального разделения памяти. Для вычислительных систем перспективным представляется развитие гарвардской архитектуры вычислительных машин с разделением памяти на память данных и память команд. Такая архитектура применяется, в частности, в процессорном кэше 1-го уровня (L1 Кэш). Гарвардская архитектура имеет ограниченное применение ввиду необходимости большого числа шин и неоптимального использования объема памяти.

Увеличение скорости передачи данных реализуется за счет сокращения длины каналов связи между элементами (включая память) вычислительной или когнитивной системы. В естественных когнитивных системах эта цель, очевидно, достигнута: обработка данных и их хранение осуществляется в одном органе — мозге (например, человеческом), в результате при реализации интеллектуальной деятельности длина каналов связи минимальна. Длина всех каналов связи мозга с периферийными системами в силу объективных причин не может быть уменьшена без ущерба общей конкурентоспособности вида и индивидов, однако объем передаваемых по этим каналам данных в большинстве случаев невелик. Там, где этот объем существенен (например, передача данных по главному нерву), соответствующий орган размещается предельно близко к мозгу.

Для технических систем частичное решение задачи сокращения длины каналов связи обеспечивается использованием вычислений вблизи памяти (NMC — near-memory computing [22]). Для реализации таких вычислений

осуществляется перемещение устройств обработки данных как можно ближе к памяти, чтобы уменьшить затраты энергии и задержки, связанные с перемещением данных. Устройства обработки данных (процессоры) помещаются рядом с памятью, например, через 3D-укладку или посредством размещения логических слоев поверх памяти.

Наибольший эффект увеличения скорости передачи данных достигается при использовании память-ориентированной архитектуры обработки данных. Основным вариантом реализации данной архитектуры является вычисление в памяти (PIM — processing-in-memory или processor-in-memory [22]). При реализации вычисления в памяти данные в процессе обработки не перемещаются между устройством обработки (например, процессором в вычислительных системах) и памятью, что требует затрат времени, энергии и ограничено пропускной способностью каналов связи, а остаются в памяти, в которую интегрируется устройство для их обработки.

Альтернативным вариантом реализации память-ориентированной архитектуры являются вычисления с памятью (computing with memory [23]). Такие вычисления предполагают использование массивов памяти (обычно одно- или двумерных) в качестве таблиц просмотра (lookup tables, LUT), в которых хранятся результаты сложных вычислений, что позволяет быстрее и с меньшими затратами энергии извлекать значения вместо их повторного вычисления.

К числу вариантов память-ориентированной архитектуры, широко используемым в когнитивных системах на базе искусственных нейронных сетей, относят аналоговые вычисления в памяти (analog in-memory computing, AIMC) [24], [25], основанные на использовании специальных типов энергонезависимой памяти, таких как мемристоры (резистивная память, resistive random-access memory, RRAM) или память с фазовым переходом (phase-change memory, PCM). Веса нейронной сети хранятся в этих ячейках памяти в виде значений проводимости. Операции умножения матриц на векторы (основа глубинного обучения) выполняются с использованием физических законов (закона Ома и правил Кирхгофа). Вычисления происходят одновременно по всему массиву памяти, что обеспечивает их очень высокую производительность и низкое энергопотребление, недостижимые для вычислительных и искусственных когнитивных систем с традиционной архитектурой.

Как показал наш анализ, развитие память-ориентированной архитектуры определяется двумя основными тенденциями: во-первых, приближением устройств обработки данных к памяти, пределом которого является вычисление в памяти, и, во-вторых, заменой обработки данных извлечением из памяти готовых решений. Из этих двух тенденций центральной, по мнению авторов, является вторая тенденция, а первая должна стать важным фактором ее успешной реализации. Для наглядного представления приоритизации тенденций развития память-ориентированной архитектуры необходимо ввести два понятия, определяющие характер доступа к памяти: обработка данных «по эту сторону памяти» и обработка данных «по ту сторону памяти».

Обработка данных «по эту сторону памяти» — это формирование прямых запросов со стороны устройства обработки данных к блокам связанных данных (например, файлам) в памяти. Именно это имеет место при использовании адресуемой памяти. Наличие индексации принципиально не меняет характер доступа к памяти, поскольку роль индексации для этого вида памяти — вспомогательная.

Обработка данных «по ту сторону памяти» предполагает обращение к «библиотекарию» — аналоговому, цифровому (аппаратному, программному) или гибриднему (цифровому, включающему аналоговые компоненты) устройству, систематизирующему блоки связанных данных исходя из их синтаксического или семантического содержания согласно установленным правилам. «Библиотекарь» является программным или аппаратным расширением механизма внимания (одного из ключевых элементов механизма внимания в рамках моделей рабочей памяти), выполняющим наряду с активацией данных (извлечением требуемых данных из долговременной памяти в кратковременную, в которой они могут быть использованы), функцию предварительной систематической индексации данных в долговременной памяти. Посылаемый через «библиотекаря» запрос может быть не только адресным, но и текстовым (синтаксическим) или семантическим, в ответ на который будет сформирована выборка соответствующих запросу адресов блоков связанных данных и/или будет извлечено их полное содержимое. При обработке данных «по ту сторону памяти» возможно радикальное повышение скорости работы, поскольку основная, наиболее ресурсоемкая часть обработки осуществляет не вычислительной или искусственной когнитивной системой, а (находящимся «по ту сторону памяти») специально созданным «библиотекарем», который непрерывно работает с памятью и использует высокопроизводительный инструментарий (в том числе аналоговый и гибридный), пригодный для решения прямой задачи обработки данных (сортировка, структурирование, классификация, определение связей, подготовка массивов с готовыми решениями типовых задач и т.д.), но неприменимый для решения обратной задачи (поиск блоков связанных данных, соответствующих заданным критериям).

Задача устранения очередей решается путем разделения общей памяти вычислительной или искусственной когнитивной системы на локальные области, связанные с функциональными или структурными подсистемами [26]. Для связи между подсистемами также необходима общая память, однако если подсистемы квазиавтономные и преобладающая часть операций обработки данных происходит локально (в пределах подсистем с собственной памятью), то емкость общей памяти может быть небольшой, а требования к ее скорости (ввиду малых объемов передаваемых данных) — невысокие. В естественных когнитивных системах лишь высшая нервная деятельность обладает высокой централизацией, все прочие функциональные и структурные системы обладают высокой автономией и наделены собственными локальными подсистемами памяти. Некоторые из этих подсистем (например, сенсорные) работают с очень большими объемами данных.

На практике рассмотренные выше задачи уменьшения объема обрабатываемых данных и увеличение скорости их передачи целесообразно решать не на уровне всей вычислительной или искусственной когнитивной системы, а на уровне функциональных или структурных подсистем, что, очевидно, существенно снижает сложность указанных задач, в том числе нивелирует проблему размещения большого числа элементов в пределах жестко ограниченного пространства.



Итак, резюмируя обозначенные подходы к совершенствованию архитектуры обработки данных, можно констатировать, что она должна обеспечивать многопоточность обработки данных (при этом потоки не обязательно гетерогенны, равноправны и реализованы сходным инструментарием), быть память-ориентированной (наибольшие перспективы имеет обработка данных «по ту сторону памяти») с памятью, разделенной на локальные области, привязанные к квазиавтономным функциональным или структурным подсистемам вычислительной или искусственной когнитивной системы.

Заключение

1. Ключевым фактором, определяющим эффективность обработки информации, является память. Задача повышения скорости получения данных из памяти, записи данных, а также реализации других операций по обработке информации решается в зависимости от организации памяти. Определение возможных вариантов организации памяти и путей ее совершенствования — приоритетная в контексте развития информационных технологий.

2. По мнению авторов, можно выделить три основных критерия функциональной классификации памяти: длительность хранения данных, степень формализации данных и осознанность данных. По длительности хранения данных память разделяется на кратковременную, кратковременную и рабочую. По степени формализации данных — на информативную и неинформативную. По осознанности — на имплицитную и эксплицитную. Последняя, в свою очередь, делится на семантическую, лично-синтаксическую (эпизодическую) и безлично-синтаксическую.

3. По организации хранения данных и используемым принципам доступа к ней память делится на адресуемую, стековую, ассоциативную и семантическую. Перечисленные виды памяти различаются характером используемой в них индексации. Для ассоциативной и семантической памяти индексация является необходимой составляющей их работы. Реализация вычислений с использованием этих видов памяти отнесена авторами к вычислениям «по ту сторону памяти», т.е. требующим операций обработки данных со стороны дополнительного устройства — «библиотекаря», упорядочивающего (индексирующего) данные в памяти.

4. Существующая проблема одновременного повышения сложности объектов управления, влекущей за собой снижения скорости системы управления, и необходимости повышения быстродействия управления по мере усложнения объектов управления, может быть решена только совершенствованием архитектуры обработки данных. Цель совершенствования архитектуры формализуется в виде трех задач: необходимо уменьшить объем обрабатываемого потока данных, увеличить скорость передачи данных и устранить очереди при одновременном обращении к одной памяти нескольких функциональных или структурных подсистем.

5. Решением поставленных задач является архитектура обработки данных, обеспечивающая многопоточность обработки данных, являющаяся память-ориентированной (с обработкой данных «по ту сторону памяти») с памятью, разделенной на локальные области, привязанные к квазиавтономным функциональным и структурным подсистемам.

Основными научными результатами, полученными авторами в рамках данного исследования, являются формализация критериев определения функциональных подсистем памяти (по длительности хранения данных, по формализации и осознанности данных), классификация цифровой памяти по организации хранения данных и используемым алгоритмам доступа к ним (адресуемая, стековая, ассоциативная, семантическая). Также в процессе исследования сформулирован и аргументирован методологический подход к представлению обработки данных в памяти через описание процесса рабочей памяти и его составляющих, предложена оригинальная авторская концепция память-ориентированной архитектуры обработки данных «по ту сторону памяти», основанная на использовании «библиотекаря» и позволяющая многократно повысить производительность вычислений (в том числе производительность в «жестком реальном времени»). Можно констатировать, что в результате исследования достигнуто существенное расширение знаний об организации памяти вычислительных и искусственных когнитивных систем.

Финансирование

Исследование выполнено при поддержке Российского научного фонда по гранту № 24-19-00692, <http://rscf.ru/project/24-19-00692/>.

Конфликт интересов

Не указан.

Рецензия

Шайхулова А.Ф., Уфимский университет науки и технологий, Уфа Российская Федерация
DOI: <https://doi.org/10.60797/IRJ.2026.165.70.1>

Funding

This research was supported by the Russian Science Foundation under grant No. 24-19-00692, <http://rscf.ru/project/24-19-00692/>.

Conflict of Interest

None declared.

Review

Shaykhulova A.F., Ufa University of Science and Technology, Ufa Russian Federation
DOI: <https://doi.org/10.60797/IRJ.2026.165.70.1>

Список литературы / References

1. Buschman T.J. Working Memory Is Complex and Dynamic, Like Your Thoughts / T.J. Buschman, E.K. Miller // *Journal of Cognitive Neuroscience*. — 2022. — Vol. 35. — № 1. — P. 17–23. — DOI: 10.1162/jocn_a_01940.
2. Khosla S. Survey on Memory-Augmented Neural Networks: Cognitive Insights to AI Applications / S. Khosla, Z.Z. Zhu, Y. He // *arXiv*. — 2023. — DOI: 10.48550/arXiv.2312.06141.
3. Mienye I.D. Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications / I.D. Mienye, T.G. Swart, G. Obaido // *Information (Switzerland)*. — 2024. — Vol. 15. — № 9. — 517 p. — DOI: 10.3390/info15090517.



4. Krichen M. Long Short-Term Memory Networks: A Comprehensive Survey / M. Krichen, A. Mihoub // AI. — 2025. — Vol. 6. — № 9. — 215 p. — DOI: 10.3390/ai6090215.
5. Furizal F. Long Short-Term Memory vs Gated Recurrent Unit: A Literature Review on the Performance of Deep Learning Methods in Temperature Time Series Forecasting / F. Furizal, A. Fawait, H. Maghfiroh [et al.] // International Journal of Robotics and Control Systems. — 2024. — Vol. 4. — № 3. — P. 1506–1526. — DOI: 10.31763/ijrcs.v4i3.1546.
6. Привалова И.Л. Кратковременная память: роль в обучении и физиологические механизмы (обзор литературы) / И.Л. Привалова, Е.В. Черных, Л.Н. Шульгина // Ученые записки Крымского федерального университета имени В.И. Вернадского. Биология. Химия. — 2023. — Т. 9. — № 4. — С. 191–203. — EDN KZRХВЕ.
7. Грибков А.А. Концептуализация памяти в рамках теории когнитивных систем / А.А. Грибков, А.А. Зеленский // Философская мысль. — 2025. — № 11. — С. 17–35. — DOI: 10.25136/2409-8728.2025.11.76544. — EDN JVPJJU.
8. Wang S. Capacity and Allocation across Sensory and Short-Term Memories / S. Wang, S.P. Tripathy, H. Ögmen // Vision. — 2022. — Vol. 6. — № 1. — 15 p. — DOI: 10.3390/vision6010015.
9. Tatsumi S. Relationship between autonomic nervous function and brain functions such as memory and attention / S. Tatsumi, D. Kuratsune, H. Kuratsune // Physiology & Behavior. — 2025. — Vol. 288. — DOI: 10.1016/j.physbeh.2024.114721.
10. Циркин В.И. Нейрофизиология: Физиология памяти : учебник для вузов / В.И. Циркин, С.И. Трухина, А.Н. Трухин. — Москва : Юрайт, 2021. — 407 с.
11. Sridhar S. Cognitive neuroscience perspective on memory: overview and summary / S. Sridhar, A. Khamaj, M.K. Asthana // Frontiers in Human Neuroscience. — 2023. — Vol. 17. — DOI: 10.3389/fnhum.2023.1217093.
12. Грибков А.А. Онтологизация познания: уровни онтологичности, границы и средства онтологизации / А.А. Грибков // Общество: философия, история, культура. — 2024. — № 5 (121). — С. 15–21. — DOI: 10.24158/fik.2024.5.1.
13. Грибков А.А. Эмпирико-метафизическая общая теория систем : монография / А.А. Грибков. — Москва : Академия Естествознания, 2024. — 360 с. — DOI: 10.17513/np.607.
14. Steinberg J. Associative memory of structured knowledge / J. Steinberg, H. Sompolinsky // Scientific Reports. — 2022. — Vol. 12. — № 1. — P. 1–15. — DOI: 10.1038/s41598-022-25708-y.
15. Yuan Z. Text Semantics-Driven Data Classification Storage Optimization / Z. Yuan, X. Lv, Y. Gong [et al.] // Applied Sciences (Switzerland). — 2024. — Vol. 14. — № 3. — DOI: 10.3390/app14031159.
16. Зеленский А.А. Реализуемость управления движением промышленных роботов, станков с ЧПУ и мехатронных систем. Часть 1 / А.А. Зеленский, А.П. Кузнецов, Ю.В. Илюхин [и др.] // Вестник машиностроения. — 2022. — № 11. — С. 43–51. — DOI: 10.36652/0042-4633-2022-11-43-51.
17. Зеленский А.А. Онтологические аспекты проблемы реализуемости управления сложными системами / А.А. Зеленский, А.А. Грибков // Философская мысль. — 2023. — № 12. — С. 21–31. — DOI: 10.25136/2409-8728.2023.12.68807.
18. Родина О.Н. О параллельных процессах в мышлении при решении задач / О.Н. Родина, П.Н. Прудков // Известия Российской академии образования. — 2022. — № 3 (59). — С. 147–161. — DOI: 10.51944/20738498_2022_3_147.
19. Khalil R. Adaptive Decision-Making "Fast" and "Slow": A Model of Creative Thinking / R. Khalil, M. Brüne // European Journal of Neuroscience. — 2025. — Vol. 61. — № 5. — DOI: 10.1111/ejn.70024.
20. Солдатова Г.У. Медиамногзадачность: от когнитивных функций к цифровой повседневности / Г.У. Солдатова, Е.Ю. Никонова, А.Г. Кошева [и др.] // Современная зарубежная психология. — 2020. — Т. 9. — № 4. — С. 8–21. — DOI: 10.17759/jmfp.2020090401.
21. Зеленский А.А. Основы формальной теории систем реального времени / А.А. Зеленский, А.А. Грибков // Информационно-управляющие системы. — 2025. — № 5 (138). — С. 2–10. — DOI: 10.31799/1684-8853-2025-5-2-10.
22. Khan A.A. The Landscape of Compute-near-memory and Compute-in-memory: A Research and Commercial Overview / A.A. Khan, J.P. Lima, H. Farzaneh [et al.] // arXiv. — 2024. — DOI: 10.48550/arXiv.2401.14428.
23. Khabbazan B. Towards Efficient LUT-based PIM: A Scalable and Low-Power Approach for Modern Workloads / B. Khabbazan, M. Riera, A. González // arXiv. — 2025. — DOI: 10.48550/arXiv.2502.02142.
24. Bowen P. Analog, In-memory Compute Architectures for Artificial Intelligence / P. Bowen, G. Regev, N. Regev [et al.] // arXiv. — 2023. — DOI: 10.48550/arXiv.2302.06417.
25. Leroux N. Analog in-memory computing attention mechanism for fast and energy-efficient large language models / N. Leroux, P.P. Manea, C. Sudarshan [et al.] // Nature Computational Science. — 2025. — Vol. 5. — № 9. — P. 813–824. — DOI: 10.1038/s43588-025-00854-1.
26. Зеленский А.А. Память-центрические модели систем управления движением промышленных роботов / А.А. Зеленский, Ю.В. Илюхин, А.А. Грибков // Вестник Московского авиационного института. — 2021. — Т. 28. — № 4. — С. 245–256. — DOI: 10.34759/vst-2021-4-245-256.

Список литературы на английском языке / References in English

1. Buschman T.J. Working Memory Is Complex and Dynamic, Like Your Thoughts / T.J. Buschman, E.K. Miller // Journal of Cognitive Neuroscience. — 2022. — Vol. 35. — № 1. — P. 17–23. — DOI: 10.1162/jocn_a_01940.
2. Khosla S. Survey on Memory-Augmented Neural Networks: Cognitive Insights to AI Applications / S. Khosla, Z.Z. Zhu, Y. He // arXiv. — 2023. — DOI: 10.48550/arXiv.2312.06141.
3. Mienye I.D. Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications / I.D. Mienye, T.G. Swart, G. Obaido // Information (Switzerland). — 2024. — Vol. 15. — № 9. — 517 p. — DOI: 10.3390/info15090517.



4. Krichen M. Long Short-Term Memory Networks: A Comprehensive Survey / M. Krichen, A. Mihoub // AI. — 2025. — Vol. 6. — № 9. — 215 p. — DOI: 10.3390/ai6090215.
5. Furizal F. Long Short-Term Memory vs Gated Recurrent Unit: A Literature Review on the Performance of Deep Learning Methods in Temperature Time Series Forecasting / F. Furizal, A. Fawait, H. Maghfiroh [et al.] // International Journal of Robotics and Control Systems. — 2024. — Vol. 4. — № 3. — P. 1506–1526. — DOI: 10.31763/ijrcs.v4i3.1546.
6. Privalova I.L. Kratkovremennaya pamyat': rol' v obuchenii i fiziologicheskie mekhanizmy (obzor literatury) [Short-term memory: role in learning and body machinery (literature review)] / I.L. Privalova, E.V. Chernykh, L.N. Shulgina // Uchenye zapiski Krymskogo federal'nogo universiteta imeni V.I. Vernadskogo. Biologiya. Khimiya [Scientific Notes of the V.I. Vernadsky Crimean Federal University. Biology. Chemistry]. — 2023. — Vol. 9. — № 4. — P. 191–203. — EDN KZR XBE. [in Russian]
7. Gribkov A.A. Konceptualizaciya pamyati v ramkakh teorii kognitivnykh sistem [Conceptualization of memory within the framework of cognitive systems theory] / A.A. Gribkov, A.A. Zelensky // Filosofskaya mysl' [Philosophical Thought]. — 2025. — № 11. — P. 17–35. — DOI: 10.25136/2409-8728.2025.11.76544. — EDN JVPJJU. [in Russian]
8. Wang S. Capacity and Allocation across Sensory and Short-Term Memories / S. Wang, S.P. Tripathy, H. Ögmen // Vision. — 2022. — Vol. 6. — № 1. — 15 p. — DOI: 10.3390/vision6010015.
9. Tatsumi S. Relationship between autonomic nervous function and brain functions such as memory and attention / S. Tatsumi, D. Kuratsune, H. Kuratsune // Physiology & Behavior. — 2025. — Vol. 288. — DOI: 10.1016/j.physbeh.2024.114721.
10. Tsirkin V.I. Nejrofiziologiya: Fiziologiya pamyati [Neurophysiology: Physiology of Memory] : textbook for universities / V.I. Tsirkin, S.I. Trukhina, A.N. Trukhin. — Moscow : Yurait, 2021. — 407 p. [in Russian]
11. Sridhar S. Cognitive neuroscience perspective on memory: overview and summary / S. Sridhar, A. Khamaj, M.K. Asthana // Frontiers in Human Neuroscience. — 2023. — Vol. 17. — DOI: 10.3389/fnhum.2023.1217093.
12. Gribkov A.A. Ontologizaciya poznaniya: urovni ontologichnosti, granicy i sredstva ontologizacii [Ontologization of cognition: levels, directions and means of implementation] / A.A. Gribkov // Obshchestvo: filosofiya, istoriya, kultura [Society: Philosophy, History, Culture]. — 2024. — № 5 (121). — P. 15–21. — DOI: 10.24158/fik.2024.5.1. [in Russian]
13. Gribkov A.A. Empiriko-metafizicheskaya obshchaya teoriya sistem [Empirical-metaphysical general systems theory] : monograph / A.A. Gribkov. — Moscow : Akademiya Estestvoznaniya, 2024. — 360 p. — DOI: 10.17513/np.607. [in Russian]
14. Steinberg J. Associative memory of structured knowledge / J. Steinberg, H. Sompolinsky // Scientific Reports. — 2022. — Vol. 12. — № 1. — P. 1–15. — DOI: 10.1038/s41598-022-25708-y.
15. Yuan Z. Text Semantics-Driven Data Classification Storage Optimization / Z. Yuan, X. Lv, Y. Gong [et al.] // Applied Sciences (Switzerland). — 2024. — Vol. 14. — № 3. — DOI: 10.3390/app14031159.
16. Zelensky A.A. Realizuemost' upravleniya dvizheniem promyshlennykh robotov, stankov s ChPU i mekhatronnykh sistem. Chast' 1 [Feasibility of motion control of industrial robots, CNC machine tools and mechatronic systems. Part 1] / A.A. Zelensky, A.P. Kuznetsov, Yu.V. Ilyukhin [et al.] // Vestnik mashinostroeniya [Bulletin of Mechanical Engineering]. — 2022. — № 11. — P. 43–51. — DOI: 10.36652/0042-4633-2022-11-43-51. [in Russian]
17. Zelensky A.A. Ontologicheskie aspekty problemy realizuemosti upravleniya slozhnymi sistemami [Ontological aspects of the problem of realizability of control of complex systems] / A.A. Zelensky, A.A. Gribkov // Filosofskaya mysl' [Philosophical Thought]. — 2023. — № 12. — P. 21–31. — DOI: 10.25136/2409-8728.2023.12.68807. [in Russian]
18. Rodina O.N. O parallel'nykh processakh v myshlenii pri reshenii zadach [On concurrent processes in thinking during problem solving] / O.N. Rodina, P.N. Prudkov // Izvestiya Rossijskoj akademii obrazovaniya [Proceedings of the Russian Academy of Education]. — 2022. — № 3 (59). — P. 147–161. — DOI: 10.51944/20738498_2022_3_147. [in Russian]
19. Khalil R. Adaptive Decision-Making "Fast" and "Slow": A Model of Creative Thinking / R. Khalil, M. Brüne // European Journal of Neuroscience. — 2025. — Vol. 61. — № 5. — DOI: 10.1111/ejn.70024.
20. Soldatova G.U. Mediamnogozadachnost': ot kognitivnykh funkcij k cifrovoj povsednevnosti [Media multitasking: from cognitive functions to digital] / G.U. Soldatova, E.Yu. Nikonova, A.G. Koshevaya [et al.] // Sovremennaya zarubezhnaya psikhologiya [Journal of Modern Foreign Psychology]. — 2020. — Vol. 9. — № 4. — P. 8–21. — DOI: 10.17759/jmfp.2020090401. [in Russian]
21. Zelensky A.A. Osnovy formal'noj teorii sistem real'nogo vremeni [Fundamentals of the formal theory of real-time systems] / A.A. Zelensky, A.A. Gribkov // Informacionno-upravlyayushchie sistemy [Information and Control Systems]. — 2025. — № 5 (138). — P. 2–10. — DOI: 10.31799/1684-8853-2025-5-2-10. [in Russian]
22. Khan A.A. The Landscape of Compute-near-memory and Compute-in-memory: A Research and Commercial Overview / A.A. Khan, J.P. Lima, H. Farzaneh [et al.] // arXiv. — 2024. — DOI: 10.48550/arXiv.2401.14428.
23. Khabbazan B. Towards Efficient LUT-based PIM: A Scalable and Low-Power Approach for Modern Workloads / B. Khabbazan, M. Riera, A. González // arXiv. — 2025. — DOI: 10.48550/arXiv.2502.02142.
24. Bowen P. Analog, In-memory Compute Architectures for Artificial Intelligence / P. Bowen, G. Regev, N. Regev [et al.] // arXiv. — 2023. — DOI: 10.48550/arXiv.2302.06417.
25. Leroux N. Analog in-memory computing attention mechanism for fast and energy-efficient large language models / N. Leroux, P.P. Manea, C. Sudarshan [et al.] // Nature Computational Science. — 2025. — Vol. 5. — № 9. — P. 813–824. — DOI: 10.1038/s43588-025-00854-1.
26. Zelensky A.A. Pamyat'-centricheskie modeli sistem upravleniya dvizheniem promyshlennykh robotov [Memory-centric models of industrial robot motion control systems] / A.A. Zelensky, Yu.V. Ilyukhin, A.A. Gribkov // Vestnik Moskovskogo aviacionnogo instituta [Bulletin of the Moscow Aviation Institute]. — 2021. — Vol. 28. — № 4. — P. 245–256. — DOI: 10.34759/vst-2021-4-245-256. [in Russian]