Гибридные подходы и моделирование активности человеческого мозга

БогдановА.В.[1], Гущанский Д.Е.,[1] Дегтярев А.Б.,[1] Лысов К.А.[1], Ананьева[2] Н. И., Незнанов Н.Г.,[2] Залуцкая Н.М.[2]
[1] Санкт-Петербургский государственный университет
[2] ФГБУ «Санкт Петербургский научно-исследовательский психоневрологический институт им. В.М. Бехтерева»

Резюме. Традиционный подход к моделированию человеческого мозга предполагает использование современных вычислительных устройств с последовательным наращиванием их мощностей до допустимого уровня. Альтернативой ему служит гибридное решение, основывающееся на концепции нейроморфных вычислений и представляющее собой комбинацию искусственных нейронных сетей, работающих на специально сконструированных для задачи аппаратных решениях. Особенности конструкции предполагают воспроизведение механизмов работы человеческого мозга, а созданные на их основе вычислительные устройства обеспечивают поддержку работы нейронных сетей. Существующие вычислительные модели мозга требуют значительного времени на обработку данных даже при запуске на суперкомпьютерах и в настоящее время не способны работать в режиме реального времени. Поскольку человеческий мозг состоит из двух полушарий, работающих и выполняющих разные функции, подход на основе комбинирования аналоговых и цифровых систем в единое архитектурное решение выглядит перспективно. Описанию результатов исследований человеческого мозга и его активности как основы для построения гибридных вычислительных систем и методам работы с ними и посвящена данная работа.

Ключевые слова: нейроинформатика, вычислительная нейробиология, высокопроизводительные вычисления, моделирование человеческого мозга.

Hybrid approaches and human brain activity modelling

Bogdanov A.V.[1], Gushchanskiy D.E.[1], Degtyarev A.B.[1], Lysov K.A.[1], Ananyeva N.I.[2], Neznanov N.G.[2], Zalutskaya N.M.[2],
[1] St. Petersburg University
[2] St.Petersburg V.M. Bekhterev Psychoneurological Research Institute

Summary. The traditional approach to human brain modeling suggests modification of modern systems and microcircuits as long as their performance reaches a permissible limit. A different hybrid approach is based on neuromorphic computing. The idea we utilize is combination of artificial neural networks with specialized microcircuits. The architecture of the microchip needs to reproduce the mechanisms of the human brain and to be a kind of hardware support for neural networks. Existing models of the brain even on powerful supercomputers require significant computation time and are not yet able to solve problems in real time. Since the human brain consists of two parts with different functions and different data processing principles, there is a very promising approach which suggests combining digital and analog systems into single one. In current collaboration we incorporate some results of study of activity of human brain as a base of building of hybrid computational system and foundation to the approach of running it.

***Key words*:** neuroinformatics, computational neuroscience high performance computing, brain modelling.

До недавнего времени одним из крупнейших споров нейробиологии велся вокруг того, как нейроны кодируют информацию. Было неясно, посылается ли информация в цифровой или аналоговой форме, либо мозг использует оба средства одновременно. Оказалось, что оба [1]. Открытыми остаются следующие вопросы:

1. Как классифицировать сигнал (в качестве дискретного или аналогового)?
2. Как определить, какую информацию он несет (провести декодирование)?

На данный момент нет единого мнения, как интерпретировать информацию при декодировании. Большинство ученых разделилось на приверженцев частотного кодирования (rate coding) [2] и приверженцев временного кодирования (temporal coding) [3]. Частотное кодирование подразумевает, что сигнал всегда несет какую-нибудь информацию и ее интерпретация зависит только от частоты сигнала. Временное же кодирование предполагает, что в любой момент времени сигнал может являться либо шумом, либо некоторой информацией, требующей интерпретации. В данной работе нас больше интересовало, как сигналы классифицировать в качестве аналоговых или цифровых, чтобы заложить основу для формирования гетерогенного вычислительного комплекса, где одна составляющая генерирует аналоговые сигналы, а другая — цифровые.

Цифровые сигналы, передаваемые обычными компьютерами, нисколько не похожи на аналоговые сигналы, применяемые в старых телевизорах и радиоприемниках. Отличить их друг от друга просто, чего нельзя сказать того же о нейронных сигналах — там разделить цифровые и аналоговые сигналы довольно сложно. Нейробиологам давно

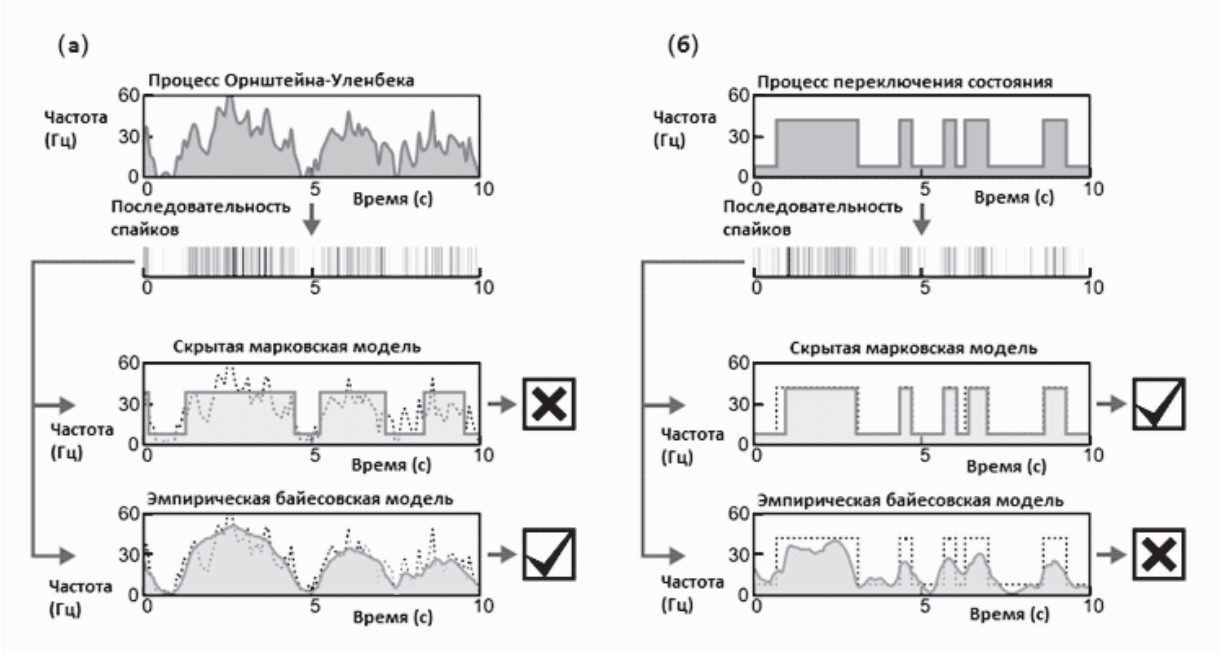

Рис. 3. (а) Последовательность спайков, сгенерированная с помощью процесса Орнштейна-Уленбека (Ornstein-Uhlenbeck Process) (синий). Эмпирическая байесовская модель (зеленый) лучше аппроксимирует сигнал, чем скрытая марковская модель (оранжевый). (б) Последовательность спайков, полученная с помощью процесса переключения состояния (Switching State process) (красный). Скрытая марковская модель лучше аппроксимирует сигнал, чем эмпирическая байесовская модель

известно, что нейроны передают сигналы в форме электрических импульсов, которые называются биоэлектрическими потенциалами или «спайками». Несколько взятых вместе спайков называется последовательностью спайков. Точный способ кодирования информации в спайках неизвестен, однако ученые открыли как минимум два протокола кодирования. В 1990 году нейробиологи обнаружили, что напряжение мышцы зависит от количества «спайков» в определенный период времени, от скорости их прибывания. Этот вид сигнала имеет лишь два состояния — включено или выключено — так что он определенно является цифровым. Однако другие нейробиологи утверждают, что информация может быть закодирована и по-другому — посредством разницы во времени между отдельными спайками при их прибытии. Это аналоговое кодирование.

Сложность заключается в разграничении этих двух сигналов, поскольку они оба зависят от характеристики спайков, которые путешествуют по нейрону. Этот вопрос вызывает частые споры среди нейробиологов, поскольку отсутствует согласия относительно того, когда сигнал является цифровым, а когда аналоговым.

Не так давно японские физики Ясухиро Мотидзуки (Yasuhiro Mochizuki) и Сигеру Синомото (Shigeru Shinomoto) из Университета Киото разработали способ автоматического определения вида кодирования [6]. Способ основан на идее о том, что некоторые статистические модели лучше выражают цифровой код, чем аналоговый код, и наоборот.

Метод довольно прямолинеен. Ученые анализируют сигнал нейрона и затем стараются повторить его сначала с помощью эмпирической байесовской модели, а затем — с помощью скрытой марковской модели. Далее на основе модели, которая лучше отражает характеристики первоначального сигнала, они определяют, является ли сигнал аналоговым или цифровым. Получается, что если эмпирическая байесовская модель лучше отражает сигнал, тогда сигнал, вероятно, аналоговый, если же скрытая марковская модель подходит лучше, тогда сигнал, скорее всего, цифровой (рис. 3) [6].

Данный подход был проверен на сигналах, которые возникали в разных частях мозга длиннохвостых макак, и подтвердил, что разные части мозга используют разные формы кодирования. Это дает повод проверить на практике обратную ситуацию, когда будет построена система, комбинирующая аналоговые и цифровые сигналы для генерации нейронных спайков.

Резистивные процессорные устройства

В последние годы в рамках DARPA SyNAPSE [4] и Human Brain Project [5] был разработан ряд нейроморфных (грубо повторяющих структуру нейронов и синапсов в мозгу человека) архитектур, реализующих концепцию резистивных про-

Таблица 1. Сравнение CPU Power8, Nvidia Tesla K40 и различных архитектур систем RPU.

Система	Производительность, тераопс	Энергопотребление, Ватт	Энергоэффективность, гигаопс/Ватт	Размер сети, число весов	Коэффициент ускорения (по сравнению с CPU)
CPU Power8 12 ядер	0.676	250	2.7	-	1
GPU NVidia Tesla K40	4.3	242	17.8	-	6.4
Архитектура 1	5000	250	20100	200 млн.	7400
Архитектура 2	21000	250	83800	840 млн.	31000
Архитектура 3	420	22	19000	1680 млн.	620

цессорных устройств (Resistive Processing Unit, RPU) [7]. RPU — вычислительный элемент, аналоговый по своей природе, небольшой по размерам и способный восстанавливать свою историю, чтобы обучаться. Он получает множество аналоговых данных, в форме напряжений, и на основе прошлого опыта использует взвешенную функцию из них, чтобы решить, какой результат передавать на следующий слой вычислительных элементов. Синапсы имеют озадачивающее и пока неполностью понятное положение в мозге человека, но чипы из RPU организованы в двумерные массивы.

Одно ядро IBM Power8 CPU может достичь пиковой производительности порядка 50 гигафлопс, чего должно быть достаточно для поддержки одного тайла RPU. Однако, предельная мощность будет достигнута уже при 12 тайлах при расходе 20 Ватт на одно ядро. Энергоэффективность этого решения (Архитектура 1 в табл. 1) будет равняться 20 тераопс/Ватт. Аналогичные вычислительные ресурсы могут быть обеспечены 32 ядрами GPU, но с большей энергоэффективностью, позволяя тем самым параллельную работу до 50 тайлов. Энергоэффективность такого решения (Архитектура 2 в таблице 1) оценивается как 84 тераопс/Ватт. Дальнейшее увеличение числа тайлов, способных работать параллельно, может быть осуществлено посредством создания энергоэффективных цифровых микросхем с минимальной занимаемой площадью, оперирующих числами с плавающей точкой с ограниченным битовым разрешением. Альтернативный подход (Архитектура 3 в таблице 1) может быть основан на нескольких вычислительных ядрах, обрабатывающих данные тайлы последовательно. Последовательная обработка необходима для поддержки большего числа тайлов, что, в свою очередь, позволяет работать с сетями большего размера. Например, микросхема с 100 тайлами и одно вычислительное ядро с производительностью 50 гигаопс будут способны работать с сетью, содержащей более чем 1.6 миллиарда весов, потребляя при этом около 22 Ватт — 20 Ватт на поддержку работы процессора и передачу данных по шине, а остальные 2 Ватт — на блок RPU, поскольку в один момент времени будет активен только один тайл. Это дает энергоэффективность порядка 20 тераопс/Ватт, что на 4 порядка лучше, чем CPU и GPU.

Для глубоких нейронных сетей, содержащих более миллиарда весов, архитектура RPU с массовым параллелизмом может достигать 30000-кратного ускорения по сравнению с высокоэффективными микропроцессорами, обладая при этом энергоэффективностью в 84000 гигаопс/Ватт. Задачи, требующие многодневной тренировки сети на кластерах с тысячами машин, могут быть разрешены за часы с использованием только одного RPU-ускорителя. Система из нескольких RPU-ускорителей будет способна обрабатывать задачи «Больших Данных» с триллионами параметров, которые невозможно успешно решать на современной технике. К таким задачам, например, относятся распознавание речи с одновременным переводом на мировые языки, анализ в реальном времени больших потоков научных или финансовых данных, интеграция и анализ разнородных потоков данных, снятых с сенсоров значительного количества устройств IoT (Internet of Things, Интернет Вещей).

Из-за того, что RPU специализированы и не требуют преобразования аналоговой информации в цифровую или доступа к какой-либо памяти, кроме своей собственной, они могут быть быстры и поглощать мало энергии. Поэтому, теоретически, сложная нейронная сеть может быть напрямую смоделирована путем выделения одного RPU к одному программному нейрону. К сожалению, RPU неточен из-за своей аналоговой природы и обилия шума в схемах, поэтому алгоритм должен иметь устойчивость к «врожденным» неточностям в RPU.

Программирование

Нейрокомпьютер NS16e работает с нейронными сетями в режиме реального времени. Для работы он соединяется с сервером на основе архитектуры x86 (цифровой подсистемой) через шину PCI Express (рис. 4). Сервер может загружать на нейрокомпьютер и выгружать с него большие объемы данных. При наличии на сервере GPU он может тренировать нейронные сети большого объема, которые сразу же могут быть запущены на NS16e. Процесс тонко настраивается: изменения можно вносить в рамках подготовки тренировочных данных, механизма обучения сети, ее оптимизации под конкретное аппаратное обеспе-

чение. При этом весь процесс может быть запущен одной командой.

Базовые вычисления проводятся на массиве 4x4 TrueNorth плат, взаимодействующих друг с другом посредством асинхронного протокола без использования дополнительных интерфейсных плат.

Через PCI Express сервер может собирать и отправлять данные со скоростью 500Мб/с. Микросхемы ППВМ на NS16e служат в роли моста между сервером и аппаратными нейронными схемами, таким образом выступая в роли переводчика между фон-неймановской и нейронной архитектурами, оперирующими разными понятиями. Компьютер на основе архитектуры фон Неймана оперирует инструкциями и бинарными данными, в то время как нейронная архитектура — пиковыми сигналами между нейронами. При вводе команды сервер отправляет нейронную модель через шину и загружает ее на микросхемы NS16e. В терминале на сервере можно следить за процессом загрузки, а после, при проведении спайки нейронов — и за тем, как много генерируется спайков и как обновляются нейроны.

Часть нашего мозга отвечает за восприятие, в то время как остальная часть — за моторные функции. Подобного разбиения можно достичь и в нейрокомпьютерах: отдельной микросхеме можно сопоставить для обработки конкретный участой нейронной сети. Например, каждой микросхеме можно сопоставить слой нейронной сети, или набор слоев, ответственных за распознавание конкретного фрагмента. Работа с этой особенностью называется задачей размещения ядра (core placement problem) [8]. В общем случае, в нее входят попытки перенести ядра в наилучшее с точки зрения топологии место, чтобы повысить внутреннюю скорость передачи данных внутри системы, строящейся из сетей на чипах. Из-за этой особенности две идентичные сети могут работать с разной скоростью, поскольку, например, одна из них использует передачу информации внутри микросхемы, а другая передает ту же информацию между микросхемами.

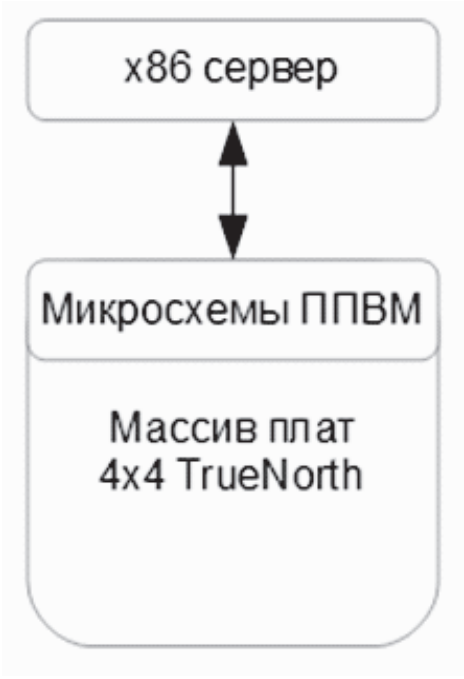

Рис. 4. Схема вычислительной системы на основе нейрокомпьютера

Написание программ для NS16e осуществляется посредством интегрированной системы и входящими в нее утилитами разработки — DevKit. В программировании для нейрокомпьютера можно выделить три ключевых этапа:

1. Анализ данных и предобработка. Входные данные конвертируются в стандартный формат, понятный всем средствам разработки, и трансформируются в наборы признаков. Данные рекомендуется размещать в Lighting Memory-Mapped Database (LMDB) [9], высокопроизводительной встроенной транзакционной базе данных формата «ключ-значение». Данный формат популярен в среде глубокого машинного обучения, поскольку быстрота чтения данных LMDB позволяет быстро переносить их на GPU для обработки. Для импорта и предобработки данных существует консольная утилита tn-signal-processor, написанная на языке программирования C++. Она способна импортировать JPG и PNG файлы в LMDB и применять к ним различные преобразования: кадрирование, вращение, фильтрация, кодирование и декодирование в пики, визуализация и другие.

2. Обучение. Обучение строится на предварительно подготовленных и отформатированных данных. Сверточные сети TrueNorth (TrueNorth Convolutional Networks, TNCN) [10] — это инструментарий для создания и обучения нейронных сетей, удовлетворяющих архитектурным особенностям TrueNorth. Использование программной среды Corelet Programming Environment (CPE) вместе с TNCN освобождает программиста от части работы с абстракциями: последовательностями слоев, размерами фильтров, точностью данных и т.п.

3. Сопоставление логических ядер файла модели и физического массива микрочипов в нейрокомпьютере. В случае системы с несколькими схемами задача приобретает нетривиальный характер, поскольку передача данных между элементами может критически ограничить производительность аппаратной платформы.

Для эвристического размещения логических ядер файла модели на физическую архитектуру используется утилита Neuro Synaptic Core Placer (NSCP). Она минимизирует пересечения микросхем в ядре графа связности. Идентификация биологических нейронных сетей осуществляется посредством анализа функциональной магнитно-резонансной томографии (фМРТ), в основе которой лежит оксигенация синхронного сигнала активации (BOLD-сигнал) в определенных областях мозга в состоянии покоя и при предъявлении определенных парадигм [11].

Результаты

При анализе данных фМРТ состояния покоя (fMRI resting state) в системе SPM 12.0 с последующим анализом методом независимых компонентов было выбрано 8.нейрональных сетей. По сравнению с остальными эти компоненты давали яркое визуальное синхронной для каждой сети отображение основных биологических нейронных сетей. Они идентифицируются как:

1) левая передняя теменная, левая боковая, левая нейронная сеть «рабочая память», левая нейронная сеть «внимание»;

2) правая передняя теменная, правая боковая нейронная сеть, правая нейронная сеть «рабочая память», правая нейронная сеть «внимание»;

3) спинная передняя временная нейронная сеть, нейронная сеть «осуществление управления»;

4) передняя теменная вентральная нейронная сеть, языковая нейронная сеть;

5) визуальная нейронная сеть;

6) слуховая нейронная сеть;

7) сенсорная моторная нейронная сеть;

8) нейронная сеть «по умолчанию» [12].

Далее проводился сравнительный анализ трудности функциональной связности для ключевых узлов «стандартных» биологических нейронных сетей. Анализ показывает нарушение структуры и синхронности операций в нейронной сети «по умолчанию» (№ 8) у пациентов с аффективными и когнитивными нарушениями при ряде заболеваний мозга (рис. 5) [13].

К примеру, в сравнении со здоровыми добровольцами схожего возраста и пола, у пациентов с депрессией в сочетании с когнитивными нарушениями было найдено изменение в работе нейронной сети (рис. 6, 7). Компьютерное моделирование мозговой активности предоставляет возможность оценки влияния ошибок на результаты исследований, ограничивая на их основе потенциальные парадигмы оперирования модели.

Исследования переноса возбуждения в человеческом мозге делают возможным использование

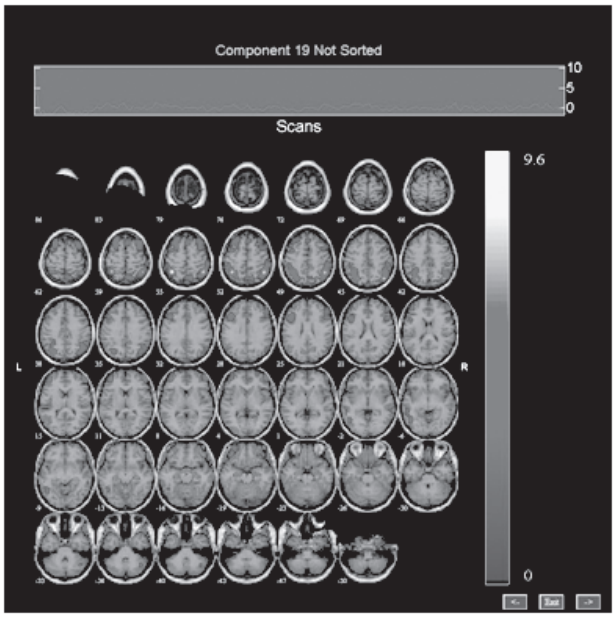

Рис. 5. Сеть в состоянии покоя

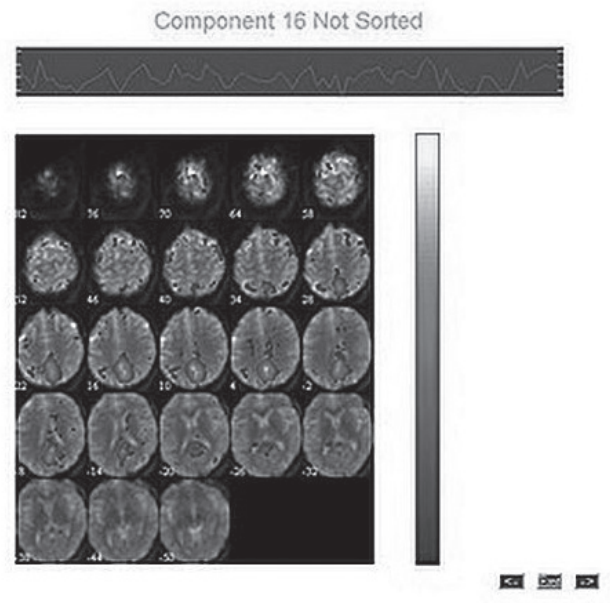

Рис. 6. МРТ пациента с депрессией и когнитивными нарушениями

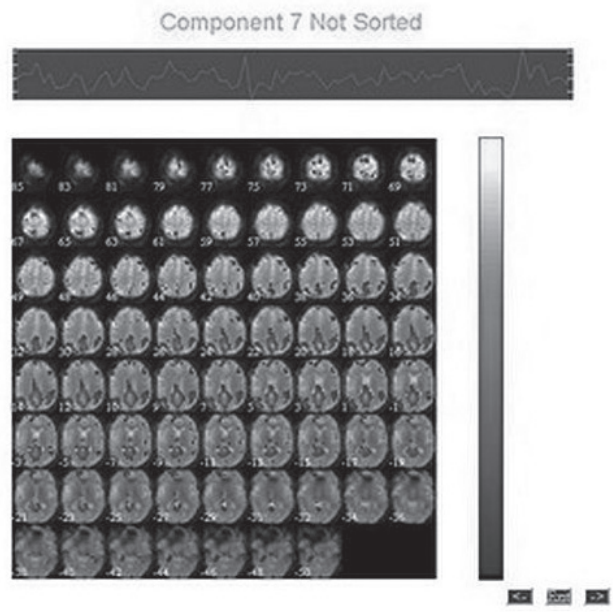

Рис. 7. МРТ здорового человека

реальных биологических данных для выработки набора команд и подготовки сценариев работы систем, спроектированных по виду, представленному на рис. 4.

Заключение

Интеграция цифровых и аналоговых подсистем в единый вычислительный комплекс представлялась сложной задачей из-за значительной разницы в скорости работы цифровых и аналоговых устройств. Ситуация радикально изменилась с релизом компанией IBM ряда продуктов — от вычислительных чипов до программного обеспечения — направленных на разработку нейрокомпьютеров. Технологии IBM позволили увеличить скорость нейрокомпьютеров и соединять их с цифровыми системами. Особое внимание в дальнейших исследованиях следует уделять возможности вычислительных объединения ресурсов с разнообразнейшими характеристиками в единый пул, а также их интеграции в гибридные распределенные среды многопоточных процессоров.

В рамках данной совместной работы были показаны результаты исследований активности человеческого мозга в качестве основы для построения гибридных вычислительных систем и основы для метода работы с ними.

Благодарности

Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда (грант 17-06-23047) и Российского фонда фундаментальных исследований (гранты 16-07-00886 и 17-04-00147).

Литература

1. *Thorpe S.J. Spike arrival times: A highly efficient coding scheme for neural networks In Eckmiller, R.; Hartmann, G.; Hauske, G. Parallel processing in neural systems and computers. North-Holland, 1990. pp. 91–94.*
2. *Kandel, E.; Schwartz, J.; Jessel, T.M. Principles of Neural Science (3rd ed.). Elsevier, 1991.*
3. *Dayan, Peter; Abbott, L. F. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. Massachusetts Institute of Technology Press, 2001.*
4. *Broad Agency Announcement. Systems of Neuromorphic Adaptive Plastic Scalable Electronics. DARPA-BAA, 2008. — https://www.fbo.gov/download/0b6/0b62b2149395d4bd8a28dff1b9046944/BAA08-28.doc*
5. *The Human Brain Project. A Report to the European Commission. — https://ec.europa.eu/research/participants/portal/doc/call/h2020/fetflag-1-2014/1595110-6pilots-hbp-publicreport_en.pdf*
6. *Yasuhiro Mochizuki, Shigeru Shinomoto. Analog and digital codes in the brain. Department of Physics, Kyoto University, Kyoto 606-8502, Japan, 2013. — http://arxiv.org/pdf/1311.4035v1.pdf*
7. *Tayfun Gokmen, Yurii Vlasov. Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices. IBM T. J. Watson Research Center. — https://arxiv.org/ftp/arxiv/papers/1603/1603.07341.pdf*
8. *Jonas Gomes Filho, Marius Strum, and Wang Jiang Chau. Using Genetic Algorithms for Hardware Core Placement and Mapping in NoC-Based Reconfigurable Systems — International Journal of Reconfigurable Computing, vol. 2015.*
9. *Lightning Memory-Mapped Database (LMDB) — https://en.wikipedia.org/wiki/Lightning_Memory-Mapped_Database*
10. *Arnon Amir, Pallab Datta, William P. Risk, Andrew S. Cassidy, Jeffrey A. Kusnitz, Steve K. Esser, Alexander Andreopoulos, Norm Pass, Dharmendra S. Modha. Cognitive Computing Programming Paradigm: A Corelet Language for Composing Networks of Neurosynaptic Cores. IBM Research -http://www.research.ibm.com/software/IBMResearch/multimedia/IJCNN2013.corelet-language.pdf*
11. *Wasserman L., Ananiev N., Wasserman E., Ivanov M., Mazo G., Neznanov N., Gorelik A., Yezhova R., Ershov B., Sorokina A., Yanushko M. Neurocognitive Deficits and Depressive Disorders: Structural-Functional Approach in Comparative Multivariate Researches. V.M. Bekhterev Revue of Psychiatry and Medical Psychology. 2013. № 4. P. 58-67.*
12. *Wasserman L., Ananieva N., Gorelik A., Yezhov R., Ershov B., Lipatov L., Folomeeva K., Chuikova A. Affective-Cognitive Disorders: Research Methology Of Structural And Functional Relationship On Temporal Lobe Epilepsy Model. Bulletin of South Ural State University. Serie: Psychology. 2013. T. 6. № 1. P. 67-71.*
13. *Kissin M., Ananieva N., Shmeleva L., Yezhov R. Features of Neuromorphology of Anxiety and Depressive Disorders in Temporal Lobe Epilepsy. V.M. Bekhterev Revue of Psychiatry and Medical Psychology. 2012. № 2. P. 11-17.*

Сведения об авторах

Ананьева Наталья Исаевна — доктор медицинских наук, профессор, руководитель отделения клинико-диагностических исследований ФГБУ «Санкт Петербургский научно-исследовательский психоневрологический институт им. В.М. Бехтерева». E-mail: ananieva_n@mail.ru.

Богданов Александр Владимирович — д. ф.-м. н., профессор кафедры компьютерного моделирования и многопроцессорных систем СПбГУ. E-mail: bogdanov@csa.ru

Гущанский Дмитрий Евгеньевич, ассистент кафедры Компьютерного моделирования и многопроцессорных систем, СПбГУ, dmitriy.guschanskiy@spbu.ru

Дегтярев Александр Борисович, д. т. н., профессор кафедры Компьютерного моделирования и многопроцессорных систем, СПбГУ, deg@csa.ru

Залуцкая Наталья Михайловна, кандидат медицинских наук, доцент, ведущий научный сотрудник отделения гериатрической психиатрии ФГБУ «Санкт Петербургский научно-исследовательский психоневрологический институт им. В.М. Бехтерева», e-mail: nzalutskaya@yandex.ru

Лысов Кирилл Александрович, студент СПбГУ, thereis9000@gmail.com

Незнанов Николай Григорьевич — доктор медицинских наук, профессор, директор ФГБУ «Санкт-Петербургский научно-исследовательский психоневрологический институт им. В.М. Бехтерева», научный руководитель отделения гериатрической психиатрии ФГБУ «Санкт-Петербургский научно-исследовательский психоневрологический институт им. В.М. Бехтерева», e-mail: spbinstb@bekhterev.ru