Санкт-Петербургский Государственный Университет
Кафедра компьютерного моделирования и многопроцессорных систем

Лысов Кирилл Александрович

Выпускная квалификационная работа бакалавра

Использование IBM Bluemix – инструментария для создания систем управления большими данными

Направление 010300
Фундаментальная информатика и информационные технологии

Научный руководитель,
кандидат физ.-мат. наук,
Академик Европейской академии Евросайенс и РАЕН, профессор
Богданов А. В.

Санкт-Петербург
2016

Оглавление

ВВЕДЕНИЕ

Сегодня Facebook, Google и другие технологические компании решают множество интеллектуальных задач, применяя традиционные компьютеры и чипы. Рано или поздно они сталкиваются с «бутылочным горлышком» в пропускной способности и производительности своих систем, особенно если существует острая необходимость в обучении нейронных сетей (время обучения может отнимать дни и недели). Вся индустрия электроники ищет пути решения данной проблемы.

Традиционным подходом принято считать эксплуатацию и модификацию современных систем и чипов до тех пор, пока их производительность не будет достигать предела. Другой подход, нетрадиционный, ищет возможности в вычислениях, вдохновленных человеческим мозгом, **нейроморфных вычислениях**.

Идея нейроморфных вычислений (neuromorphic computing) была предложена инженером Карвером Мидом (Carver Mead) [1] в 80-х гг. прошлого века и заключалась в применении искусственных нейронных сетей в комбинации со специализированными чипами, архитектура которых напоминала бы структуру человеческого мозга и являлась бы аппаратной поддержкой нейронных сетей.

Такие чипы должны резко увеличить производительность нейронных сетей и буквально совершить прорыв в областях их применения:

- **Финансы**: предсказание кредитного рейтинга, банкротства; определение мошенничества; определение платежеспособности клиентов; прогноз цен; прогноз экономических показателей.

- **Медицина**: постановка диагноза; определение стоимости лечения.
- **Промышленность**: контроль процессов; контроль качества; предсказание температуры и силы.
- **Анализ данных**: предсказание; классификация; обнаружение отклонений; анализ временных рядов; извлечение знаний.
- **Маркетинг**: предсказание продаж; целевой маркетинг.
- **Операционный анализ**: оптимизация инвентаря; оптимизация расписаний; принятие решений.
- **HR (Human Resources)**: отбор кандидатов; составление расписаний; профилирование персонала.
- **Энергетика**: предсказание нагрузки на электросети; предсказание потребностей в энергии; предсказание цены на бензин/уголь; системы контроля энергии; мониторинг гидроэлектростанций.
- **Наука**: определение паттернов; определение веществ; моделирование физических систем; оценка экосистем; ботаническая классификация; обработка сигналов; анализ биологических систем.
- **Образование:** обучающие нейронные сети; оценка работ обучающихся; предсказание производительности студентов.
- **Другое:** спортивные ставки; разработка видеоигр; транспортные проблемы.

Ключевым аспектом нейроморфных вычислений является понимание того, как комбинация индивидуальных нейронов, микросхем, приложений, архитектур создает желаемые вычисления, влияет на

представление информации, устойчивость к повреждениям, сочетает обучение и разработку, адаптируется к локальным изменениям (свойство пластичности) и совершает эволюционные преобразования.

Нейроморфные вычисления являются междисциплинарным предметом, сочетающим биологию, физику, математику, компьютерные науки, электронику, чтобы создавать нейронные системы, системы компьютерного зрения, слуховые процессоры, автономных роботов, физическая архитектура и дизайнерские решения которых основаны на принципах биологических систем. [2]

ПОСТАНОВКА ЗАДАЧИ

Целью данной работы является сравнительный анализ существующих решений среди нейроморфных архитектур с последующим выбором лучшего в качестве основы для рабочей станции разработки нейронных сетей, а так же формирование такой конфигурации оборудования, которое в комбинации с рабочей станцией позволило бы достичь максимальной производительности.

Решение данной задачи потребует ознакомления с предметной областью нейроморфных вычислений и обзора существующих программ, в рамках которых ведется работа над нейроморфными технологиями.

ОБЗОР ЛИТЕРАТУРЫ

При написании данной работы были использованы научные статьи, блоги и материалы представителей Корнелльского

университета, Университета Мичигана, Калифорнийского технологического института, Национальной лаборатории имени Лоуренса в Беркли, лаборатории HRL, IBM J. T. Watson Research Center, IBM Research – Almaden, DARPA и Human Brain Project.

История появления нейроморфных вычислений описана в интервью Ширли Коэн с инженером Карвером Мидом от 17 июля 1996 г. [1], начиная с трех университетских курсов (искусственные нейронные сети Джона Хопфилда, нейроморфные аналоговые схемы Карвера Мида, физика и вычисления Ричарда Фейнмана) и заканчивая кооперацией специалистов из различных сфер знаний (когнитивная и поведенческая биология, физика, математика, компьютерные науки и т. д.) и созданием Центра Инженерии Нейроморфных систем (Center for Neuromorphic Systems Engineering , CNSE).

Концепция нейроморфного эволюционирующего аппаратного обеспечения (neuromorphic evolvable hardware) предложена в работе Боду и Галлагера «Качественный анализ эволюционирующих нейроморфных контроллеров полета» [2], в которой было продемонстрировано преимущество гибридного контроллера, основанного на рекурретных нейронных сетях (RNN), относительно традиционного в плане общей устойчивости и эффективности реакции к непредвиденным обстоятельствам в произвольном окружении.

Концепция резистивного вычислительного устройства (resistive processing unit, RPU) предложена в работе Тайфана Гокмена и Юрия Власова «Ускорение обучения глубоких нейронных сетей резистивными связными устройствами» [3]. В работе описано устройство, которое может хранить и обрабатывать данные локально, а сеть из таких

устройств ускоряет обучение глубокой нейронной сети (DNN). Проведено сравнение разных вариантов системы на RPU с системами на CPU и GPU, система на RPU продемострировала большее число синаптических связей при меньшем потреблении энергии.

Архитектура на основе спайковых нейронов (spiking neuron) рассмотрена в работе Эндрю Нера, Умберто Олчезе, Дэвида Бэлдуззи, Гулио Тонони «Нейроморфная архитектура с временной пластичностью для распознавания объектов и движений» [4]. Архитектура устойчиво выполняла задачи по распознаванию объектов и движений даже при наличии отвлекающих объектов и шума.

Первый удачно спроектированный конфигурируемый чип, эффективно реализующий нейроморфную архитектуру описан в работе Джае-сун Сео, Бернарда Бреззо, Йонг Лу, Бенджамина Паркера, Стивена Эссера, Роберта Монтои, Бипина Ражендрана, Хосе Тиерно, Лиланда Чанга, Дармендры Моды «Нейроморфный 45nm CMOS чип с масштабируемой архитектурой для обучения сетей спайковых нейронов». Чип продемонстрировал высокую эффективность в задачах распознавания образов и задач с ассоциативной памятью.

Программа SyNAPSE описана в одном из широковещательных объявлений (Broad Agency Announcement, BDA) агенства DARPA «Системы нейроморфной адаптивной пластичной масштабируемой электроники» [6]. В ней речь идет о широкомасштабной поддержке агенством разработок систем, воспроизводящих биологические процессы и функционал мозга млекопитающего.

Мозг кота с 10^9 нейронов и 10^{13} был симулирован в работе Раджагопала Анантанарайнана, Стивена Эссера, Хорста Саймона,

Дармендры Моды «Кот из мешка: кортикальные симуляции 10^9 нейронов, 10^{13} синапсов» [7]. В Ливерморской национальной лаборатории был построен суперкомпьютер с 147,456 CPU и 144TB основной памяти, проведено две симуляции: одна с 1.6 млрд. нейронов и 8.87 трлн. синапсов с экспериментальной таламокортикальной связностью серого вещества, вторая – с 900 млн. нейронов и 9 трлн. синапсов с вероятностной связностью. Были продемострированы практически идеальные слабая и сильная масштабируемости.

Критика симуляции мозга кота была описана в открытом письме руководителя программы Blue Brain Project Генри Маркрама «Заявление IBM - обман» [9]. Основная идея – симуляция слишком упрощена, отсутствует биологический реализм.

Еще один, более сложный чип с массивом 16 x 16 (256) нейронов, реализующий метод «интегрировать-и-сработать» с утечками и 262,144 синапсов рассмотрен в работе Пола Мероллы, Джона Артура, Филиппа Акопяна, Набила Имама, Раджита Маноара, Дармендры Моды «Цифровое нейросинаптическое ядро, использующее поперечную память с 45pJ на спайк в 45nm» [12]. Был сконструирован и протестирован чип с «поперечной» (crossbar) памятью, решающей проблему бутылочного горлышка при обращении к памяти вне чипа.

Система, воспроизводящая громерулярный слой обонятельной колбы млекопитающего описана в работе Набила Имама, Томаса Клиланда, Раджита Манохара, Пола Мероллы, Джона Артура, Филиппа Акопяна, Дармендры Моды «Реализация громерулярного слоя обонятельной колбы с помощью нейросинаптического ядра» [13]. Система применяет нейросинаптическое ядро [12] для симуляции связи

между митральными клетками, перигломерулярными клетками, клетками с короткими аксонами внутри обонятельной колбы.

Концепция мемристора (memory + resistor = memristor) подробно описана в работе Кук-Кван Кима, Сидхарта Габы, Даны Вилер, Хосе Круза-Альбрехта, Тахира Хуссейна, Нараяна Ширивазы «Функционирующий гибридный мемристорный массив с поперечной памятью для хранения данных и нейроморфных приложений» [15]. Был построен первый в мире функционирующий мемристорный массив. Особенность мемристора заключается в том, что этот элемент может изменять и сохранять свое сопротивление в зависимости от протекшего через него заряда.

Программа Compass, позволяющая симулировать нейроморфную архитектуру на любом суперкомпьютере описана в работе Роберта Прейсла, Теодора Вонга, Паллаба Датты, Мирона Фликнера, Рагвендры Синга, Стивена Эссера, Вильяма Риска, Хорста Саймона, Дармендры Моды «Compass: масштабируемый симулятор для архитектур когнитивных вычислений» [16]. Было продемонстрировано множество приложений TrueNorth без наличия самого чипа с такой архитектурой: классификация аудиозаписей и изображений, распознавание символов и т. д.

Язык для программирования в парадигме нейроморфных архитектур Corelet описан в работе Мирона Фликнера, Родриго Альвареза-Иказы, Эммета МакКуинна, Бена Шоу, Норма Пасса, Дармендры Моды «Парадигма программирования когнитивных вычислений: язык Corelet для формирования сетей нейросинаптических ядер» [22]. Corelet был разработан для эффективного наложения

логической структуры нейронной сети на железо, поскольку последовательное выполнение команд фон Неймановской архитектуры не подходит для программирования сетей нейросинаптических ядер.

Цели и задачи проекта Human Brain Project описаны в докладе для Европейской Комиссии [27]. Сам проект схож с DARPA SyNAPSE [6], но более объемен в плане инвестиций и количестве участников.

ГЛАВА 1. Программы по развитию нейроморфных вычислений

1.1. DARPA SyNAPSE

SyNAPSE – программа, финансируемая DARPA (Рис. 1-2.) для разработки нейроморфных машинных технологий, воспроизводящих биологические процессы. Говоря проще, это попытка создать новый вид компьютера со схожей формой и функционалом, что и мозг млекопитающего. Такой искусственный мозг будет использован для создания роботов с интеллектом как у мышей или котов.

Рис. 1-2. Логотипы: слева – программа SyNAPSE, справа - DARPA

SyNAPSE – бэкроним для Systems of Neuromorphic Adaptive Plastic Scalable Electronics (Системы из Нейроморфной Адаптивной Пластичной Масштабируемой Электроники). Начался проект в 2008 г. и на январь

2013 г. получил $102.6 млн. финансирования. Программа связана в основном с IBM и HRL, которые, в свою очередь, предоставляют результаты своих исследований разным американским университетам.

Главной целью является построение микропроцессорной системы, которая повторяет мозг млекопитающего по функционалу, размеру и потреблению энергии. Она должна воспроизводить 10 миллиардов нейронов, 100 триллионов синапсов, потреблять один киловатт и занимать меньше, чем 2 литра объема.

1.1.1. Предпосылки создания

На протяжении шести десятилетий современная электроника прошла через серию крупных разработок (транзисторы, интегрированные микросхемы, память, микропроцессоры), что привело к повсеместному распространению программируемых электронных машин. Из-за ограничений по аппаратной части и архитектуре эти машины ограничены в сложных условиях реального мира, требующих интеллекта, который еще не встречался в алгоритмически-вычислительной парадигме. Программа SyNAPSE ставит целью перелом в парадигме программируемых машин и ищет новый путь для создания полезных интеллектуальных машин. [6]

Виденье программы DARPA SyNAPSE заключается в создании доступной электронной нейроморфной машинной технологии, воспроизводящей биологические процессы. Программируемые машины ограничены не только их вычислительной мощностью, но и архитектурой, требующей написанные человеком алгоритмы для описания и обработки информации из их окружения. Биологические

нейронные системы, в свою очередь, автономно обрабатывают информацию в сложных окружениях и автоматически узнают релевантные и вероятностно стабильные свойства и ассоциации. Поскольку системы реального мира имеют множество составляющих с бесконечной комбинаторной сложностью, то нейроморфные электронные машины были бы предпочтительными в качестве ведущих приложений, но полезных практических реализаций пока не существует.

Ключом к достижению виденья программы SyNAPSE является беспрецедентный мультидисциплинарный подход, который может координировать агрессивные технологические разработки в следующих областях: 1) аппаратное обеспечение; 2) архитектура; 3) симуляция; и 4) окружение.

- **Аппаратное обеспечение** – реализация будет включать CMOS устройства, новые синаптические компоненты и комбинации проводных и программных/виртуальных связностей. Они будут поддерживать техники обработки критической информации, наблюдаемые в биологических системах спайковое кодирование (spike encoding) и временную пластичность (spike-timing-dependent plasticity).
- **Архитектура** – будет поддерживать критические структуры и функции, наблюдаемые в биологических системах, такие как: связность, иерархическая организация, схема основного компонента, соревновательная самоорганизация, модулирующие/усилительные системы. Как в биологических системах, обработка будет максимально распределенной, нелинейной, с врожденной толерантностью к шумам и дефектам.

- **Симуляция** – цифровые симуляции схем и систем в больших масштабах будут использованы для подтверждения функциональности отдельных компонент так и систем в целом, а так же для информирования общей разработки системы наперед, благодаря нейроморфной аппаратной реализации.
- **Окружение** – эволюция, виртуальные платформы для обучения, оценка и бенчмаркинг интеллектуальных машин по различным параметрам восприятия, познания и реакции.

Понимание этой амбициозной цели потребует сотрудничества множества технических дисциплин, таких как: вычислительная нейронаука, искусственные нейронные сети, высокопроизводительные вычисления, нейроморфная интегральная схемотехника (VLSI), информатика, когнитивные науки, науки о материалах, неконвенциональная наноэлектроника, дизайн и производство CMOS.

1.1.2. Результаты

Симуляция мозга кота

IBM разработали мощный параллельный кортексный симулятор под названием C2. Он работает на Blue Gene/P (Рис. 3.) суперкомпьютере под назвонием Dawn в Ливерморской национальной лаборатории. Суперкомпьютер имел 147,456 CPU и 144 террабайта основной памяти. Самая большая кортексная симуляция состояла из 1.6 млрд. нейронов и 8.87 трлн. синапсов. Это соответствует уровню кошачьего кортекса и 4.5% человеческого. Симуляция повторялась в 643 раз медленнее, чем в реальном мире. Симуляцие объединяла спайковые нейроны, STDP, аксонные задержки. Шаг симуляции был 0.1 msec.

Рис. 3. суперкомпьютер Dawn – Blue Gene/P.

Архитектура и связность симулированной сети была вдохновлена биологическими процессами (Рис. 4.). Она включала зрительную кору, сопутствующие секции таламуса и ретикулярное ядро. Регионы симулированного кортекса были сконструированы из таламокортикальных модулей. Каждый модуль имел 10,000 кортикальных нейронов, 334 таламных нейронов, 130 ретикулярных ядер. В каждом модуле кортикальные нейроны были разделены в 4 слоя (настоящий мозг млекопитающего имеет 6 слоев). Уровень возбудимости ингибиторных нейронов был так же смоделирован на экспериментальных данных. Самая большая модель имела 278 x 278 модулей, что соответствует 1.6 млрд. нейронов.

SpikeStream – фреймворк для получения сенсорной информации, закодированной в спайках. Спайки были закодированы для представления геометрических визуальных объектов и слуховых высказываний алфавита.

BrainCam – фреймворк для записи срабатываний всех нейронов и записи в видео для убедительной визуализации – схож по принципу с EEG trace. Видео (150 MB mpeg) показывает как раздражитель в форме

букв “IBM” распространяется. Скорость и паттерн распространения соответствует наблюдениям у животных. Симуляция так же воспроизводит альфа волны (8-12 Hz) и гамма волны (>30 Hz), которые часто наблюдаются у млекопитающих.

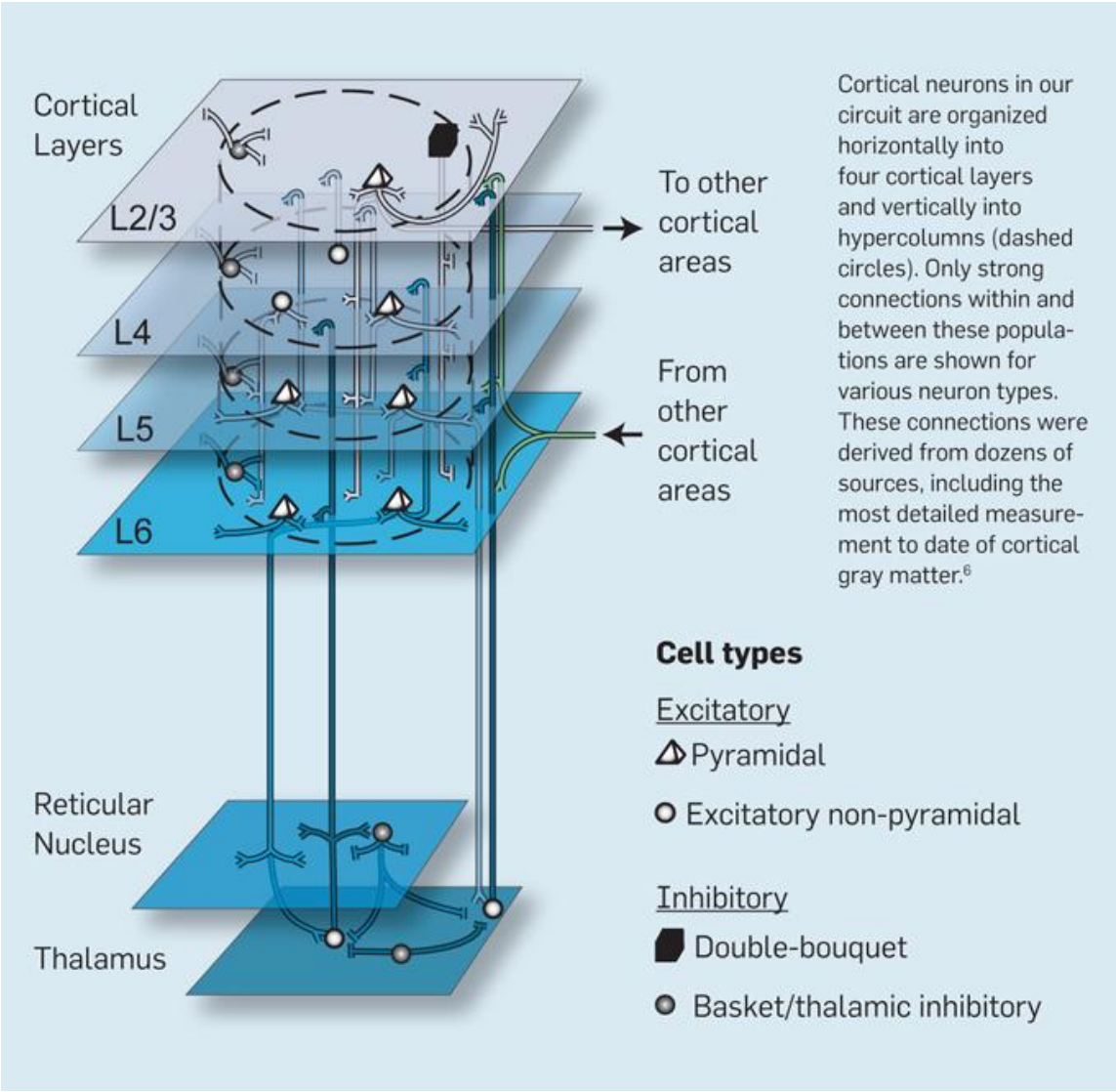

Рис. 4. Кортикальная модель, используемая в симуляции.

Будущие планы заключаются в обогащении модели длинной связностью между кортикальными зонами, повышении разрешения путем уменьшения размера каждого модуля от 10 тыс. нейронов до 100 тыс. Предсказано, что для 100% человеческого мозга симуляция

потребует 4 петабайта памяти и суперкомпьютер >1 экзафлопс. Это должно быть достигнуто до 2018 г., если суперкомпьютеры будут развиваться с той же скоростью, что и в последние десятилетия. [7]

Критика симуляции мозга кота

После анонса IBM симуляции мозга кота Генри Маркрам из Blue Brain Project [8] опубликовал сильную критику данного заявления. Он назвал это «публичный мегатрюк – чистейший случай научного обмана публики». Маркрам написал в открытом письме [9], что эти симуляции не соответствуют сложности даже мозга муравья.

Первый аргумент Маркрама заключался в том, что число симулированных нейронов грубо соответствует мозгу кота, модель для каждого нейрона слишком проста. Нейроны как отдельные точки не имеют биологического реализма. Подходящая симуляция реальных нейронов требует решения в миллион раз большего решения уравнений, чем было у IBM. И даже в этом случае не будет симулирована и миллионная часть мозга кота.

Второй аргумент заключался в том, что большие симуляции тривиальных нейронов были проведены годами ранее. Действительно, Юджин Ижикевич [10] (сейчас CEO Brain Corporation [11]) провел 100-миллардную нейронную симуляцию в 2005 г. Нерецензированная статья, опубликованная IBM не внесла ничего нового и интересного.

Цифровое нейросинаптическое ядро

В августе 2011 г. IBM объявили, что они они построили цифровое нейросинаптическое ядро. Микропроцессор реализует 256 нейронов

метода «интегрировать-и-сработать» с утечками (Рис. 5, 6.). Нейроны выстроены в массив 16 x 16. Каждый нейрон соединен с другими с помощью 1,024 синапсов, составляя 262,144 синапсов на ядро.

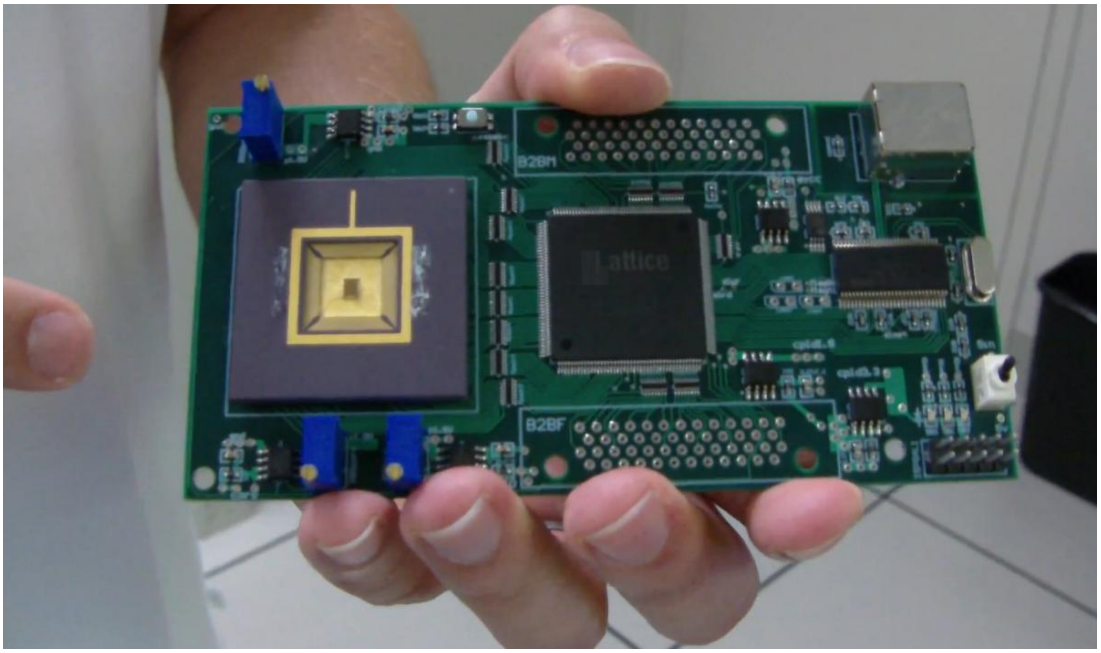

Рис. 5. Нейросинаптическое ядро и плата.

Был применен 45 nm SOI процесс сборки. Это был новейший тип сборки для коммерческих ПК в 2008 г. В августе 2012 г. они поставлялись уже по 22 nm сборке. Целое ядро имеет 3.8 млн. транзисторов и помещается в 4.2 mm2. Каждый нейрон занимает 35 μm x 95 μm, реальный же нейрон занимает от 4 до 100 μm в диаметре.

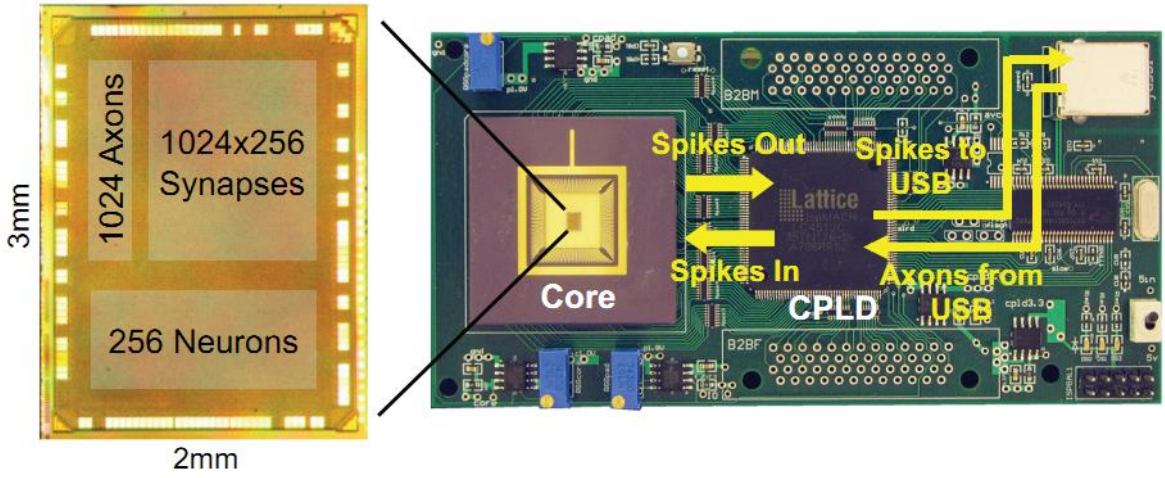

Рис. 6. Схема нейросинаптического ядра и платы.

Ядро (Рис. 7.) было поставлено на специально отпечатанную плату и соединено с ПК через USB. Таким путем оно может контактировать с разными виртуальными и реальными окружениями. Ядро научилось распознавать рукописный текст и играть в пинг понг.

Ядро полностью детерминистичное. Это отличает его от предыдущего аналогового нейроморфного аппаратного обеспечения, чувствительного к разным конструкциям и окружающим температурам. Чип имел такт ~1 kHz, отвечающем ~1 ms биологического шага. Внутри такт ~1kHz был использован и для других вычислений.

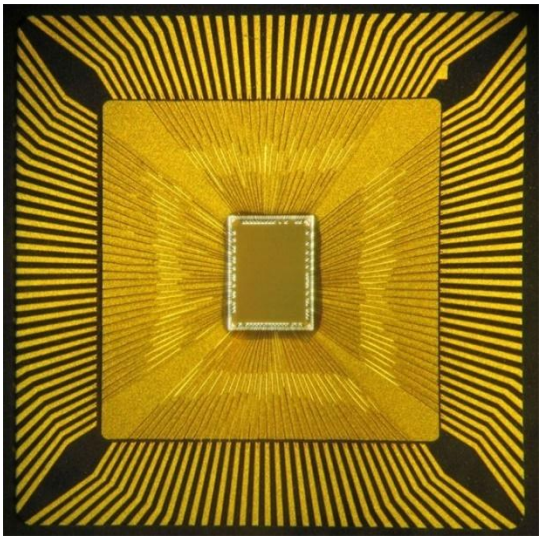

Рис. 7 Микроснимок нейросинаптического ядра

В отличие от традиционной Фон Неймановской архитектуры, вычислительные элементы и элементы памяти сильно интегрированы. Это ускоряет высокопараллельные вычисления, снижает потребляемую энергию. Теоретически возможно построить большую сеть из этих ядер, создавая энергоэффективную «нейронную ткань» для широкого спектра приложений. Конечной целью является построение человекоподобной системы с 100 трлн. синапсов. [5] [12]

Реализация громерулярного слоя обонятельной колбы

В июне 2012 команда SyNAPSE презентовала систему, которая использовала приведенный выше нейроморфный чип, чтобы воспроизвести главные функциональные свойства громерулярного слоя обонятельной колбы млекопитающего. Нейронные схемы в чипе отображают связи между митральными клетками, перигломерулярными клетками, клетками с внешним хохолком, клетками с короткими аксонами внутри обонятельной колбы.

Схемы потребляют только 45 pJ энергии за спайк с напряжением 0.85 V и могут быть использованы для обработки в химически чувствительных устройствах с малым энергопотреблением. [13]

IBM Brain Wall

Так называемая «стена мозгов» (Рис. 8.) является инструментом визуализации, построенным IBM в Almaden research center в Калифорнии. Она позволяет ученым обозревать состояния активации нейрона в большой нейронной сети. Паттерны нейронной активности могут быть осмотрены при перемещении по сети.

Массив 4 x 4 мониторов может отображать 262,144 нейрона одновременно. Каждый нейрон представляется одним серым пикселем. Большие сети могут быть визуализированы в будущем с помощью группировки нескольких нейронов в один пиксель. Инструмент может быть использован для визуализации суперкомпьютерных симуляций и активности внутри нейросинаптического ядра.

Рис. 8. IBM Brain Wall.

Мемристор

HRL анонсировали в декабре 2011 г., что они построили мемристорный массив поверх CMOS чипа. Это был первый в мире функционирующий мемристорный массив.

Из-за высокой плотности и низких энергозатрат мемристорная технология является важной для продолжения закона Мура. Чип HRL имеет мультибитовую полноадресную память с плотностью 30 Gbits/cm^2. Такая плотность является беспрецедентной в микроэлектронике.

Одновременное хранение и логическая обработка делает мемристоры подходящими для нейроморфных вычислений. Память и логические устройства являются одним и тем же, нечтом вроде нейронных схем мозга.

Гибридная система от HRL (Рис. 9.) может надежно хранить 1,600-пиксельные изображения используя новую схему программирования. Команда планирует масштабировать чип для поддержки эмуляции миллионов нейронов и миллиардов синапсов. Работа финансировалась программой SyNAPSE и Национальным научным фондом (NSF).

В будущем возможно, что эта мемристорная технология может быть использована для реализации разных вариантов нейросинаптического ядра, описанного выше. Используя мемристоры, эти ядра могут быть уменьшены в размерах и энергопотреблении, что дает возможность для построения больших массивов ядер с необходимым числом нейронов для симуляции мозга человека. [14] [15]

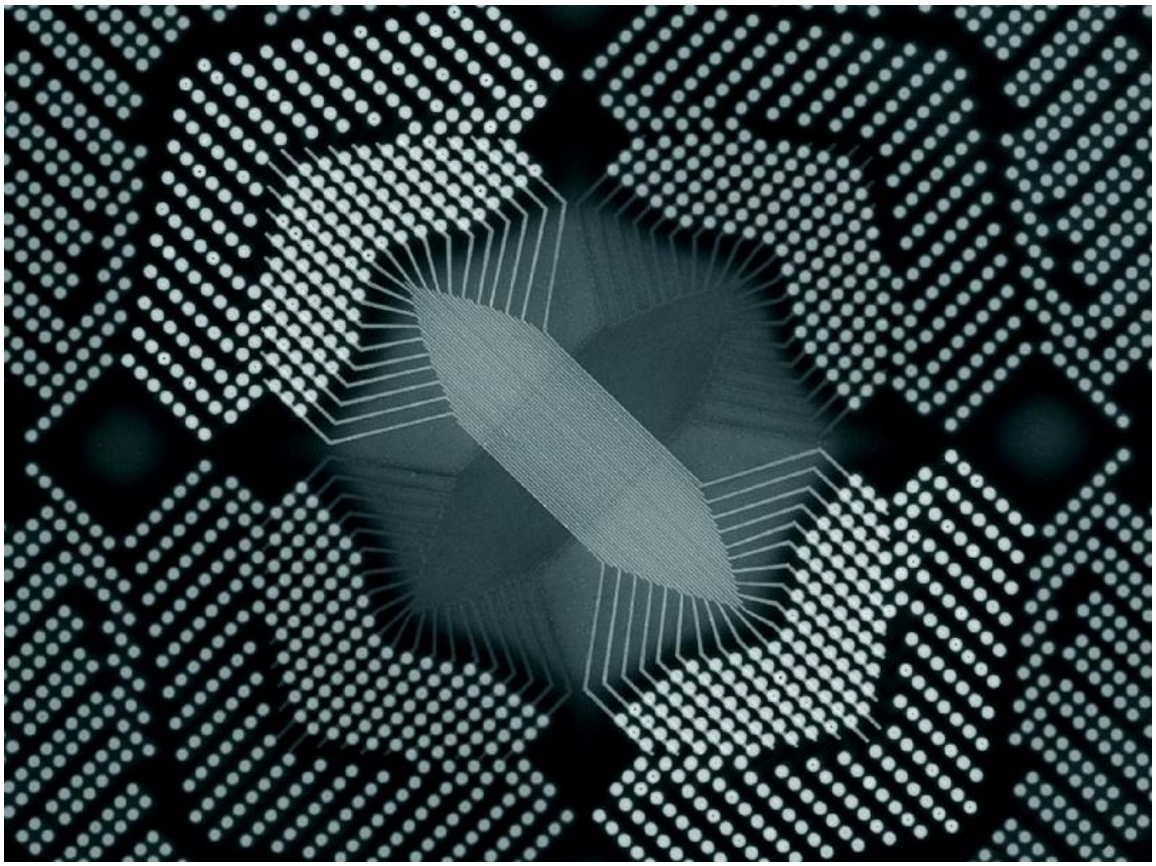

Рис. 9. Мемристорный поперечный массив от HRL.

Нейроморфная архитектура

Нейроморфная архитектура (Рис. 10) содержит 766 спайковых искусственных нейронов, организованных в слои по подобию иерархии в человеческом мозгу. Так же она использует нейроны с методом «интегрировать-и-саботировать» с утечками (LIF) и простые бинарные синапсы, и это позволяет устойчиво распознавать образы, движения, внимание к важным предметам и управлять двигателем. Это было доказано при тестировании сети в симуляции на стандартном компьютере. Сеть утилизирует взрывной STDP и синаптическую гомеостатическую ренормализацию, две относительно новые идеи в области спайковых нейронных сетей.

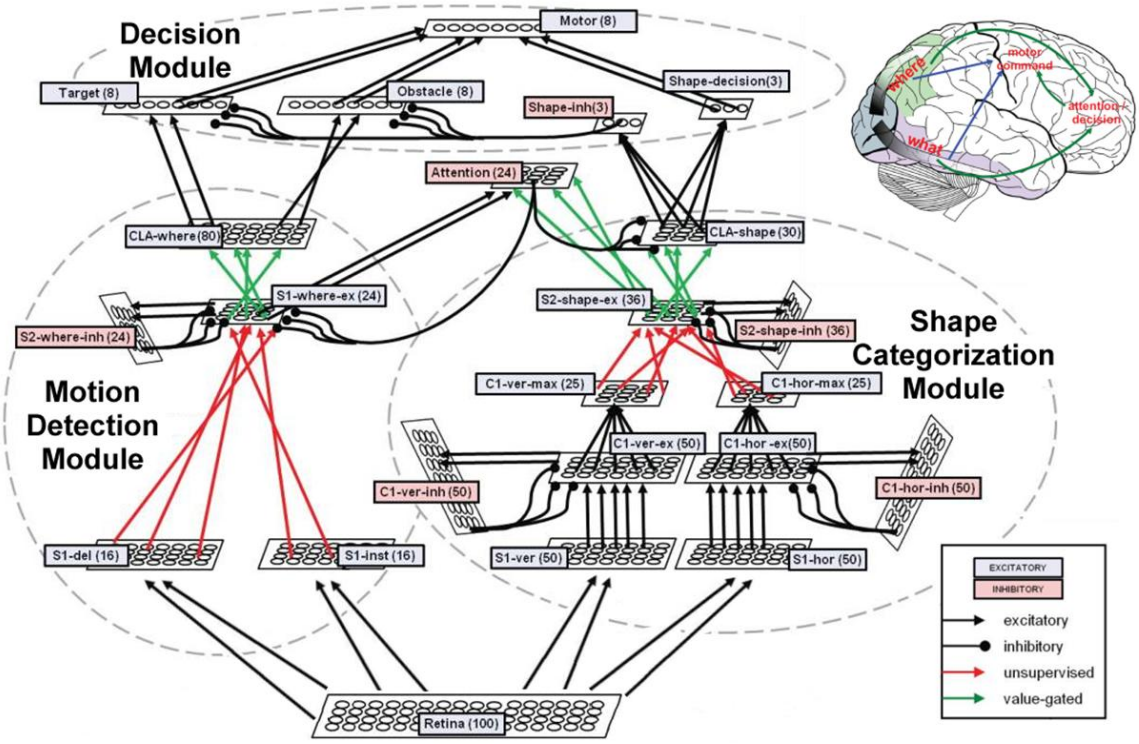

Рис. 10. Минимальная нейроморфная архитектура.

Архитектура была создана с возможностью разворачивания на цифровых нейросинаптических ядрах, описанных выше. IBM работают

над внутриядерными коммуникациями для построения большой сети из этих ядер на одном чипе. С появляением данного аппаратного обеспечения станет возможным построение архитектуры, описанной выше. Схема из 766 нейронов является только минимальным прототипом. Станет возможным масштабирование до тысяч, сотен тысяч нейронов с появлением данного аппаратного обеспечения. [4]

TrueNorth и Compass

TrueNorth – новая модульная, масштабируемая, не фон Неймановская, с низким энергопотреблением, когнитивная вычислительная архитектура, разрабатываемая IBM как часть программы SyNAPSE. Она содержит масштабируемую сеть из нейросинаптических ядер, каждое ядро содержит нейроны, дендриты, синапсы и аксоны.

Compass, так же разработанное IBM, программное обеспечение, симулирующее архитектуру TrueNorth. Позволяет тестировать архитектуру на обычном суперкомпьютере перед непосредственным встраиванием в специализированное нейроморфное аппаратное обеспечение. Помимо того, что Compass - мультитредовый и высокопараллельный функциональный симулятор, он так же является параллельным компилятором, который может отображать сеть из долгих нейронных путей в мозгу макаки на TrueNorth (Рис. 11.).

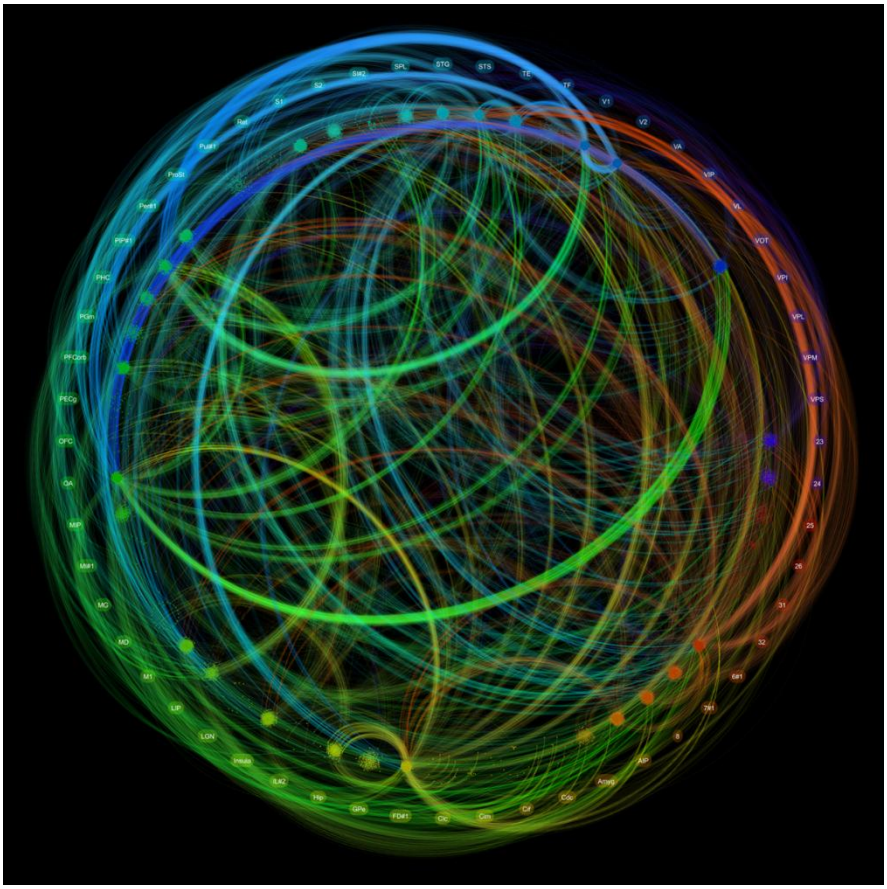

Рис. 11. Визуализация нейронных пути в мозгу макаки, симулированных TrueNorth архитектурой.

IBM и Национальная лаборатория им. Лоуренса в Беркли запускают Compass на 96 Blue Gene/Q стойках суперкомпьютера Sequoia Ливерморской национальной лаборатории. Sequoia является одним из мощнейших суперкомпьютеров (место в TOP 500 [17]). 96 стоек содержат 1,572,864 ядер и 1.5 петабайт памяти. Система была способна симулировать TrueNorth архитектуру до масштаба 2.084 млрд. нейросинаптических ядер, содержащих $53x10^{10}$ нейронов и $1.37x10^{4}$ синапсов. Нейроны имели средний уровень спайкования в 8.1 Hz, но они работали в 1.542 раза медленнее, чем в реальном времени. Система демонстрировала почти идеальную слабую масштабируемость.

Для сравнения, конечной целью DARPA SyNAPSE является когнитивная архитектура с 10^{10} нейронов и 10^{14} синапсов. Это аппроксимирует примерное число нейронов и синапсов в мозгу человека.

Но несмотря на примерно схожее число нейронов и синапсов модель TrueNorth не является реалистичной биологической симуляцией челвоеческого мозга. Вычисления (нейроны), память (синапсы), коммуникации (аксоны, дендриты) математически абстрагированы от биологических деталей для инжинерных целей максимизации функционала, минимизации стоимости, минимизации сложности аппаратной реализации.

Compass использовался для демонстрации множества приложений архитектуры TrueNorth: оптический поток, механизмы внимания, классификация аудио и изображений, распознавание символов, навигация роботов. [16]

Мультиядерный нейросинаптический чип

Нейроморфная вычислительная лаборатория IBM и Корнельского университета [18] работает над вторым поколением нейросинаптических процессоров. Нейросинаптические ядра, как и в первом поколении, будут имеют по 256 нейронов. Внутриядерные коммуникации были переработаны и новые процессоры содержат по ~4,000 ядер, по миллиону нейронов на процессор.

1.2. The Human Brain Project

The Human Brain Project – это исследовательский проект, целью которого является моделирование мозга человека с помощью суперкомпьютеров для лучшего понимания функционирования мозга [27]. Одними из конечных целей проекта являются способность создания компьютерной имитации человеческого мозга и способность лучшей диагностики различных медицинских проблем мозга.

Главной целью проекта является конструирование реалистических имитаций мозга человека – это потребует информации на молекулярном и клеточном уровне, которая даст возможность смоделировать и понять биологические и медицинские процессы. Кроме того, эту информацию можно будет использовать для проектирования и реализации новых типов компьютеров и роботов.

Основная цель проекта - создать единую открытую платформу для экспериментов с имитацией функций человеческого мозга, некий единый открытый формат исследований. Можно будет разрабатывать и новые компьютерные модели имитации мозга, и тестировать новые методы лечения болезней мозга.

Начальная стадия проекта рассчитана на период 10 лет и он является самым крупным и амбициозным проектом по имитации человеческого мозга. Проект HBP должен стать для исследователей мозга чем-то вроде ЦЕРН. ЦЕРН (CERN) — расположенный вблизи Женевы Европейский Центр ядерных исследований (Европейская организация по ядерным исследованиям), крупнейшая в мире лаборатория физики высоких энергий.

Мозг человека состоит примерно из 100 млрд. нейронов и 100 трлн. синаптических связей. Для моделирования подобной искусственной нейросети нужны большие организационные, аппаратные, программные ресурсы, что требует объединения усилий учёных из разных стран.

Проект Human Brain Project должен стать стандартной платформой для исследователей всего мира, где будут накапливаться экспериментальные данные из разных источников и они будут доступны всем.

Структура

Проект организован в виде 13 подпроектов:

1. Мозг мыши – Стратегические данные / SP1 - Strategic Mouse Brain Data
2. Мозг человека – Стратегические данные / SP2 - Strategic Human Brain Data
3. Когнитивные архитектуры / SP3 - Cognitive Architectures
4. Математические и теоретические основы исследований мозга / SP4 - Mathematical and Theoretical Foundations of Brain Research
5. Нейроинформатика (платформа) / SP5 - Neuroinformatics (Neuroinformatics Platform)
6. Моделирование мозга (платформа) / SP6 - Brain Simulation (Brain Simulation Platform)
7. Сверхпроизводительные вычисления (платформа) / SP7 - High Performance Computing (High Performance Computing Platform)

8. Медицинская информатика (платформа) / SP8 - Medical Informatics (Medical Informatics Platform)
9. Нейроморфные вычисления (платформа) / SP9 - Neuromorphic Computing (Neuromorphic Computing Platform)
10. Нейроробототехника (платформа) / SP10 - Neurorobotics (Neurorobotics Platform)
11. Применение / SP11 - Applications
12. Этика и общество / SP12 - Ethics and Society
13. Менеджмент / SP13 - Management

Помимо отдельных подпроектов стратегической целью проекта одновременно является разработка общедоступных платформ информационно-коммуникационных технологий в шести областях исследований:

- Нейроинформатика / Neuroinformatics
- Моделирование мозга /Brain simulation
- Сверхпроизводительные вычисления / High-performance computing
- Медицинская информатика / Medical informatics
- Нейроморфные вычисления / Neuromorphic computing
- Нейроробототехника / Neurorobotics

Для всего проекта и каждого подпроекта и платформы определены свои координаторы и участники.

ГЛАВА 2. Сравнительный анализ нейроморфных архитектур

2.1. HiCANN

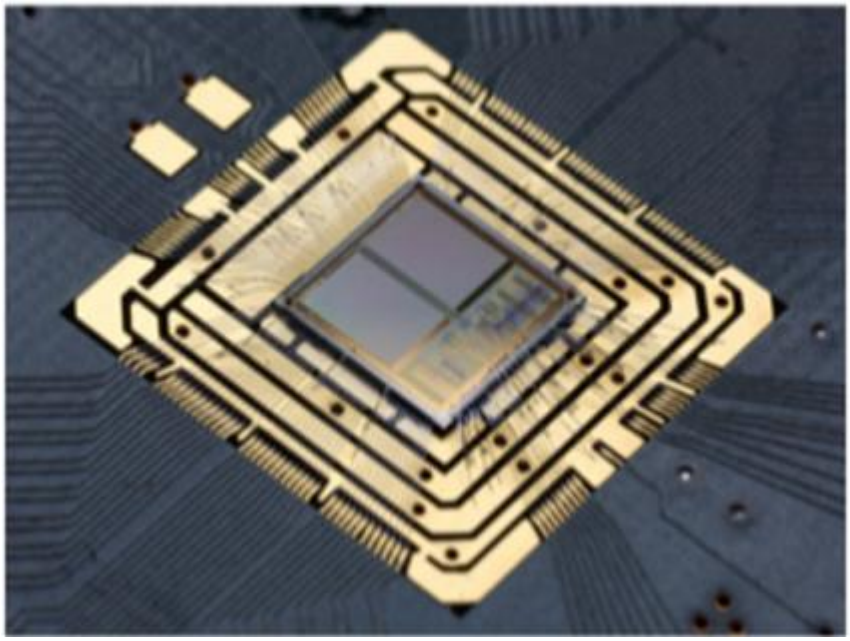

Рис. 12. Чип архитектуры HiCANN.

Разработчики: Гейдельбергский университет, Human Brain Project.

Спецификация чипа: 512 нейронов, 114,688 синапсов.

Потребляемая мощность: 3,000 mW/cm^2.

Особенности: симулирует мозговую активность, ускоряя ее в 10,000 раз, чтобы упаковать сутки в 10 секунд; фокус на обучении мозга и пластичности.

Цифровой или аналоговый: гибридный.

Тех. процесс: 180 nm (цель – 65nm).

Самая большая конфигурация на данный момент: 6 полупроводниковых пластин; 1.2 млн. нейронов; 300 млн. синапсов.

След. конфигурация: 20 полупроводниковых пластин; 4 млн. нейронов; 1 млрд. синапсов.

Конечная конфигурация: 5,000 полупроводниковых пластин; 5 млрд. нейронов; 1.3 трлн. синапсов.

2.2. SpiNNaker

Рис. 13. Чип архитектуры SpiNNaker.

Разработчики: Университет Манчестера, Human Brain Project.

Спецификация чипа: 16,000 нейронов, 16 млн. синапсов.

Потребляемая мощность: 1,000 mW/cm^2.

Особенности: позволяет симулировать мозг в большом масштабе при низком энергопотреблении; помогает улучшить модели болезней мозга.

Цифровой или аналоговый: цифровой.

Тех. процесс: 130 nm.

Самая большая конфигурация на данный момент: 1,152 чипа: 20 млн. нейронов; 20 млрд. синапсов.

След. конфигурация: 5,750 чипов: 100 млн. нейронов; 100 млрд. синапсов.

Конечная конфигурация: 1 млрд. нейронов; 1 трлн. синапсов.

2.3. Нейроморфная архитектура HRL

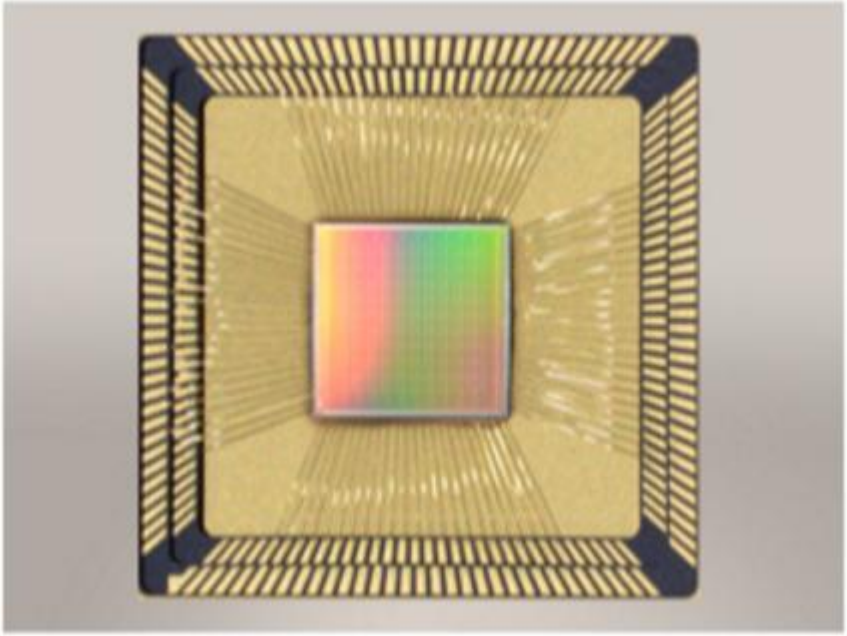

Рис. 14. Чип архитектуры HRL.

Разработчики: HRL Laboratories, DARPA SyNAPSE.

Спецификация чипа: 576 нейронов, 73,000 синапсов.

Потребляемая мощность: 120 mW/cm^2.

Особенности: усиливает биологический реализм с помощью своих возможностей к обучению; гибкость программирования.

Цифровой или аналоговый: гибридный.

Тех. процесс: 90 nm.

Самая большая конфигурация на данный момент: 4 полупроводниковых пластин; 2,304 нейрона; 292,000 синапсов.

След. конфигурация: 4 чипа: 4,096 нейронов; 520,000 синапсов.

Конечная конфигурация: информация недоступна.

2.4. Neurogrid

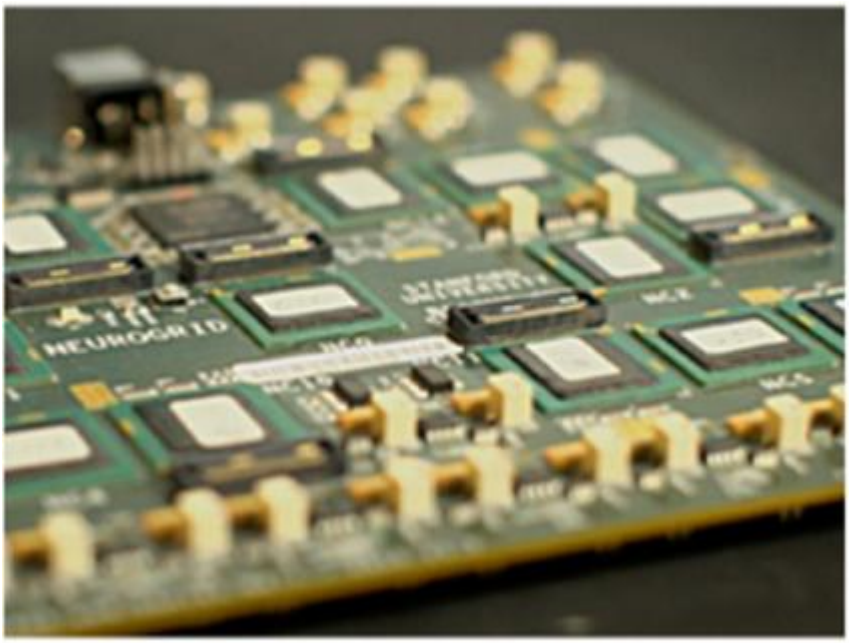

Рис. 15. Чип архитектуры Neurogrid.

Разработчики: Стенфордский университет.

Спецификация чипа: 65,536 нейронов, 500 млн. синапсов.

Потребляемая мощность: 50 mW/cm^2.

Особенности: позволяет симулировать реалистичные биологические модели мозга; облегчает разработку автономных роботов.

Цифровой или аналоговый: гибридный.

Тех. процесс: 180 nm.

Самая большая конфигурация на данный момент: 16 чипов: 1 млн. нейронов; 8 млрд. синапсов.

След. конфигурация: информация недоступна.

Конечная конфигурация: информация недоступна.

2.5. TrueNorth

Рис. 16. Чип архитектуры TrueNorth.

Разработчики: IBM, DARPA SyNAPSE.

Спецификация чипа: 1 млн. нейронов, 256 млн. синапсов.

Потребляемая мощность: 20 mW/cm^2.

Особенности: энергоэффективный нейроморфный чип, который найдет применение в мобильных устройствах, облачных вычислениях и т. д.

Цифровой или аналоговый: цифровой.

Тех. процесс: 28 nm.

Самая большая конфигурация на данный момент: 16 чипов: 16 млн. нейронов; 4 млрд. синапсов.

След. конфигурация: 4,096 чипов: 4 млрд. нейронов; 1 трлн. синапсов.

Конечная конфигурация: 10 млрд. нейронов, 100 трлн. синапсов.

Очевидно, что именно TrueNorth является лучшей из представленных архитектур. Он имеет 1 млн. нейронов против 65 тыс. у ближайшего конкурента Neurogrid, а так же является самой энергоэффективной: 20 mW/cm^2 против 50 mW/cm^2 у Neurogrid.

ГЛАВА 3. Рабочая станция разработки нейронных сетей

В предыдущей главе мы выяснили, что лучшим решеним будет компьютер, основанный на чипе с архитектурой TrueNorth. На данный момент к выходу на рынок готовится только одна модель, о ней подробно пойдет речь в данной главе.

3.1. Нейросинаптический суперкомпьютер IBM NS16e

Не так давно вышла в свет первая в мире платформа с 1 млн. нейронов для мобильных приложений (NS1e), основанный на чипах IBM TrueNorth (TN). Затем появилась масштабируемая система из 16 млн. нейронов (NS1e-16 [19]), которая представляла собой 16 связанных NS1e вместе с поддерживающей периферией, включавшей в себя хост-сервер, роутер, контролеры питания и др. компоненты.

Нынешний этап – **NS16e** (не путать с NS1e и NS1e-16) [26], состоящая из специализированной платы 4x4, подложки (interposer board) и готовой системы на кристалле (SOC) Avnet Zynq AES-MMP-7Z045-G. Три платы собраны в единую NS16e структуру с помощью вертикальных коннекторов, чтобы питать все платы и для обмена данными и управляющими сигналами.

Плата 4x4 содержит 16 программируемых TrueNorth чипов, способных реализовывать большие (до 16 млн. нейронов) нейронные модели для большого числа приложений. Чтобы максимизировать пропускную способность с TrueNorth чипами на 4x4 плате, используются расширители портов данных на входных/выходных портах чипа.

Аналогичный набор микросхем используется для настройки всех чипов с приминением цепей сканирования (scan chains). Плата 4x4 помимо чипов содержит регулировщики домена питания, контролер питания, контуры с датчиками и несколько SPI/I2C программируемых устройств.

Подложка (interposer board) предоставляет высокоскоростные интерфейсы и домены питания для всех системеных компонентов NS16e. Она содержит PCIe x4 коннектор для быстрого общения с хост-машиной и SFP+ Ethernet гнездо (пока не активен). На подложке есть контролер питания и ряд предохранителей, выключающих систему в случае неполадки с электричеством. Для отладки установлены светодиоды, отображающие неполадки в доменах питания. Так же есть JTAG коннектор для программирования Zynq модуля.

Модуль Zynq предоставляет ППВМ (программируемая пользователем вентильная матрица, FPGA) для реализации произвольной клейкой логики (glue logic) для контроля и потенциальной модификации (если есть необходимость) входящих/исходящих данных в NS16e. Zynq может быть использован для прямого доступа к памяти (DMA), трансдукции данных в спайк (data-to-spike transduction), фильтрации и т. д. Высокоскоростные трансиверы системы на кристалле (SOC) используются для PCIe моста между NS16e системой и хост-сервером. Zynq SOC так же содержит два ARM ядра, которые пока не активированы в текущей версии NS16e для минимизации энергопотребления.

Команда IBM Brain-inspired Computing создала спецификации, схемы и произвела физическое проектирование системы. Они

изготовили и собрали платы с применением передовых электронных компонентов.

На рисунках 17-18 показана плата 4x4 (без чипов), на рисунках 19-20 – подложка (interposer board):

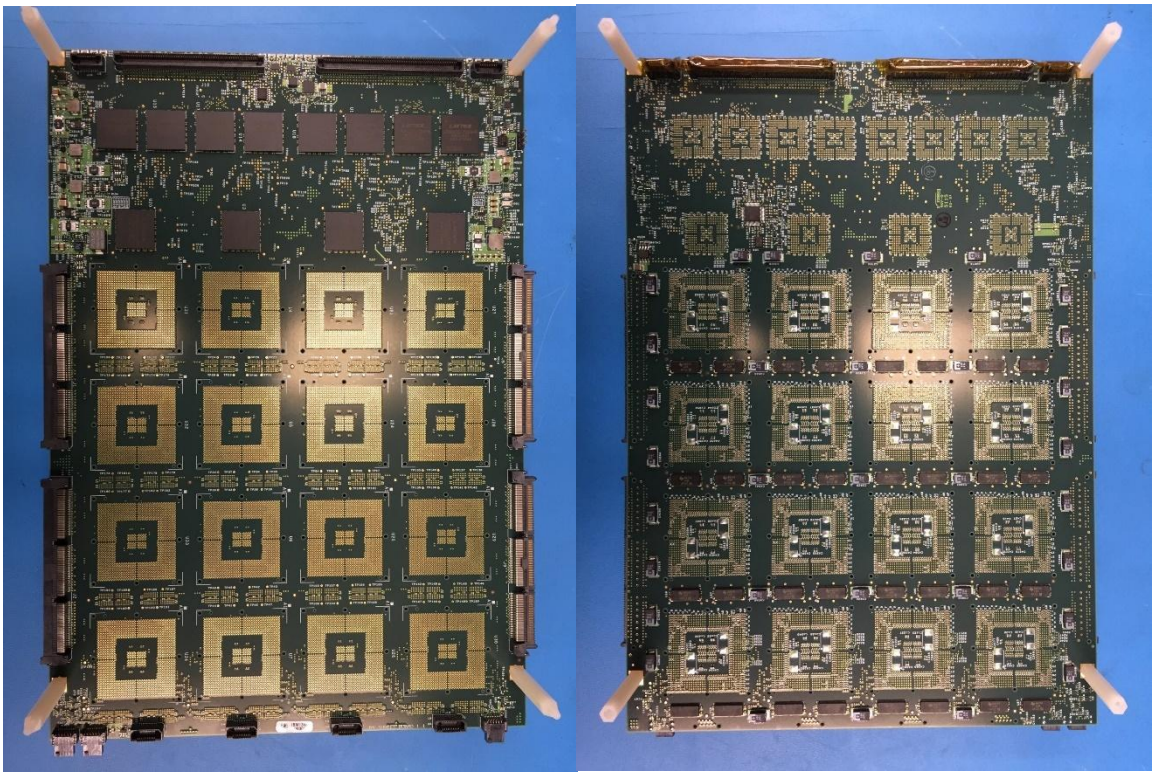

Рис. 17-18. Плата 4x4: слева - вид сверху, справа – вид снизу.

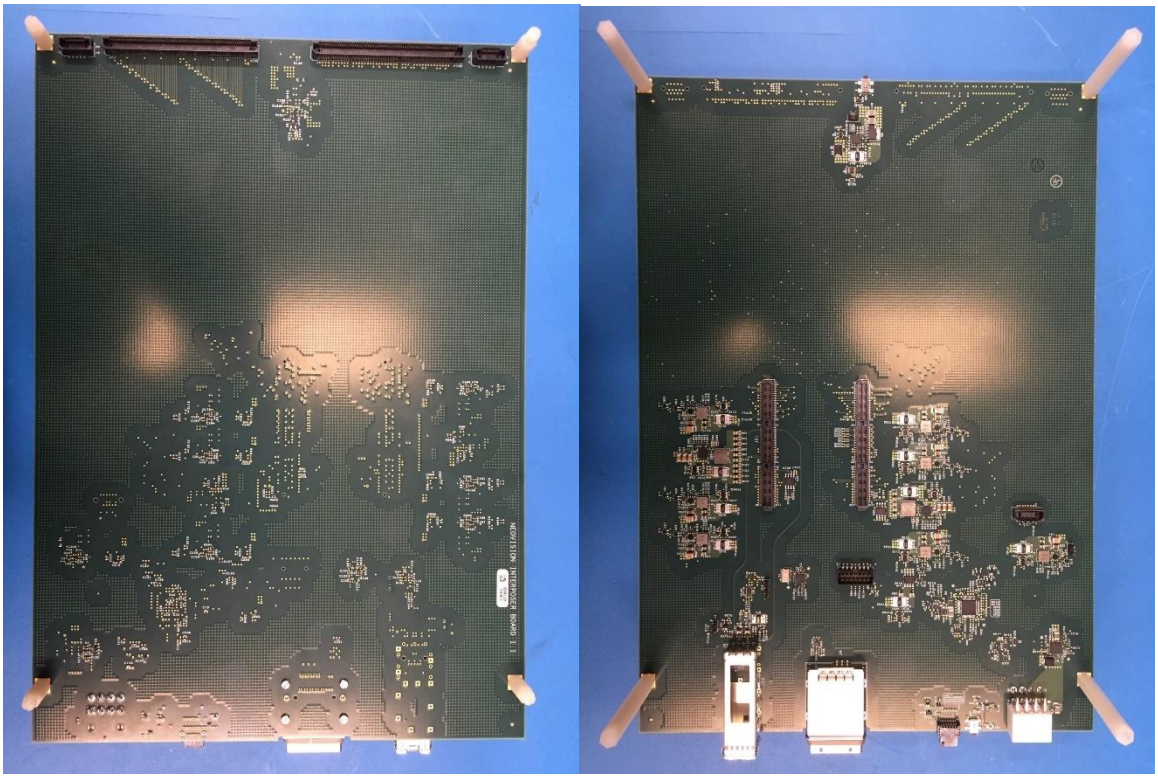

Рис. 19-20. Подложка: слева - вид сверху, справа – вид снизу.

Поскольку на данный момент число чипов TrueNorth ограничено, то было решено применить технику сборки с минимизацией рисков и сделать несколько систем без впаивания чипов. Это дало возможность протестировать полный дизайн системы без, собственно, самих чипов. На рисунке 21 изображены две платы, готовые к соединению (модуль Zynq не изображен на фотографии):

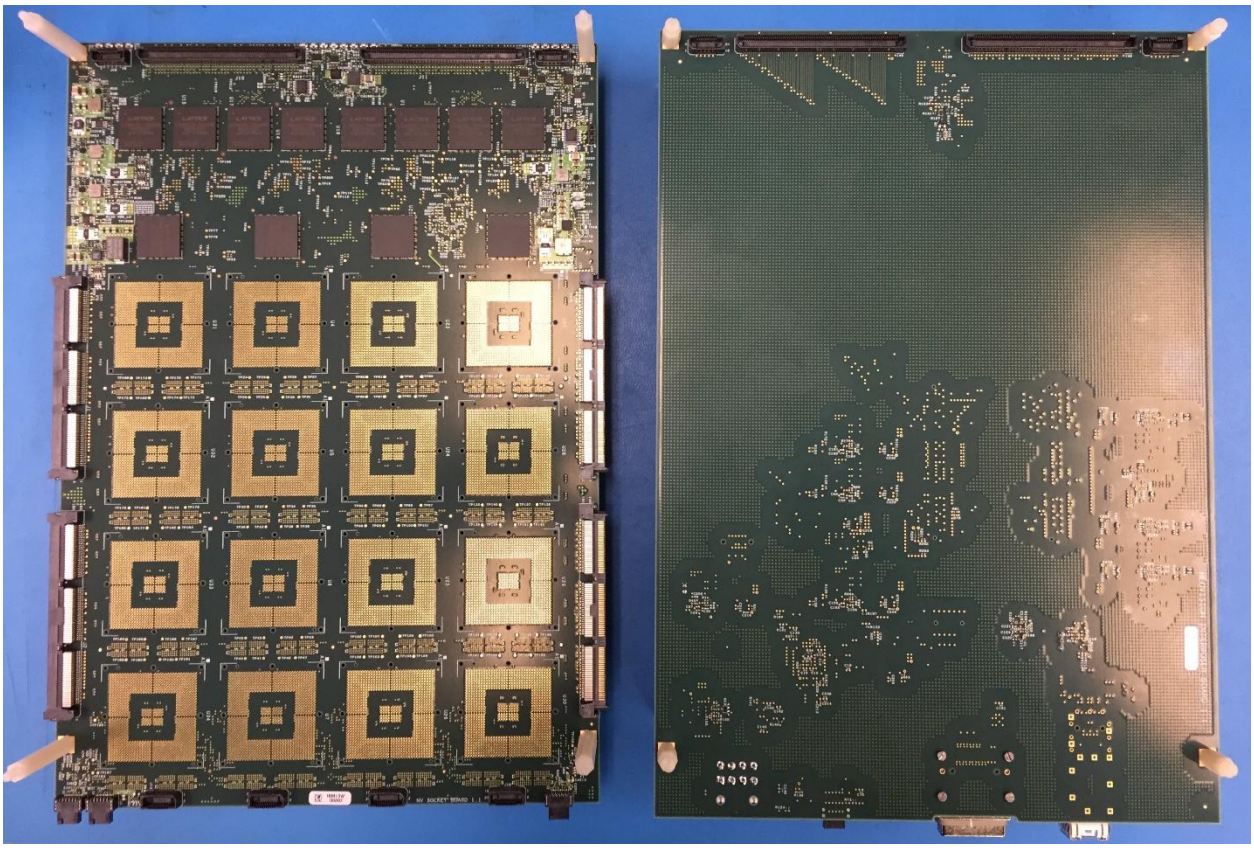

Рис. 21. 4x4 и подложка готовы для соединения.

Можно догадаться, что создание такой системы – задача нетривальная. Было проведено множество hardware, firmware, software тестов для проверки функциональности системы.

На первых прототипах были предусмотрены специальные сокеты, которые могут быть установлены после сборки платы для самих чипов. После тестирования первых двух плат (без чипов) чипы все-таки были установлены 4x4 с помощью вышеупомянутых сокетов (см. рис. 22):

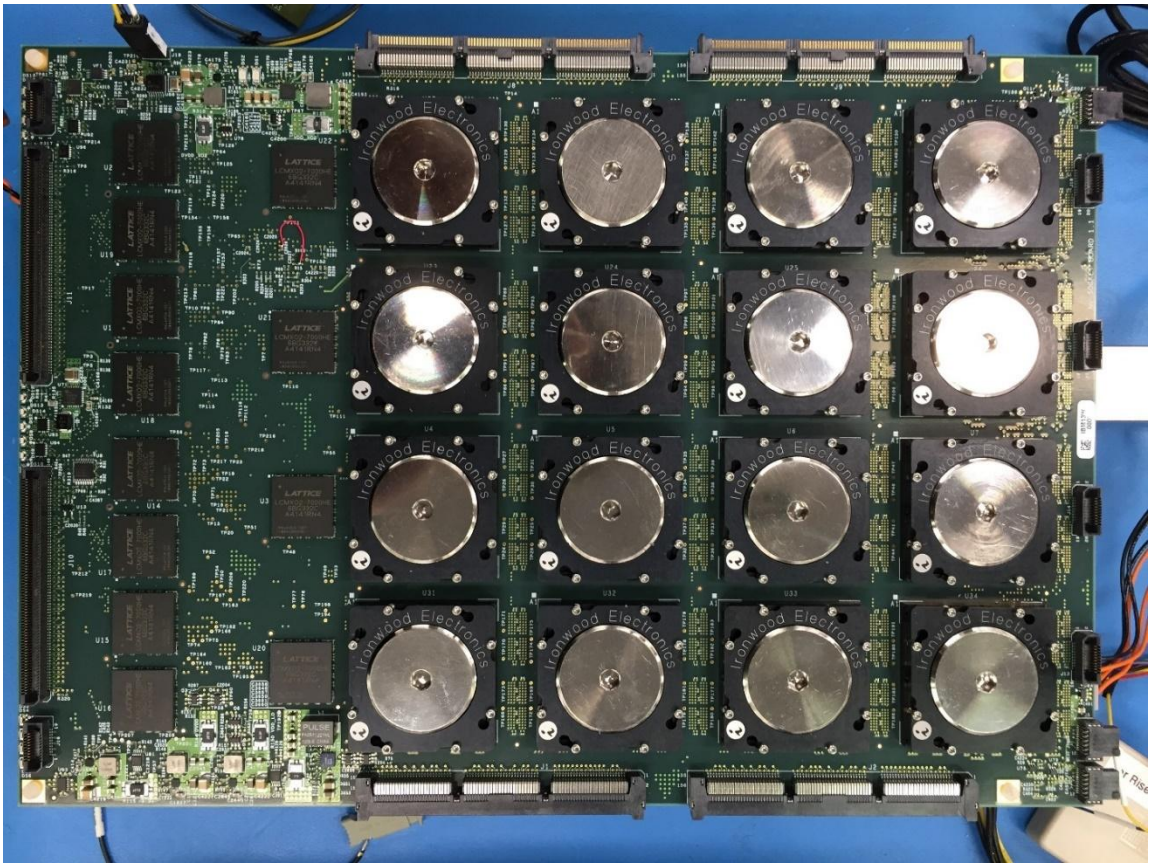

Рис. 22. Система с TrueNorth чипами, установленных с применением сокетов.

Усиление этой платформы должно происходить шаг за шагом, чтобы ток в системе был в пределах нормы; все ключевые интерфейсные сигналы и клейкая логика (glue logic) отслеживаются с помощью осциллографов для полной операционной и сигнальной целостности. Низкоуровневая сборка системы показана на рис. 23:

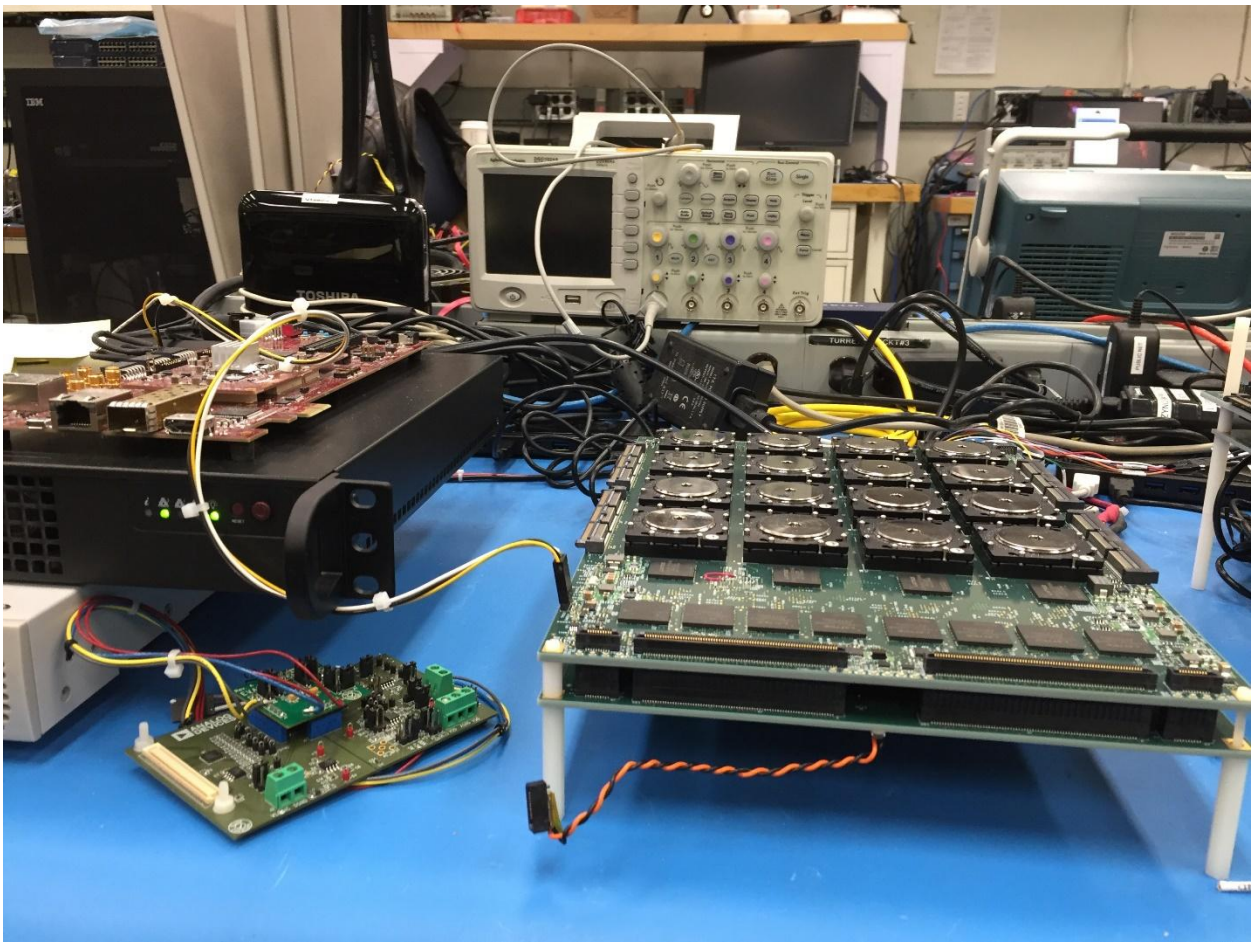

Рис. 23. Низкоуровневая сборка с TN чипами в сокетах.

Сборка и тестирования комплексных плат занимает несколько дней (если все будет удачно). Эта платформа не является исключением. После того как все низкоуровневые тесты дали положительный результат было решено сделать первые прототипы с впянными чипами, первый показан на рис. 24:

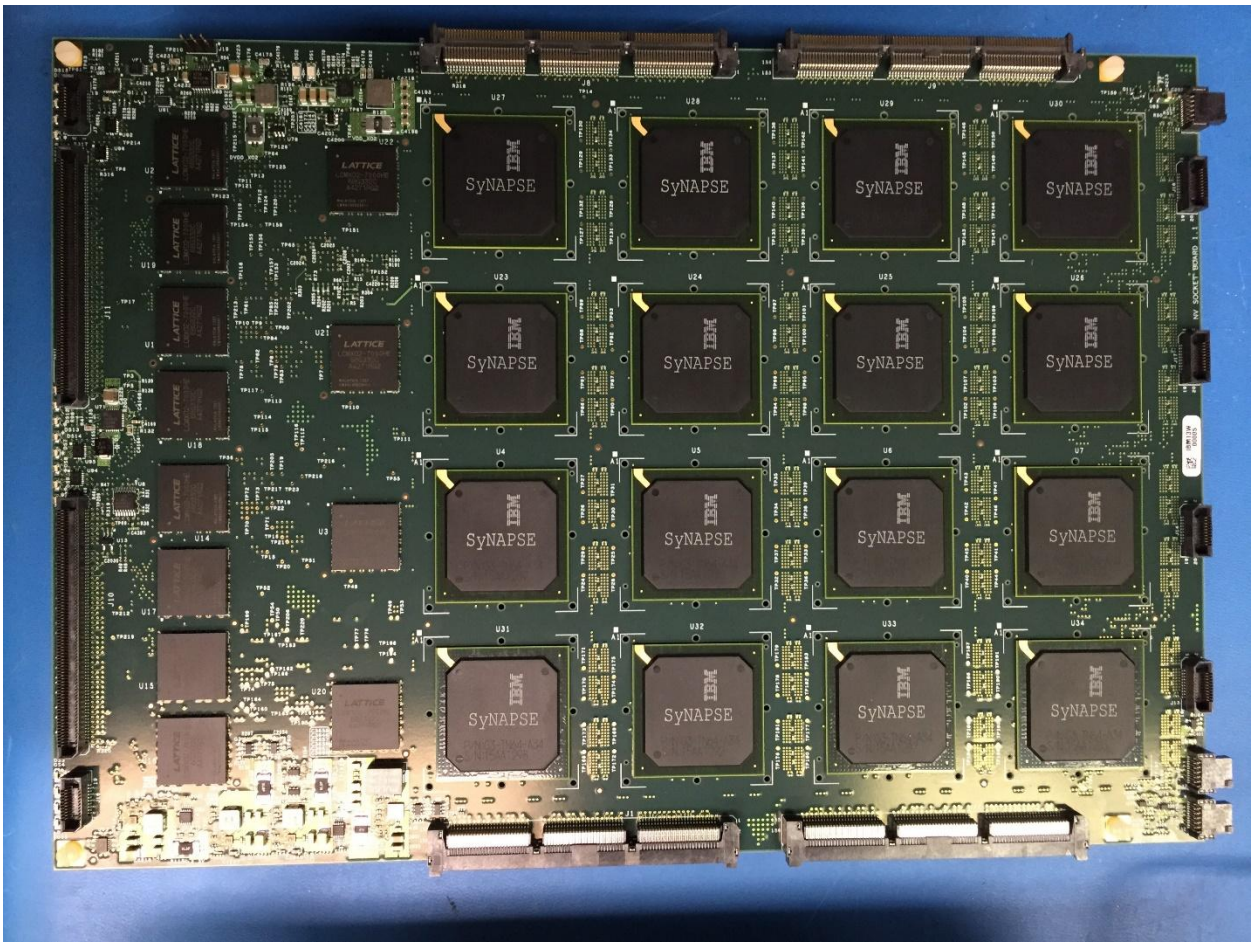

Рис. 24. Первый 4x4 прототип с впаянными TN чипами.

После успешного завершения низкоуровневых тестов NS16e системы были соединены с хост-серверами (рис. 25, 26.) с помощью нескольких интерфейсов: Xilinx JTAG для программирования Zynq, Lattice JTAG для контролера питания и программирования FPGA, PCIe 2.0 для обмена данными между хост-сервером и TN чипами.

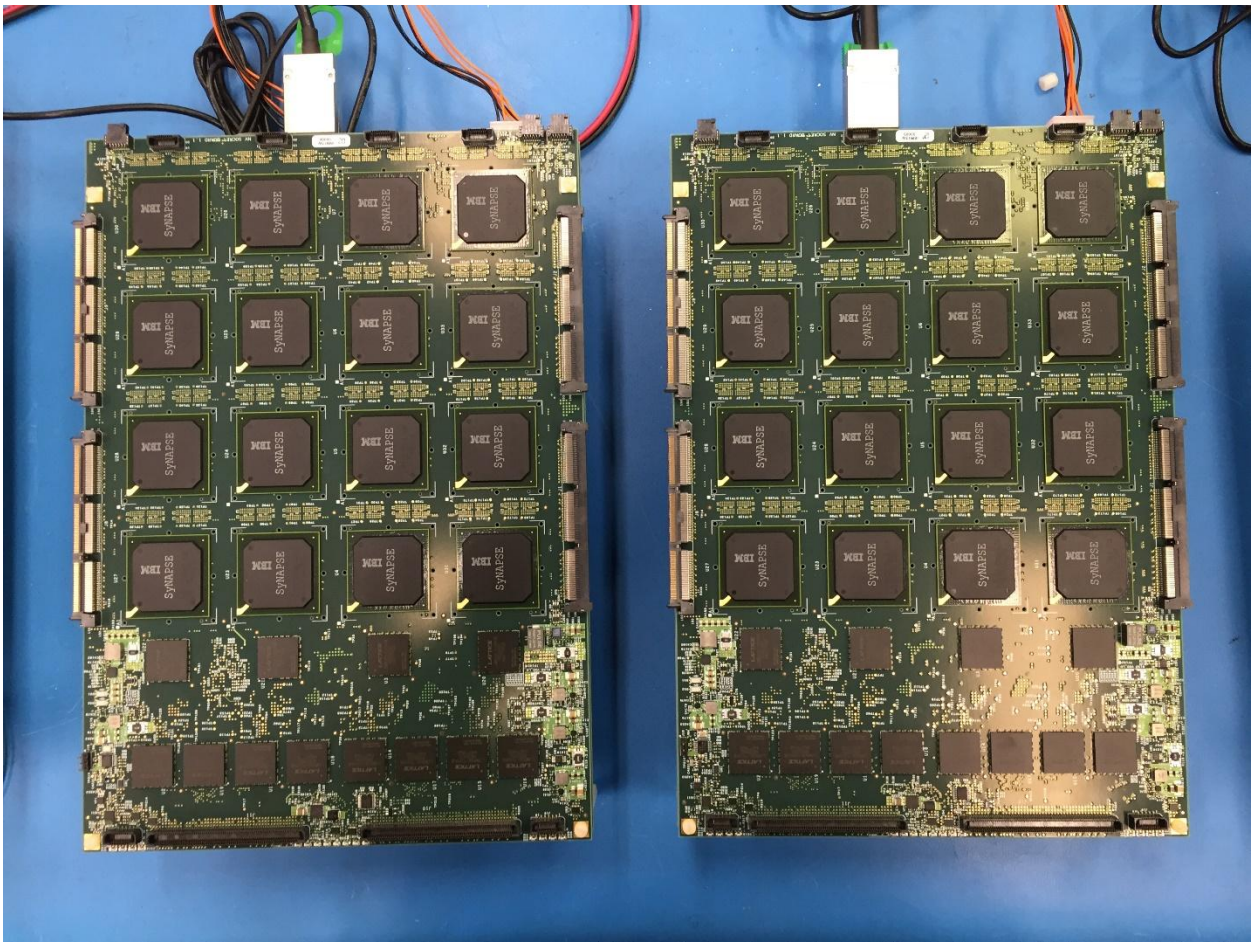

Рис. 25. NS16e-системы соединены с хост-машинами.

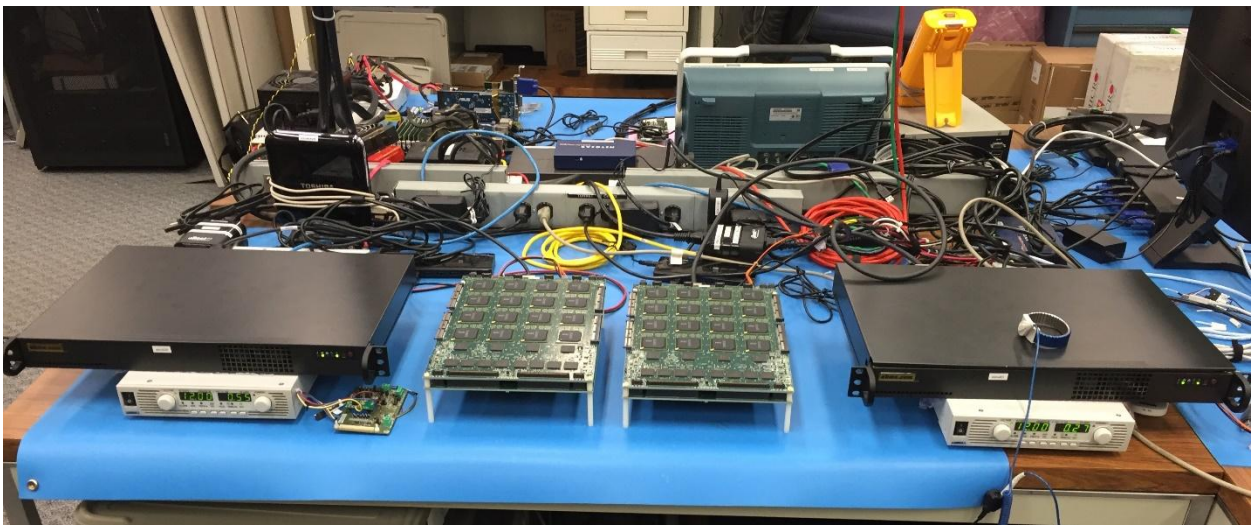

Рис. 26. NS16e-системы соединены с хост-машинами.

С этого момента начинается высокоуровневое тестирование нейросетевых алгоритмов и приложений прямо на свежей системе.

Некоторые члены команды работали по выходным, поскольку в лаборатории было тихо, не было уборщиков и регулярных совещаний:

Рис. 27. Скотт Лекатч (Scott Lekutch) принимает пищу и тестирует приложение на NS16e одновременно.

В конце концов, были сделаны первые прототипы NS16e. Теперь на ней можно запускать сложные нейронные приложения, например, продвинутое распознавание звука/изображений и классификацию в реальном времени на 16 млн. спайковых нейронов.

3.2. Создание корпуса NS16e

Для разворачивания в Ливерморской национальной лаборатории (LNLL) необходимо было упаковать стек из трех плат (описанный выше) в защитный корпус. Поскольку NS16e является первым в своем роде нейросинаптическим суперкомпьютером, то необходимо было отобразить его новизну в новом корпусе и при этом защитить платы и предоставить доступ к нужным разъемам, переключателям и индикаторам.

Т. к. платы были уже собраны воедино, то для достижения всех вышеперечисленных задач нужно было решить много сложных дизайнерских проблем. Для этого на протяжении нескольких лет велась серьезная кооперация с дизайнерской командой IBM, которые ранее предложили модели потенциальных приложений TrueNorth (Cognitive Apps). Ааарон Кокс (Aaron Cox) занимался промышленным дизайном и создал золотистую крышку для чипа (см. рис. 28). Дизайн с выпирающими в разные направления ушками отображает ключевую особенность чипа TrueNorth – быть объединенным с другими такими же чипами в один двумерный массив. Поскольку NS16e – первая система, эксплуатирующая такую функцию, то было решено сделать сам массив видимым (рис. 29).

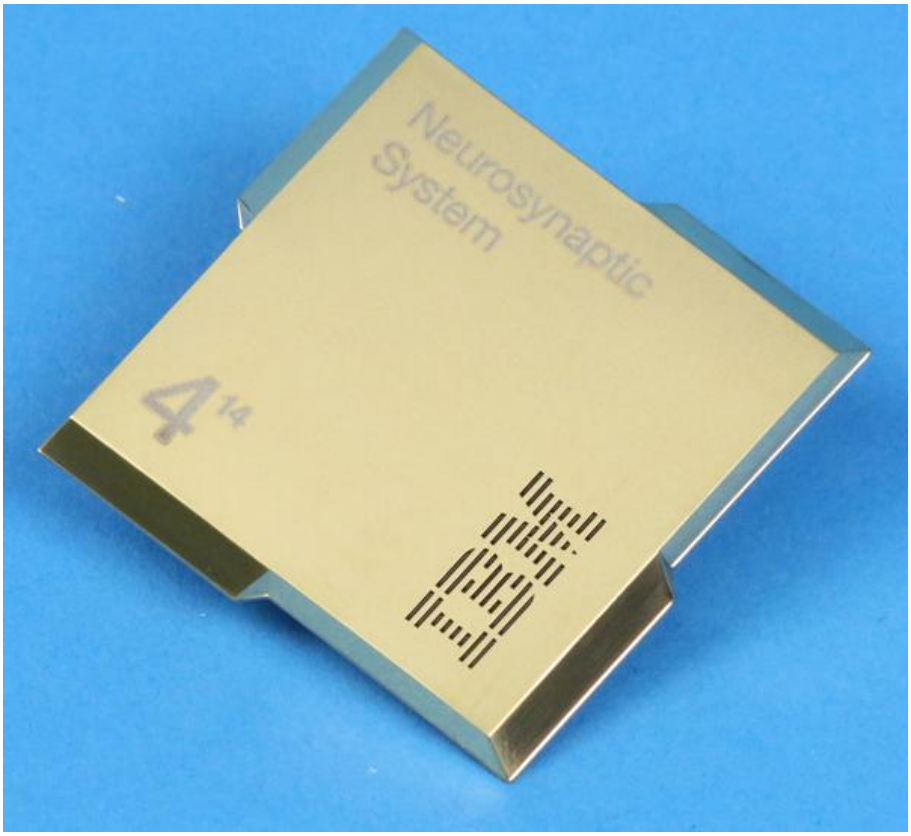

Рис. 28. Крышка TN чипа. 4^{14} – число синапсов на одном чипе (4^{14} = 256 млн.)

Рис. 29. корпус NS16e.

Для адаптации корпуса к стеку из плат был напечатан на 3D-принтере слайдер (slide mechanism) (рис. 30), который позволяет открыть корпус нажатием кнопки, а так же рокер (rocker mechanism) (рис. 31) для другой кнопки. Сам рокер сделан из прозрачного материала, чтобы помимо открывания корпуса он выдавал цвет внутренних светодиодов.

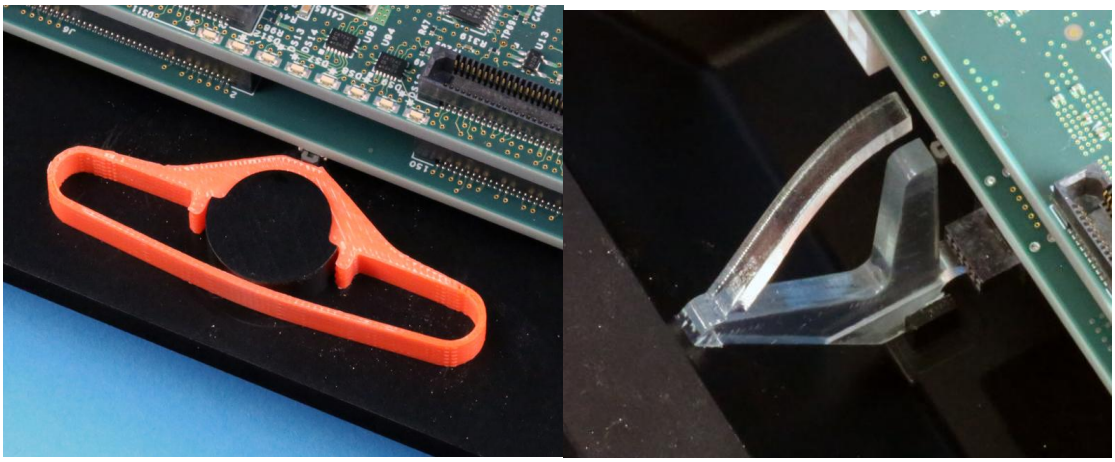

Рис. 30-31. Слева – слайдер, справа - рокер.

Было решено поддерживать 2 режима корпуса: как отдельную станцию на рабочем столе (рис. 32) и для закрепления на серверной 2U-стойке (рис. 33)

Рис. 32-33. Слева - режим станции, справа – монтаж в стойку.

3.3. NS16e как рабочая станция разработки нейронных сетей

В отличие от систем с одночипными платами, которые строились ранее (NS1e-16), эта машина работает с более мощными нейронными сетями в реальном времени. Стандартная комбинация состоит из NS16e, соединненой с x86 хост-сервером по PCI Express. Хост -сервер может закачивать/выкачивать большие объемы данных в NS16e. Хост-сервер с GPU может создавать и обучать большие неронные сети и NS16e может немедленно их запускать. Одной командой запускается весь процесс подготовки тренировочных данных, создания нейронной сети, обучения, оптимизации сети для железа и запуска на NS16e. Очевидно, что процесс очень быстр и комбинация мощного сервера и NS16e – машина мечты для разработки больших нейронных сетей.

Условная схема конечной системы показана на рис. 34. Основные нейронные вычисления проходят на массиве 4x4, где чипы соединены проводами и общаются друг с другом с помощью спайков. Все коммуникации регулируются асинхронным протоколом без дополнительным интерфейсных чипов.

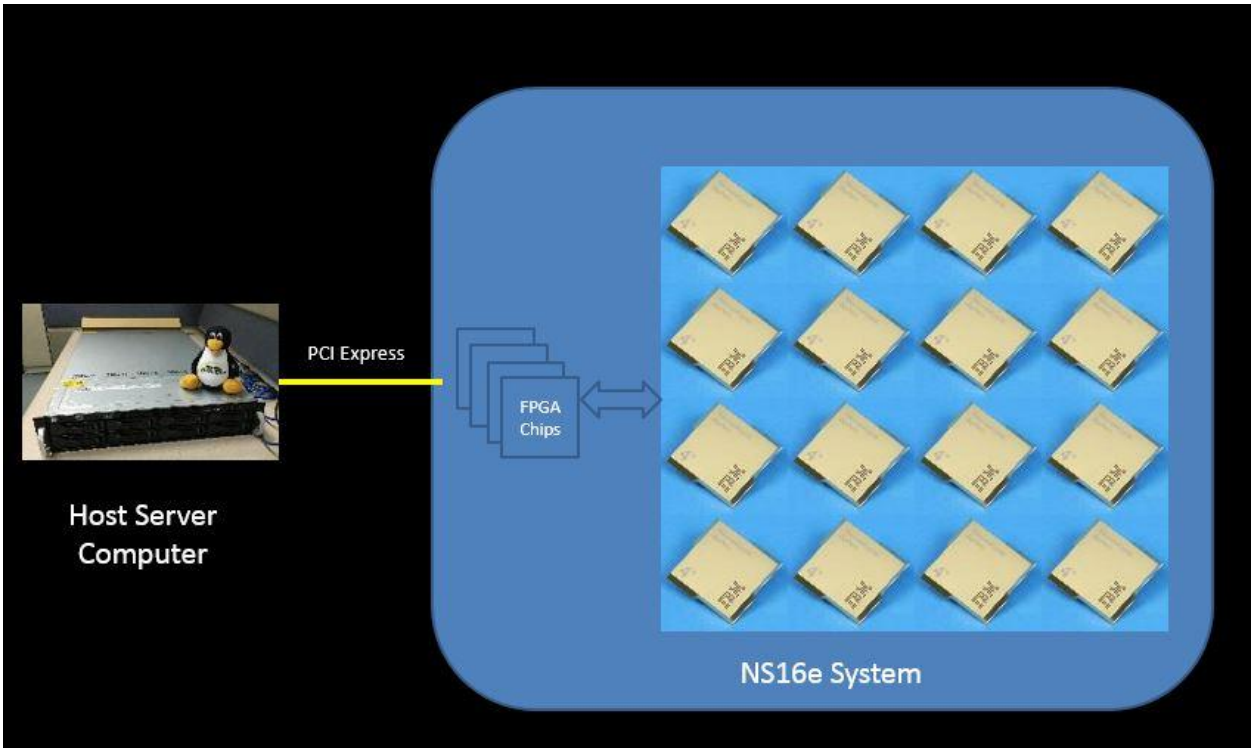

Рис. 34. условная схема конечной системы.

Хост-сервер соединен с NS16e с помощью PCI Express и со скоростью 500Mb/sec получает/выдает данные. FPGA и на NS16e работают как мост между сервером и чипами, являясь транслятором между фон-Неймановской и нейроморфной архитектурами, общающихся на разных языках. Цифровой фон-Неймановский компьютер оперирует инструкциями и бинарными данными, в то время как нейросинаптический компьютер – спайковыми сигналами между нейронами.

Вот что происходит, когда мы вводим комманду для запуска нейронной модели в хост-машину. Хост отсылает нейронную модель по PCI Express и загружает ее на чипы в NS16e. Терминал на компьютере показывает процесс загрузки. Примерно через 30 секунд (в зависимости от размера нейронной сети) загрузка завершена и сеть начинает выдавать спайки. На хосте можно увидеть только как много спайков

генерируется и как часто обновляются нейроны. Нейронные спайки кодируются не так просто для человеческого восприятия. И тут в дело вступают визуализаторы, которые расшифровывают спайки из NS16e и визуализируют то, что NS16e пытается сказать.

Допустимы мы распознаем образы на NS16e. Мы загружаем нейронную сеть в NS16e, отсылаем спайки, кодирующие изображение (A) (Рис. 35.), получаем ответ в спайках и визуализируем их. Получаем (B).

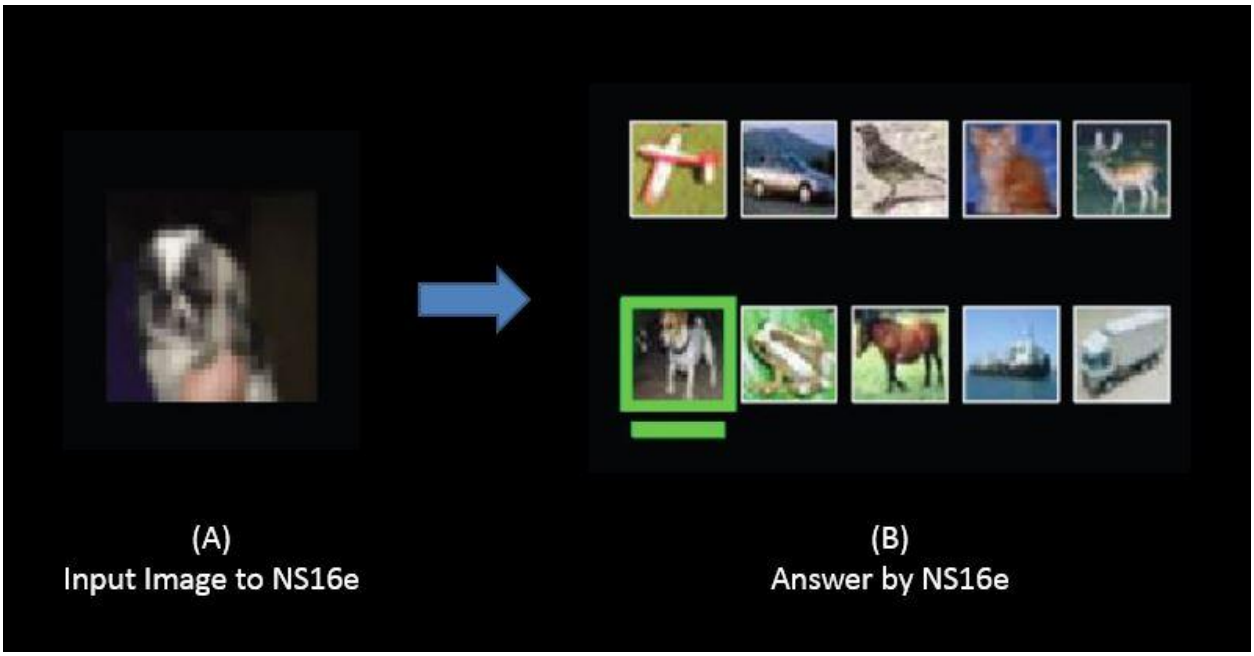

Рис. 35. распознавание образов с помощью NS16e.

Система развивается, создаются и тестируются новые алгоритмы, новые техники генерации моделей и новые алгоритмы оптимизации. Уникальная деталь этой машины заключается в том, что необходима оптимизация логического отображения нейронных сетей в физические схемы. В нашем мозгу часть коры отвечает за распознавание образов, а остальные – за двигательную функцию. NS16e имеет подобное свойство. Можно назначить конкретному чипу часть большой нейронной сети.

Например мы можем назначить каждому чипу по одному слою сети или назначить кусочек многослойной сети, отвечающий за распонавание конкретной части изображения. Такое явление называется **«проблемой размещения ядер»** (core placement problem) и оно уникально для машин вроде NS16e. Оно заставляет некоторые сети работать быстрее и эффективнее, потому что внутричиповые коммуникации быстрее, чем межчиповые. Это то место, над которым разработчики по-прежнему трудятся.

NS16e – это еще один шаг к более масштабным системам, которые будут потреблять только маленькую толику энергии относительно CPU и GPU.

3.4. Как программировать нейросинаптический суперкомпьютер

Для программирования NS16e используется интегрированная экосистема со специальными средствами разработки - **DevKit**.

Одним из тестовых приложений в DevKit является простой классификатор изображений, оно использует GPU для обучения сверточной нейронной сети (convolutional neural network, CNN) на стандартных данных и разворачивает обученную сеть на NS16e.

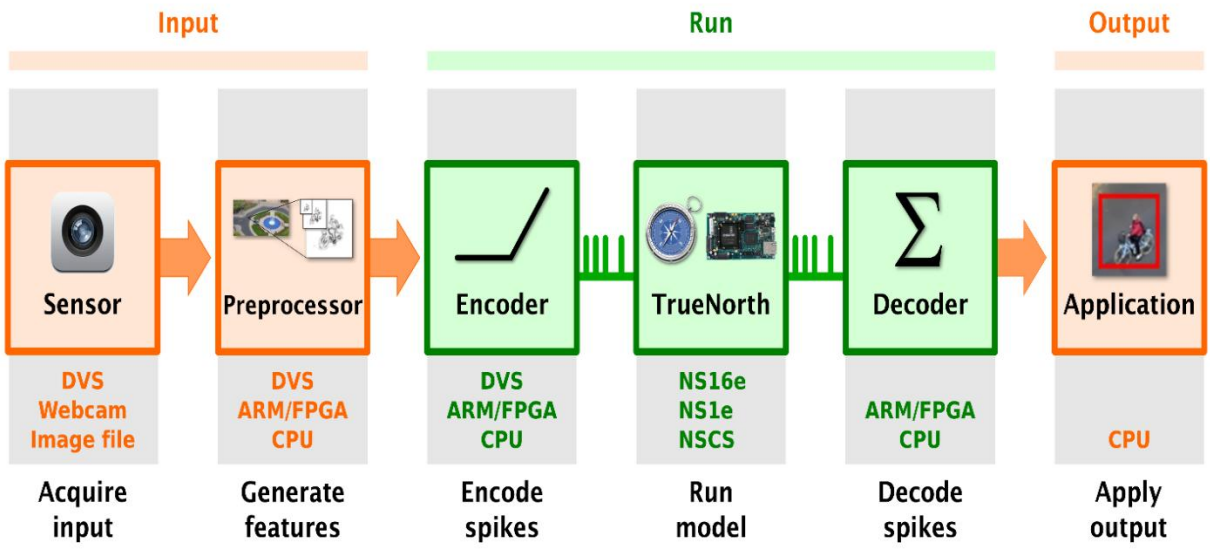

Рис. 36. схема процесса классификации изображений.

Базовая структура классификатора показана на рис. 36. Вначале он получает поток данных, который может быть набором файлов изображений с диска, вебкамеры или даже набором спайков с сенсора динамического зрения (DVS) [20]. Сырые данные могут быть предобработаны в признаки как карты выделенных границ перед тем как закодироваться в спайковые потоки на вход TrueNorth. В зависимости от того как приложение распределено по системным компонентам, предобработка и кодирование может производиться в серверном CPU, встроенными в плату ARM ядрами или FPGA или буквально в спайкинговом сенсоре. Входные спайки могут быть обработаны на NS1e или NS16e или NSCS функциональном симуляторе. Выходные спайки декодируются и отправляются на визуализацию.

В этом примере сеть учится классифицировать изображения из CIFAR-10 и CIFAR-100 данных [21]. Задача упрощается с помощью предобработки и кодирования полного набора данных в оффлайн режиме. Входной файл спайка зациклен через TrueNorth, чтобы

генерировать предсказания классов для изображений в тестовых данных, которые подаются на рабочую станцию для визуализации.

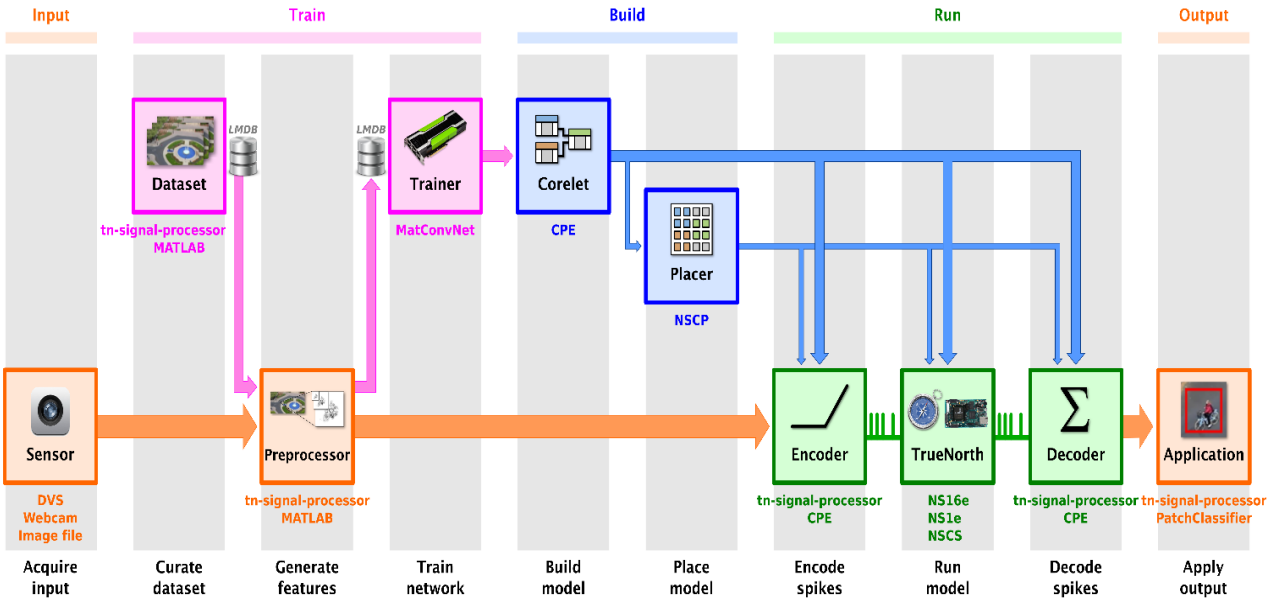

Рис. 37. поток разработки для TrueNorth классификатора изображений.

Цепь действий, необходимая для обучения и развертывания классификатора показана на рис. 37. DevKit предоставляет как минимум один инструмент для заполнения каждого блока в этой цепи, но каждый из инструментов может быть заменен на произвольный, если он реализует тот же интерфейс. К примеру, DevKit включает два эквивалентных инструмента для чтения/записи **LMDB** [23] баз данных: консольную утилиту tn-signal-processor или набор MATLAB MEX файлов, но если данные блоки получили LMDB в правильном формате, то им все-равно какой инструмент использован.

3.4.1. Набор данных и предобработка

Первые шаги заключаются в базовом анализе данных. Сначала импортируется начальный набор данных в стандартом формате, чтобы

все инструменты его понимали, затем преобразуется в набор признаков. Данные хранятся в LMDB, популярном формате в области глубинного обучения, поскольку быстрое чтение LMDB позволяет быстро перекачивать данные в GPU.

Для импорта и предобработки данных предоставляется C++ консольная утилита – tn-signal-processor. Она как швейцарский нож, может импортировать JPG и PNG файлы в LMDB формат, резать, поворачивать, накладывать фильтры, кодировать/декодировать в спайки, визуализировать данные и много чего еще.

3.4.2. Обучение и Corelet

Теперь, полученные в нужном формате данные можно использовать для обучения. **TrueNorth Convolutional Networks** (TNCN) [22] – фреймворк для создания и обучения нейронных сетей, которые соответствую ограничениям в дизайне TrueNorth. С помощью **Corelet Programming Environment** (CPE) TNCN освобождает от фокуса на абстрактных данных: последовательность слоев, размер фильтра, точность данных и т. д.

Для ускорения обучения с GPU используется фреймворк глубинного обучения **MatConvNet** [24], который обрачивает **cuDNN** [25] примитивы NVIDIA в бинарные MEX файлы, которые легко исполняются в MATLAB.

3.4.3. Плейсер

Последний шаг перед разворачиванием нейросети на железо – назначение логических ядер в файле модели к физическим в NS16e

массиве чипов. В мультичиповых система это может быть непростой задачей из-за бутылочного горлышка в интерфейсах между чипами.

Поэтому применяется утилита, которая эвристически назначает логические ядра из TNCN файла модели к физическому местоположению в NS16e массиве ядер (рис. 38.), используя минимизацию пересечения чипов в графе связности ядро-к-ядру. Neuro Synaptic Core Placer (NSCP) является критическим элементом, чтобы большие модели успешно работали на железе.

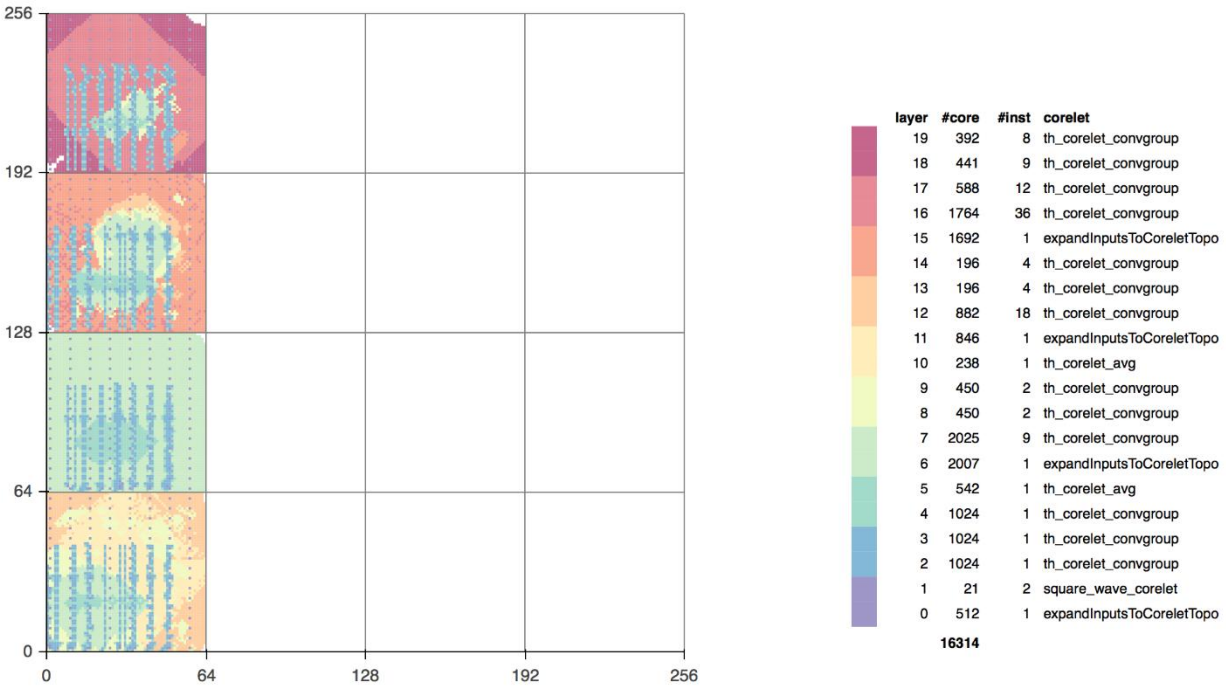

Рис. 38. Физическое размещение ядер для CIFAR-10 на NS16e.

3.4.4. Тестовое приложение

Теперь обученный классификатор можно прогнать на железе. Для расшифровки, оценки и визуализации меток тестовых изображений существует консольная утилита PatchClassifier, которая выдает массив иконок категорий для данного набора изображений, если изображение

выделено зеленым, значит предсказание верно, красным – неверно. Длина полосы под изображением – уверенность классификации (рис. 39.).

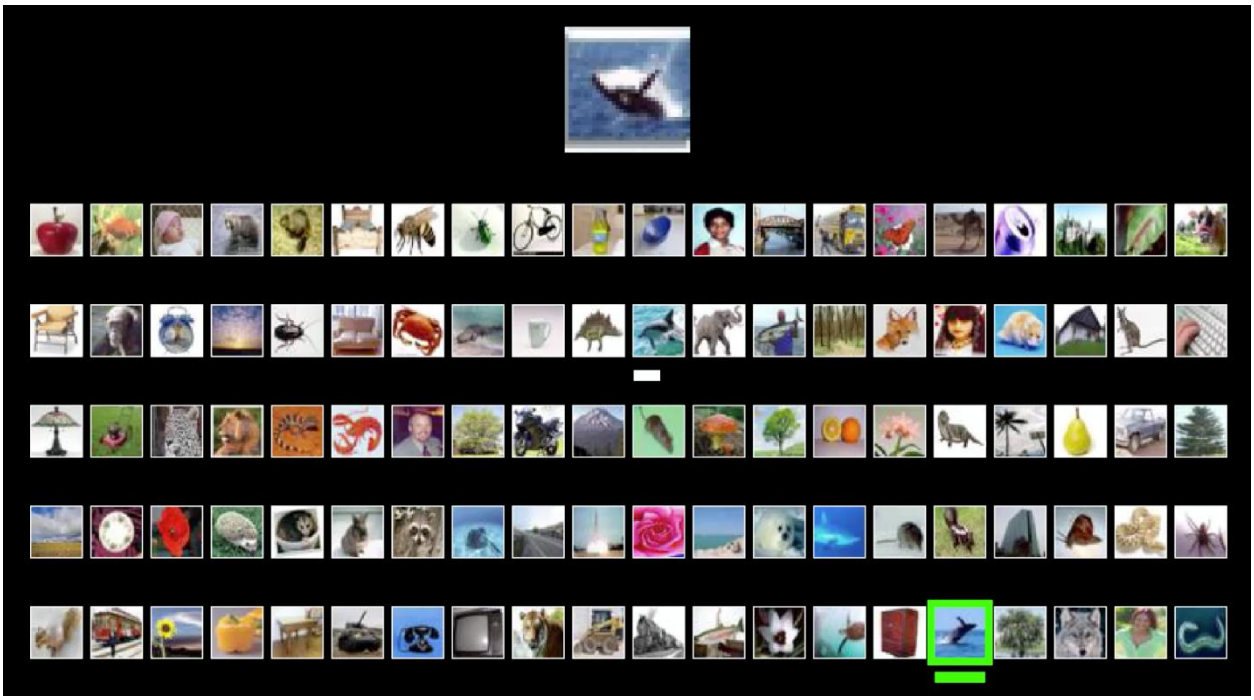

Рис. 39. визуализатор PatchClassifier, настроенный на CIFAR-100 данные (зеленый – верное предсказание, белая полоска – следующий подходящий выбор).

3.5. Конечная конфигурация оборудования

Для эффективной работы NS16e нужен хост-сервер с GPU. Для GPU лучшим решением на данный момент является NVIDIA Tesla P100 [28].

На основе P100 к выходу на рынок готовится платформа NVIDIA DGX-1 (Рис. 40.) [29], предназначенная специально для задач искусственного интеллекта и машинного обучения.

Рис. 40. NVIDIA DGX-1.

Спецификация DGX-1:

Число GPU: 8x Tesla GP100

TFLOPS (GPU/CPU): 170/3

Память GPU: 16 GB на GPU

CPU: Dual 20-core Intel Xeon E5-2698 v4 2.2 GHz

Ядра NVIDIA CUDA: 28672

Системная память: 512 GB 2133 MHz DDR4 LRDIMM

Хранилище: 4x 1.92 TB SSD RAID 0

Сеть: Dual 10 GbE, 4 IB EDR

ПО: Ubuntu Server Linux OS, DGX-1 Recommended GPU Driver

Вес: 134 lbs

Максимальные требования к энергии: 3200W

Операционная температура: 10 – 35 ^{o}C

Ориентировочно, форм-фактор у DGX-1 будет 2U как у NS16e. GPU на DGX-1 сможет быстро обучать нейронную сеть, NS16e – ее запускать, а большой размер хранилища (~8TB) работать с большими данными.

Поскольку NS16e может эмулировать 4 млрд. синаптических

связей, а мозг человека имеет порядка 100 млрд., то понадобится около 25 станций NS16e и столько же DGX-1.

ВЫВОДЫ

В ходе работы были рассмотрены самые крупные существующие программы поддержки нейроморфных вычислений: DARPA SyNAPSE и The Human Brain Project. Больше внимания было уделено первой, поскольку The Human Brain Project, несмотря на свои масштабы (около 112 участников) оказалась существенно менее результативной. Стоит так же отметить, что в работе не рассматривались менее масштабные проекты как Google Brain или Blue Brain Project.

Было рассмотрено пять архитектур: HiCANN, SpiNNaker, HRL, Neurogrid, TrueNorth. Последняя существенно опережает своих конкурентов по количеству искусственных нейронов на одном чипе и в плане энергоэффективности.

Была выбрана рабочая станция для разработки нейронных сетей, NS16e, а так же рассмотрен процесс разработки приложений с ее помощью. Она является единственным компьютером на основе чипа с TrueNorth архитектурой.

Для работы с NS16e была выбрана платформа DGX-1 с самой производительной на данный момент GPU, NVIDIA Tesla P100, для быстрого обучения нейронных сетей и работы с большим объемом данных, что дает возможности для симуляции работы человеческого мозга.

К сожалению, на момент написания данной работы, платформы NS16e и DGX-1 находятся в процессе подготовки к выходу на рынок, поэтому у автора не было возможности взаимодействовать с ними непосредственно.

ЗАКЛЮЧЕНИЕ

В России, к сожалению, нет масштабных проектов по поддержке нейроморфных вычислений.

Последняя нейроморфная архитектура под названием NeuroMatrix была разработана еще в 90-х годах, но тем не менее чипы на ее основе по-прежнему производятся.

В данной работе были рассмотрены программы и разработки в области нейроморфных архитектур. Был проведен сравнительный анализ пяти самых распространенных архитектур и на основе самой эффективной было подобрано соответствующее решение в качестве рабочей станции для разработки нейронных сетей.

Список литературы и источников

1. Carver A. Mead. Interviewed by Shirley K. Cohen. CALIFORNIA INSTITUTE OF TECHNOLOGY ARCHIVES. Pasadena, California. July 17, 1996. - http://oralhistories.library.caltech.edu/133/2/OH_Mead.pdf
2. Boddhu, S. K.; Gallagher, J. C. (2012). "Qualitative Functional Decomposition Analysis of Evolved Neuromorphic Flight Controllers". Applied Computational Intelligence and Soft Computing 2012: 1–21. - http://www.hindawi.com/journals/acisc/2012/705483/

3. 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
4. Andrew Nere, Umberto Olcese, David Balduzzi, Giulio Tononi. A Neuromorphic Architecture for Object Recognition and Motion Anticipation Using Burst-STDP. May 2012 - http://dx.doi.org/10.1371/journal.pone.0036958
5. Jae-sun Seo, Bernard Brezzo, Yong Liu, Benjamin D. Parker, Steven K. Esser, Robert K. Montoye, Bipin Rajendran, Jose A. Tierno, Leland Chang, Dharmendra S. Modha, Daniel J. Friedman. A 45nm CMOS Neuromorphic Chip with a Scalable Architecture for Learning in Networks of Spiking Neurons. IBM T. J. Watson Research Center, IBM Research – Almaden. 2011 - http://www.modha.org/papers/013.CICC2.pdf
6. Broad Agency Announcement. Systems of Neuromorphic Adaptive Plastic Scalable Electronics. DARPA-BAA 08-28. 9 April 2008. - https://www.fbo.gov/download/0b6/0b62b2149395d4bd8a28dff1b9046944/BAA08-28.doc
7. Rajagopal Ananthanarayanan, Steven K. Esser, Horst D. Simon, Dharmendra S. Modha. The Cat is Out of the Bag: Cortical Simulations with 10^9 Neurons, 10^{13} Synapses. IBM Almaden Research Center, Lawrence Berkeley National Laboratory. November 2009. - http://www.modha.org/C2S2/2009/11182009/content/SC09_TheCatIsOutofTheBag.pdf
8. Blue Brain Project profile on artificialbrains.com - http://www.artificialbrains.com/blue-brain-project

9. Neuroscience Expert Dr. Henry Markram on the IBM "Cat Brain" Simulation: "IBM's claim is HOAX" - http://technology-report.com/2009/11/neuroscience-expert-dr-henry-markram-on-the-ibm-cat-brain-simulation-ibms-claim-is-a-hoax/
10. Eugene M. Izhikevich homepage - http://www.izhikevich.org/
11. Brain Corporation profile on artificialbrains.com - http://www.artificialbrains.com/brain-corporation
12. Paul Merolla, John Arthur, Filipp Akopyan, Nabil Imam, Rajit Manohar, Dharmendra S. Modha. A Digital Neurosynaptic Core Using Embedded Crossbar Memory with 45pJ per Spike in 45nm. IBM Research – Almaden, Cornell University. - http://www.modha.org/papers/012.CICC1.pdf
13. Nabil Imam, Thomas A. Cleland, Rajit Manohar, Paul A. Merolla, John V. Arthur, Filipp Akopyan, Dharmendra S. Modha. Implementation of olfactory bulb glomerular-layer computations in a digital neurosynaptic core. Cornell University, IBM Research – Almaden. June 2012. - http://journal.frontiersin.org/article/10.3389/fnins.2012.00083/full
14. ARTIFICIAL SYNAPSES COULD LEAD TO ADVANCED COMPUTER MEMORY AND MACHINES THAT MIMIC BIOLOGICAL BRAINS. HRL Laboratories, LLC. 23 March 2012. - http://www.hrl.com/hrlDocs/pressreleases/2012/prsRls_120323.html
15. Kuk-Hwan Kim, Siddharth Gaba, Dana Wheeler, Jose M. Cruz-Albrecht, Tahir Hussain, Narayan Srinivasa, Wei Lu. A Functional Hybrid Memristor Crossbar-Array/CMOS System for Data Storage and Neuromorphic Applications. The University of Michigan, HRL Laboratories LLC. - http://pubs.acs.org/doi/pdf/10.1021/nl203687n

16. Robert Preissl, Theodore M. Wong, Pallab Datta, Myron Flickner, Raghavendra Singh, Steven K. Esser, William P. Risk, Horst D. Simon, Dharmendra S. Modha. Compass: A scalable simulator for an architecture for Cognitive Computing. IBM Research – Almaden, Lawrence Berkeley National Lab. November 2012. - http://www.modha.org/blog/SC12/SC2012_Compass.pdf
17. SEQUOIA – BLUEGENE/Q, POWER BQC 16C 1.60 GHZ, CUSTOM. TOP500. - http://www.top500.org/system/177556
18. Asynchronous VLSI and Architecture. Cornell University. - http://vlsi.cornell.edu/bio.php
19. Revealed: A Scale-Out Synaptic Supercomputer (NS1e-16). Dharmendra S Modha's Brain-inspired Computing Blog. 15 December 2015. - http://p9.hostingprod.com/@modha.org/blog/2015/12/
20. Dynamic Vision Sensor. DVS Overview. - http://inilabs.com/products/dynamic-vision-sensors/
21. The CIFAR-10 and CIFAR-100 datasets. - https://www.cs.toronto.edu/~kriz/cifar.html
22. Arnon Amir, Pallab Datta, William P. Risk, Andrew S. Cassidy, Jeffrey A. Kusnitz, Steve K. Esser, Alexander Andreopoulos, Theodore M. Wong, Myron Flickner, Rodrigo Alvarez-Icaza, Emmett McQuinn, Ben Shaw, Norm Pass, Dharmendra S. Modha. Cognitive Computing Programming Paradigm: A Corelet Language for Composing Networks of Neurosynaptic Cores. IBM Research – Almaden. - http://www.research.ibm.com/software/IBMResearch/multimedia/IJCNN2013.corelet-language.pdf

23. Symas Lightning Memory-mapped Database. - https://symas.com/products/lightning-memory-mapped-database/
24. MatConvNet: CNNs for MATLAB. - http://www.vlfeat.org/matconvnet/
25. NVIDIA CUDNN. GPU Accelerated Deep Learning. - https://developer.nvidia.com/cudnn
26. A Scale-up Synaptic Supercomputer (NS16e): Four Perspectives. Dharmendra S Modha's Brain-inspired Computing Blog. March 2016. - http://modha.org/blog/2016/03/a_scaleup_synaptic_supercomput.html
27. 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
28. The Most Advanced Datacenter GPU Ever Built - http://www.nvidia.com/object/tesla-p100.html
29. NVIDIA DGX-1 DEEP LEARNING SYSTEM. The World's First Deep Learning Supercomputer in a Box. - http://images.nvidia.com/content/technologies/deep-learning/pdf/61681-DB2-Launch-Datasheet-Deep-Learning-Letter-WEB.pdf

Рекомендуемые энциклопедические статьи

1. https://en.wikipedia.org/wiki/Carver_Mead
2. https://en.wikipedia.org/wiki/Artificial_neural_network
3. https://en.wikipedia.org/wiki/Neuromorphic_engineering
4. https://en.wikipedia.org/wiki/Hebbian_theory

5. https://en.wikipedia.org/wiki/Spike-timing-dependent_plasticity
6. https://en.wikipedia.org/wiki/Human_Brain_Project
7. https://en.wikipedia.org/wiki/SyNAPSE
8. https://en.wikipedia.org/wiki/Static_random-access_memory
9. https://en.wikipedia.org/wiki/Very-large-scale_integration
10. https://en.wikipedia.org/wiki/CMOS
11. https://en.wikipedia.org/wiki/Blue_Gene
12. https://en.wikipedia.org/wiki/Lawrence_Livermore_National_Laboratory
13. https://en.wikipedia.org/wiki/Terabyte
14. https://en.wikipedia.org/wiki/Human_brain
15. http://www.scholarpedia.org/article/Axonal_conduction_delays
16. https://en.wikipedia.org/wiki/Visual_cortex
17. https://en.wikipedia.org/wiki/Thalamus
18. https://en.wikipedia.org/wiki/Thalamic_reticular_nucleus
19. https://en.wikipedia.org/wiki/Petabyte
20. https://en.wikipedia.org/wiki/FLOPS
21. https://en.wikipedia.org/wiki/Henry_Markram
22. https://en.wikipedia.org/wiki/Biological_neuron_model
23. https://en.wikipedia.org/wiki/45_nanometer
24. https://en.wikipedia.org/wiki/Silicon_on_insulator
25. https://en.wikipedia.org/wiki/USB
26. https://en.wikipedia.org/wiki/Deterministic_system
27. https://en.wikipedia.org/wiki/Von_Neumann_architecture
28. https://en.wikipedia.org/wiki/Glomerulus_(olfaction)
29. https://en.wikipedia.org/wiki/Olfactory_bulb

30. https://en.wikipedia.org/wiki/Olfactory_bulb_mitral_cell
31. https://en.wikipedia.org/wiki/IBM_Research_-_Almaden
32. https://en.wikipedia.org/wiki/Memristor
33. https://en.wikipedia.org/wiki/Moore%27s_law
34. https://en.wikipedia.org/wiki/Neuromorphic_engineering
35. https://en.wikipedia.org/wiki/National_Science_Foundation
36. https://en.wikipedia.org/wiki/IBM_Sequoia
37. https://en.wikipedia.org/wiki/Scalability
38. https://en.wikipedia.org/wiki/Optical_flow
39. https://en.wikipedia.org/wiki/System_on_a_chip
40. https://en.wikipedia.org/wiki/Scan_chain
41. https://en.wikipedia.org/wiki/Serial_Peripheral_Interface_Bus
42. https://en.wikipedia.org/wiki/I%C2%B2C
43. https://en.wikipedia.org/wiki/PCI_Express
44. https://en.wikipedia.org/wiki/Small_form-factor_pluggable_transceiver
45. https://en.wikipedia.org/wiki/Light-emitting_diode
46. https://en.wikipedia.org/wiki/JTAG
47. https://en.wikipedia.org/wiki/Field-programmable_gate_array
48. https://en.wikipedia.org/wiki/Glue_logic
49. https://en.wikipedia.org/wiki/Direct_memory_access
50. https://en.wikipedia.org/wiki/ARM_architecture
51. https://en.wikipedia.org/wiki/Convolutional_neural_network
52. https://en.wikipedia.org/wiki/Edge_detection
53. https://en.wikipedia.org/wiki/Information_and_communications_technology
54. https://en.wikipedia.org/wiki/CERN

55. https://en.wikipedia.org/wiki/Heidelberg_University
56. https://en.wikipedia.org/wiki/Power_density
57. https://en.wikipedia.org/wiki/Wafer_(electronics)
58. https://en.wikipedia.org/wiki/University_of_Manchester
59. https://en.wikipedia.org/wiki/Stanford_University
60. https://en.wikipedia.org/wiki/IBM
61. https://ru.wikipedia.org/wiki/NeuroMatrix