УДК 004.383.8.032.26

НЕЙРОМОРФНЫЙ ЧИП «АЛТАЙ», ОРИЕНТИРОВАННЫЙ НА ПРИМЕНЕНИЕ В СИСТЕМАХ ТЕХНИЧЕСКОГО ЗРЕНИЯ, РТК И БЕСПИЛОТНЫХ ТРАНСПОРТНЫХ СРЕДСТВАХ

В.М. Канглер (*vkangler@motivnt.ru*)
К.Е. Панченко (*kpanchenko@motivnt.ru*)
ООО “Мотив”, Новосибирск

Аннотация. Рассмотрен проект создания нейроморфного чипа не фон Неймановской архитектуры, моделирующего импульсные нейронные сети (spiking neural networks) и ориентированного на использование во встраиваемых системах, в системах технического зрения и управления робототехническими комплексами. Обозначены ключевые проблемы современных вычислительных систем и недостатки фон Неймановской архитектуры. Продемонстрирована неэффективность существующих вычислительных архитектур в решении ряда задач и обоснована необходимость новой вычислительной архитектуры, которая должна вдохновляться примером человеческого мозга. Показана перспективность использования нейроморфной архитектуры. Перечислены типы задач, которые эффективно решаются с помощью искусственных нейронных сетей. Выражена фактическая безальтернативность использования аппарата искусственных нейронных сетей в решении неформализуемых и слабо формализуемых задач. Представлен нейроморфный чип “Алтай” и обозначены его ключевые преимущества. Обоснованы выбранный поход к обучению моделируемых нейронных сетей, а также целесообразность цифровой реализации по сравнению с аналогово-цифровой. Описаны используемая модель нейрона, архитектурные принципы, структура всего чипа, а также структура и функционирование отдельного нейроядра. Приведен ряд моделей применения разрабатываемого чипа в робототехнических комплексах и системах специального назначения.

Ключевые слова: нейроморфные технологии; импульсные нейронные сети; архитектура фон Неймана; нейроморфная архитектура; аппаратная реализация нейронных сетей; нейроморфный чип.

Введение

Наступающая новая эра интеллектуальных систем, интернета вещей, роботов и беспилотного транспорта нуждается в новых вычислительных средствах, способных в реальном режиме времени обрабатывать гетерогенные "зашумленные данные" большого объема. Для того чтобы эффективно решать задачи по распознаванию образов, глубокому анализу данных и автономному управлению, эти новые вычислительные системы должны сочетать в себе производительность суперкомпьютеров, низкое энергопотребление и высокую надежность. Из этого следует, что новые вычислительные системы должны иметь новую не фон Неймановскую высокопараллельную архитектуру. В природе такие системы уже существуют — это биологические нервные системы, самой продвинутой из которых является человеческий мозг.

1 Ограничения, накладываемые традиционной фон Неймановской архитектурой при реализации аппарата искусственных нейронных сетей (ИНС)

Наращивание количества ядер в микропроцессорах является основным способом повышения их производительности. Производители уже отказались от классической трактовки “закона Мура” ввиду увеличения сложности в одноядерных процессорах и интерпретируют этот закон по-новому — количество процессорных ядер на кристалле удваивается примерно каждые 18 месяцев. Если ориентироваться на такие формулировки, то в ближайшие десятилетия будут получены кристаллы с миллионами процессорных ядер. Хотя вопрос автоматического распределения кода по вычислителям все еще является “ненайденным Святым Граалем” и остается одним из ключевых вопросов вычислительных наук. В то же время биологический мозг демонстрирует высочайший уровень параллелизма и эффективности. Это и вдохновляет ученых и инженеров на создание нейроморфных устройств [Furber et al., 2013].

Традиционно аппарат искусственных нейронных сетей (ИНС) использовался для решения следующих задач: распознавание образов, реализация ассоциативной памяти, аппроксимация функций, управление процессами и автономными системами, фильтрация, сглаживание, прогнозирование. Разнообразие этих задач свидетельствует в пользу универсальности нейронных сетей как систем обработки информации [Хайкин, 2006]. В связи с развитием робототехники и увеличением потребности в быстрой обработке растущих объемов данных и мультимедиа информации растет и заинтересованность в эффективном

решении упомянутых задач. Это подстегивает интерес к системам, моделирующим большие нейронные сети. Моделирование нейронных сетей требует существенных вычислительных ресурсов. Для моделирования используются нейроускорители на сигнальных процессорах (DSP), графических процессорах (GPU) или ПЛИС (FPGA). Однако, параллельная и управляемая событиями природа нейронных сетей не является естественной для модели последовательной обработки традиционной компьютерной архитектуры (архитектура фон Неймана). В архитектуре фон Неймана необходима высокая пропускная способность для пересылки состояния входных и выходных значений нейронов между физически разделенными процессорами и памятью. Это приводит к высокому энергопотреблению и ограничивает возможности масштабирования таких систем [Imam et al., 2012]. Нейроморфные архитектуры изначально лишены этих недостатков и являются более естественными для моделирования нейронных сетей.

2 Актуальность применения нейроморфных технологий в системах технического зрения, навигации и управления

Актуальность нейроморфных технологий также подтверждается цитатой из книги вице-президента IBM по стратегическим решениям и исследованиям Джона Келли. "Архитектура фон Неймана так долго используется, потому что она обеспечивает мощное средство решения многих вычислительных задач. Большинство программ, написанных для сегодняшних компьютеров, основаны на этой архитектуре. Она имеет недостаток, который делает ее неэффективной — это так называемое "узкое горлышко архитектуры фон Неймана". В этой архитектуре каждый этап обработки требует нескольких шагов, где данные и команды перемещаются туда-сюда между памятью и процессором. Это требует огромного количества перемещений данных и обработки. Это также означает, что шаги обработки должны выполняться последовательно. И хотя здесь можно ввести некоторый параллелизм, но этого явно недостаточно. В течение последних десятилетий разработчики могли увеличивать возможности процессоров, делая их меньше и быстрее. В таком подходе мы почти достигли предела наших возможностей, и как раз в то время, когда мы нуждаемся в еще большей вычислительной мощности, чтобы справиться с требуемой сложностью и обработкой больших данных. Это приводит к невыполнимым требованиям для сегодняшних компьютерных технологий, в основном потому что современные компьютеры требуют колоссальной энергии для их выполнения.

Что необходимо, так это новая вычислительная архитектура, которая

должна вдохновляться примером человеческого мозга.

Обработка данных должна быть распределена по всей вычислительной системе, а не быть сосредоточенной в центральном процессоре. Обработка и память должны быть тесно интегрированы, чтобы уменьшить челночную пересылку данных и команд туда-сюда. И этапы обработки должны выполняться одновременно, а не последовательно. Когнитивный компьютер, используя такую архитектуру, будет реагировать на запросы быстрее, чем сегодняшние компьютеры, потребует меньшего количества перемещений данных; будет использоваться меньше энергии. Это не означает, что архитектура фон Неймана не будет использоваться. Она будет использоваться вместе с новой архитектурой в единых гибридных системах." [Kelly, 2013]

Под нейроморфными, как правило, понимают устройства аппаратно моделирующие работу импульсных нейронных сетей (Spiking Neural Networks, SNN).

Нейроморфные приборы будут востребованы в робототехнике: в системах технического зрения, навигации и управления. Развитие “традиционной” микроэлектроники позволило робототехнике (БПЛА, промышленные роботы, беспилотные транспортные средства) постоянно увеличивать функциональность и гибкость. Несмотря на это, интеллектуальные функции по-прежнему выполняются человеком — оператором. Обычно для этого требуются двунаправленные каналы связи с высокой пропускной способностью между оператором и роботом. Несмотря на то, что это приемлемо для многих ситуаций, повышение интеллектуальности и автономности управления за счет применения нейроморфных приборов кардинально изменит сценарии использования роботов в большинстве областей применения. Такие качества как энергоэффективность, функционирование в реальном режиме времени и размещение высокоуровневой системы управления на борту очень важны для робототехнических систем, в которых длительность автономной работы, удаленность от центра управления и короткое время реакции являются критичными [NICE, 2015]. Ярким примером таких систем могут быть космические робототехнические системы.

3 «АЛТАЙ» — разрабатываемый цифровой КМОП нейроморфный чип, моделирующий работу импульсных нейронных сетей

Разрабатываемый нейроморфный чип “Алтай” — «вычислительное» устройство, функционирующее на принципах схожих с биологическими нейронными системами.

В качестве основных моделей применения рассматриваются:

- Обработка изображений, в том числе кадров видеоряда в различных спектрах, например, в видимом, тепловом и других, с целью обнаружения, выделения и категоризации значимых объектов.
- Обработка акустических, в том числе и гидроакустических, сигналов с целью выделения интересующих акустических паттернов.
- Обработка физиологических сигналов. Например, анализ электроэнцефалограммы с целью организации эргономичного взаимодействия человека с робототехническими системами.

К ключевым преимуществам нейрочипа “Алтай” можно отнести:

- Высокая эффективность по энергопотреблению, производительности и размерам.
- Возможность решения неформализуемых и плохо формализуемых задач.
- Высокая масштабируемость, ограниченная только требованиями по энергопотреблению и массогабаритным параметрам.

3.1 Основополагающие принципы построения нейрочипа

“Алтай” должен обеспечивать эффективное функционирование произвольной импульсной нейронной сети.

Настройка параметров нейронов сети, в том числе и обучение, задание структуры сети, связей между нейронами, должны выполняться вне чипа с использованием инструментальных программных средств.

Вынесение задач обучения за пределы кристалла расширяет возможности выбора подходов и способов формирования параметров сети, с тем чтобы получить требуемую выходную функцию, зависящую от входных данных, как для отдельных подсетей, так и для всей моделируемой сети в целом. Также это повышает эффективность использования нейрочипа на этапе функционирования обученной сети.

3.2 Архитектура нейрочипа «Алтай»

3.2.1 Ключевые архитектурные решения

В качестве основополагающих были выбраны следующие архитектурные решения:

- Полная цифровая реализация на КМОП технологии [Imam et al., 2012].
- Достаточно универсальная, но простая в реализации на кристалле цифровая модель нейрона. Модель должна поддерживать возможность выполнять широкий набор вычислительных

функций одним нейроном или группой нейронов: арифметические операции; логические операции; классическое поведение нейронов; обработку сигналов и вероятностное вычисление. [Cassidy et al., 2013].

- Параметры функционирования нейронов формируются вне кристалла [Arthur et al., 2012]. Это позволит использовать различные алгоритмы обучения нейронной сети. Кроме того, параметры нейронов могут быть получены не только в результате выполнения процедуры обучения, но и жестко заданы, если нужно реализовать на нейронной сети какую-то конкретную функцию, например, когда часть сети выполняет фильтрацию изображения функцией свертки.
- Многоядерная архитектура. Сам нейрочип представляет собой масштабируемую сеть нейроядер. Ядро — обособленная небольшая группа импульсных нейронов. Коммуникация спайков внутри ядра реализуется матрицей связей (crossbar), которая обеспечит высокую пропускную способность сигналов и исключит их взаимную блокировку. Сам нейрочип представляется как двумерная сеть из ядер. Для коммуникации между ядрами используются разделяемые шины и относительная AER (Address Event Representation) маршрутизация. Такой подход позволит масштабировать нейронные сети, объединяя множество нейрочипов в единый вычислительный комплекс.
- Нейрочип проектируется по модели GALS (Globally Asynchronous Locally Synchronous, глобально асинхронная локально синхронная схема) [Wang, 2008]. Ядра являются синхронными схемами, каждое из которых функционирует в своем домене синхронизации. Все коммуникационные блоки нейрочипа являются асинхронными. Такой подход является более сбалансированным по показателям энергопотребления, производительности и эффективности использования площади кристалла.

3.2.2 Используемая модель нейрона

В проекте используется модификация “классической” модели нейрона — LIF (Leaky Integrate-and-Fire) с постоянной величиной утечки. Модель несколько упрощена, в ней используется только целочисленная арифметика.

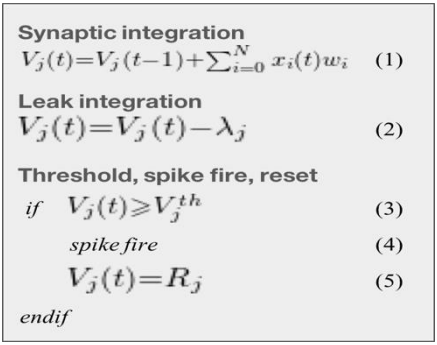

Рис. 1. Используемая модель нейрона.

Для *j*-го нейрона в момент времени *t* мембранный потенциал $V_j(t)$ является суммой потенциала на предыдущем шаге $V_j(t-1)$ и синаптических входов. Для каждого из *N* синапсов, синаптический вход равен произведению входного спайка на синапсе $x_i(t) \in \{0;1\}$ с его синаптическим весом w_i. Затем из накопленного потенциала вычитается величина утечки λ_j. Если накопленная величина $V_j(t)$ превышает порог α_j, то нейрон испускает импульс и мембранный потенциал "сбрасывается" до значения R_j.

3.2.3 Архитектура нейрочипа и принцип функционирования

Нейрочип "Алтай" включает в себя следующие функциональные блоки:

- Двумерную решетку нейроядер (32x32, 1024 ядра).
- 4 последовательных двунаправленных асинхронных порта ввода-вывода.
- Блок конфигуратора, предоставляющий I2C интерфейс для загрузки параметров моделируемой импульсной сети в нейроядра.
- Блок мониторинга, который позволяет контролировать работу нейрочипа.
- Блок автоматического регулирования внутренней частоты (ФАПЧ — фазовая автоподстройка частоты).

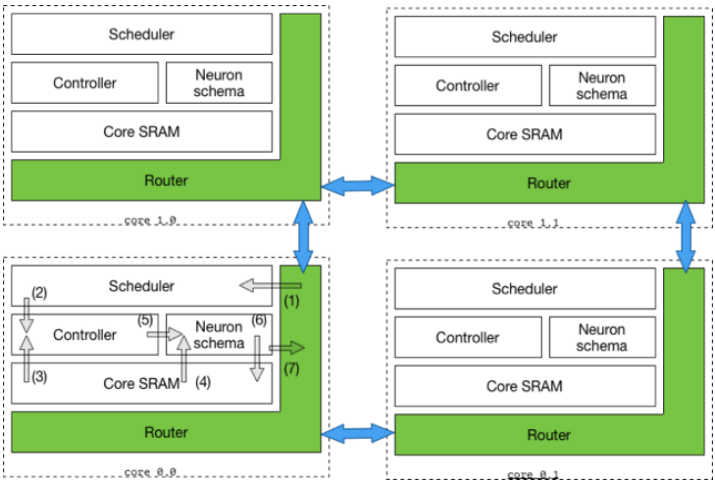

Рис. 2. Фрагмент решетки нейроядер.

Структурно ядро состоит из пяти функциональных блоков:

- Router — коммуницирует с соседними нейроядрами, принимая и маршрутизируя спайки.
- Scheduler — упорядочивает по времени и по входам (аксонам) приходящие спайки.
- Core SRAM — хранилище параметров функционирования нейронов ядра.
- Controller — реализует алгоритм управления последовательностью вычислений.
- Neuron schema — схема, реализующая логику функционирования нейрона.

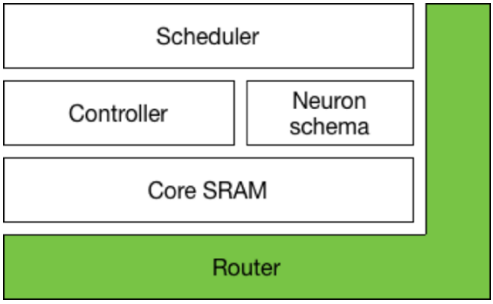

Рис. 3. Структура нейроядра.

Ядро моделирует 2^n нейронов и 2^m входов (аксонов).

Когда спайк достигает целевого ядра, маршрутизатор направляет его планировщику (1), где спайк буферизируется.

В каждый такт планировщик передает контроллеру все спайки, соответствующие по порядку текущему такту (2). Получив спайки, контролер организует последовательность вычислений для $\mathbf{2^n}$ нейронов по одному (один за другим).

В течение одного такта обсчитываются все $\mathbf{2^n}$ нейронов.

3.3 Оценки параметров энергопотребления и производительности

Проект по разработке нейроморфного чипа "Алтай" находится на начальной стадии и в данный момент проходит этап верификации HDL-кода нейроядер и всего чипа на “золотой модели”. Кроме того, получены оценки параметров энергопотребления для варианта, ориентированного на отечественный 90нм техпроцесс производства.

Ниже в таблице приведены планируемые параметры.

Таблица. 1. Планируемые характеристики нейрочипа «Алтай»

Количество нейроядер на чип	1024
Количество моделируемых нейронов	262144
Количество синапсов	67108864
Производительность — количество синаптических операций в секунду (Sops)	$67*10^9$
Период шага обработки (сек.)	$1*10^{-3}$
Внутреняя частота работы нейроядер (МГц)	70
Напряжение питания (В)	1
Площадь кристала ($мм^2$)	225
Оценка мощности одного ядра (мВт)	0.6
Оценка потребляемой мощности (Вт)	0.612

Предполагаемый тип корпуса	TQFP-144 144 вывода 22мм x 22мм

4 Сравнение с существующими решениями

Одной из самых очевидных и перспективных областей применения разрабатываемого чипа являются системы технического зрения. В настоящий момент эта область переживает революционные изменения. Системы, основанные на свёрточных глубоких нейронных сетях (CDNN), начиная с 2012 года, демонстрируют лучшие результаты в задачах распознавания образов и классификации изображений. В ряде типовых задач алгоритмы CDNN в качестве распознавания изображений сравнялись с человеком [LeCun et al., 2015], [Esser et al., 2016].

Сравним оценочные параметры разрабатываемого чипа с известными проектами на примере моделирования свёрточных нейронных сетей. Моделирование свёрточных сетей является вычислительно очень ресурсоёмкой задачей. Обучение CDNN требует миллионы циклов и по времени может занимать недели даже на мощных современных GPU. Поэтому в мире многие исследователи сосредоточены на оптимизации структур и алгоритмов функционирования CDNN, а также на том, чтобы предложить эффективные аппаратные решения, позволяющие использовать обученные CDNN во встраиваемых системах.

Для автономных и встраиваемых систем ключевыми параметрами являются производительность (количество кадров, обрабатываемых в секунду), эффективность (количество энергии необходимое для обработки одного кадра) и точность распознавания. Если сравнивать популярные микропроцессоры, графические процессоры, FPGA и ASIC (application-specific integrated circuit, интегральная схема специального назначения), то по этим параметрам решения, основанные на ASIC, вне конкуренции. Ниже приводится таблица отражающая параметры моделирования AlexNet [Krizhevskyet al., 2012] на общедоступных CPU и GPU. Видно, что GPU/mGPU значительно опережает CPU как по производительности, так и по энергоэффективности [Nvidia, 2015].

Таблица. 2. Основные параметры существующих конкурентных решений

	Intel Xeon E5-2698 v3	Intel Core i7-6700K	Nvidia Titan X	Nvidia Tegra X1

Тип	CPU	CPU	GPU	mGPU
Призводительность (fps)	476	242	3216	258
Потребляемая мощность (W)	149	62.5	227	5.7
Энергоэффективность (J/f)	0.313	0.258	0.071	0.022

Рассмотрим наши результаты в сравнении с последними достижениями схожей тематики. Конечно, напрямую сравнивать не совсем корректно, так как нужно сравнивать на одних и тех же задачах/тестах, но сравнение на близких задачах позволит понять хотя бы порядки величин. В таблице (#) приводим результат сравнения наших расчетных показателей с информацией из ряда последних публикаций [Nvidia, 2015], [Han et.al, 2016] , [Esser et al., 2016], [Chen et al., 2016].

Таблица. 2. Сравнения существующих решений с нейрочипом «Алтай»

	Nvidia Titan X	Nvidia Tegra X1	A-Eye	IBM True-North	MIT Eyeriss	Altai (our)
Тип	GPU	mGPU	FPGA	ASIC	ASIC	ASIC
Год	2015	2015	2015	2016	2016	2016
Призводитель-ность (fps)	3216	258	8	1738	17	980
Потребляемая мощность (W)	227	5.7	9.63	0.208	0.1175	0.5212
Энергоэф-фективность (J/f)	$7.1 * 10^{-2}$	$2.2 * 10^{-2}$	1.203	$1.2 * 10^{-4}$	$6.9 * 10^{-3}$	$5.3 * 10^{-4}$

Проведенный сравнительный анализ с существующими передовыми решениями аппаратного ускорения свёрточных сетей показал, что не смотря на то, что чип Алтай ориентирован не на самый современный технологический процесс производства (техпроцесс 90нм, Микрон), он превосходит перспективный ускоритель Eyeriss (техпроцесс 65нм, TSMC)

по энергоэффективности и по производительности. При этом Алтай не сильно уступает самому большому из существующих нейроморфному чипу IBM TrueNorth (техпроцесс 28нм, Samsung).

Заключение

Нейрочипы сейчас находятся на той же стадии, на которой микропроцессоры были в 1980-х. Эту область в ближайшее время ожидает бурный рост [Markets and Markets, 2015]. Нейрочипы будут активно использоваться в системах зрения, управления и навигации робототехнических комплексов.

Проект по разработке нейроморфного чипа "Алтай" находится на стадии верификации HDL-кода, и в данный момент получены только расчетные данные по энергопотреблению и вычислительной мощности. Но уже сейчас чип "Алтай" можно рассматривать как один из кандидатов для использования во встраиваемых системах, где требуется реализовывать решение неформализуемых или плохо формализуемых задач (например, задачи распознавания в системах технического зрения) с высоким быстродействием и низким энергопотреблением. Низкое энергопотребление и небольшие габариты позволяют использовать его в системах технического зрения даже в небольших устройствах, например, в "умных" боеприпасах, малых БПЛА и т.п. Кроме систем технического зрения данный чип может использоваться в бортовых подсистемах управления и диагностики.

Список литературы

[Зверев 2007] Зверев Г.Н. К обобщенной теории обработки наблюдений // Нефтепромысловая геофизика. – М.: ИГ и РГИ, 2007.

[Хайкин, 2006] Хайкин С. Нейронные сети: полный курс, 2-е издание, Пер. с англ. - М.: Издательский дом "Вильямс", 2006. - 1104с. ISBN 5-8459-0890-6(рус.)

[Arthur et al., 2012] Arthur J., Merolla P., Akopyan F., Alvarez R., Cassidy A., Chandra S., Esser S., Imam N., Risk W., Rubin D, Manohar R., Modha D. Building block of a programmable neuromorphic substrate: A digital neurosynaptic core // Neural Networks (IJCNN), The 2012 International Joint Conference on Year: 2012 Pages: 1 - 8, DOI: 10.1109/IJCNN.2012.6252637

[Cassidy et al., 2013] CassidyA., Merolla P., Arthur J., Esser S., Jackson B., Alvarez-Icaza R., Datta P., Sawada J., Wong T., Feldman V.,Amir A.,Rubin D, Akopyan F., McQuinn E., Risk W., Modha D. Cognitive computing building block: A versatile and efficient digital neuron model for neurosynaptic cores // Neural Networks (IJCNN), The 2013 International Joint Conference on Year: 2013 Pages: 1 - 10, DOI: 10.1109/IJCNN.2013.6707077.

[Chen et al., 2016] Yu-Hsin Chen, Tushar Krishna, Joel Emer, Vivienne Sze Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural

Networks // International Solid-State Circuits Conference 2016.

[Esser et al., 2016] Steven K. Esser, Paul A. Merolla, John V. Arthur, Andrew S. Cassidy, Rathinakumar Appuswamy, Alexander Andreopoulos, David J. Berg, Jeffrey L. McKinstry, Timothy Melano, Davis R. Barch, Carmelo di Nolfo, Pallab Datta, Arnon Amir, Brian Taba, Myron D. Flickner, Dharmendra S. Modha "Convolutional Networks for Fast, Energy-Efficient Neuromorphic Computing" // arXiv preprint arXiv:1603.08270, 2016.

[Furber et al., 2013] Furber S, Lester D., Plana L., Gaside J., Painkras E., Temple S., Brown A. Overview of the SpiNNaker System Architecture // Computers, IEEE Transactions on Year: 2013, Volume: 62, Issue: 12, pp 2454 - 2467, DOI: 10.1109/TC.2012.142

[Imam et al., 2012] Imam N., Akopyan F., Arthur J., Merolla P., Manohar R., Modha D. A Digital Neurosynaptic Core Using Event-Driven QDI Circuits // Asynchronous Circuits and Systems (ASYNC), 2012 18th IEEE International Symposium on Year: 2012 Pages: 25 - 32, DOI: 10.1109/ASYNC.2012.12.

[Han et.al, 2016] S. Han et.al, EIE: Efficient Inference Engine on Compressed Deep Neural Network // ISCA 2016.

[Kelly, 2013] Kelly J., Hamm S. Smart Machines: IBM's Watson and the Era of Cognitive Computing // Columbia University Press (September 24, 2013)

[Krizhevskyet al., 2012] Krizhevsky, A., Sutskever, I. and Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks // NIPS 2012: Neural Information Processing Systems

[LeCun et al., 2015] Yann LeCun, Yoshua Bengio, Geoffrey Hinton Deep learning // Nature Vol 521, 2015, doi:10.1038/nature14539

[Markets and Markets, 2015] Neuromorphic chip market global forecast to 2022 // Markets and Markets, 2015.

[NICE, 2015] Summary Report from 2015 Neuro-Inspired Computational Elements (NICE) Workshop.

[Nvidia, 2015] GPU-Based Deep Learning Inference: A Performance and Power Analysis // Nvidia Whitepaper, November 2015.

[Wang, 2008] Xin Wang "Designing Globally-Asynchronous LocallY-Synchronous On-Chip Communication Networks" // Tempere University of Technology, 2008, ISBN 978-952-15-2005-1.