Ускорители вычислений. Возможности нейронных сетей, применение на практике

В статье рассказано об особенностях, возможностях и сферах применения ускорителей вычислений. Вместе со статьей опубликовано интервью с В. Райчевым, менеджером проектов отдела промышленных компьютеров компании «Ниеншанц-Автоматика», которая представляет на российском рынке ускорители вычислений известных мировых производителей.

Компания «Ниеншанц-Автоматика», г. Санкт-Петербург

Ускорителями вычислений обычно называют платы расширения с микросхемой FPGA (field-programmable gate array, к русскому эквивалентному названию мы вернемся чуть позже) и с необходимой для работы этой микросхемы «обвязкой», примерно как графические ускорители – это платы расширения с графическим сопроцессором и необходимыми вспомогательными микросхемами и навесными элементами. Если сразу перейти к техническим характеристикам, то назначение и полезность этих плат будут неочевидны даже для специалистов, поэтому начать нам придется с теории и истории развития FPGA. А чтобы эта теория не казалась оторванной от практики, сразу отметим, что в 2015 году одного из основных разработчиков FPGA, компанию Altera («Алтера»), купила корпорация Intel за немалую сумму – 16,7 млрд долларов США – и сегодня на рынке именно Intel наиболее активно продвигает ускорители вычислений.

Дословный перевод FPGA таков: массив вентилей (логических элементов) с программированием «в поле» (за воротами завода), что обычно называют программируемой пользователем вентильной матрицей. В свою очередь, ASIC (application-specific integrated circuit – интегральная схема специального назначения), напротив, программируется на заводе-изготовителе для определенной области применения из общих одинаковых заготовок, что позволяет снизить затраты в средне- и мелкосерийных партиях. В обоих случаях под программированием понимается загрузка в такую микросхему прошивки.

Последнее время большие надежды возлагают на аппаратную реализацию нейронных взаимодействий на основе FPGA, причем ASIC с их жесткими внутренними взаимосвязями уже не подходят в принципе.

Приведем несколько цитат на эту тему из англоязычной брошюры компании Intel “Efficient Implementation of Neural Network Systems Built on FPGAs, and Programmed with OpenCL™” («Эффективность реализации систем нейронных сетей, построенных на микросхемах FPGA и программированных в OpenCL™»): «Системы нейронных сетей с глубоким обучением сегодня предоставляют наилучшие способы решения многих задач с большим объемом вычислений в областях распознавания изображений и обработки естественных языков <...>, причем наиболее часто для идентификации и распознавания изображений используются сверточные нейронные сети CNN (convolutional neural network)». В них применяется функция свертки из математического функционального анализа, позволяющая оценить «схожесть» двух математических функций после отражения или сдвига по оси. На практике это означает оценку схожести/подобия изображений после поворота головы или изменения угла съемки. Продолжим цитирование: «Микросхемы FPGA служат естественным выбором для реализации нейронных сетей благодаря объединению в одном устройстве вычислительных и логических ресурсов вместе с памятью. Однако применение FPGA непрактично для широкомасштабного использования в сложных алгоритмических системах из-за традиционных для этих микросхем сред разработки на низкоуровневых языках аппаратного программирования. Эту проблему устраняет комплект разработчика Intel FPGA SDK for OpenCL <...>, позволяющий разработчикам обращаться из языка программирования уровня С к FPGA в виде ускорителей для обычных ЦП».

Остается добавить, что OpenCL (Open Computing Language – открытый язык вычислений) – это каркас разработки (фреймворк) компьютерных программ с параллельными вычислениями для обычных центральных процессоров, графических процессоров и FPGA. Параллелизм OpenCL отлично подходит для программирования путей в матрице

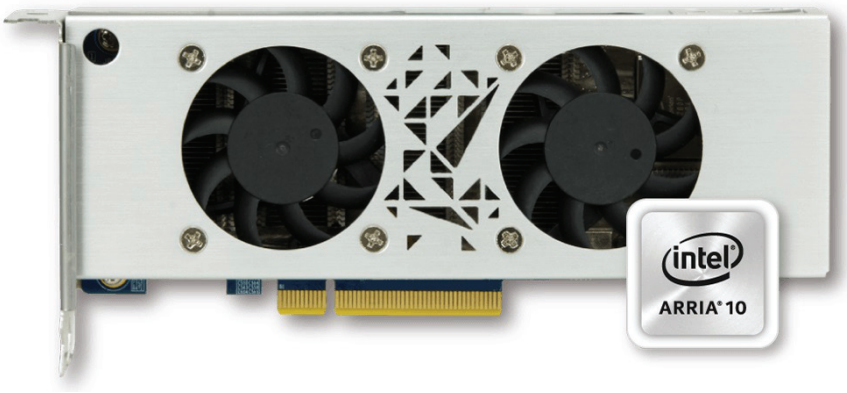

Рис. 1. Плата Mustang-F100 от IEI Integration

Рис. 2. Ускоритель вычислений Mustang-V100, разработанный компанией IEI Integration

FPGA, причем сами эти микросхемы лучше встроить в систему в виде платы расширения, то есть ускорителя вычислений. А применение ускорителей вычислений не ограничивается только обработкой изображений. Мощный FPGA-ускоритель вычислений от IEI Integration представлен платой Mustang™-F100-A10 (рис. 1), которая оснащена FPGA-чипом Intel Arria 10 GX1150 и памятью DDR4 объемом 8 ГБ. Ускоритель может быть установлен в слот PCI Express x8 и занимает два слота в ширину. Потребляемая мощность – менее 60 Вт.

С верхним сегментом рынка и компанией Intel мы познакомились: современно, солидно и очень-очень дорого. Но если стоит задача распознавания цвета лимонада на конвейере, чтобы не перепутать этикетки, а вовсе не распознавания лиц в плотном потоке пассажиров метро, то можно выбрать намного более экономичную альтернативу, сохранив некоторую совместимость с рассмотренными выше «тяжелыми» системами. В этом смысле интерес представляют VPU (Vision Processing Unit, не перепутайте с Video/Visual Processing Unit, используемыми для формирования, а не анализа изображений), каковым является Intel® Movidius™ Myriad™, которые могут быть распаяны на платах расширения различного формата, причем лучше сразу иметь несколько таких микросхем для распараллеливания задачи. Именно так построен ускоритель вычислений Mustang-V100-MX8 (рис. 2) упомянутой выше компании IEI Integration: восемь VPU-процессоров Intel Movidius Myriad X MA2485, плата устанавливается в слот PCI Express x4 и имеет компактные габариты: половинная высота и половинная длина, потребляемая мощность менее 30 Вт.

Для оптимизации выполнения алгоритмов, использующих нейронные сети, компания Intel предлагает набор средств разработки (toolkit) под названием OpenVINO (Open Visual Inference & Neural Network Optimization). Набор включает в себя ряд библиотек, средств оптимизации и информационных ресурсов для разработки софта, использующего машинное зрение и глубокое обучение.

Иными словами, средства позволяют оптимизировать/распараллеливать вычислительную нагрузку при обсчете нейронных сетей по аппаратным средствам Intel, в том числе исполняемых на обычных ЦП, графических процессорах GPU, VPU и FPGA. К тому же в состав комплекта входит более 40 готовых к развертыванию демонстрационных алгоритмов.

Все серии ускорителей вычислений компании IEI Integration в нашей стране предлагает петербургская компания «Ниеншанц-Автоматика», которая специализируется на поставках и технической поддержке оборудования для промышленной автоматизации. Среди заказчиков этой компании ПАО «Газпром», ПАО «Транснефть», ПАО «Ростелеком», ОАО «РЖД», ГУП «Московский метрополитен», ОАО «Северо-Западный Телеком». Несомненно, компания «Ниеншанц-Автоматика» поможет в любом проекте с ускорителями вычислений IEI Integration при решении любых вопросов: от предпродажных консультаций до монтажа и сдачи в эксплуатацию на объекте, от регулярных инспекций после продажи до поставки запчастей и обучения персонала заказчика. Чтобы подробнее узнать о таком популярном сегодня решении, как ускорители вычислений, мы обратились к представителю компании «Ниеншанц-Автоматика», специализирующемуся на продвижении ускорителей.

Интервью с Владимиром Райчевым, менеджером проектов отдела промышленных компьютеров компании «Ниеншанц-Автоматика»

ИСУП: Владимир! С помощью ускорителей вычислений можно решать большой круг задач: это и машинное обучение, и транскодирование видеопотоков, и финансовая аналитика, и многое другое, поэтому они в настоящее время вызывают огромный интерес. Кто сейчас главный игрок на этом рынке?

В. Райчев: Пожалуй, главного игрока сегодня назвать сложно, его просто нет. На мой взгляд, наиболее популярны решения от NVIDIA, например бытовые видеокарты. Так исторически сложилось, что корпорация NVIDIA первой начала осваивать эту область. Также популярны платы-ускорители Intel, внедряя которые, эта компания заняла довольно агрессивную позицию. У Intel два решения: одно – на чипах Movidius, которые перешли в портфель Intel вместе с из-

раильской компанией, изначально их разрабатывающей. Второе решение – на ПЛИС (FPGA), которые Intel приобрела вместе с компанией Altera и теперь выпускает под своим брендом. Вроде бы, два похожих решения, но у них очень много различий, в первую очередь по решаемым задачам.

Чипы Movidius™ предназначены для задач, связанных с видеообработкой. Допустим, требуется распознавать людей, попавших в зону работы системы видеонаблюдения. Мы можем также задать дополнительные параметры: пол, возраст, настроение. Контентов много! Вплоть до того, что можно определить, носит ли человек в производственных помещениях специальную антистатическую обувь или ходит в обычной уличной. Причем алгоритмы можно обучить не только выполнению таких задач, более сложные им тоже под силу. Существует интернациональный проект Botkin.AI, который нацелен на то, чтобы обучить алгоритм прогнозировать развитие рака легких по рентгеновским снимкам. Это проект с колоссальным объемом работы, в котором принимают участие не только программисты, но и врачи. В том числе над ним работают и наши соотечественники. С одной из компаний-участниц мы налаживаем сотрудничество.

ИСУП: Скажите, исходя из своего опыта, как один из основных дистрибьюторов в России, предлагающих подобные решения: для каких целей чаще всего приобретаются ускорители вычислений?

В. Райчев: В России однозначно лидирует направление Big Data, а основными заказчиками выступают IT-компании. Интересный факт: около года назад голосовой помощник от «Яндекса», «Алиса», научилась искать информацию по фотографиям с камеры или любым другим изображениям. Это стало возможным благодаря ускорителям, разработанным компанией Intel.

ИСУП: Если рассматривать номенклатуру изделий, которые вы поставляете на отечественный рынок, то на каком из чипов построено больше решений? На Arria 10?

В. Райчев: В нашей номенклатуре преобладают решения от Intel. Одно

▲ В. Райчев, менеджер проектов компании «Ниеншанц-Автоматика»

решение с Arria 10, Mustang-F100-A10, предназначено для задач, связанных с нейронными сетями и искусственным интеллектом. И несколько решений для интеллектуальной видеоаналитики, использующих визуальный процессор Intel Movidius Myriad X. На сегодняшний день доступно одно решение – Mustang-V100-MX8. Однако во время проведения Computex 2019 будет анонсировано еще четыре платы различных форм-факторов.

ИСУП: А с NVIDIA вы тоже работаете?

В. Райчев: Да, и решения от NVIDIA у нас тоже есть, причем в защищенном исполнении IP67.

ИСУП: Давайте поговорим о конкретных моделях, которые вы предлагаете своим покупателям, например об ускорителе вычислений Mustang-V100-MX8 от IEI Integration.

В. Райчев: Mustang-V100-MX8 – не первая ласточка. Я имею в виду, что с данной технологией уже можно было познакомиться в устройстве Movidius Neural Compute Stick, которое пользователи прозвали «нейрофлешкой» из-за внешнего вида. Наше решение выполнено в формате платы расширения для шины PCI-express, имеет «на борту» восемь чипов Myriad X и предназначено для машинного зрения – обработки графической информации с помощью нейронных сетей. Мы с нетерпением ждем появления плат в форм-факторах mPCIE и M.2 для использования их во встраиваемых компьютерах.

ИСУП: Владимир, а для каких задач могут понадобиться встраиваемые компьютеры с предустановленными платами Mustang?

В. Райчев: Это, вероятно, самая интересная часть – применение и возможности таких компьютеров при децентрализованных вычислениях. Так как мы говорим про решения на базе визуальных процессоров Intel Movidius Myriad X, то основную сферу можно легко обозначить: видеоаналитика с помощью нейронных сетей (машинное зрение), для работы которых требуются большие вычислительные мощности. Визуальный процессор как раз и предоставляет такие мощности при малом потреблении энергии. То есть теперь мы можем, к примеру, использовать алгоритм, который будет получать изображение с камеры и распознавать на нем объекты. У нас есть рабочий демостенд с распознанием людей в кадре и определением их пола, возраста и настроения.

ИСУП: Звучит интересно, однако как машинное зрение может быть применено в автоматизации технологических процессов?

В. Райчев: Наш второй демо-стенд как раз для промышленности, на нем представлена оцифровка данных стрелочного индикатора, например манометра, с помощью встраиваемого компьютера и веб-камеры. Принцип работы следующий: камера устанавливается напротив манометра и подключается к компьютеру со специально обученным алгоритмом. При первом запуске оператору необходимо произвести калибровку с помощью графического интерфейса и указать цену деления шкалы. После чего алгоритм сможет отслеживать изменение положения стрелки и предоставлять оцифрованные значения манометра. Эти данные уже можно отправлять в SCADA-систему. Ускоритель в данном примере позволяет провести обработку данных почти без задержек. Другим направлением для машинного зрения в промышленности является контроль качества выпускаемой продукции.

ИСУП: Как вы считаете, почему совсем недавно, только в 2018 году, вокруг ускорителей вычислений на рынке начался настоящий ажиотаж?

В. Райчев: Я считаю, что это было вполне ожидаемо, алгоритмы с использованием нейронных сетей получили достаточную популярность, чтобы крупные компании увидели потенциальную прибыль. А так как основным замедляющим фактором была недостаточная производительность при обработке данных алгоритмами, то вполне логично, что появились специализированные аппаратные решения. К тому же многие компании уделяют большое внимание оптимизации алгоритмов и средств их разработки. В итоге мы уже подобрались к такому уровню оптимизации и производительности, что практически каждый может провести свой собственный эксперимент и понять, насколько это ему интересно, как это применимо к его задачам. Выяснилось, что задачи, которые можно решить с помощью машинного зрения, буквально повсюду. Самое простое – это аналитика видеоданных. Но также они помогают строить более точные модели прогнозирования на основе данных, которые раньше невозможно было собрать или обработать. Глубокое обучение (Deep lerning) – это новый виток развития теории о принятии решений, что позволяет поручить компьютерам новые задачи, на которые раньше был способен только человек.

ИСУП: А конкретно где это сейчас применяется? Например, в ЦОД, где анализируются данные с дорожных камер видеонаблюдения, применяются ли возможности нейронной сети?

В. Райчев: Насколько мне известно из открытых источников, камеры фотовидеофиксации в Москве уже подключены к ЦОД, использующему искусственную нейронную сеть для распознавания марки и модели автомобиля.

ИСУП: А в нефтяной промышленности?

В. Райчев: Да, конечно. В нефтяной промышленности в первую очередь ставится вопрос о безопасности периметра и контроля доступа: кто из персонала подходил к тому или иному объекту, какую территорию пересекал. Искусственная нейронная сеть позволяет «отфильтровывать» данные, получаемые с камер, применить к их классификации сложные правила и тем самым повысить эффективность системы видеонаблюдения: избежать ложных срабатываний даже при неблагоприятных погодных условиях. А также алгоритм может сигнализировать, например, носит ли персонал каску, находясь на объекте.

ИСУП: А нейрофлешки? Например, Intel предлагает своим потребителям нейрофлешку Movidius NCS с протоколом USB 3.0. Я знаю, что ее достаточно активно используют. У нас нет, но, например, в Польше ее ставят на дроны. Летящий дрон способен в режиме реального времени вести обработку данных, фиксировать утечку газа, лесные пожары и т. д.

В. Райчев: Нейрофлешки хорошо подходят для прототипирования, но не всегда – для решения реальных задач. Mustang-V100-MX8 по производительности сопоставим с восемью Intel NCS и вместо восьми NCS можно установить лишь одну карту Mustang.

Беседовал С. В. Бодрышев,
главный редактор журнала «ИСУП».

Компания «Ниеншанц-Автоматика»,
г. Санкт-Петербург,
тел.: +7 (812) 326-5924,
e-mail: ipc@nnz.ru,
сайт: www.nnz-ipc.ru