Заключение

Сервисы FusionStorage предназначены для упрощения хранения корпоративных данных с одновременной минимизацией рисков и возможных проблем при переходе на облачные вычисления. Среди основных преимуществ FusionStorage:

- *<u>сверхдолгое управление жизненным циклом данных:</u> глобальное пространство имен, формируемое по регионам и кластерам. Доступ к данным осуществляется в непосредственной близости от пользователей, а службы клиенттов глобализируются и мобилизуются для создания единого пула облачных ресурсов. FusionStorage устраняет «горячие точки» кластера и обеспечивает масштабируемость на уровне EB (поддержка расширения емкости от терабайт до эксабайт и распределение ресурсов хранения по требованию), а также обеспечивает автоматическую балансировку нагрузки для удовлетворения требований управления сверхдолгого жизненного цикла (от 15 до 20 лет) для таких приложений, как резервное копирование и архивирование;*
- *<u>наиболее рентабельный пул ресурсов облачного хранилища:</u> для бОльшей части хранилища поддерживается помехоустойчивое кодирование (erasure code, EC). По сравнению с режимом трех копий коэффициент использования пространства улучшен как минимум на 30%. Производительность и емкость на единицу объема улучшены, что обеспечивает высокую экономическую эффективность. Кроме того, система поддерживает 10 миллиардов объектов для удовлетворения требований по чтению и записи в одном сегменте новых приложений/служб клиентов;*
- *<u>безопасность и надежность данных:</u> система развернута в режиме полного резервирования, гарантируя, что сервисы не будут прерываться при сбое нескольких узлов или шкафов. Система может выдерживать сбои максимум четырех шкафов без прерывания обслуживания и автоматически обнаруживает неисправности, запуская восстановление данных со скоростью до 2 ТБ/час. Функция WORM предотвращает фальсификацию данных и обеспечивает их безопасность и точность;*
- *<u>преимущества EC</u>:*
  - *расширенная гибкость: поддерживаются схемы: +2, +3, +4, +2: 1, +3: 1 и +4: 1. Большинство поставщиков поддерживают только схемы +1 и +2. Таким образом, использование FusionStorage в EC является самым высоким, которое может достигать 80% и более, в то время как продукты других поставщиков обеспечивают только 66%;*
  - *зрелые решения: рекомендуются для коммерческого использования;*
  - *высокая производительность: TPS и пропускная способность – одни из самых высоких для данного класса хранилищ;*
- *<u>уменьшенные санкционные риски</u> за счет возможности реализации аппаратного уровня FusionStorage на серверах Huawei, работающих на ARM-процессорах собственной разработки и производства.*

***Денис Дубинин**, компания Huawei.*

Микросхемы ускорения вычислений нейросетей

Интервью с Юрием Панчулом – старшим инженером по проектированию интегральных схем для ускорения вычислений нейросетей стартапа Кремниевой Долины компании Wave Computing.

Юрий Панчул – старший инженер по проектированию интегральных схем, Wave Computing.

SN. Чем занимается ваша компания и с чем вы едете на форум Skolkovo Robotics 2019?

Ю.П. На форуме в «Сколково» я презентую наш проект Triton, который представляет собой комбинацию трех типов вычислительных устройств для ускорения вычислений нейросетей:

- *первый* – кластер из классических процессоров общего назначения;
- *второй* – процессор потоков данных (dataflow processor) на основании архитектуры крупнозернистого реконфигурируемого массива CGRA (Coarse Grained Reconfigurable Array);
- *третий* – матричный умножитель на основе систолического массива из умножителей-сумматоров (multiply-add). Эти устройства представляют собой три разных способа организации вычислений с разным балансом гибкости и производительности.

Если говорить о них подробнее, то классические процессоры – самые гибкие. Они могут вычислить все множество нейросетей, определяемых стандартами типа TensorFlow и ONNX.

Процессоры потоков данных могут вычислять в 10 раз быстрее, чем кластеры классических процессоров, но накладывают ограничения на типы узлов нейросети. При этом они могут вычислять многое из того, что не могут вычислять матричные ускорители, например, необычные активационные функции (activation function).

Матричные умножители ориентированы на вычисления узкого подмножества и ориентированы на сверточные сети (CNN – Convolutional neural network). Зато они вычисляют по порядку в 10 раз быстрее, чем процессоры потоков данных, и в 100 раз быстрее, чем кластеры из процессоров общего назначения.

Если мерить по плотности производительности (по количеству операций, которые можно выполнить на структуре размеров в один квадратный миллиметр микросхемы), то процессоры потоков данных на порядок больше по вычислительной плотности, чем классические. А матричные – на порядок больше по вычислительной плотности, чем процессоры потоков данных.

Сейчас мы работаем над платформой, где используем варианты комбинирования. Данная платформа, по сути, – блоки, на которые мы продаем лицензии компаниям, производящим чипы. В эти блоки входит *кластер из классических процессоров общего назначения* – MIPS AI Processor, он же – процессорный кластер MIPS I6500 с кодовым названием Daimyo, который состоит из процессоров MIPS I6400 Samurai (значение имени Daimyo – это “начальник над самураями”). У процессора есть векторные расширения – команды, которые могут сразу выполнять операции не над парой чисел, а над двумя группами чисел (векторами). С помощью векторных расширений алгоритмы вычислений нейросетей лучше оптимизируются. Каждый процессор является 64-битным, суперскалярным, с аппаратной поддержкой многопоточных вычислений (SMT – simultaneous multithreading), что повышает его пропускную способность.

Также в этот пакет входит *процессор потоков данных (dataflow processor)* на основании архитектуры крупнозернистого реконфигурируемого массива CGRA (Coarse Grained Reconfigurable Array), он называется WaveFlow. Состоит он из переменного массива процессорообразных элементов (от нескольких до десятков тысяч), которые соединены между собой сетью из переключателей на одном большом кристалле с парой миллиардов транзисторов.

Третья составляющая – матричный умножитель на основе систолического массива из умножителей-сумматоров (multiply-add) – называется WaveTensor и похож по организации на Google TPU.

Вся эта комбинация продается как semiconductor IP – код конфигурируемых блоков на языке описания аппаратуры Verilog. Пакет Triton у нас покупают производители конечных микросхем, и с помощью технологии логического синтеза превращают его в файл в формате GDSII,

по которому на фабрике изготавливают микросхемы.

SN. Расскажите об областях применений ваших разработок.

Ю.П. Главная область приложений, на которую ориентируется наша компания, – видеоаналитика. Например, камеры в шоппинг-центрах, мимо которых проходят люди. Встроенный компьютер распознает их настроение, возраст, пол и потом показывает им подходящую рекламу или направляет к ним продавца.

Другая область применений – это автомобильная электроника. Процессоры MIPS I6500 сертифицированы для автомобильной индустрии и лицензированы, в частности, компанией DENSO, производителем электроники для Toyota.

Нейроускоритель может масштабироваться от датацентров до локальных устройств – от умных сенсоров до автомобилей, камер и мобильных устройств. Помимо этого нейроускорителя из трех частей, наша компания лицензирует разнообразные процессорные ядра, на основе которых можно делать не только распознавание нейросети, но и другие компоненты для роботов, в частности, небольшие процессоры для актуаторов (которые, например, запускают электромоторы, двигающие руки робота) и главного контролирующего компьютера робота. У нас есть совместный проект с компанией NediaTek на процессор для 5G мобильных сетей.

Робот состоит из нескольких вычислительных устройств, в центре робота находится процессор общего назначения, на котором может работать, к примеру, встроенный Linux, и рядом с этим процессором может стоять специализированное устройство для распознавания образов. Например, робот может видеть разные объекты, распознавать их, после чего главный процессор может на основе этого отдавать команды более малым процессорам, которые называются актуаторами. Это процессоры, которые запускают двигатели в манипуляторах робота или отдают команды разным устройствам.

Для всех трех главных частей робота – главного процессора, распознающих образы чипов и актуаторов – у нас есть решения. И мы выдаем лицензии различным компаниям на процессорные ядра, которые могут использоваться во всех видах этих чипов. Мы лицензируем данные решения для компаний, которые делают чипы для машин, для роботов, и они уже делают чипы на основе наших ядер.

В том числе к нашим лицензиатам сейчас относятся японская компания Denso, которая делает автомобильную электронику для Toyota, а также компания Mobileye, которая сейчас стала частью Intel. Они используют наши процессорные ядра, стоявшие в устройствах Mobileye, Volvo и других машинах.

В число наших типичных клиентов входят компании, делающие рекламу, системы распознавания ситуаций, а также нейросети для распознавания поведения людей в соцсетях или на web-сайтах. Раньше эти устройства мы делали в качестве прототипа, а теперь с отдельными компаниями мы начинаем делать разные варианты комбинаций из трех типов вычислителей. Но мы можем работать и в других областях. Одно из наших перспективных направлений – самоуправляемые автомобили, так как 80% современных устройств для помощи водителю во время вождения, так называемые ADAS (advanced driver assistance system), используют наши процессорные ядра.

Кроме того, на конференции Skolkovo Robotics мы анонсировали новую инициативу, зародившуюся только недавно. Исторически та часть нашей компании, которая вышла из Стэнфорда, – MIPS и которая была частью Silicon Graphics, частью Imagination Technologies, жила за счет лицензирования процессорных ядер и лицензирования архитектуры. То есть за счет продажи другим компаниям права делать собственные процессоры, совместимые по программному обеспечению с процессорами нашей компании. Это то же самое, чем была занята компания ARM, которая является в настоящее время лидером встроенных процессоров. Но в последнее время возникло движение RISC/V – это открытая архитектура. И мы хотим двигаться в этом направлении. Мы решили открыть нашу архитектуру, чтобы и другие компании могли делать решения на ее основе, без уплаты нам лицензионных отчислений. Мы считаем, что это расширит экосистему наших процессоров.

Архитектура процессора – это то, как видит процессор программист. В нее входит система команд и видимые программисту регистры. Микроархитектура – это организация процессора с точки зрения электронного инженера – стадии конвейера и вычислительные блоки. В России есть компании, в частности, ЭЛВИС и НИИСИ, имеющие процессоры с собственной микроархитектурой, совместимой с архитектурой MIPS и, следовательно, совместимым со всем написанным для нее в мире программным обеспечением. Мы считаем, что открытие архитектуры MIPS дает возможность этим и другим компаниям лучше маркетировать свои процессоры. В целом все это обогащает нашу с ними общую экосистему.

SN. Какие еще бывают типы ускорителей нейросетей и в чем их отличия?

Ю.П. Можно сказать, что существует пять типов ускорителей: классические нейропроцессоры, матричные умножители, ускорители на основе графических процессоров, dataflow-процессоры на архитектуре потоков данных, и есть разные специальные решения, например, матрицы малых классических процессоров.

Первый тип ускорителей нейросетей – это просто использование классических процессоров с так называемыми векторными расширениями. В этом случае ставятся несколько обычных процессоров в процессорный кластер, и каждый из процессоров может выполнять не только простые команды, типа сложения пары чисел, но и делать операции с векторами, с массивами из чисел. Также в обычные процессоры входит многопоточность. Но все это – виды оптимизации обычных процессоров.

Второй тип – матричный ускоритель. Это то, что сейчас делает, в частности, Google. Данное специальное устройство выглядит как массив из блоков, которые умеют умножать и производить сложение. Этот массив выстраивается таким образом, чтобы была возможность очень быстро умножать матрицу. Матричное умножение – это хорошо оптимизируемая операция, и 90 с лишним процентов вычислений в нейросети происходит с использованием матричного умножения. Но, к сожалению, кроме матричных умножений есть и другие операции, поэтому приходится использовать и другие типы ускорителей.

Третий тип – ускорители на основе так называемых графических процессоров, главным производителем которых является NVIDIA. Графические процессоры возникли в 90-е годы для ускорения игр. Когда в 90-е годы стали возникать трехмерные шутеры, для качественной картинки стали строить графические процессоры. Чем он, собственно, отличается от обычного процессора? В обычном процессоре делается выборка вычисления каждой команды или инструкции, а в графическом процессоре делается выборка команды или инструкции, которая используется во многих потоках. То есть одна команда распадается на несколько – именно таким способом вычисляются картинки на экране, с помощью так называемых шейдеров.

Для вычисления цветов пикселей на экране можно использовать такую структуру, когда одна инструкция используется в большом количестве арифметических устройств для вычисления находящихся рядом частей картинки. Впоследствии оказалось, что данные структуры можно использовать и для вычислений другого типа – математических суперкомпьютерных вычислений. Также их удобно использовать для вычисления нейросети более гибким способом, чем это делается с помощью матричных ускорителей.

К безусловным преимуществам графических процессоров относится и то, что в компании NVIDIA (лидер в этой сфере) работает много инженеров. В связи с этим у компании много ресурсов, чтобы сделать разные оптимизации этих устройств для разных вычислений и, в частности, для вычисления нейросети. В этом они очень сильны. Но они не особенно сильны в лицензировании своих технологий в форме IP-блоков. Это уже наш бизнес. Wave (а точнее поглощённый ею MIPS) занимается разработкой и продажей процессоров в виде IP-блоков уже 20 лет, с 1999 года. IP-блок, или IP-ядро (блок интеллектуальной собственности – intellectual property) – это схема в представлении на языке описания аппаратуры Verilog, которую можно встраивать в различные чипы других компаний. В то же время NVIDIA занялась продажей своих технологий в виде IP-блоков только в последнее время. Пример – открытое ядро NVDLA. До этого NVidia использовала свои блоки только в составе собственных же чипов.

Четвертый тип ускорителей нейросетей – наше устройство – dataflow-процессор «на основе крупнозернистого реконфигурируемого массива (Coarse Graoned Reconfigurable Array – CGRA). Это архитектура потока данных, т.е. процессор, который состоит из большого количества

процессорообразных элементов. Причем, наше преимущество в том, что данный процессор может находиться и в маленьких устройствах (как всего лишь несколько таких элементов), и в больших устройствах (как десятки тысяч таких элементов). Это называется scalability – масштабирование.

Dataflow-процессор вытягивает из памяти целый «тензор» (матрицу данных) через сеть переключателей, рассылающих эти данные между кластерами для обработки. В каждом кластере находится специальная небольшая программа, которая получает данные нейросети извне и делает с ними различные операции – не только умножение со сложением, как это происходит в матричных умножителях, но и более сложные операции.

Dataflow-процессор по вычислительной мощности и гибкости находится между классическим процессором и матричным умножителем и при этом конкурирует с графическим процессором. Хотя графические процессоры тоже находятся между классическим процессорами и матричными умножителями, но графические процессоры изначально были созданы для оптимизации графики, поэтому они не настолько хорошо подходят для типов вычислений, которые делаются для нейросетей.

Кроме этих четырех типов ускорителей есть еще и *особые решения*, например, стартап под названием Esperanto, делающий матрицу из большого количества малых классических процессоров. Это решение тоже имеет смысл, но оно требует большей площади чипа, больше энергии, чем dataflow-процессор.

SN. Расскажите об интеграции и масштабировании ускорителей нейросетей в составе решений.

Ю.П. Масштабирование – это сила нашей компании. То, что мы сейчас делаем, это именно IP-блоки, которые могут масштабироваться. Этот блок вместе с процессором может стоять как на миниатюрном чипе, который может сочетаться с каким-нибудь сенсором (и тогда можно распознавать с очень низким энергопотреблением), но может быть представлен и в виде большого устройства (например, в сервере для датацентра). Несколько вариантов конфигурации процессоров позволяют создавать решения с разными параметрами площади, энергопотребления и производительности, что очень важно.

SN. Можете подробнее рассказать о рынке ускорителей нейросетей?

Ю.П. Главный игрок на рынке ускорителей нейросетей – NVidia с особым доминированием на рынке training. Затем отъел немаленький кусок рынка Google с матричным умножителем на основе систолического массива. Далее есть решения как от крупных игроков (Huawei, AMD, Xilinx), так и от стартапов (Wave, Graphcore, Habana), но уровня готовности NVidia и Google они еще не достигли. Рынок находится в процессе формирования и фрагментации.

Вообще существует большое количество стартапов, чьи проекты находятся в процессе разработки. У них нет уже выпущенных продуктов, которые можно использовать прямо сейчас. Один из таких стартапов находится намного ближе к решению, чем многие другие компании. К примеру, Graphcore, которая сделала ускоритель, отличающийся тем, что на одном чипе между процессорными элементами установлена память. Из-за этого доступ к памяти чипа становится гораздо быстрее, чем у других. Есть еще компания Habana, которая тоже находится на слуху, но ее решение еще не присутствует на рынке.

Сейчас рынок также представлен текущими решениями на основе графических процессоров, они сильны в области тренинга. Есть несколько стартапов и несколько аналогичных компаний в США, в Китае. Эти компании пробуют внедрить элементы нейросетей в традиционные продукты.

SN. Какова эффективность разных типов ускорителей нейросетей при тренировке моделей?

Ю.П. Это довольно интересный вопрос. Прежде следует сказать, чем отличаются алгоритмы тренировок от вывода (inference). Для тренировки нужна более сложная арифметика. В частности, нередко для тренировки используют арифметику с плавающей точкой (как стандартную, так и ее вариации – bfloat16, например). А для вывода применяют обычную целочисленную арифметику или арифметику с фиксированной точкой. Но разные типы ускорителей используют разные типы данных. Причем, разные ускорители применяют чипы с плавающей точкой и многочисленные чипы с фиксированной точкой.

Сейчас также рассматривается использование новых типов данных (например, bfloat16). У нас есть в компании специалист, который занят именно этим делом. Для того чтобы повысить эффективность тренировки, можно использовать разные необычные форматы чисел с плавающей точкой.

Сейчас в тренировке, конечно, лидером является NVIDIA. Мы находимся в процессе создания IP для тренировки. Дело в том, что традиционно считалось: тренировку нужно делать фиксировано на месте – в облаке, на большом мегапикселе, потому что она происходит долго. В настоящее время ставится другая задача – делать тренировку моделей локально.

Например, когда вы используете AI, действия тренировки должны приспособиться к конкретным действиям. И в таком случае лучше делать тренинг локально, а не отсылать в облако или другое место. Для локального тренинга возникает необходимость в устройствах, которые стоят не только в облаке, но и, например, в мобильных телефонах для распознавания в контексте его владельца. Эти типы лучше приспособлены.

Сейчас наша компания разрабатывает IP-коды (IP-cores, IP-ядра), которые можно использовать как для обычной работы нейросети, так и для локального процесса тренировки нейросети. Возникает необходимость использования отдельных приложений или приложений, встроенных в мобильные телефоны.

SN. Спасибо!

Платформа HPE Edgeline EL8000 Converged Edge System

Февраль 2019 г. — Компания Hewlett Packard Enterprise (HPE) анонсировала выпуск новой платформы HPE Edgeline EL8000 Converged Edge System, чтобы помочь поставщикам услуг связи извлекать выгоду из информационно-ёмких сервисов с малым значением задержки данных для доставки мультимедиа-продуктов, интернет-технологий для мобильности и умных городов. Новая система позволяет поставщикам услуг связи на основе открытых стандартов обрабатывать огромные объемы данных в режиме реального времени непосредственно на границе сети (конечных устройствах), чтобы повысить гибкость и снизить затраты. HPE также объявила о сотрудничестве с Samsung и Tech Mahindra для ускорения внедрения технологий 5G при использовании системы HPE Edgeline EL8000 для развертывания нового поколения виртуальных 5G-приложений, ориентированных на граничные вычисления.

По прогнозам IDC к 2025 году по всему миру к интернету будет подключено более 150 млрд устройств, большинство из которых будут создавать данные в режиме реального времени. К 2025 году, по прогнозам IDC, такие данные, будут составлять почти 30% от общего количества во всемирной сети). По этим же данным, в 2018 году объем глобальных данных составил 33 зеттабайта, а к 2025 году он вырастет до 175 зеттабайтов *(исследование IDC: "The Digitization of the World — From Edge to Core" – «Глобальная цифровизация: от периферии к центру»)*.

Чтобы помочь поставщикам услуг связи ускорить переход на 5G, HPE и Samsung Electronics Corporation (SEC) разрабатывают совместное решение vRAN, работающее поверх всей инфраструктуры, от ядра сети до ее периферии. Решение основано на технологиях беспроводных сетей и службах системной интеграции Samsung, а также на платформе HPE Edgeline EL8000 Converged Edge System.

В то время как некоторые поставщики услуг связи стремятся развернуть сети связи 5G с 2020 года, другие, возможно, не смогут сделать это в течение нескольких лет, потенциально оставляя целые регионы без покрытия сетей 5G. В течение этого переходного периода телеком-операторы будут использовать программное обеспечение мобильных граничных вычислений с множественным доступом (MEC), которое обеспечивает многие преимущества 5G, но с использованием инфраструктуры 4G LTE. По этой причине HPE начинает сотрудничать с Tech Mahindra, мировым лидером в области программного обеспечения для MEC, предоставляющему решения MEC на основе новой платформы HPE Edgeline EL8000 Converged Edge System.

Для предоставления новых услуг, подразумевающих значительный рост собран-