

# ЭВОЛЮЦИЯ АРХИТЕКТУРЫ SmartNIC: ПЕРЕХОД НА УСКОРИТЕЛИ И ВОЗРАСТАНИЕ РОЛИ ПЛИС

СКОТТ ШВЕЙЦЕР (SCOTT SCHWEITZER), Technology Evangelist, Xilinx

*Помимо типовой сетевой карты, в архитектуры SmartNIC входят дополнительные вычислительные ресурсы. Поскольку эти архитектуры умных сетевых карт отличаются друг от друга, мы подробно рассмотрим несколько из них, реализованных крупнейшими и самыми известными поставщиками.*

Типовая сетевая интерфейсная карта (network interface card, NIC) построена на одной микросхеме ASIC, которая функционирует как Ethernet-контроллер. В качестве наглядных примеров можно привести линейку адаптеров ConnectX от компании Mellanox, NetXtreme от Broadcom или XtremeScale от Xilinx. Часто эти контроллеры в дальнейшем оптимизируются. Так, например, контроллеры семейства ConnectX поддерживают оборудование InfiniBand, а контроллеры XtremeScale позволяют использовать адаптеры с технологией kernel bypass, когда драйвер сетевой карты работает напрямую с приложением в обход ядра операционной системы. Эти контроллеры обладают превосходными функциональными возможностями и являются лучшими в отрасли, но они не относятся к категории SmartNIC.

Уточним, что мы рассматриваем SmartNIC как сетевую интерфейсную карту (NIC), в которую можно загрузить дополнительное программное обеспечение после ее покупки, чтобы добавить новые функции или осуществить поддержку требуемых функций. Такая возможность сродни той, что имеется у пользователей новых смартфонов, на которые после их приобретения устанавливаются разные приложения.

Возможность загружать код в NIC-карту после ее приобретения позволяет назвать это устройство интеллектуальной картой SmartNIC. Она требует дополнительных вычислительных мощностей и встроенной памяти, которые обычно отсутствуют у типовых NIC-карт. Большинство реализаций SmartNIC начинается с базового контроллера Ethernet либо на кремниевом кристалле с установленной прошивкой, либо с отдельной микросхемы на адаптере.

Затем используется один из трех следующих подходов, чтобы сделать типо-

вую NIC-карту умной за счет увеличения ее вычислительной мощности путем добавления:

- кластера ядер Arm;
- специализированных сетевых процессоров (flow processing cores, FPC), например P4 (Pentium 4);
- ПЛИС.

Многие из карт SmartNIC часто используют одно или несколько ядер Arm для управления внутри NIC. Некоторые даже позволяют загружать модифицированное ядро Linux в одно или несколько из этих ядер. Arm-ядра обычно осуществляют загрузку кода в другие процессорные элементы, сбор статистики и запись в журнал, а также наблюдение за состоянием и конфигурацией SmartNIC-карт. Они не работают с сетевыми пакетами и часто являются внеполосными. Это значит, что к ним нельзя получить доступ через «обычные» сетевые интерфейсы или команды PCIe. Кроме того, эти ядра должны принимать только правильно подписанные

пакеты прошивки через уже защищенные интерфейсы. Однако эти Arm-ядра сами по себе, как правило, не увеличивают ценности набора функций, обеспечиваемых SmartNIC-картой.

## СРАВНЕНИЕ СЕТЕВЫХ КАРТ РАЗНЫХ ПРОИЗВОДИТЕЛЕЙ

Чтобы понять, чем карты SmartNIC отличаются от обычных сетевых карт, давайте познакомимся с ведущими изделиями SmartNIC от четырех крупнейших производителей сетевых карт и двух новичков. Мы выбрали шесть компаний: Broadcom, Intel, Nvidia (ранее Mellanox), Netronome, Pensando и Xilinx. Кроме того, мы рассмотрим проект стартапа Fungible.

### Broadcom

Broadcom – безусловный лидер на рынке Ethernet-контроллеров сетевых карт. Эта компания выбрала однокристальный подход для реализации SmartNIC-карты Stingray (см. рис. 1). Сто-

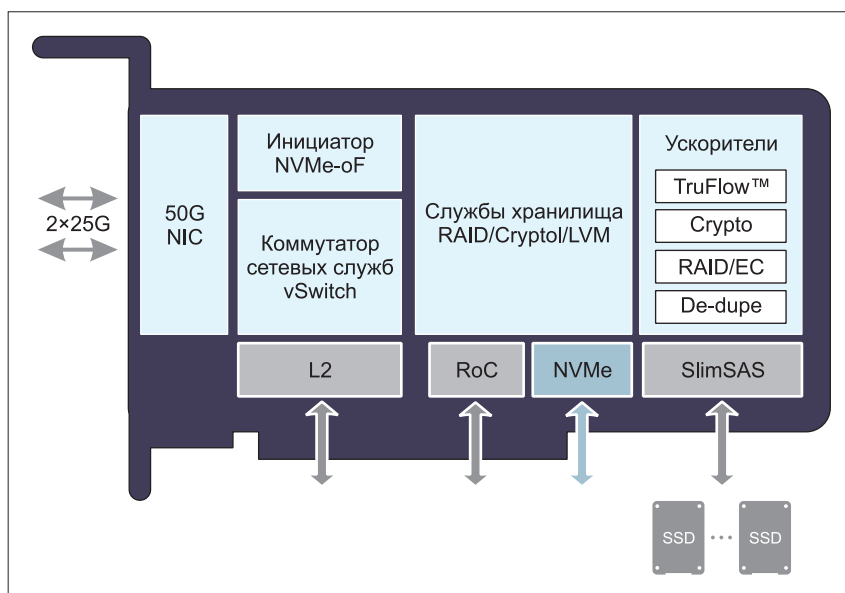


Рис. 1. Архитектура карты Stingray из презентации SDC 2019 компании Broadcom

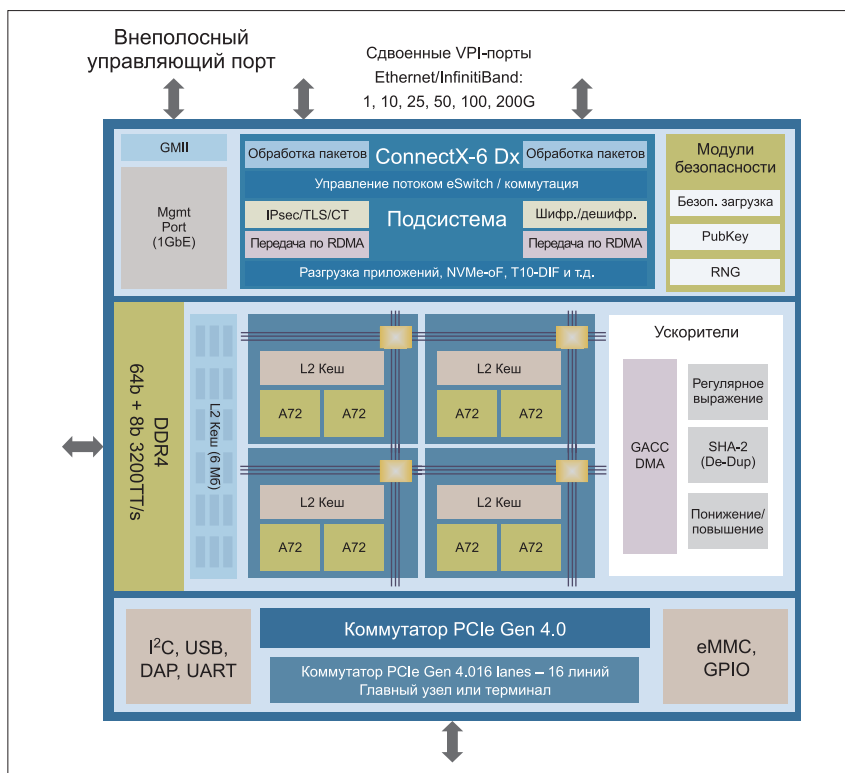


Рис. 2. Архитектура платы Bluefield 2 от компании Mellanox

имость производства однокристалльных решений SmartNIC на уровне платы всегда ниже стоимости изготовления нескольких плат с микросхемами, как это делают некоторые другие конкуренты.

Компания Broadcom разработала карту Stingray на основе ИС контроллера NetXtreme-S серии BCM58800. Восемь ядер Arm v8 A72 с тактовой частотой 3 ГГц используются в кластерной конфигурации. Возможно, на частоте 3 ГГц это самые быстрые ядра. Кроме того, в карте Stingray можно установить до 16 Гбайт памяти DDR4. Кроме того, в эту карту были добавлены логические функции, чтобы разгрузить шифрование на скорости до 90 Гбит/с и обработку данных хранилища, например осуществить удаляющее кодирование (erasure coding) и облегчить работу избыточного массива независимых жестких дисков (RAID).

Наконец, компания Broadcom задействовала технологию TruFlow, которая с помощью настраиваемого ускорителя переносит стандартные сетевые поточные процессы в оборудование. Скорее всего, с этой целью компания использует процессор P4. Такой подход позволяет ядрам Arm заниматься ресурсоемкими задачами не только на уровне потока, но и пакетов. Судя по опубликованным материалам, TruFlow аппаратно облегчает решение таких задач как Open vSwitch (OvS).

Компания также утверждает, что технология TruFlow реализует в оборудовании многие из классических концепций

программно-определяемых сетей (SDN): классификацию, сопоставление и макрокоманды. Сетевая карта Stingray имеет два программируемых компонента: TruFlow и кластер из четырех 3-ГГц двухъядерных комплексов Arm v8 A72. Broadcom готовится перевести Stingray на 7-нм техпроцесс, чтобы увеличить число ядер с восьми до 12. Для проектирования приложений SmartNIC и контроллеров хранения компания предлагает комплект разработчика Stingray.

#### Nvidia

Nvidia определила графические процессоры (GPU), которые нашли широкое применение в качестве ускорителей для высокопроизводительных вычислений (HPC). В начале 2020 г. Nvidia окончательно закрыла сделку по покупке за 7 млрд долл. Mellanox – производителя телекоммуникационного оборудования, ключевого разработчика технологии InfiniBand. Стремится завладеть рынком высокопроизводительных вычислений, компания выбрала поставщика межсоединений InfiniBand, чтобы предложить полное решение для рынка HPC. Такая стратегия очень схожа с той, которой придерживался в прошлом производитель суперкомпьютеров Cray.

Nvidia также недавно приобрела Cumulus Networks – лидера в области операционных систем (ОС) для Ethernet-коммутаторов с открытым исходным кодом. Программное обеспечение всегда было слабым местом компании

Mellanox, и Nvidia, очевидно, учла это обстоятельство. Mellanox – один из старейших участников рынка SmartNIC, но таковым он стал за счет приобретения: плата Bluefield 2 (см. рис. 2) появилась после покупки Tiler в 2016 г. компанией Mellanox через израильскую полупроводниковую компанию EZchip. Tiler одной из первых создала высокопараллельную реализацию SmartNIC с использованием собственных СФ-блоков на основе давнишнего исследовательского проекта MIT. Плату Bluefield 2 в компании именуют модулем обработки данных (Data Processing Unit, DPU).

Составной единицей процессора Tiler является ячейка, содержащая полноценный процессор с кэшами 1- и 2-го уровней и неблокируемый коммутатор, который соединяет процессор (вычислительное ядро) с сетью. Ядра могут работать и под управлением отдельной операционной системы, и в группах под многопроцессорной ОС типа SMP Linux.

Флагманское изделие Tiler еще в 2013 г. поддерживало до 72 ядер MIPS, контроллеры памяти, модули шифрования, блоки PCIe и mPipe, набор каналов для нескольких MAC с разъемами SFP+. Компания Mellanox усовершенствовала этот процессор, перейдя на ядра Arm и заменив mPipe логикой ConnectX.

На плате установлены восемь ядер Arm v8 A72, как и в случае с Broadcom, но с тактовой частотой только 2,4 ГГц. Эти ядра составлены в виде кластера из четырех пар Arm-ядер. SmartNIC-карты Bluefield реализованы по 16-нм техпроцессу Avago, но, как и Broadcom, Tiler намеревается перейти на 7-нм процесс и использовать 12 ядер вместо восьми. На плате установлены также контроллер памяти DDR4, двухпортовый сетевой Ethernet-адаптер или InfiniBand (два на 100 Гбит/с или один на 200 Гбит/с), а также специализированные ASIC-блоки для ускорения функций регулярных выражений, хэширования SHA-2 и т.д.

Хотя такое решение очень похоже на Stingray от Broadcom, в нем отсутствует параллельный процессор P4, составляющий основу архитектуры Broadcom. Не секрет, что большинство компаний намеревается использовать этот процессор в своих архитектурах, как это сделали Broadcom, Xilinx и Pensando. В создании механизма обработки пакетов с помощью P4 во внешнем интерфейсе карты Bluefield компания Nvidia сможет воспользоваться опытом программирования Cumulus Networks на P4.

#### Pensando

Одним из самых новых стартапов в сегменте SmartNIC является компа-

ния Pensando, которая была основана группой инженеров, инициировавшей разработку нескольких ключевых технологий Cisco System и создание четырех компаний, позже приобретенных Cisco.

Учитывая репутацию группы учредителей и председателя совета директоров, а также их предыдущий опыт, ожидается, что компанию Pensando приобретет Cisco. Cisco располагает стандартной технологией сетевых карт и несколькими собственными проектами SmartNIC, но ходят слухи, что они не увидят свет, и Pensando явно стремится исполнить этот продел.

Сначала у Pensando были две сетевые карты, но она оставила одну – DSC-25 Distributed Services, название которой схоже с названием карт Cisco. Карта DSC-25 располагает одним 4-Гбайт DPU-процессором P4 для обработки данных, дополнительными 4-Гбайт Arm-ядрами и аппаратными ускорителями отдельных функций (см. рис. 3).

Процессор под названием Capri представляет собой программируемый блок P4 с несколькими параллельными каскадами. Компания не указала точный объем параллельной обработки, скорость обработки пакета, значения задержки и джиттера. По замыслу Pensando, данные приложений P4 должны оставаться в кэше Capri в случае промаха кэша. Таким образом, для подачи команды требуется выборка из памяти, что снижает производительность по всем показателям. Другие дополнительные вычислительные блоки, к которым относятся Service Processing Offloads, принимают участие в шифровании, дисковых операциях и выполнении других задач. Компания Pensando утверждает, что Capri обеспечит скорость передачи, соответствующую максимальной пропускной способности канала.

### Netronome

Netronome – седебородый стартап в этом сегменте рынка. Объем финансирования компании, начавшей свою деятельность в 2003 г., к настоящему времени составил 73 млн долл. Компания активно продвигает процессоры P4 с 2015 г., когда она представила первые SmartNIC-карты. Хотя Netronome добилась значительных успехов, в последнее время ходят слухи, что она может уйти с рынка.

На рисунке 4 показана архитектура специализированного сетевого процессора NFP4000 от Netronome. Вместо одного процессора P4 компания использует программируемые ядра двух классов: 48 ядер для обработки пакетов и 60 ядер – для обработки потока. Дополнительная микросхема предназначена для классификации, модифика-

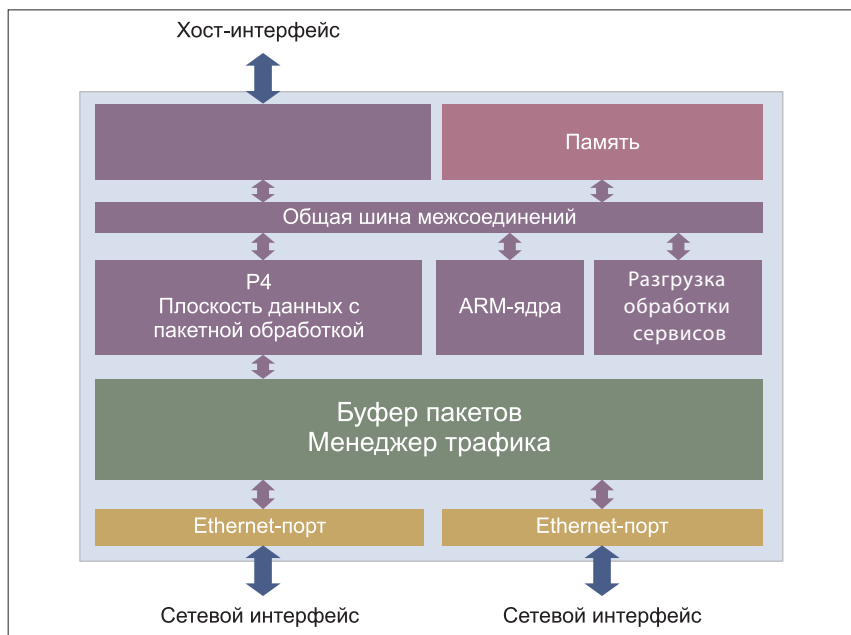


Рис. 3. Архитектура DPU-модуля от Pensando

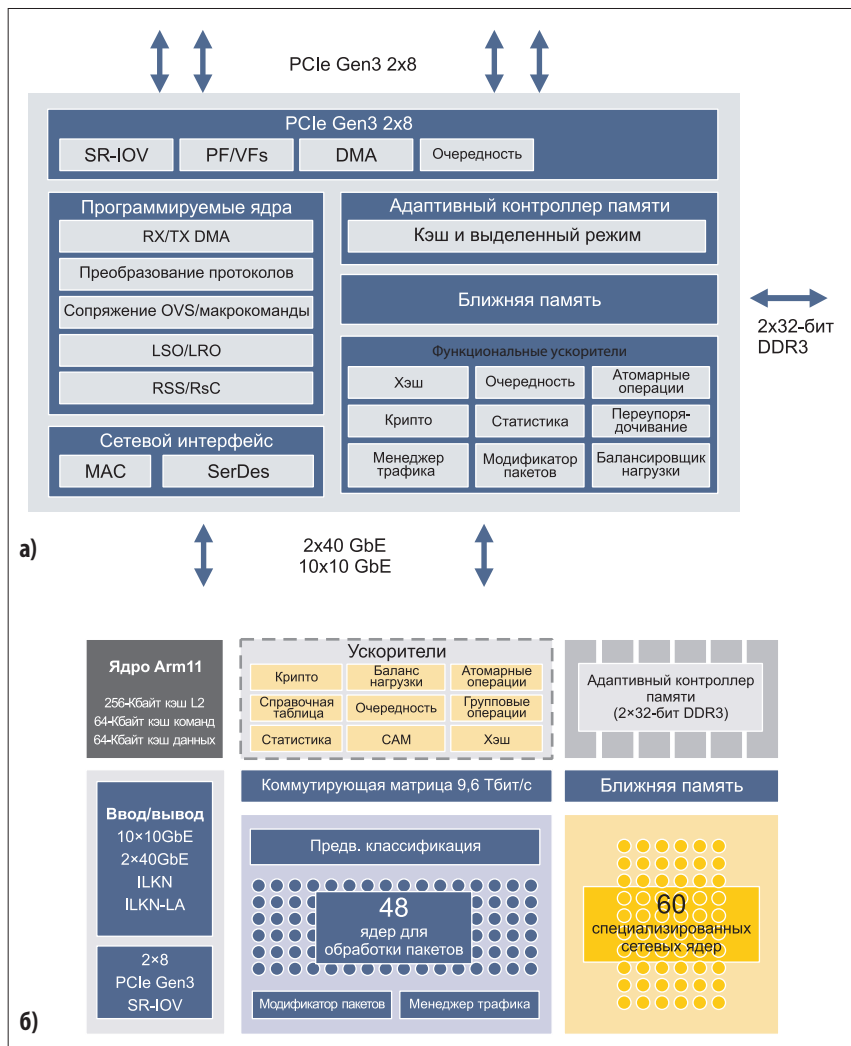


Рис. 4. Архитектура специализированного сетевого процессора NFP4000 от Netronome

ции и управления. Все эти ядра можно запрограммировать на P4.

Компания Netronome утверждает, что эти ядра могут поддерживать один 100-Гбит/с канал с максимальной

пропускной способностью 148 млн пакетов в секунду; при этом обеспечиваются миллионы точных совпадений и потоков с произвольными символами. Кроме того, ядра поддер-

## Вычисления, ориентированные на данные

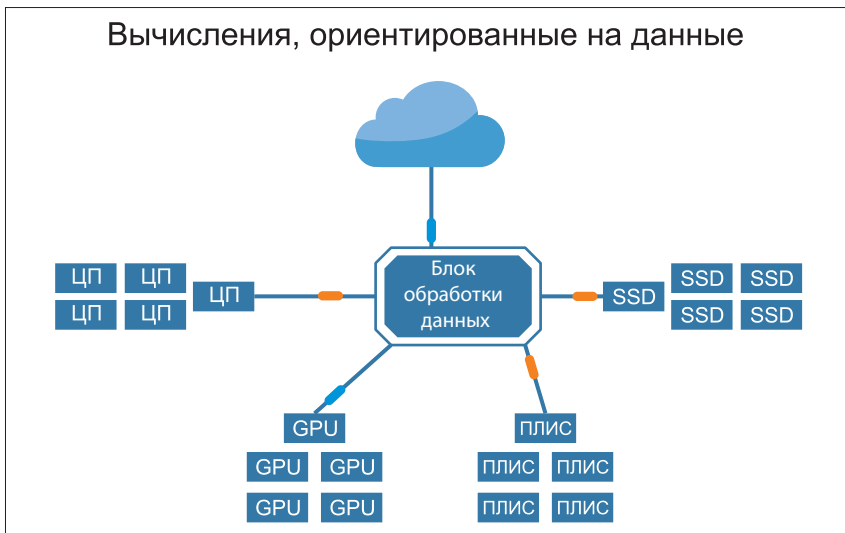


Рис. 5. Гибкая архитектура высокого уровня

живают более 100 тыс. сетевых соединений с туннелированием, требующих инкапсуляции. В очень длинный список приложений, которые поддерживает эта технология, входят системы обнаружения вторжений (IDS), системы предотвращения вторжений (IPS), межсетевой экран нового поколения (NGFW), маршрутизаторы, балансировщики нагрузки, брокеры сетевых пакетов, SDN, NFV и целый ряд других приложений.

### Fungible

К кругу рассматриваемых компаний принадлежит и стартап Fungible. В ближайшее время Fungible намеревается анонсировать свою продукцию после трех раундов финансирования на общую сумму почти 300 млн долл., из которых 200 млн были получены от Softbank Vision в прошлом году. В настоящее время в компании работают 180 человек; пока она не выпускает продукцию, не имеет доходов и известных клиентов.

Fungible утверждает, что производит блоки обработки данных (data processing unit, DPU), но фактическая архитектура ее устройств неизвестна. Пока Fungible ограничивается невразумительными демонстрациями общей схемы (см. рис. 5). Некоторые специалисты подозревают, что Fungible использует популярный термин DPU для привлечения венчурных инвестиций. Ближайшее будущее покажет, насколько эти подозрения оправданы.

Один из основателей и главный архитектор Fungible отработал 10 лет в частной технологической компании Chelsio Communications, специализирующейся на сетевых Ethernet-картах для устройств хранения. Вице-президент по разработке программного обеспечения и микропрограмм также является

бывшим сотрудником Chelsio, в которой он проработал 13 лет. Если бы Chelsio использовала архитектуру на основе ASIC, для производства новых Ethernet-контроллеров SmartNIC с нуля в настоящее время потребовалось бы не менее 50 млн долл.

Чтобы получить прибыль, компания Fungible, скорее всего, пойдет по пути наименьшего сопротивления и воспользуется для реализации своего исходного изделия платформой ПЛИС, в которую загрузит конфигурацию ASIC. В дальнейшем компания, вероятно, изменит архитектуру, чтобы привлечь как можно больше клиентов. Такая стратегия позволит

легко исправить недостатки решения и скорректировать его в соответствии с потребностями клиентов. К настоящему времени ПЛИС содержат огромное количество вентиляльных элементов, и на рынке появились полнофункциональные процессорные архитектуры, например RISC-V, которые загружаются на платформы ПЛИС.

### Intel

Более 10 лет Intel поддерживает линейку высокопроизводительных контроллеров 10-GbE. Объем поставок ее платформы XL710 исчисляется миллионами единиц, и она является основным компонентом многих серверов центров обработки данных. Для новой SmartNIC-карты N3000 корпорация Intel создала плату с пятью своими микросхемами (см. рис. 6). Это дорогостоящий подход, т. к. большинство поставщиков стремится к однокристалльной конструкции. Два своих контроллера Ethernet XL710 и ПЛИС Arria 10 компании Intel используют PEX8747-48-полосный 3-позиционный кристалл переключателя PCIe-поколения. По восемь полос используются контроллерами XL710, 16 полос – ПЛИС Arria, а 16 полос предназначены для разъема PCIe. Пятая микросхема представляет собой контроллер для управления базовой платой (BMC) ПЛИС MAX 10, что во многом похоже на то, как ядра Arm используются в других SmartNIC для работы с уровнем управления.

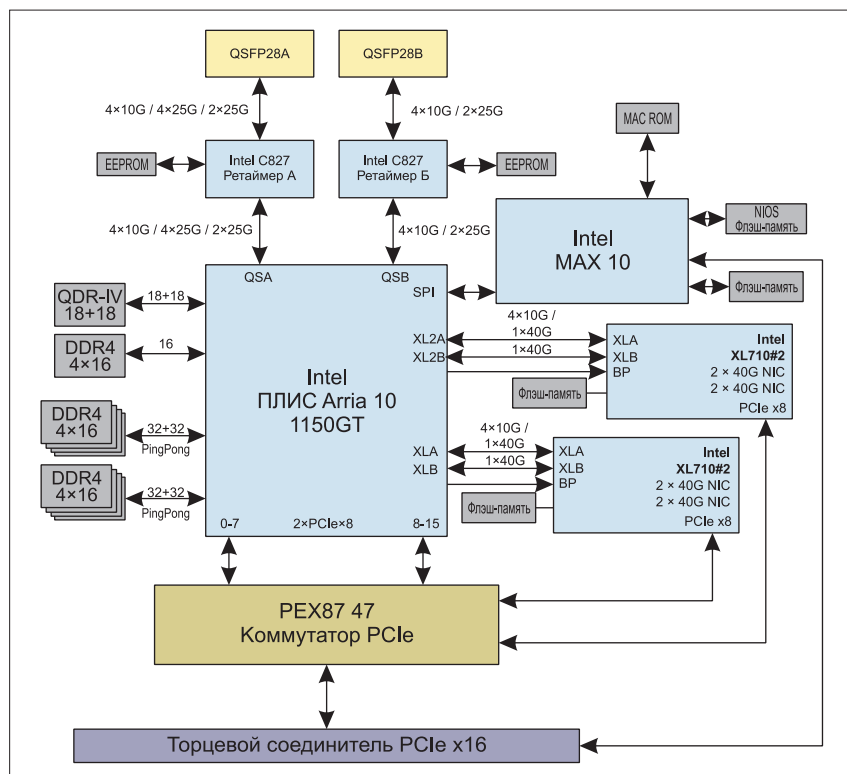


Рис. 6. Архитектура SmartNIC-карты N3000 компании Intel

На плате имеются два порта QSFP28, с которыми ПЛИС соединяется напрямую. На каждый XL710 приходится по восемь 10G-линий ПЛИС. Это классический подход проводного подключения, который позволяет ПЛИС работать с пакетами до их передачи в XL710.

В создании сетевой карты с использованием стандартного контроллера Ethernet и ПЛИС нет ничего нового. Еще в 2012 г. компания Solarflare Communications установила ПЛИС между двумя портами QSFP на сетевой карте и контроллером Ethernet, создав платформу АОЕ (Application Onload Engine).

Эта платформа стала предшественницей проекта с N3000, выбранного Intel, но она позволила компании Solarflare достичь впечатляющих результатов – задержка распространения составила 350 нс. Спустя примерно восемь лет рекорд обновился, достигнув 24,2 нс.

Intel предлагает с помощью ПЛИС выполнять обработку пакетов до контроллеров XL710. В состав ПЛИС входят 1150 тыс. программируемых логических элементов и два банка памяти DDR4 по 4 Гбайт, каждый из которых имеет достаточный объем, чтобы с помощью SmartNIC реализовать:

- виртуальный шлюз широкополосной сети (Virtual Broadband Network Gateway, vBNG);
- иерархическое качество обслуживания (HQoS);
- классификацию пакетов, определение политики, составление графика и формирование пакетов;
- виртуализованное усовершенствованное пакетное ядро (Virtualized Evolved Packet Core, vEPC);
- сеть нового поколения 5G (Next-Generation Core Network, NGCN);
- набор протоколов для защиты данных (Internet Protocol Security, IPSec);
- протокол маршрутизации SRv6 (Segment routing for IPv6);
- векторную обработку пакетов (Vector Packet Processing, VPP);
- виртуальную сеть радиодоступа (Virtual Radio Access Network, vRAN).

Хотя платформа N3000 и рассчитана на указанные рабочие нагрузки, пока не ясно, предоставила ли Intel все необходимое программное обеспечение для разгрузки каждого из этих приложений на карте SmartNIC. Как известно, дьявол кроется в программном обеспечении – оборудование всех упомянутых компаний превосходно, другое дело – подходящий софт.

## Xilinx

Другим выдающимся разработчиком и производителем ПЛИС в сегменте SmartNIC является Xilinx – первая компания, которая коммерциализировала ПЛИС в середине 1980-х гг. В настоящее время Xilinx занимает первую строчку на рынке ПЛИС, а Intel – вторую. Xilinx приобрела Solarflare Communications осенью 2019 г., которая с 2012 г. разрабатывала сетевые карты на базе ASIC и ПЛИС для электронной торговли.

В создании карт SmartNIC Alveo U25 от Xilinx приняла участие группа инженеров Solarflare из Кембриджа (см. рис. 7).

Alveo U25 подключает два порта SFP28 к микросхеме серии Zynq. Эту ИС можно считать системой-на-кристалле (СНК), поскольку в ее состав входит не только ПЛИС, но и четырехъядерный процессор Arm A53 для обработки пакетов. Zynq подключается напрямую к хост-серверу через восемь линий PCIe Gen 3 или через SerDes к микросхеме Ethernet-контроллера X2, который, в свою очередь, соединяется с хостом с помощью восьми линий PCIe Gen 3. Такой подход позволяет Zynq обрабатывать пакеты до их передачи кристаллу X2, или полностью обойти X2.

В состав Alveo U25 входит 6-Гбайт память DDR4, доступ к которой имеют ПЛИС и Arm-ядра микросхемы Zynq. В состав ПЛИС входят 520 тыс. логических элементов, малое количество которых в достаточной мере компенсируют четыре ядра Arm. Карты Alveo U25, в первую очередь, адресованы тем приложениям, которым требуется разгрузка открытого виртуального коммутатора (Open virtual Switch, OvS). Компания заявила, что в ближайшем будущем добавит разгрузку для IPsec, машинного обучения (ML), глубокой проверки пакетов (DPI), транскодирования видео и аналитики.

Как и Intel, компания Xilinx выпускает несколько линеек кристаллов для вычислительных систем: Kintex, Virtex, Zynq и Versal. Kintex и Virtex – «чистые» ПЛИС; некоторые модели этой линейки имеют почти 3 млн логических ячеек, что почти в три раза больше, чем у процессора N3000 компании Intel. Использование кремниевых интерпозеров позволило Xilinx слоями установить в кристаллы Virtex память объемом до 16 Гбайт с высокой пропускной способностью (HBM). Эта технология применяется и в других микросхемах всех четырех линеек.

В состав СНК Zynq от Xilinx входят ПЛИС, четыре ядра Arm, Arm-ядра для работы в режиме реального времени, контроллеры DDR и логика подключения для Ethernet и PCI Express. Возможности платформы адаптивного ускорения вычислений (ACAP) Versal выходят далеко за рамки СНК. Эта платформа построена на основе 7-нм кристаллов. ACAP расширяет архитектуру Zynq за счет добавления сотен ядер искусственного интеллекта (AI), блоков цифровой обработки сигналов (DSP) и многого другого. По сути, ядра искусственного интеллекта представляют собой вычислительные средства одинарной точности. Компания намеревается объединить представленную сетевую карту SoftNIC с платформой Versal.

## ТЕКУЩЕЕ СОСТОЯНИЕ РЫНКА SmartNIC

Как показывает опыт компаний Netronome и даже Solarflare Communications, сетевые карты SmartNIC появились давно. Такие крупные клиенты как Google и Amazon ушли с рынка, разработав и создав собственные решения. Тем временем Facebook и Microsoft предоставили высокоуровневые архитектуры, кото-

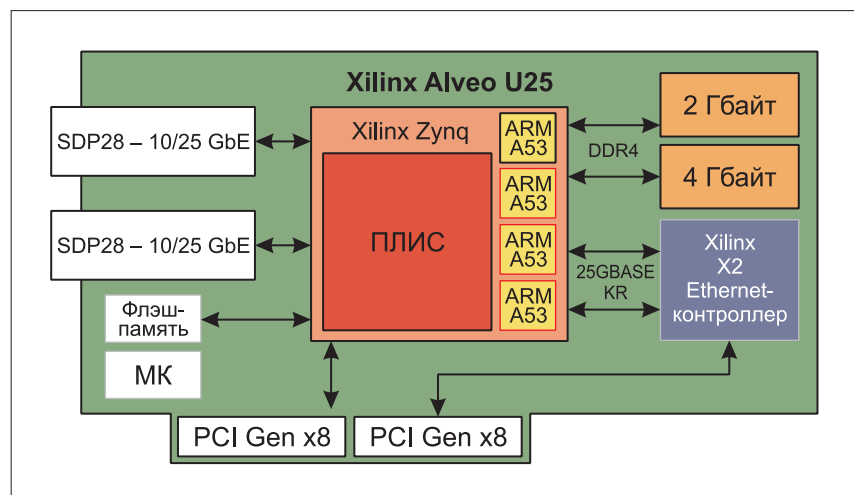


Рис. 7. Структурная схема Alveo U25 от Xilinx

рыми поспешили воспользоваться отраслевые поставщики.

С течением времени СнК и, что важнее, ПЛИС в своем развитии достигли такого уровня, когда они могут стать базовой технологией для сетевых карт SmartNIC. 10 лет назад на рынке был бум графических процессоров (GPU), ознаменовавших собой первую значительную веху на пути совершенствования технологий аппаратного ускорения. В настоящее время для расширения возможностей сетей используются ПЛИС, у которых количество логических элементов превышает три миллиона, а также другие компоненты с ПЛИС блоки обработки для функционирования сетей, памяти, хранения и вычислений. Для вычислений в данном случае применяются кластеры ядер на кристаллах в блоках СнК или даже платформ АСАР.

Такие достижения позволяют заявить о формировании второй волны аппаратного ускорения. Поскольку в свое время графическим процессорам понадобились новые программные API и инструменты для поддержки платформ, схожие требования предъявляют и аппаратные ускорители на базе ПЛИС. По мере окончательного формирования рынка SmartNIC ожидается, что он подстегнет развитие следующего поколения аппаратных ускорителей на базе ПЛИС, обеспечив реализацию еще более высокопроизводительных вычислений.

Развитие SmartNIC-карт способствуют наращиванию вычислительной мощности и, следовательно, ускорению работы на границах сети, благодаря чему освобожденные ресурсы центральных процессоров серверов

направляются на критически важную обработку данных.

Представьте себе, что функции хранения, шифрование, проверку пакетов данных и сложную маршрутизацию возьмут на себя карты SmartNIC. В результате в ЦП хоста вернется значительная часть циклов центрального процессора, обычно затрачиваемых на выполнение этих задач.

Чтобы быть впереди крупных участников рынка, новые компании, например Pensando и Fungible, будут продолжать создавать SmartNIC-карты с инновационными функциями и возможностями, а технологические лидеры, к которым относятся компании Xilinx, Intel, Broadcom и Nvidia, будут совершенствовать базовые вычислительные ядра и специализированные процессоры P4. ▢