Arm v9 – новая архитектура фирмы ARM

Ключевые слова: архитектура, безопасность, конфиденциальные вычисления, производительность, расширения.

Фирма ARM, ведущий лицензиар ядер процессоров разных типов, представила новую архитектуру – Arm v9. Она предназначена для замены архитектуры Arm v8, выпущенной на рынок 10 лет назад. В новую архитектуру добавлен ряд расширений (в частности, для задач искусственного интеллекта и машинного обучения) и функций, позволяющих увеличить производительность процессоров и обеспечить их другими преимуществами.

Фирма ARM осуществила серьезную модернизацию своей архитектуры, представив ее следующую версию – Arm v9. Новая архитектура предоставляет дополнительные функции обеспечения безопасности, конфиденциальных вычислений[6] и искусственного интеллекта (ИИ), повышает общую производительность – ожидается, что v9 обеспечит более чем 30 %-ный рост производительности двух следующих поколений мобильных устройств и инфраструктуры. Функции ИИ, которые до сих пор наиболее часто предоставляются вместе с графическими процессорами, теперь будут доступны также в сочетании с центральными и нейронными процессорами фирмы.

Предшествующая архитектура, Arm v8, была выпущена на рынок 10 лет назад. Специалисты ARM ожидают, что новая архитектура в течение следующих 10 лет будет доминировать в области вычислительных кремниевых ИС в широком диапазоне приложений – от Интернета вещей до суперкомпьютеров. Отмечается, что вскоре число процессоров на основе ядер ARM, выпущенных партнерами фирмы, превысит 200 млрд шт. Причем если первые 100 млрд процессоров были произведены за 26 лет, то вторые 100 млрд ИС будут отгружены всего за пять лет.

Руководство компании подчеркнуло, что архитектура Arm v9 будет именно десятилетним проектом с ежегодным представлением обновленных версий – v9.1, v9.2 и т. д. Ключевые функции, охватываемые первоначальной версией новой архитектуры, в основном касаются двух областей: удовлетворение глобального спроса на повсеместные специализированные вычисления и повышение безопасности каждого приложения.

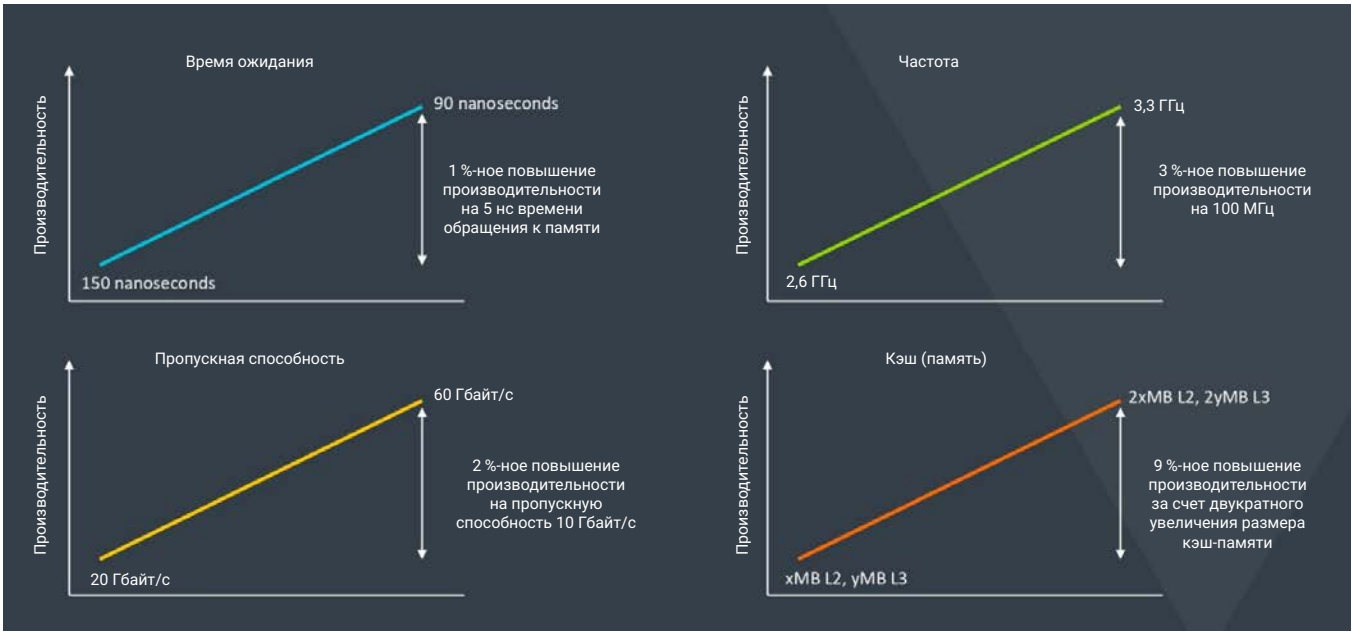
ПРОИЗВОДИТЕЛЬНОСТЬ ЦП

Проблемы разработки СФ-блоков процессоров для перспективных компьютеров связаны со все более сложными, постоянно развивающимися, гетерогенными рабочими нагрузками на рынках мобильных устройств, автомобилей и инфраструктуры. Часть проблем решается за счет использования современных техпроцессов со все меньшими проектными нормами, но этот подход связан с увеличением стоимости и сроков производства.

К новым ИС предъявляются жесткие требования с точки зрения обеспечения рентабельности инвестиций по отношению как к традиционным, так и к перспективным вычислительным рабочим нагрузкам. С учетом высокой стоимости отказа для выведенных на рынок новинок как в абсолютном стоимостном выражении, так и в воздействии на ширину рыночного окна[7], существует также требование использовать аттестованные, высококачественные СФ-блоки. Ведутся работы над технологиями, позволяющими максимизировать частоту, пропускную способность, размер кэша и уменьшить время ожидания – все это должно способствовать максимизации производительности процессора (рис. 1).

Споры о достоинствах специализированных ускорителей, видеопроцессоров, ускорителей ИИ и машинного обучения (МО) до сих пор ведутся, однако места для всех этих типов приборов на рынке в обозримом будущем пока хватает. Тем не менее требования современных коммерческих рабочих нагрузок требуют программируемости ускорителей. Это подразумевает все – от библиотек и компиляции с использованием языка С до виртуализации, чтобы программируемые ускорители ИИ можно было легко использовать в облачной среде, вплоть до отладки и анализа производительности. Если добавить требования безопасности, то конструкция ускорителя существенно приблизится к конструкции процессора.

С этой точки зрения разработчики ARM считают необходимым продолжать расширять архитектуру процессоров так, чтобы процессоры (на основе ядер ARM) могли обеспечивать еще большее ускорение рабочих нагрузок и делать это программируемым, защищенным, всеобъемлющим и проверенным образом. Сегодня невозможно игнорировать сильнейшую фрагментированность некоторых рабочих нагрузок ИИ и ЦОС-процессоров на мобильном рынке и то, какую выгоду можно извлечь из их объединения в процессорную среду. Именно здесь ARM стремится продвинуть свою архитектуру и вычислительные проекты.



Источник: ARM

Рисунок 1. Использование архитектуры Arm v9 увеличивает производительность центрального процессора на 30 %

Примечание: L2, L3 – кэш-память второго и третьего уровней; xMB, 2xMB, yMB, 2yMB – емкость кэш-памяти в Мбайт.

РАСШИРЕНИЯ ИИ И МО

В рамках архитектуры Arm v9 будет представлено несколько новых функций, связанных с ИИ, включая расширенную аппаратную поддержку ИИ по всему портфелю процессоров, графических процессоров и нейронных процессоров ARM. В основе этого лежит убеждение специалистов ARM в том, что все процессоры – от приборов для суперкомпьютеров до изделий для облачных вычислений и оконечных устройств – должны будут обрабатывать рабочие нагрузки ИИ. Разработчики считают, что ключом к инновациям во всех формах вычислительной техники станет целенаправленное проектирование систем. Разные вычислительные задачи требуют различных комбинаций вычислительных компонентов. Многие устройства Интернета вещей нуждаются в интерпретации своего «мира», и сочетание ядер M-профиля с микропроцессорами Ethos-U55 для этого хорошо подходит. В автомобильных системах партнеры ARM будут все чаще объединять множество больших и малых процессоров с графическими процессорами, нейронными процессорами и своими собственными СФ-блоками для формирования правильных вычислительных решений для подобных автономных систем.

Например, различные сочетания вычислительных компонентов могут работать в гарнитурах виртуальной реальности (VR)[8] (большой нейронный процессор и графический процессор наряду с маленькими нейронным процессором и центральным процессором), смартфонах (большой центральный процессор и графический процессор наряду с маленькими центральным процессором и нейронным процессором) и устройствах Интернета вещей (маленький центральный процессор и нейронный процессор). Для этих трех вариантов использования можно построить три разные «системы-на-кристалле» (SoC) с тремя очень разными типами и размерами процессоров. Если баланс вычислительных компонентов окажется неправильным, то на выходе получится или ИС со слишком малым быстродействием, или слишком дорогая ИС. Выбор, правильный для одного партнера ARM, одного устройства или одного варианта использования, окажется просто неприменимым в других случаях (рис. 2).

В свое время в рамках архитектуры Arm v8 была представлена поддержка арифметических операций FP16[9] и BFloat[10], популярных при обработке данных с использованием ИИ, а также функция масштабируемых векторных расширений (scalable vector extension, SVE). Функция SVE была разработана в сотрудничестве с Fujitsu и другими компаниями для суперкомпьютерных процессоров Fugaku. Она добавляет возможности векторной обработки для повышения производительности ИИ и цифровой обработки сигнала (ЦОС/DSP). При этом масштабируемость SVE обеспечивает возможность применения концепций, используемых для суперкомпьютеров, к гораздо более широкому спектру продуктов. В рамках архитектуры Arm v9 для создания SVE2 была добавлена расширенная функциональность, улучшенные масштабируемые векторные расширения, которые хорошо работают для систем 5G и многих других вариантов использования, таких как VR и дополненная реальность (augmented reality, AR – технология, накладывающая сгенерированные вычислительным устройством изображения или текст на видение пользователем реального мира и представляющая таким образом смешанную картину), а также для машинного обучения в процессоре. В течение следующих нескольких лет специалисты ARM планируют расширять этот процесс еще больше с целью существенно улучшить выполнение матричных вычислений в центральном процессоре.

Источник: ARM

Рисунок 2. ARM намерена представить в рамках архитектуры Arm v9 ряд технологий, позволяющих повысить производительность центральных процессоров на 30 %

* Эффективность данных (data efficiency) – эффективность прилагаемых к данным процессов, таких как хранение, доступ, фильтрация, распределение.
** Прюнинг (pruning) – отсечение ветвей дерева возможных решений, один из алгоритмов поисковых систем (предусматривающий исключение заведомо нерелевантных документов при поиске с целью ускорения выполнения запроса), систем ИИ и принятия решений.
*** Краевые вычисления (edge computing) – метод оптимизации облачных вычислительных систем путем переноса обработки данных на границу сети вблизи источника данных, благодаря чему снижается трафик между датчиками и ЦОД. Требует использования ресурсов, не подключенных к сети постоянно (ноутбуки, смартфоны, планшетные ПК, датчики и т. п.).
Примечание: Mali – семейство графических процессоров ARM, Cortex – семейство процессоров ARM, Ethos – семейство нейронных процессоров ARM, Neoverse – семейство серверных процессоров ARM.

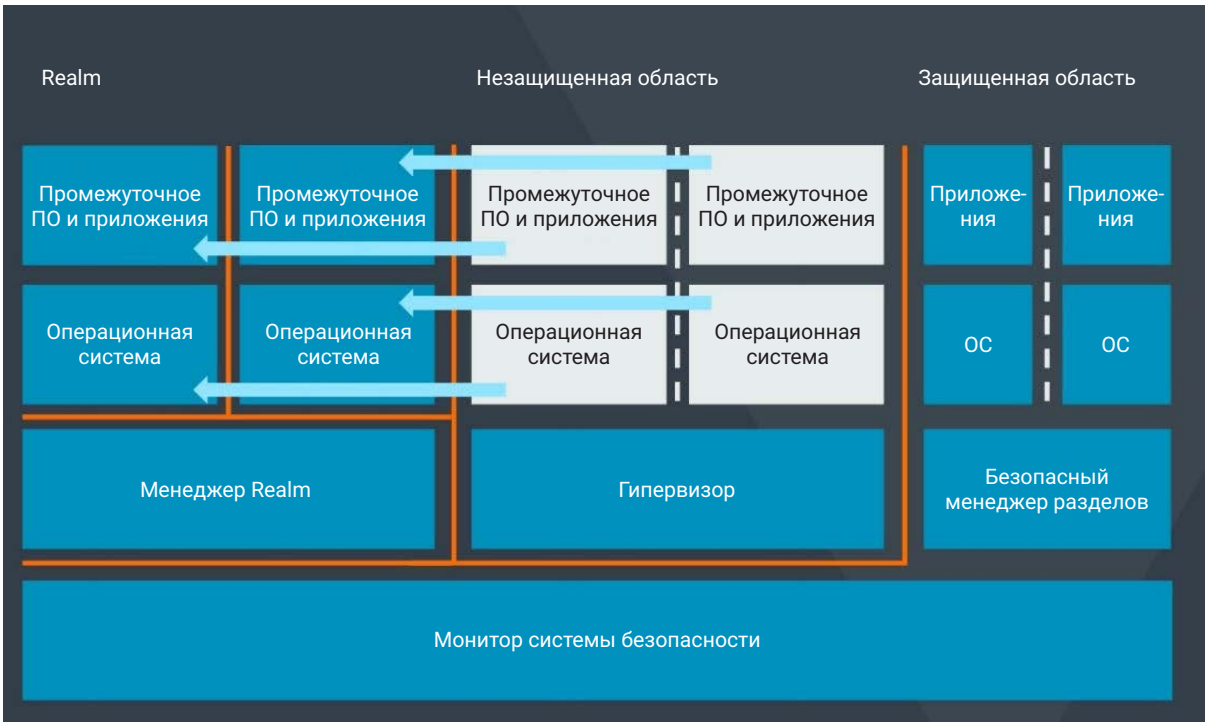
КОНФИДЕНЦИАЛЬНЫЕ ВЫЧИСЛЕНИЯ

Большое внимание в рамках архитектуры Arm v9 уделяется безопасности. В частности, за последние пять лет в сотрудничестве с Microsoft была разработана функция Realms, позволяющая использовать конфиденциальные вычисления в рамках архитектуры конфиденциальных вычислений ARM (Arm Confidential Computing Architecture, Arm CCA). Arm CCA строится на безопасных и небезопасных сферах современной технологии TrustZone[11].

Сегодня традиционная модель вычислений опирается на огромное доверие к операционным системам и гипервизорам, на которых исполняются приложения. Конфиденциальные вычисления устраняют предположение, что привилегированное ПО, ответственное за запуск вычислительной системы, должно иметь возможность видеть или манипулировать данными этих запущенных сеансов. Это значительно увеличивает доверие к вычислительной инфраструктуре.

Realms позволит запускать приложения или сервисы (услуги) таким образом, чтобы данные были защищены от проверки или вторжения базисных (централизованных) систем (host systems) или любого другого ПО, работающего на этой хост-системе. Это может быть применено как к виртуальным машинам в «облаке», так и к приложениям на смартфоне.

Обычно при аренде мощностей поставщика услуг гиперразмерных вычислений[12] клиент получает виртуальную машину, размещенную в многопользовательской системе, где клиенту выделяется некоторая доля общего адрес-



Источник: ARM

Рисунок 3. Архитектура Arm v9 вводит функцию Realms – концепцию фирм ARM и Microsoft для обеспечения безопасности данных за счет их защиты от базисных (централизованных) систем (host systems) и другого ПО

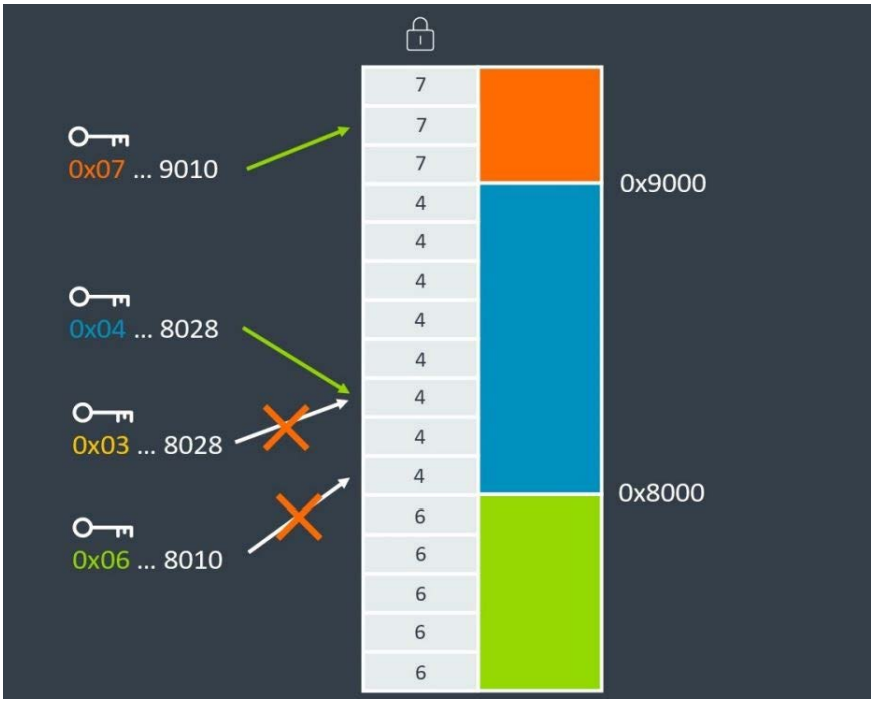
ного пространства. В экосистеме Arm CCA одно из самых простых изменений состоит в том, что виртуальная машина будет размещаться не в общем адресном пространстве, а в области, где адресное пространство защищено от других виртуальных машин, совместно использующих систему. То же самое верно и для ноутбука или ПК, когда у клиента есть вторая операционная система, совместно использующая ресурсы хост-системы.

Типичная система Android сегодня обладает сочетанием небезопасного ПО (запускающего основной стек), некоторых безопасных сервисов (работающих по технологии TrustZone) и, возможно, некоторых сервисов управления цифровыми правами (DRM)[13], работающих как виртуальные машины вместе с Android. Одна из возможностей заключается в том, что часть защищенных сервисов могут мигрировать из зоны доверия в свою собственную область, обеспечивая более динамичную среду этого сервиса. Сервис DRM также может переместиться в свою собственную область, что повысит его конфиденциальность, поскольку данные окажутся защищены от основного стека Android.

Приложения со смешанной критичностью, такие как робототехника и автомобильная электроника, также могут использовать области для разделения сервисов, которые питают критические системы безопасности, защищая их память от помех.

Функция Realms будут доступна не сразу, но станет частью будущей версии Arm v9. Прежде чем это произойдет, станет доступна еще одна новая функция безопасности – расширение тегирования памяти (MTE[14]) (рис. 4).

Функция MTE, разработанная в сотрудничестве с Google, может быть использована для поиска проблем с пространственной и временной памятью в программном обеспечении. Реальность такова, что многие из проблем безопасности в действительности сводятся к тем же старым и давно из-



Источник: ARM

Рисунок 4. Расширение тэгирования памяти может использоваться для обнаружения пространственных и временных проблем памяти в ПО

вестным проблемам безопасности памяти, которые преследовали компьютеры в течение последних 50 лет. Две наиболее распространенные проблемы безопасности памяти – переполнение буфера и использование после освобождения – проявляют чудеса постоянства на протяжении многих лет. Ситуация осложняется тем, что зачастую они присутствуют в ПО в течение многих лет, прежде чем их обнаруживают. Функция MTE позволит ПО связать указатель памяти с тегом и проверить правильность этого тега при использовании указателя. Если доступ находится вне зоны действия или использование памяти переместилось дальше, то проверка тега завершится неудачей.

MTE – это одна из первых функций, которая будет запущена в рамках архитектуры Arm v9 и будет доступна в первом поколении процессоров ARM, основанных на ней. Программная поддержка MTE будет внедрена в ОС Android 11 и openSUSE.

СТАНДАРТИЗАЦИЯ И ВНЕДРЕНИЕ

Еще одной темой, обсуждавшейся на прошедшей презентации ARM, были проблемы стандартизации, а именно – баланс между слишком жесткой стандартизацией, означающей, что клиенты ARM не могут разрабатывать дифференцированные решения, и недостаточной, снижающей совместимость программного обеспечения.

У Arm уже есть успешная программа для серверов под названием «архитектура системы на основе сервера» (server-based system architecture, SBSA) с ее программой валидации Server Ready, поощряющей то, что ARM считает правильным балансом стандартизации. Как часть архитектуры Arm v9, область действия этой программы будет расширена, чтобы вклю-

чить краевые и оконечные устройства, в рамках программы под названием System Ready, разработанной с нуля, чтобы поддерживать все потребности экосистемы ARM, от самого маленького до самого большого устройства.

Представители корпорации MediaTek на презентации ARM заявили, что первый смартфон MediaTek с процессором Arm v9 будет коммерчески доступен к концу 2021 г. Большинство партнеров ARM будут стремиться производить образцы на базе Arm v9 примерно в те же сроки. Поточно-массовое производство процессоров на основе архитектуры Arm v9 начнется в 2022 г.

Ward-Foxton Sally. Arm v9: First New Architecture in a Decade Doubles Down on AI and Security. EE Times, March 30, 2021: https://www.eetimes.com/arm-v9-first-new-architecture-in-a-decade-doubles-down-on-ai-and-security/?utm_source=newsletter&utm_campaign=link&utm_medium=EETimesDaily-20210331&oly_enc_id=5245B7817912J8Z#