Компьютер и человек – кто кого?

«Компьютер делает не то, что вы хотите, а то, что вы ему приказываете»
Автор неизвестен

Продолжение.
Начало в №№ 7, 8, 9, 10/2023, №№ 1, 2, 4, 5, 6, 7, 8, 9, 10/2024

Арсений Ворошилов, по материалам ARM

Как бы подробно ни были рассмотрены достоинства и недостатки генеративного искусственного интеллекта, каковы бы ни были выгоды и риски от его применения, очевидно одно – технология уже здесь, ее применение становится все шире, а потому имеет смысл уделять больше внимания практическим аспектам GenAI.

Быстрое и гибкое масштабное развертывание генеративного искусственного интеллекта – конкурентные преимущества за счет решения проблем и достижения высочайшей производительности

Стремление к инновациям и развитию с помощью GenAI

Гонка за разработкой и расширением возможностей генеративного искусственного интеллекта (GenAI) вызвала волну инноваций и новых испытаний для современных разработчиков. Тогда как большие языковые модели LLM становились масштабнее и сложнее, обрастая сотнями миллиардов параметров, в последнее время наметился сдвиг в сторону более компактных и эффективных моделей, характеризующихся меньшим числом параметров [1]. Типы контента, доступные моделям GenAI для обработки и создания, тоже меняются. Теперь в их состав входят полноценное видео, изображения, звук и текст.

Наметившийся рост инвестиций и развитие технологий, несомненно, продолжатся. По сути, 75% лидеров бизнеса считают, что в нынешних условиях конкурентоспособность зависит от применения AI, GenAI и машинного обучения. Так, во всяком случае, утверждается в отчете исследования Experian, проведенного компанией Forrester Consulting [2].

Чтобы сохранить конкурентоспособность и в полной мере реализовать потенциал AI, разработчикам нужны гибкие высокопроизводительные вычислительные платформы, которые поддерживают AI-нагрузку во всем диапазоне от облака до оконечного устройства. Ниже рассказывается, как индустрия может преодолевать ключевые сложности, связанные с использованием GenAI, применительно к эффективности, скорости, защите данных и экологичности.

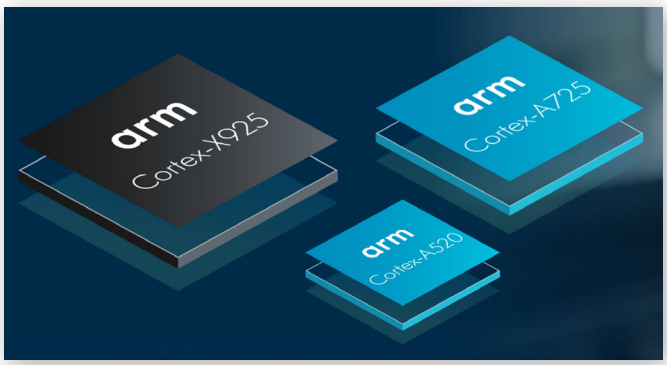
Запуск GenAI и масштабирование эффективности в зависимости от нагрузки

Эволюция моделей AI и больших языковых моделей (LLM) означает, что наблюдается экспоненциальный рост размеров обучающих моделей, параллельно с чем новые логические – инференсные (inference) – модели становятся все более и более эффективными. Нынешние наиболее распространенные LLM все еще содержат миллиарды параметров, что требует существенных вычислительных ресурсов и больших объемов памяти. Тем не менее методы логической оптимизации типа квантизации сделали модели LLM более эффективными. Это, вкратце, означает, что LLM можно применять шире, в различных центрах обработки данных (ЦОД), а разработчики ощутят существенные преимущества от масштабного применения логической оптимизации на таких оконечных устройствах, как смартфоны, ноутбуки и даже устройства IoT (Интернета вещей).

Для большинства устройств, на которые ложится нагрузка, включая и оконечные устройства, наиболее эффективным вычислительным ресурсом все чаще становится центральный процессор.

«Выполнение многих вычислительных процессов, связанных с GenAI, становится эффективнее, если нагрузку несут центральные процессоры. Процессор Arm имеет ряд достоинств, в том числе гибкость, обеспечивая при этом ключевые строительные блоки, необходимые разработчикам для упрощения развертывания, такие как модели, средства администрирования, MLOps, базы данных и материалы для тренировки, – говорит архитектор сверхкрупных решений Arm Рави Малхотра. – *Центральные процессоры Arm могут обеспечить работу широкого спектра нейросетей с использованием различных форматов данных, позволяя им справляться с перспективными нагрузками и помогая добиться масштабируемости и адаптируемости. Это то, что подходит для любого варианта применения»*.

В зависимости от масштаба нагрузки может потребоваться компромисс между стоимостью и быстродействием, а центральные процессоры обеспечивают хороший баланс для систем меньшего масштаба [3].



Процессоры Arm

Kleidi повышает эффективность AI-обработки

Для расширения возможностей генеративного искусственного интеллекта разработчикам нужен доступ к наиболее быстрым методам внедрения AI в сочетании с оптимальной производительностью для любой рабочей нагрузки, выполняемой как на устройстве, так и в облаке.

Набор библиотек и сообществ разработчиков Arm Kleidi содержит технологию Arm, интегрированную непосредственно с ведущими AI-платформами, чтобы ускорить работу моделей на центральных процессорах Arm. Это поддерживается ресурсами разработчиков и динамичной экосистемой поставщиков программного обеспечения для машинного обучения (ML), а также проектов с открытым исходным кодом.

Назначение Arm Kleidi состоит в том, чтобы позволить разработчикам достигать максимальной производительности GenAI на процессорах Arm, обеспечивая ускорение миллиардов рабочих процессов на ежедневно используемых устройствах.

Для рабочих процессов, нуждающихся в дальнейшем ускорении, процессоры Arm обеспечивают опции настройки. Они хорошо сочетаются с графическими картами GPU и способны задействовать широкие возможности параллельных вычислений, предоставляемые такими картами. Вычислительная платформа Arm обеспечивает разработчикам гетерогенные вычисления, давая возможность выбрать оптимальное устройство обработки, что является важнейшим фактором в достижении максимальных производительности и эффективности.

Сокращение сроков выхода на рынок

Искусственный интеллект развивается стремительно. *«Мы становимся свидетелями драматических изменений, и многим организациям трудно поддерживать такой темп. Вот почему эффективность программной платформы Arm является ключом к тому, чтобы AI-экосистемы развивалась без отставания»*, – отмечает директор по управлению продукцией Arm Ронан Нотон.

Наблюдается приток новых моделей различных размеров и назначения. Их перенос на оконечные устройства или в облако требует, чтобы они оказывали меньшую нагрузку, что достигается с помощью эффективных методов, таких как агрессивная квантизация, спекулятивное декодирование и контекстная разреженность.

Нотон отмечает: *«С выпуском новых моделей искусственного интеллекта мы видим, как экосистема с открытым исходным кодом набирает обороты, чтобы запустить их на вычислительной платформе Arm всего за 24 часа. Это очень важно для разработчиков, стремящихся как можно скорее применить новейшие модели в своих приложениях»*.

«Мы все чаще наблюдаем, что вычисления, связанные с AI, обычно выполняются в центральных процессорах. Например, когда в 2023 году впервые стала доступна 4-разрядная квантизация, выполнять ее можно было только силами центральных процессоров, поскольку графические процессоры в то время были способны только на 8-разрядную квантизацию, – говорит Малхотра. – *Теперь же людям нужны еще более совершенные методы, вплоть до понижения разрядности до 2 бит и применения разреженного кодирования. Хотя сегодня это невозможно даже на графических картах и ускорителях последнего поколения, все это можно сделать на базе центральных процессоров»*.

И, наконец, успешное масштабирование требует портативности. *«Вычислительная платформа Arm является самой распространенной в мире и поддерживает глобальное внедрение AI»*, – добавляет Нотон. Интеграция со всеми основными операционными системами и AI-решениями позволяет разработчикам своевременно встраивать машинное обучение в приложения по мере того, как модели и технологии эволюционируют и обеспечивают портативность в рамках всех экосистем.

Защита данных

Организации, внедряющие AI и LLM, сталкиваются с различными задачами по защите данных, как тех, что используются для тренировки AI-моделей, так и пользовательских данных. Тем не менее 41% профессионалов сферы кибербезопасности обладают весьма ограниченным опытом AI-безопасности, а 21% из них не уверены в своей способности справиться с подобными вызовами [4].

Разработчики и отделы кибербезопасности нуждаются в доступных технологиях, которые защищают полный стек вычислительного ПО, не создавая при этом препятствий разработке. Потребность в повышенной защите вычислений – не какой-то новый феномен. По мере роста вычислительной мощности расширяется и потенциальный фронт для кибератак. Параллельно растет ценность данных, моделей и приложений, что делает их привлекательными целями для хакеров.

К счастью, Arm работала над проблемой защиты с самого начала, внедряя фундаментальные функции безопасности для процессоров Arm при каждой итерации развития архитектуры Arm.

«В течение трех десятилетий Arm инвестировала в функции защиты для процессоров Arm, внедряя такой функционал, как аутентификация по указателю (PAC), аутентификация целевой ветви (BTI) и расширения для создания меток в памяти (MTE), чтобы добиться уверенности в том, что данные на устройстве надежно защищены», – говорит Нотон.

Функция Arm MTE позволяет динамически выявлять пространственные и временные проблемы безопасности памяти, на долю которых приходится 70% всех серьезных сложностей обеспечения защиты данных [5].

Между тем, одна из функций архитектуры Arm CCA (Confidential Compute Architecture – конфиденциальной архитектуры вычислений) позволяет разработчикам создавать области, которые защищают AI-приложения и используемые ими данные на устройствах. Содержимое области и протекающие в ней процессы не могут стать жертвами атак из ПО с более высокими уровнями привилегий. К тому же области сохраняют целостность при взаимодействии с ускорителем.

Из соображений безопасности разработчики изучают возможность развертывания приложений на периферийных устройствах для определенных рабочих нагрузок с целью повышения общей безопасности. Результаты опросов показывают, что 85% тех, кто принимает решения в сфере глобальных технологий, считают, что угрозы безопасности приведут к тому, что больше AI-приложений будут развертываться в оконечных устройствах [6]. Это позволяет обрабатывать данные на пользовательском устройстве вместо того, чтобы отправлять их куда-либо. Участники опроса также признали дополнительные достоинства такого подхода – 51% из них отметили повышение эффективности, а 41% – ускорение обработки данных и уменьшение задержки.

Экологичность

Неудержимый рост GenAI принесет и острую потребность в вычислительной мощности, а также в используемой для нее энергии. На обработку запроса в ChatGPT расходуется, в среднем, почти в 10 раз больше электричества, чем на поиск в Google [7]. Тогда как центры обработки данных по всему миру потребляют 1...2% от общего количества энергии, ожидается, что к концу десятилетия эта доля вырастет до 3…4% [7]. К 2027 году на каждый доллар, потраченный на AI-сервер, организации придется закладывать в бюджет порядка 35 центов на оплату электроэнергии [8].

Организации, разрабатывающие и внедряющие искусственный интеллект, должны применять меры устойчивого развития, чтобы достигать своих целей по энергоэффективности и идти вровень со своими конкурентами по параметрам экологичности. Уже сейчас эти организации могут предпринять шаги для сокращения потребления энергии за счет применения правильной аппаратной инфраструктуры для каждой рабочей нагрузки и использования GenAI на оконечных устройствах [8].

«Имеющиеся у нас данные говорят о том, что для обеспечения прогнозируемого энергопотребления AI-вычислений исключительно для тренировочных рабочих нагрузок может потребоваться утроение глобальной выработки электроэнергии, и это сложно себе представить. К счастью, мы уже видим свидетельство того, что экосистема внедряет AI-вычисления повсюду, от облаков до периферии, и масштабно использует логическую рабочую нагрузку», – говорит Нотон.

«Наравне с экономически эффективными периферийными рабочими процессами, серверы на базе Arm обеспечивают не только повышенную производительность, но и существенную экономичность в плане потребления энергии и эксплуатационных расходов по сравнению с другими существующими серверами, – отмечает Малхорта. – *Это касается не только нагрузок, связанных с AI и ML, но и других облачных нагрузок в такой же степени»*.

Революция в области искусственного интеллекта

Технологическим лидерам требуется эффективная, высокопроизводительная платформа, позволяющая запускать генеративный искусственный интеллект в центрах обработки данных, на смартфонах, компьютерах и других оконечных устройствах. Для достижения масштабируемости рабочие нагрузки GenAI должны быть встроены во все слои стека машинного обучения. Каждая рабочая нагрузка предъявляет собственные требования к вычислениям, и только посредством гетерогенного подхода к вычислениям эти нагрузки могут быть назначены на соответствующие вычислительные ресурсы. Центральные процессоры Arm способны обеспечить соответствие меняющимся потребностям рабочих нагрузок для AI, а также безопасность, необходимую разработчикам. Для рабочих нагрузок, требующих дополнительного ускорения, процессоры Arm станут оптимальным дополнением к AI-акселераторам, таким как GPU и NPU, обеспечивая неизменную производительность.

Arm помогает разработчикам преодолеть сложности, ограничивающие масштабируемость GenAI, открывая возможности по внедрению инновационных решений по всему миру и предоставляя различным организациям конкурентные преимущества, необходимые им на завтрашнем рынке искусственного интеллекта.

Источники

1. Kamal, A., Ramzan, A., Algunaibet, R. Embracing the evolution : Navigating the landscape of large and small language models. PwC. Feb. 9, 2024. https://www.pwc.com/m1/en/publications/ embracing-the-evolution.html

2. Mukherjee, V. 75% of business leaders say AI key to their competitive advantage: Report. Business Standard. Oct. 3, 2024. https://www.business-standard.com/industry/ news/75-business-leaders-say-ai-key-to-their-competitive-advantage-report-1241003006171. html

3. Duranton, S. The Bermuda Triangle of generative AI: Cost, latency and relevance. Forbes. Feb. 5, 2024. https://www.forbes.com/sites/ sylvainduranton/2024/02/05/the-bermuda-triangle-of-generative-ai-cost-latency-and-relevance/

4. Poremba, S. Is the cybersecurity industry ready for AI? Cybersecurity Dive. June 24, 2024. https://www.cybersecuritydive.com/news/ generative-AI-cyber-threats/719624/

5. Grisenthwaite, R. Business as usual? Computing security in the age of AI. Arm. Jan. 22, 2024. https://newsroom.arm.com/blog/ai-security

6. Scaling Security in the Age of AI: How AI Innovation Relies on Investment and Best Practice. May 2024. https://report.psacertified.org/

7. AI is poised to drive 160% increase in data center power demand. Goldman Sachs. May 14, 2024. https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand

8. Forecast Analysis: AI-Optimized Servers, Worldwide. Gartner Research. Nov. 30, 2023. https://www.gartner.com/en/documents/4984931