DOI: 10.22184/1992-4178.2024.241.10.138.147

Искусственный интеллект: новые архитектуры ИИ процессоров и расширение роли в проектировании ИС

М. Макушин[1]

УДК 621.37 | ВАК 2.2.2

В Стэнфордском университете (шт. Калифорния) 25–27 августа 2024 года прошла конференция Hot Chips 2024 (далее – конференция). В ней приняло участие большое число как известных фирм, так и стартапов. Большое внимание, в частности, было уделено новым архитектурам ИИ процессоров, позволяющих сбалансировать их быстродействие и эффективность. Кроме того, отмечалось, что роль ИИ в проектировании ИС постоянно расширяется – по мере усложнения инструментальных средств САПР.

На конференции отмечалось, что использование в рамках развития ИИ больших языковых моделей (LLM)[i] повышает потребность в устойчивых вычислениях и гетерогенной (разнородной) интеграции, при этом ключевым отличительным признаком становится управление данными. Устойчивые вычисления (sustainable computing, также используется термин green computing) рассматриваются как практика максимальной эффективности использования энергии и минимизации воздействия на окружающую среду при проектировании и использовании ИС, систем и ПО (охватываются все аспекты цепочки поставок – от используемого для изготовления сырья до переработки по окончании жизненного цикла).

Ведущие разработчики систем ИИ отказываются от создания максимально быстродействующих ИИ процессоров и переходят к более сбалансированному подходу. Этот подход предполагает использование высокоспециализированных (узкоспециализированных), разнородных вычислительных элементов, передачу данных с большей скоростью и существенное снижение потребляемой мощности.

Частично этот сдвиг связан с внедрением чиплетов[ii] в 2,5D- и 3,5D-модулях, лучше адаптирующихся к различным рабочим нагрузкам и типам данных, а также способствующих увеличению удельной (в пересчете на 1 Вт) производительности. Технология 3,5D считается следующим шагом совершенствования методик перспективного проектирования и является гибридным подходом, включающим в себя этажирование чиплетов и их отдельное соединение с подложкой, используемой совместно с другими компонентами [1].

Эта технология позволяет значительно повысить производительность и устранить ряд самых сложных проблем гетерогенной интеграции. 3,5D-корпусирование можно считать «золотой серединой» между 2,5D-модулями, широко используемыми в центрах обработки данных (ЦОД), и полноценными 3D-модулями, коммерциализацию которых полупроводниковая промышленность ведет более 10 лет.

К основным преимуществам 3,5D-архитектуры относятся:

- удовлетворительное разделение на физическом уровне для эффективного решения проблем отвода тепла и шумовых эффектов;
- увеличение числа (емкости) СОЗУ в быстродействующих конструкциях;
- сокращение расстояния между обрабатывающими элементами и памятью также уменьшает путь

---

[1] НОБ «Военные науки и оборонная промышленность» БРЭ, ведущий научный редактор.

[i] **LLM (Large Language Model)** – большая языковая модель, одно из основных направлений развития ИИ. Являются нейросетевыми моделями, использующими алгоритмы машинного обучения. Позволяют обобщать, прогнозировать, генерировать человеческие языки на основе больших наборов текстовых данных. Основное применение: чат-боты; написание статей, маркетинговых текстов, электронных писем; переводы текстов; поисковые системы и т.д.

[ii] **chiplet** – чиплет, специализированные микросхемы (блоки), обладающие минимальной вычислительной мощностью и рядом других функций, позволяющие чиплетам стать малым микропроцессором, устройством хранения данных, сложной логической схемой или частью MEMS, выполняющих различные функции. На основе совмещения чиплетов можно создавать более сложные ИС, формирование которых иным образом неэкономично и/или не позволяет достичь нужных проектных норм в каждом блоке.

прохождения сигналов, что позволяет значительно повысить скорость обработки данных по сравнению с планарными конструкциями [2].

Второй пункт важен тем, что СОЗУ широко используется в качестве кэш-памяти процессора и остается важным элементом для ускорения обработки данных. Но СОЗУ в отличие от цифровых ИС не масштабируются ниже 40-нм уровня и поэтому занимают все большую долю площади современных конструкций. Решение проблемы – вертикальное этажирование чиплетов.

НОВЫЕ АРХИТЕКТУРЫ

Использование ИИ в полупроводниковой промышленности только начинается и существует ряд вопросов, требующих решения. С точки зрения регулирования – это недопущение установления монополии в области ИИ и поддерживающих его средств. С точки зрения разработчиков – первична устойчивость ИИ, что затрагивает архитектуры, создание более эффективного ПО и совершенствование микроархитектур, повышение уровня интеграции чиплетов различных поставщиков [1].

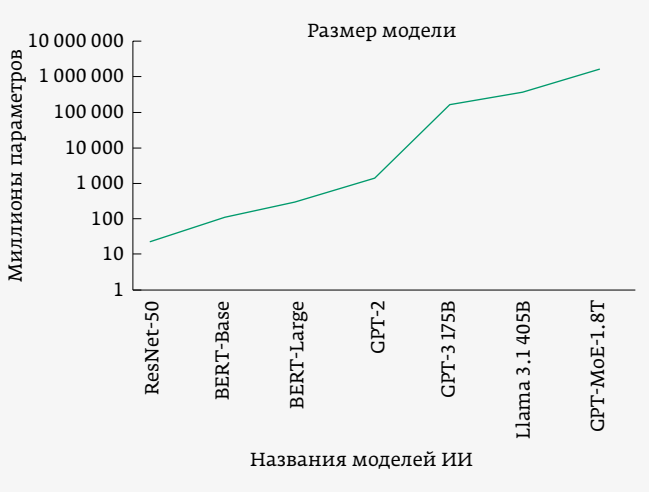


Рис. 1. Размер моделей искусственного интеллекта (ИИ) за десятилетие вырос более чем в 70 тыс. раз – за счет добавления новых возможностей и параметров. *Источник: корпорация NVIDIA (конференция Hot Chips 24)*

Станет ли корпорация Nvidia монополистом на рынке ИИ?

Появление многих новых разработок приводит к становлению корпорации Nvidia в качестве почти монополиста в мире ИИ – из-за распространения недорогих графических процессоров и моделей на базе CUDA, создаваемых на их основе. Ни один процессор общего назначения не может сравниться по энергоэффективности со специализированными ускорителями. Поэтому не случайно, что большинство многочиплетных архитектур, представленных на конференции в 2024 году, содержат не один, а несколько типов процессоров, память большей емкости и конфигурации устройств ввода-вывода, позволяющие ограничить число узких мест и обеспечить более эффективное управление данными.

Специалисты корпорации Nvidia хорошо осведомлены о ситуации и возможных конкурентных угрозах. Ее новая ИС Blackwell, представленная на конференции, сочетает в себе графические процессоры (GPU), центральные процессоры (CPU) и процессоры обработки данных (data processing unit, DPU). Схема квантования Blackwell открывает возможности создания средств ИИ низкой точности, обладающих рекордными на данный момент по скорости возможностями обучения. В свою очередь это позволяет работать с гораздо бóльшими моделями данных. С учетом увеличения размеров моделей ИИ в последнее время (рис. 1) это очень важно [3].

Изменения в центрах обработки данных

Одним из важных изменений к подходам создания конструкций процессоров в этом году стало повышенное внимание к управлению данными. С появлением ИИ речь все больше идет не только о создании огромных массивов избыточных обрабатывающих элементов и их максимально быстром запуске. Все чаще цель состоит в том, чтобы разумно расставить приоритеты – и данных, и их типов, которых становится все больше. Но этот подход не новый. На самом деле он появился в 1980 году, когда корпорация Intel представила свой сопроцессор 8087 с плавающей запятой. В 2011 году компания Arm Holdings пошла дальше, представив концепцию big-little. Она основана на разнородной вычислительной архитектуре, предполагающей задействование двух ядер (или двух типов ядер при конфигурации процессора, содержащей более двух ядер) – высокопроизводительного и с малой потребляемой мощностью. Такая концепция позволит оптимизировать энергопотребление процессоров типа «система-на-кристалле» (SoC) в зависимости от реальной рабочей нагрузки. Первый процессор big.LITTLE содержал ядра Cortex-A7 и Cortex-A15.

С тех пор эта стратегия была усовершенствована за счет более сложного разделения и расстановки приоритетов, но она обычно не ассоциируется с интегральными схемами ИИ, работающими в крупных ЦОД. Именно там происходит основная часть обучения средств ИИ, и, вероятно, такая ситуация сохранится еще некоторое время. Это обусловлено тем, что разработка LLM и многократные запросы к ним требуют значительных вычислительных мощностей. Тем не менее, не каждый вычислительный цикл требует больших затрат на обработку, и к генеративным моделям ИИ и в дальнейшем необходимо будет обращаться так же часто, как и сегодня [1].

Даже корпорация IBM, утверждающая, что обрабатывает 70% всех финансовых транзакций в мире с помощью своих суперкомпьютеров, сменила направление, сосредоточившись не только на числе триллионов операций в секунду (TOPS), но и на удельной производительности (на ватт, в пДж/с). Это особенно примечательно, поскольку, в отличие от крупных системных компаний, на долю которых в настоящее время приходится около 45% всех разработок перспективных ИС, IBM продает свои системы конечным заказчикам, а не просто предоставляет вычисления как услугу.

Представленный корпорацией новый процессор Telum для ускорения операций ввода/вывода (то есть, по сути, для передачи данных туда, где они будут обрабатываться и храниться) использует DPU, а также инновационное кэширование. В целом в него входят 8 ядер, работающих на частоте 5,5 ГГц, десять 36-Мбайт блоков кэш-памяти второго уровня (L2) и новый ускоритель на основе чиплета.

Отмечается, что DPU широко используются в промышленности для высокоэффективной обработки огромных объемов данных. Суперкомпьютеры, такие как полностью сконфигурированный IBM z16, способны обрабатывать 25 млрд зашифрованных транзакций в день. Это больше, чем подаваемое за это же время число запросов в Google, сообщений в Facebook и Tweeter вместе взятых. Для таких объемов требуются возможности устройств ввода/вывода, значительно превосходящие возможности обычных компьютерных систем. К таким возможностям относятся специальные протоколы ввода/вывода для минимизации задержек, постоянная поддержка виртуализации и способность обрабатывать десятки тысяч запросов ввода/вывода в любое время.

Новый процессор корпорации IBM также позволяет снизить энергопотребление 8-ядерного комплекса CPU на 15%, отчасти благодаря улучшенному предсказанию ветвлений. В последние пару лет эта тема постоянно обсуждалась на конференциях Hot Chips, так как более точное предсказание ветвлений и ускорение восстановления после ошибок предварительной выборки могут повысить производительность. При этом добавление в конструкцию DPU позволило сделать еще один шаг вперед, сделав его «интеллектуальным регулировщиком» трафика передачи данных. DPU монтируется непосредственно на кристалл процессора и позволяет снизить энергопотребление, необходимое для управления операциями ввода/вывода, на 70% [4].

Корпорация Intel также представила свой ускоритель нового поколения для обучения ИИ – Gaudi 3 (рис. 2). Он оснащен четырьмя ядрами глубокого обучения (DCORE), восемью этажерками встроенных ДОЗУ с высокой пропускной способностью (HBM2e) емкостью 16 Гбайт и блоками умножения матриц (MME), которые можно настраивать, а не программировать. Помимо этого, он содержит 64 ядра тензорного процессора (TPC) и подсистему памяти, включающую в себя объединенное пространство кэш-памяти второго и третьего уровней (L2 и L3). В качестве мостового соединения между кристаллами 1 и 0 использован интерпозер. В ускорителе ИИ корпорации Intel реализован подход «вычисления в непосредственной близости к памяти» (near-memory compute), предполагающий расположение схемы/кристалла/кристаллов памяти и логических приборов в составе одного модуля, созданного с использованием перспективных методов корпусирования (2,5D/3D packaging, fan-out). Наконец, в нем используется интегрированный пакет ПО, позволяющий заказчикам подключать заказные ядра TPC.

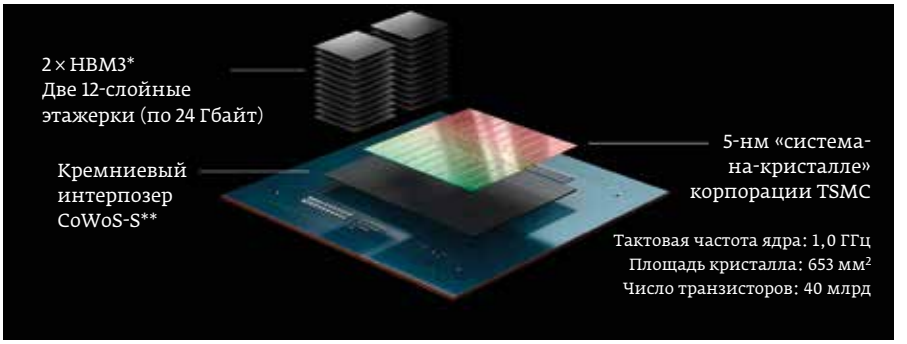
Подход Intel к управлению данными по своей концепции аналогичен подходу IBM. Для распределения работы по назначенным подразделениям Intel использует

* RDMA NIC (Remote Direct Memory Access Network Interface Card) контроллер сетевого интерфейса удаленного прямого доступа к памяти

** 16GB HBM2e – встроенное ДОЗУ 2-го поколения с высокой пропускной способностью емкостью 16 Гбайт

*** Шина PCI-Express 5-го поколения

Рис. 2. Блок-диаграмма ИС ускорителя для обучения модели ИИ Gaudi 3 корпорации Intel. *Источник: корпорация Intel (конференция Hot Chips 2024)*



* HBM3 (High Bandwidth Memory DRAM) – ДОЗУ с высокой пропускной способностью 3-го поколения
** CoWoS-S (chip on wafer on substrate) – «кристалл-на-пластине-на-подложке», технология сборки 2,5-3-мерных ИС корпорации TSMC (вариант S)

Рис. 3. Устойчивая архитектура ИС формирования краевых логических выводов компании Furiosa. *Источник: компания Furiosa (конференция Hot Chips 24)*

диспетчер синхронизации и драйвер среды выполнения (Runtime Driver). Такой подход максимизирует использование ресурсов в системе и позволяет избежать любых узких мест за счет асинхронной передачи событий с использованием диспетчера прерываний [1].

Представленный корпорацией AMD процессор MI300X основан на распределенной системе ИИ, состоящей из 12 чиплетов с 4 кристаллами ввода/вывода и 8 кристаллами ускорителей. Он, аналогично разработкам IBM и Intel, направляет данные туда, где их лучше всего обрабатывать. В процессоре используются межсоединения Infinity fabric 4-го поколения, шина PCI Express 5-го поколения, ДОЗУ 3-го поколения (HBM3) и архитектура множественного доступа с кодовым разделением 3-го поколения (CDMA3). Благодаря этому MI300X обеспечивает сбалансированное масштабирование в подсистемах вычислений, памяти и устройств ввода/вывода [5].

Изменения в краевых вычислениях

Ранее практика обработки данных с помощью ИИ была в значительной степени разделена между обучением в гиперразмерных ЦОД и формированием логических выводов на гораздо более компактных, часто мобильных устройствах. Перенесение обучения и формирование логических выводов в область краевых вычислений становится все более востребованным из-за роста затрат на перемещение больших объемов данных и времени, необходимого для получения результатов по запросам. Несмотря на то, что применение и размеры LLM продолжают увеличиваться, они не являются единственными обучаемыми моделями ИИ. Более предметно-ориентированные модели меньшего размера можно обучать с использованием менее интенсивной вычислительной инфраструктуры, а формирование логических выводов может осуществляться на устройствах с аккумуляторным питанием.

Это открывает совершенно новый рынок для гетерогенных конструкций с использованием чиплетов. При этом не обязательно, что все чиплеты конструкции будут созданы одним и тем же разработчиком или произведены на одном и том же кремниевом заводе[iii]. Схемы HBM 1–3-го поколений – первый крупный успех в этом направлении. Но чиплеты разрабатываются для самых разных приложений, аналогично тому, как в течение последних двух десятилетий использовалась интеллектуальная собственность, не включающая патенты (soft IP – авторские права, торговые марки, коммерческие секреты, а также другие активы, которые сложнее классифицировать, например общие знания о продукте или конфиденциальную информацию, принадлежащую компании). Как и в случае с ИС ИИ ЦОД, ключевым моментом является управление перемещением данных и памятью [1].

Одним из подобных решений для мобильных и настольных компьютеров можно считать SoC Lunar Lake корпорации Intel. Разработчики ставили перед собой четыре основные цели: увеличение энергоэффективности, улучшение графики и повышение производительности ядра и общей производительности всей платформы – до 120 TOPS. Подход Intel заключается в разделении логики за счет использования вычислительного чиплета и чиплета контроллера платформы в конфигурации 2,5D-интеграции со встроенной памятью (память-на-модуле).

Вычислительный чиплет изготовлен корпорацией TSMC по технологическому процессу N3B, в качестве базового кристалла используется процессор 1227 корпорации Intel, печатная плата поставляется TSMC (изготовлена по процессу N6). Для объединения всех элементов использована технология Foveros – технология 3D-корпусирования центральных процессоров, разработанная корпорацией Intel, позволяющая осуществлять сов-

[iii] **foundry** – кремниевый завод, производство ИС по спецификациям заказчика с предоставлением заказчику широкого спектра услуг использования инструментальных средств фирм-союзников из числа поставщиков САПР для проектирования собственных ИС с использованием базы библиотек стандартных элементов различных fabless- и IDM-фирм (по контрактам foundry с последними), платформ и сложнофункциональных блоков (на тех же условиях). Кремниевые заводы могут заниматься разработкой новейших технологических процессов, но разработкой собственных конструкций ИС, как правило, не занимаются.

местное (вертикальное) размещение разнородных процессорных ядер, ядер графического процессора, ускорителя ИИ. Использование памяти-на-модуле обеспечивает два основных преимущества. Во-первых, это обеспечивает возможность оптимизации устройства на физическом уровне под низкую потребляемую мощность вследствие малого числа межсоединений, а также оптимизацию собственно памяти-на-модуле. Во-вторых, площадь, занимаемая материнской платой, сократилась до 250 мм^2 [6].

Корпорация Qualcomm представила на конференции собственную SoC Oryon, разработанную по тем же принципам. Она включает в себя три процессорных кластера, каждый из которых содержит 4 ядра. Два кластера ориентированы на высокую производительность, а один – на энергоэффективность. Отраслевые обозреватели отмечают, что во многих из осуществленных на конференции презентаций выделялась микроархитектура, в основном определяющая, как выполняются команды на уровне аппаратного обеспечения. Как и в случае с гораздо более крупными системами, во многих из этих проектов центральное место занимает то, как и где обрабатываются и хранятся данные.

В SoC Oryon встроено восемь основных декодеров, подготавливающих команды для исполнительных блоков, блока сохранения загрузки и блока векторного исполнения. Сами команды поступают в буфер переупорядочения микроопераций в процессоре (re-order buffer). В нем содержится около 600 записей, что дает представление о том, сколько команд будет выполнять машина «на лету». С точки зрения выхода из строя, на каждый невыполненный цикл работы придется восемь невыполненных команд.

Особое внимание в SoC Qualcomm привлек модуль управления памятью. Он поддерживается очень большим унифицированным буфером трансляции 2-го уровня, и это сделано в первую очередь для обработки большого объема данных. Буфер предназначен для работы со всеми виртуализированными структурами, уровнями безопасности, и эта структура намного больше, чем обычно. Разработчики считают, что это должно свести задержку трансляции данных к абсолютному минимуму [7].

Многие из участников конференции – хорошо известные компании, но были и новички. Например, южнокорейский стартап FuriosaAI, разрабатывающий ИС ИИ для краевых вычислений. Он представил «процессор тензорного сжатия для устойчивых вычислений с использованием ИИ».

Представленная в 2021 году первоначальная конструкция была оптимизирована для моделей масштаба BERT (Bidirectional Encoder Representation Transformers, двунаправленная нейронная сеть-кодировщик, модель представления языка, предназначенная для предварительного обучения глубоких двунаправленных представлений на простых немаркированных текстах путем совмещения левого и правого контекстов во всех слоях). Впоследствии были выпущены базовые модели на основе GPT3 (Generative Pre-trained Transformer, мощная языковая модель, разработанная компанией OpenAI, способная генерировать текст, трудно отличаемый от написанного человеком; модель обучается на большом объеме текстовых данных и может выполнять разнообразные задачи, связанные с обработкой естественного языка). Эти модели в пять раз крупнее BERT. Разработчики считают, что развитие идет в сторону наиболее эффективных моделей ИИ, предлагающих больше преимуществ. При разработке представленной на конференции архитектуры конструкции (рис. 3) в качестве наиболее эффективного средства формирования логических выводов ставка была сделана на демон генератора псевдослучайных чисел (Random Number Generator Daemon, RNGD/PRNGD).

Центральное место в этой архитектуре, предназначенной для краевых ЦОД, занимает быстрое перемещение данных в память и обратно. Утверждается, что пропускная способность памяти составляет 1,5 Тбайт/с. RING также оснащен двумя стеками HBM3 общей емкостью 48 Гбайт и 256-МБайт СОЗУ [8].

РАСШИРЕНИЕ РОЛИ ИИ ПРИ ПРОЕКТИРОВАНИИ ИС

Использование ИИ в системах автоматизированного проектирования придает новый импульс развитию индустрии этих инструментальных средств. Ведущие поставщики САПР обновляют свои системы функциями ИИ и машинного обучения (МО), а также привлекают стартапы и университеты, пытающиеся разработать дифференцированные подходы для решения ряда проблем с помощью новых инструментов и методологий.

На ранних этапах разработки инструментальных средств САПР возникали вопросы, связанные с разработкой аппаратного обеспечения, что привело к автоматизации сложных, трудоемких задач, которые продолжают лежать в основе закона Мура о масштабировании в различных измерениях. В отличие от этого, ИИ добавляется на зрелом этапе разработки САПР, когда инновации способствуют повышению эффективности и продуктивности известных процессов и приложений.

Дополнение возможностей инструментальных средств САПР средствами ИИ основано на предположении, что в отличие от других отраслей, где работники опасаются их замены на ИИ, перегруженные работой специалисты по проектированию и верификации ИС ждут от ИИ помощи в избавлении от рутинных операций (снижение трудоемкости ряда задач проектирования). Это принципиально

ОТЕЧЕСТВЕННЫЙ ПРОИЗВОДИТЕЛЬ

Специальное оборудование

для электронной промышленности

- Разработка и производство технологического оборудования
- Разработка и внедрение современных технологий
- Поставка зарубежного оборудования и комплексных технологий
- Модернизация и сервисное обслуживание технологического оборудования

смотреть каталог

8 (499) 729-77-51 | nppesto.ru

не изменит квалификацию данных специалистов, но позволит им уделять больше времени инновациям. При этом приоритетное внимание должно уделяться сокращению времени разработки ИС с целью повышения эффективности инноваций [9].

Использование ИИ в инструментальных средствах САПР позволит осуществить «сдвиг влево» (shift left), то есть переместить некоторые операции проектирования со своих традиционных «мест» ближе к начальным этапам технологического процесса проектирования. Речь идет о том, чтобы добиться фактического пересечения друг с другом этапов проектирования, относящихся как к начальным, так и к завершающим этапам процесса проектирования в целом. Это позволит получить представление о возможном результате на ранней стадии проектирования, а затем быстро довести весь процесс до конца. Также возникает возможность ускорения получения многими специалистами информации, которую в противном случае им пришлось бы ждать три месяца или более.

Пространство проектирования ИС настолько огромно, что трудно выявить причинно-следственные связи. Это увеличивает время и затраты на решение проблем. Рост сложности проектирования в последние годы многократно вырос (рис. 4), поэтому применение ИИ в инструментальных средствах САПР становится безальтернативным – для снижения времени проектирования и расходов на него.

При внедрении в процессы проектирования ИС технологий ИИ, помогающих в оптимизации, используется обучение с подкреплением (reinforcement learning), чтобы быстрее и с меньшими усилиями находить решения. Данный подход дополняется генеративным ИИ, что помогает быстрее получать информацию, обучать людей и обмениваться информацией в команде, превращая пользовательские интерфейсы в интерфейсы на естественном языке. Так как модели воспринимаются на естественном языке, появляется возможность легкого получения информации автоматизированным способом для определения того, что имеет значение. Это ключевой фактор сокращения времени, помимо ускорения реализации этапов проектирования [10].

Что касается традиционных подходов к верификации, ИИ поможет ответить на вопросы о том, что нужно протестировать, а что еще не тестировалось. В случае «сдвига влево» можно использовать ИИ для раннего определения рабочих нагрузок, требований, моделирования и т.д. Таким образом, вместо разрозненного потока теперь возникает взаимосвязанный подход – от идеи до реализации с использованием поддерживаемого ИИ цифрового двойника конструкции ИС.

На практике это означает, что большинство текущих проектов стартапов в области инструментальных средств САПР сосредоточены на разработке виртуальных ИИ-помощников (AI copilots), основанных на естественном языке. Эти помощники представляют собой нейросети, способные переводить инструкции на английском языке в пригодный для использования код быстрее и точнее,

Рис. 4. С ростом сложности проектирования увеличивается потребность в инструментальных средствах САПР, оснащенных ИИ. *Источник: корпорация Synopsys (конференция Hot Chips 2024)*

АКЦИОНЕРНОЕ ОБЩЕСТВО
ЭЛЕКОНД

РАЗРАБОТКА И ПРОИЗВОДСТВО КОНДЕНСАТОРОВ

Оксидно-электролитические алюминиевые конденсаторы К50-...

Номинальное напряжение, $U_{ном}$, В,	3,2 ... 485
Номинальная емкость, $C_{ном}$, мкФ,	1,0 ... 470 000
Диапазон температур среды при эксплуатации, $T_{ср}$, °С	-60 ... 125

Объемно-пористые танталовые конденсаторы К52-...

Номинальное напряжение, $U_{ном}$, В,	3,2 ... 200
Номинальная емкость, $C_{ном}$, мкФ,	1,5 ... 60 000
Диапазон температур среды при эксплуатации, $T_{ср}$, °С	-60 ... 175

Оксидно-полупроводниковые танталовые конденсаторы К53-...

Номинальное напряжение, $U_{ном}$, В,	2,5 ... 63
Номинальная емкость, $C_{ном}$, мкФ,	0,033 ... 2 200
Диапазон температур среды при эксплуатации, $T_{ср}$, °С	-60 ... 175

Суперконденсаторы К58-...

Номинальное напряжение, $U_{ном}$, В,	2,5 ... 2,7
Номинальная емкость, $C_{ном}$, Ф,	1,0 ... 4 700
Диапазон температур среды и эксплуатации, $T_{ср}$, °С	-60 ... 65

Накопители электрической энергии на основе модульной сборки суперконденсаторов НЭЭ, МИК, МИЧ, ИТИ

Номинальное напряжение, $U_{ном}$, В,	5,0 ... 48
Номинальная емкость, $C_{ном}$, Ф,	0,08 ... 783
Диапазон температур среды при эксплуатации, $T_{ср}$, °С	-60 ... 65

Россия, 427968, Удмуртская Республика, г. Сарапул, ул. Калинина, д. 3
Тел.: (34147) 2-99-53, 2-99-89, 2-99-77, факс: (34147) 4-32-48
e-mail: elecond-market@elcudm.ru; www.elecond.ru

Система менеджмента качества сертифицирована на соответствие требованиям ISO 9001

чем генеративный ИИ[iv] общего назначения, и с большей степенью интерактивности [11].

Генеративные ИИ помощники проектирования

AI copilots входят в класс цифровых помощников проектирования на основе генеративного ИИ, в который также входят чат-боты, предоставляющие технические рекомендации (руководства), такие как анализ кода и краткие сведения о верификации. AI copilots функционируют как инструменты для завершения кода (как ПО, помогающее разработчикам писать код более эффективно) на основе ИИ, а более перспективные подходы позволяют вводить интерактивные предложения. Google Trends, публичное веб-приложение корпорации Google, основанное на ее поисковике, показывает, что термин AI copilot впервые появился в июне 2021 года, после того, как на платформе GitHub было представлено средство Hub Copilot, первоначально предназначенное для Visual Studio корпорации Microsoft. Hub Copilot стал шаблоном для других подобных инструментов.

Возможные камни преткновения

Использование ИИ в инструментальных средствах САПР может столкнуться с другими проблемами, помимо сокращения времени и затрат на проектирование. На первый взгляд, такие инструментальные средства САПР, особенно в таких областях, как синтез, размещение элементов ИС на кристалле и маршрутизация, а также верификация конструкции ИС на логическом уровне, являются очевидными бенефициарами от использования методов ИИ. Инструментальное средство САПР, использующее алгоритмы размещения логических ячеек на основе ИИ с учетом предыдущего опыта работы с тысячами ранее реализованных проектов разработки ИС, скорее всего, гораздо быстрее придет к приемлемой компоновке/размещению, чем в случае начала работы каждый раз с основных принципов. Но возникают серьезные вопросы относительно использования данных для обучения ИИ и их принадлежности:

- сохраняет ли крупная полупроводниковая компания, обладающая сотнями предыдущих разработок, свои данные в тайне и использует ли их только для улучшения процесса проектирования на физическом уровне, не допуская до них конкурирующий стартап в тени?
- имеет ли фирма-разработчик САПР право «изучать» все ранние разработки клиентов и «передавать» это ноу-хау другим пользователям?

Аналогичные вопросы могут возникнуть при использовании генеративного ИИ для написания проверочных тестов на основе предыдущих наборов верификации, которые ИИ «видел» ранее.

Также к серьезным проблемам относятся согласование обмена данными, обучение модели и выбор интерфейсов прикладного программирования. Создание успешных инструментальных средств САПР на основе ИИ требует их открытости, независимости от поставщика и обеспечения безопасного обмена данными без нарушения прав интеллектуальной собственности. Компании, внедряющие закрытые решения, будут ограничивать подобные возможности, что приведет к увеличению барьеров на пути внедрения инноваций [9].

Что после помощников проектирования?

Отраслевые специалисты, представляющие известных разработчиков САПР, активно обсуждают вопрос: станут ли AI copilots образцом для будущих ИИ САПР или же появятся совершенно новые инструментальные средства САПР ИИ, которые перевернут эту индустрию с ног на голову. Первый путь – эволюционный, второй – революционный. В первом случае речь идет о повышении эффективности работы пользователя, возможно на порядки. Во втором случае разговор о появлении новых решений, коренным образом изменяющих аспекты проектирования и верификации на каждом этапе процесса. Здесь, в частности, ИИ будет использоваться для прогнозирования результатов, а не просто для облегчения работы проектировщиков и повышения ее эффективности [9, 11]. На данный момент однозначного ответа не существует. Вполне возможна реализация обоих вариантов.

***

Искусственный интеллект только начинает приносить пользу полупроводниковой промышленности, предстоит решить еще много задач. Прежде всего, ИИ должен быть устойчивым, и это хорошо понимают крупные фирмы-изготовители ИС, а также стартапы, о чем свидетельствуют представленные на конференции Hot Chips 2024 архитектуры. Но ИС – только часть решения.

Устойчивость ИИ также требует более эффективного ПО, совершенствования микроархитектур. Это необходимо для того, чтобы запросы к большим языковым моделям выполнялись реже, а ответы LLM были более точными, чтобы можно было доверять им. Кроме того, потребуется более тесная интеграции специализированных обрабатывающих

[iv] **Generative AI** – генеративный ИИ, ориентированный на творческий потенциал алгоритмов. В отличие от традиционных систем, выполняющих задачи с заранее определенными правилами, дает машинам возможность создавать новый и неординарный контент. Основа – генеративные модели, разработанные для изучения и воспроизведения шаблонов из предоставленного набора данных. Это позволяет создавать совершенно новый контент, соответствующий стилям, структурам и нюансам, которые средства генеративного ИИ усвоили в процессе обучения.

элементов в виде чиплетов, способных быстрее и эффективнее обрабатывать различные типы данных.

Что касается индустрии инструментальных средств САПР, прошедшей через различные этапы развития, то она находится в совершенно новом пространстве. Понятно, что ИИ окажет существенное влияние на повышение производительности этих инструментальных средств, но сейчас невозможно предсказать точный результат или влияние, которое ИИ окажет на проектирование ИС.

Возможно, использование ИИ в инструментальных средствах САПР – только начало решения многих задач полупроводниковой промышленности в целом. Проектирование ИС постоянно усложняется, а освоение 3D ИС еще больше ускорит этот процесс. При этом не только сам ИИ нуждается в новых решениях для внедрения инноваций. Многим из новых «программно-определяемых продуктов на основе кремния» также необходимы аналогичные подходы. И, опять-таки, многие из этих новых решений будут основаны на ИИ. ИИ может использоваться: для повышения производительности; создания новых абстрактных моделей, ускоряющих проектирование, для объединения многих областей, нуждающихся во взаимодействии. Во многих отношениях проектировщики избалованы выбором, и это становится проблемой расстановки приоритетов. Один из основных вопросов на сегодня: какие решения в области инструментальных средств САПР на основе ИИ принесут наибольшую пользу при минимальных сбоях в работе?

Итак, ИИ – данность, которая уже никуда не денется. Но для полной реализации его потенциала потребуются усилия всей экосистемы полупроводниковой промышленности.

ЛИТЕРАТУРА

1. **Sperling Ed.** New AI Processors Architectures Balance Speed With Efficiency // Semiconductor Engineering. September 4th, 2024.
2. **Sperling Ed.** 3.5D: The Great Compromise // Semiconductor Engineering. August 21st, 2024.
3. **Ren Mark.** Introduction to AI for Chip Design // Hot Chips 2024 conference. August 25, 2024.
4. **Berry Chris.** IBM Next Generation Processor and AI Accelerator // Hot Chips 2024 conference. August 26, 2024.
5. **Smith Alan, Vamsi Krishna Alla.** InstinctTM MI300X Generative AI Accelerator and Platform Architecture // Hot Chips 2024 conference. August 26, 2024.
6. **Gihon Arik.** Lunar Lake: Powering the Next Generation of AI PCs // Hot Chips 2024 conference. August 26, 2024.
7. **Gerard Williams.** Snapdragon X Elite Qualcomm Oryon CPU: Design & Architecture Overview // Hot Chips 2024 conference. August 26, 2024.
8. **Paik June.** FuriosaAI RNGD: A Tensor Contraction Processor for Sustainable AI Computing // Hot Chips 2024 conference. August 26, 2024.
9. **Heyman Karen.** AI's Role In Chip Design Widens, Drawing In New Startups // Electronics Engineering. August 29th, 2024.
10. **Stelios Diamantidis.** AI Driven Optimization // Hot Chips 2024 conference. August 25, 2024.
11. **Hans Bouwmeester.** LLM and Chip Design // Hot Chips 2024 conference. August 25, 2024.