

ТЕНДЕНЦИИ РАЗВИТИЯ МИКРОЭЛЕКТРОННЫХ ТЕХНОЛОГИЙ

Г. Я. Красников

Российская академия наук, г. Москва, Российская Федерация

Аннотация. В статье изложены общие сведения о конструкции, принципе действия, и эволюции полевых транзисторов, являющихся базовым элементом современных электронных микросхем. Приведена актуальная «дорожная карта» развития микроэлектронных технологий до 2037 года, свидетельствующая о продолжающейся тенденции уменьшения топологических размеров и, как следствие, повышения плотности размещения транзисторов на микросхеме, что ведет к росту вычислительной производительности. Описана цифровая модель нейрона и раскрыта связь стремительного развития технологии нейронных сетей («искусственного интеллекта») с общим ростом вычислительной производительности современных компьютеров. Описаны основные предпосылки дальнейшего развития технологии: появление новых моделей нейронов, реализация «не-фон-Неймановской» архитектуры процессоров, появление новых типов памяти, увеличение плотности элементов на микросхеме за счет минимизации топологических размеров и применения 3D-технологий. Проанализированы возможные альтернативы традиционным микроэлектронным технологиям в виде квантовых компьютеров и фотонных тензорных процессоров. Раскрыты особенности квантовых и фотонных технологий, их отличия от традиционной микроэлектроники и существующие проблемы, ограничивающие их применение и требующие решения. Приведены классы задач, которые могут эффективно решаться квантовыми и фотонными вычислителями. Сделаны выводы о сохранении тренда на развитие микроэлектронных технологий на ближайшие 20 лет, расширении их возможностей для отдельных применений, а также технической возможности применения квантовых и фотонных вычислений в качестве дополнения к вычислительным системам, выполненным по классическим микроэлектронным технологиям.

Ключевые слова: микроэлектронные технологии; транзисторы; нейронные сети; машинное обучение; «искусственный интеллект»; квантовые компьютеры; фотонные тензорные процессоры.

TRENDS IN THE DEVELOPMENT OF MICROELECTRONIC TECHNOLOGIES

G. J. Krasnikov

Russian Academy of Sciences, Moscow, Russian Federation

Abstract. *The article provides general information about the design, principle of operation, and evolution of field-effect transistors, which are the basic element of modern electronic microcircuits. An up-to-date “road map” for the development of microelectronic technologies until 2037 is presented, indicating a continuing trend of decreasing topological dimensions and, as a consequence, increasing the density of transistors on a microcircuit, which leads to an increase computing productivity. A digital model of a neuron is described and the connection between the rapid development of neural network technology (“artificial intelligence”) and the general increase in the computing productivity of modern computers is revealed. The main prerequisites for further development of the technology are described: the emergence of new models of neurons, the implementation of “non-von Neumann” processor architecture, the emergence of new types of memory, an increase in the density of elements on a microcircuit by minimizing topological dimensions and the use of 3D technologies. Possible alternatives to traditional microelectronic technologies in the form of quantum computers and photonic tensor processors analyzed. The features of quantum and photonic technologies, their differences from traditional microelectronics and existing problems that limit their application and require solutions are revealed. Classes of problems that can be effectively solved by quantum and photonic computers are given. Conclusions are drawn about the continuation of the trend for the development of microelectronic technologies for the next 20 years, the expansion of their capabilities for certain applications, as well as the technical feasibility of using quantum and photonic computing as an addition to computing systems based on traditional microelectronic technologies*

Keywords: *microelectronic technologies; transistors; neural networks; machine learning; «artificial intelligence»; quantum computers; photonic tensor processors.*

Ни одна из отраслей не изменила наш мир так, как это сделала микроэлектроника. Благодаря ее развитию возникли и продолжают развиваться новые технологии в самых разных областях. Сегодня уже очевидно, что микроэлектроника и дальше будет развиваться в области вычислений и обработки данных, совершенствуя взаимодействие электронных вычислительных и подключенных технических систем. Этому взаимодействию посвящено основное внимание в контексте промышленного применения компонентов микроэлектроники.

Базовым компонентом электронной микросхемы является полевой транзистор, на принципе работы которого построены современные вычислительные системы. Транзистор управляется приложением напряжения к проводящему каналу и работает в режиме

«ключа», пропуская или не пропуская ток. При наличии порогового напряжения проводящего канала транзистор начинает пропускать ток между выводами истока и стока, что дает логический «ноль». В отсутствии напряжения транзистор закрывается и потенциал на стоковых электродах указывает на логическое значение «един». Таким образом транзистор обеспечивает реализацию двоичной логики, по которой работают все цифровые вычислительные устройства.

На рисунке 1 показаны транзисторы технологического уровня 28 нм в разрезе.

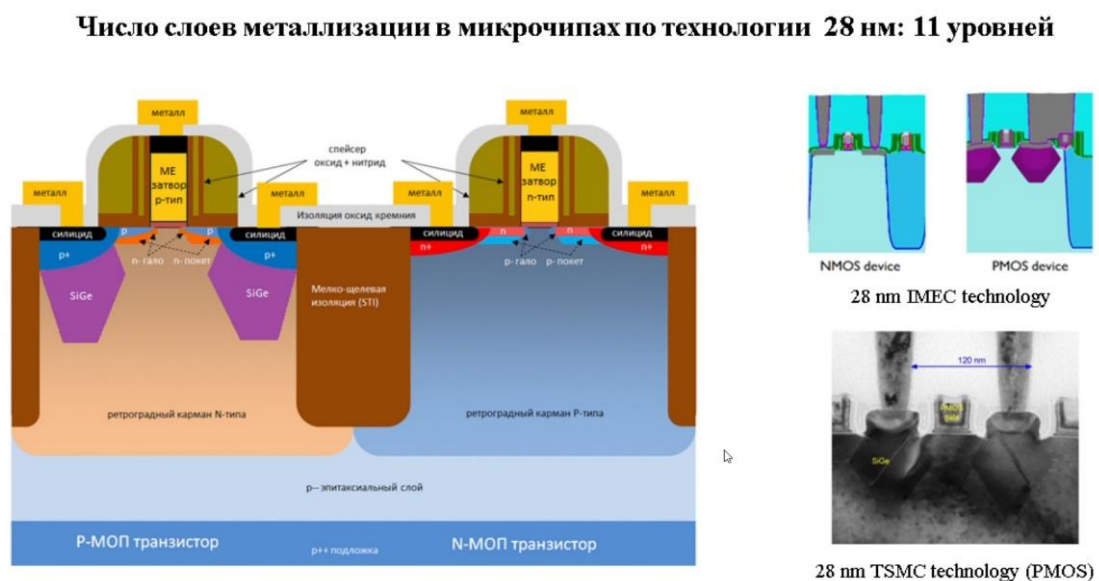


Рисунок 1 – Структура транзистора технологического уровня 28 нм

Производство транзисторов идет в так называемых «чистых комнатах». При размере «чистой комнаты» в 10 тыс. м², инфраструктура, которая обеспечивает высокие требования по влажности, температуре, запыленности занимает около 120 тыс. м². В настоящее время требования к обозначенным факторам снижены, так как пластины подаются в рабочую зону в стерильных контейнерах и не имеют контакта с открытой средой. Однако остается проблема молекулярного загрязнения: например, наличие в «чистых комнатах» больших молекул аммиачной группы, удаление которых требует применения специальных технологий.

Вторая проблема – это обеспечение чистоты всех используемых в производстве веществ: деионизированной воды, газов, различных химических веществ, используемых мишеней. Чистота

сегодня оценивается в PPT (part per trillion), то есть содержание одной частицы постороннего вещества на 10^{-12} основного вещества. Для обеспечения такой чистоты требуется целая индустрия с инфраструктурой, в которую входит методика измерения и производства особых чистых материалов, включая тару для хранения и транспортировки.

Для производства микросхемы необходимо осуществить более пяти тысяч микроопераций и провести в процессе около тысячи различных электрофизических измерений. При этом точность, с которой сегодня создаются микросхемы, составляет один атомный слой по вертикали. Если минимальный топологический размер увеличить до 1 миллиметра, то готовая микросхема будет размером 3 на 2 километра в длину и ширину и 400 метров высотой.

Микросхемы совершенствуются с момента создания первых биполярных транзисторов в 1949-м году. В 1956-м году за открытие полупроводников и транзисторного эффекта Нобелевскую премию получили три американских физика: Джон Бардин, Уолтер Браттейн и Уильям Б. Шокли. Джон Бардин один из немногих, кто получил еще и вторую Нобелевскую премию за создание теории сверхпроводимости. На всем протяжении развития микроэлектроники совершенствование микросхем происходит в соответствии с правилом Гордона Мура, сформулированным в 1964 г., согласно которому каждые 24 месяца количество транзисторов на единицу площади удваивается. Сейчас постоянно идут дискуссии, когда это правило прекратит свое действие, когда топологические размеры уменьшатся настолько, что минимизировать топологический размер будет невозможно.

На рисунке 2 приведен процесс эволюции транзисторов, где показан затвор (центральная область), сток-истоковые области в поликремнии и канал.

С развитием технологий произошел переход от планарной схемы строения транзистора к трехмерной – так называемым «наношифтам» с трехмерными каналами. Таким образом транзисторы стали объемными, что вызвало существенное увеличение плотности их размещения.

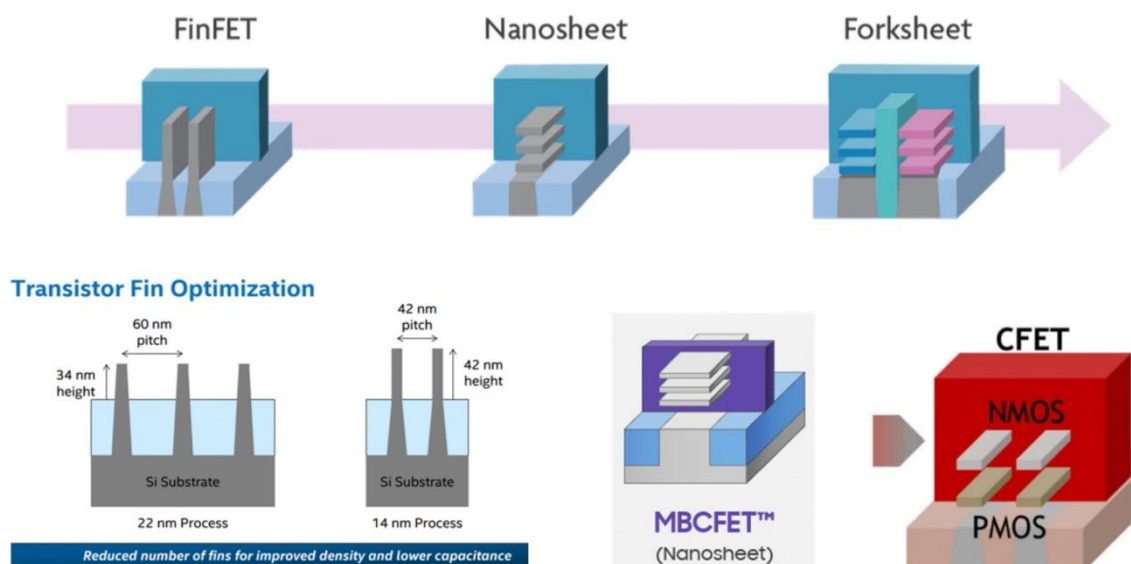


Рисунок 2 – Технологии транзисторов forksheet (вилочно-листовые FET), FinFET и GAAFET

Технология 3 нм, по которой сегодня производится интегральные микросхемы на нескольких фабриках в мире, позволяет обеспечить размещение 290 млн транзисторов на 1 мм². При среднем размере современной микросхемы в несколько квадратных сантиметров, на ней содержится порядка десяти миллиардов транзисторов. Для реализации такой технологии применяется атомарно-слоевое осаждение, атомарно-слоевое травление, позволяющее контролировать производство с точностью одного атомного слоя.

В таблице 1 представлена дорожная карта развития микроэлектроники, где на 15 лет вперед показано, как будут меняться топологические размеры и какой сложности будут создаваться микросхемы. На протяжении последних 40 лет она в целом подтверждала свою точность, однако каждые 2-3 года в нее вносились корректировки в основном за счет более агрессивного развития технологий и колоссального прогресса, произошедшего в индустрии за последние годы.

Таблица 1 – Дорожная карта развития технологий в микроэлектронике

| Дорожная карта развития технологий | Год выхода в производство (* – прогноз) | | | | | | | |
|------------------------------------|---|------|------|--------|-------|--------|--------|--------|
| | 2019 | 2022 | 2024 | 2025* | 2028* | 2032* | 2035* | 2037* |
| Наименование техпроцесса | 5 нм | 3 нм | 2 нм | 1,5 нм | 1 нм | 0,7 нм | 0,5 нм | 0,2 нм |

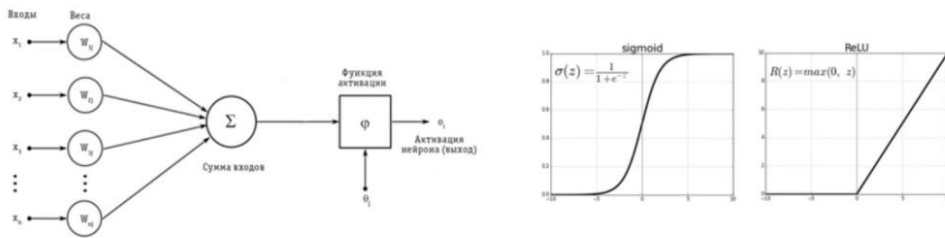
К примеру, еще два года назад предполагалось, что технология 2 нм будет освоена только в 2025 году, однако в сентябре 2023 года президент корпорации Intel объявил об освоении двухнанометровых технологий и начале выпуска продукции в 2024 году с переходом на технологию 1,8 нм к концу 2024 года.

В целом дорожные карты развития микроэлектронных технологий очень подробные и затрагивают не только топологический размер, но и характеристики интегральных схем, количество занятых в отрасли людей, объемы продаж. Сегодняшняя дорожная карта показывает, что в 2037 году топологический размер уменьшится до 0,2 нм, то есть 2 ангстрема. При этом количество транзисторов в микросхеме вырастет многократно.

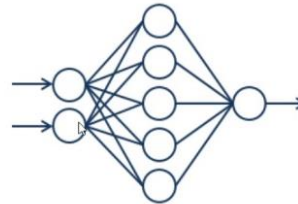
Человеческий мозг в среднем содержит 70...80 млрд нейронов. Топологический размер в 3 нм позволяет разместить на чипе 100 млрд транзисторов – больше, чем нейронов в головном мозге человека. К 2035 году микросхема средних размеров при технологии 0,5 нм позволит разместить чуть более трех триллионов транзисторов. Если мы посмотрим на 2037 год, то, соответственно, транзисторов будет уже под 6 триллионов.

К чему это приводит? Сегодня очень много говорят про «искусственный интеллект», но это больше философское название. Специалисты называют эту технологию «нейронные сети» или «машинное обучение». В последние годы мы все почувствовали их бурное развитие. Почему это произошло? Цифровая модель нейрона довольно простая (рисунок 3) и не изменилась за последние 40 лет. Различные сигналы на входе суммируются, и если достигают больших пороговых значений, то нейрон открывается и пропускает сигнал. Если же сигнал оказывается меньше порога, то он не пропускается. Все остальное связано лишь с различной архитектурой нейронных сетей.

Благодаря развитию микроэлектроники, каждые 10 лет вычислительная производительность увеличивается в 1000 раз. Сейчас мы знаем эксофлопные машины, потом будут зетафлопные (рисунок 4). Если за 30-летний период каждые 10 лет производительность увеличивалась в 1000 раз, то общее увеличение производительности за этот период произошло в миллиард раз. То есть 30 лет назад решение определенной задачи теми мощностями заняло бы 10 лет, тогда как сегодня ее решение заняло бы всего 0,3 секунды.



- **Нейроны**
- **Связи и их веса** (сумма взвешенных входов – активация нейрона)
- **Функция активации нейрона**
- **Последовательность слоев**

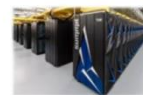


Применение GPR

{входной, скрытый, выходной}

Рисунок 3 – Нейросетевой подход в основе решения задач искусственного интеллекта

1. Развитие суперкомпьютеров – рост производительности каждые 10 лет в среднем в 1 000 раз.



2. Степень сложности интегральных схем – постоянное увеличение плотности размещения транзисторов на примере процессоров:



Рисунок 4 – Прогресс в области вычислений

Нейронные сети сегодня начали активно входить в нашу жизнь потому, что время, необходимое для проводимых ими вычислений, уже сопоставимо с реальным временем. То есть мы в реальном времени уже можем взаимодействовать с нейронными сетями и получать от них мгновенный ответ. Это и называют сегодня «искусственным интеллектом». Однако на самом деле мы находимся еще только на начальном этапе его развития. Сегодня развитие нейронных сетей идет сразу по многим направлениям, и мы уже можем выделить следующие предпосылки нового мощного рывка в этой области:

1. За счет появления новых моделей нейронов будет усложняться работа нейронных сетей.

Сегодня наиболее популярна так называемая «спайковая» модель нейрона (рисунок 5), когда получаемый сигнал должен попасть в определенный временной интервал. Если сигнал приходит в определенное время, то он пропускается, а если не вовремя, то он не пропускается. Это более современная модель нейрона.



Рисунок 5 – Реализация нейросети на спайковой модели нейрона

2. Реализация «не-фон-Неймановской» архитектуры процессоров.

Сегодня все выпускаемые процессоры основаны на «фон-Неймановской» архитектуре и не предназначены для работы с большими базами данными, так как в рамках этой архитектуры данные и команды хранятся в одной области и поступают в процессорную часть по одной шине, из-за чего возникает конфликт и задержки по времени. Поэтому разрабатывается около 20 различных вариантов новой «не-фон-Неймановской» архитектуры процессоров, которые обходили бы эту проблему и были более оптимизированы для работы с большими базами данных.

3. Мемристоры: замена классической памяти.

В микропроцессорах значительную площадь (до 40 %) занимает память, в которой хранятся данные. На один бит памяти необходимо шесть комплементарных транзисторов, потребляющих много энергии. Кроме того, обращение к памяти требует постоянной синхронизации по такту. Сегодня всеми отраслевыми исследовательскими организациями ведутся колоссальные работы над

новыми видами памяти (EFRAM, MRAM, RERAM и др.), которые энергонезависимы, занимают меньше площади и к которым можно обращаться вне зависимости от тактов работы процессора.

4. Минимизация интегральных схем. Увеличение плотности за счет 3D-технологий.

Для того, чтобы сделать микросхемы более компактными, помимо уменьшения топологического размера применяются еще 3D-технологии, позволяющие размещать элементы не только на плоскости, но и «в объеме». Это приводит к увеличению плотности их размещения транзисторов и кратному росту их количества.

Решение этих задач приведет к тому, что производительность нейронных сетей за это десятилетие увеличится не в тысячу раз, а, по нашим оценкам, в несколько десятков тысяч раз. Мы сегодня находимся на пороге настоящего бума применения нейронных сетей или машинного обучения для наших нужд.

Сегодня в качестве возможной альтернативы традиционным микроэлектронным технологиям довольно часто рассматриваются возможности квантовых и фотонных вычислений. Для оценки возможности и эффективности их применения необходимо всесторонне рассмотреть их особенности.

При применении квантовых технологий в первую очередь опираются на результаты и достижения в области атомной физики, квантовой оптики, физики конденсированного состояния, физики и техники измерений, информатики и теории алгоритмов, технологий наноэлектроники, наук о материалах. Отличие квантовых технологий от традиционной микроэлектроники характеризуется необходимостью учитывать некоторые особенности квантовых состояний: суперпозиции, запутанности и принципиально вероятностного статистического характера квантовых измерений.

Под квантовыми технологиями мы сегодня понимаем, во-первых, особые квантовые состояния, а во-вторых, возможность управлять квантовым объектом. Если в микроэлектронике мы имеем дело с большим количеством сигналов, передаваемых электронами, которые находятся в некогерентном состоянии, то в квантовых вычислениях должно быть когерентное состояние – только в этом случае мы можем управлять квантовым объектом. Для квантовых вычислений существуют специализированные платформы:

ионные, сквиды, на нейтральных атомах, фотонные и ряд других, которые активно развиваются.

Однако в реализации квантовых вычислений существует одна очень принципиальная проблема, ограничивающая их эффективность – коррекция ошибок, которая происходит не как в классической микроэлектронике.

Хотя квантовый вычислитель также, как и классический, использует Булеву алгебру, где значениями переменных являются значения истинности, обычно обозначаемые «1» и «0» соответственно, в квантовых технологиях используется такое понятие, как кубит, который может быть в состоянии ноль или единица, или в обоих сразу. Поэтому для правильной коррекции ошибок необходимо придавать запутанное состояние как можно большему количеству кубитов.

До 1994 года, когда впервые были сформулированы коды коррекции ошибок, считалось, что алгоритма исправления ошибок не существует. Сегодня появилось несколько подходов к решению этой проблемы:

1. Случайный разброс параметров кубитов.
2. Точность измерения базисных проекций кубитов.
3. Взаимодействие кубитов с окружением:
 - а) Квантовые вакуумные флуктуации;
 - б) Тепловые флуктуации;
 - в) Образование запутанности компьютерного кубита и состояния окружения.

Запутывание большого количества кубитов приводит к тому, что на один логический кубит необходимо запутывать сотни тысяч физических кубитов, держа при этом всю систему в когерентном состоянии, а это очень сложно и на текущий момент технически недостижимо. Поэтому сегодня, когда говорится о квантовых вычислениях, речь идет о так называемых «шумных компьютерах», где для коррекции ошибок используется другой подход: либо делается много повторений вычислений и затем по корреляционным полям ищут ошибку, либо применяется так называемый метод «смягчения ошибок».

Специалисты компании Microsoft, в рамках Ассоциации вычислительной техники провели анализ практических задач, в которых возможно реальное достижение квантового преимущества

(quantum advantage). В качестве классического чипа был выбран процессор NVidia A100 (технология 7 нм – TSMC, 54,2 млрд транзисторов, площадь кристалла в 826 мм²). В качестве реперной точки для квантовых вычислений исходили из оптимистического предположения о гипотетическом квантовом компьютере, который может быть доступен в ближайшие десятилетия с 10 000 логических кубитов, с исправлением ошибок, временем логической операции 10 мкс., возможностью одновременного выполнения логических гейтов на всех кубитах и реализацией связи каждого кубита с каждым (all-to-all connectivity) для обеспечения отказоустойчивости двухкубитных вентиляей.

Детальное рассмотрение работы такого квантового компьютера показало, что в ситуации наличия большого количества данных на входе, особенно характерных для нейронных сетей, у квантовых компьютеров нет никакого преимущества перед классическими. То есть при работе с большими базами данных никакого «квантового превосходства» получить не удастся. Таким образом было установлено, что если квантовые компьютеры и будут иметь преимущество, то только в случае малого количества данных и большого количества вычислений.

По данным Microsoft, «квантовое ускорение» выглядит следующим образом: время, необходимое для решения определенных задач с помощью квантовых алгоритмов, увеличивается медленнее, чем у любого известного классического алгоритма, по мере увеличения размера задачи N (рисунок 6). Однако на практике необходимо нечто большее, чем просто асимптотическое ускорение. Время перехода, при котором реализуется квантовое преимущество, должно быть достаточно коротким, а размер задачи перехода – не слишком большим.

Результаты моделирования также показали, что любую задачу, ограниченную доступом к классическим данным (например, задачу поиска в базе данных), классические компьютеры будут решать заведомо быстрее. Точно так же потенциально экспоненциальное квантовое ускорение в задачах линейной алгебры, например, при решении систем линейных уравнений, исчезает, когда соответствующая матрица должна быть загружена из классических данных или, когда должен быть прочитан полный вектор решения.

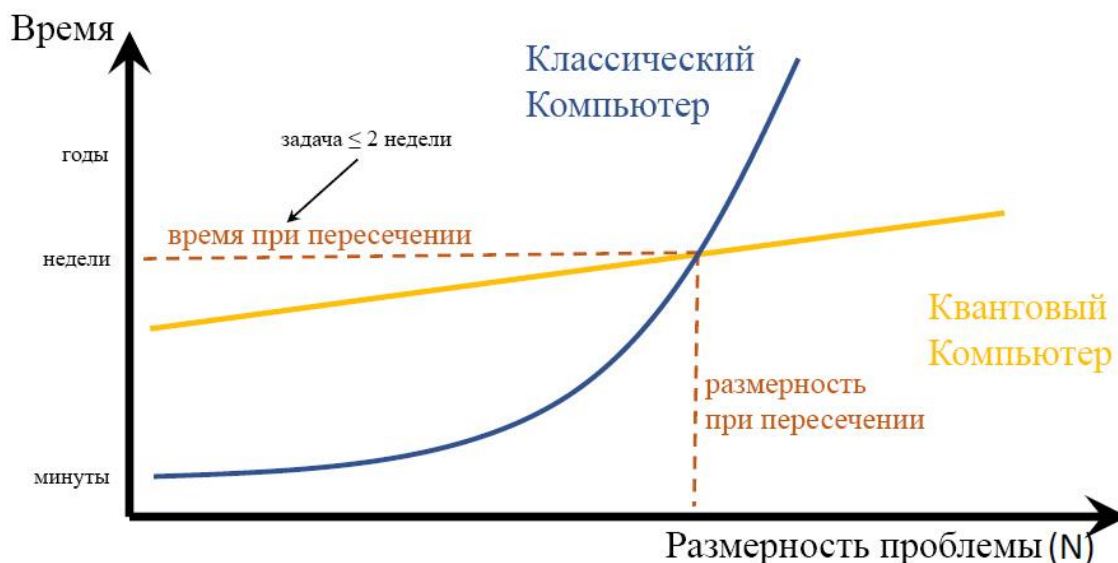


Рисунок 6 – Визуализация квантового ускорения [5]:

(ось времени масштабирована так, что квантовый алгоритм представляет собой прямую линию)

Таким образом можно сделать следующие выводы:

- большинство современных квантовых алгоритмов могут не достичь практического уровня;
- материаловедение и химия имеют огромный потенциал, поэтому можно надеяться, что будут изобретены практичные квантовые алгоритмы;
- из-за ограничений входной и выходной пропускной способности квантовые компьютеры будут оптимальны для задач больших вычислений (big compute) с небольшим количеством данных, но не для задач с большим количеством данных (big data).

Еще одно направление вычислений – фотонный тензорный процессор. Это процессор, оперирующий тензорами, представленными в виде многомерных массивов чисел для выполнения матричного умножения и свертки матриц. Благодаря высокой скорости передачи данных с помощью фотонов света появилась идея использовать этот эффект для вычислений в фотонных процессорах. Этот метод отличается от квантовой платформы тем, что здесь не нужно иметь фотон в когерентном состоянии, что позволяет избежать многих ошибок. Также благодаря этой особенности фотонные вычислители потребляют на несколько порядков меньше

энергии. В этой связи основной областью применения тензорных фотонных процессоров могут стать:

- нейронные сети для поиска объектов на фото и видео, распознавания образов, генерации и дорисовки изображений;
- реализация метода опорных векторов для задач машинной классификации данных;
- линейная и полиномиальная регрессия для задач прогнозирования, анализа стоимости ценных бумаг, машинной медицинской диагностики.

В числе преимуществ тензорных процессоров можно выделить:

- большую частоту выполняемых операций MAC (сравнение электронных и фотонных процессоров проводят по «плотности вычислений» (compute density) – числу операций MAC в секунду (multiply-accelerate, умножение-накопление), отнесенного к единице площади – MACs/mm²);
- менее жесткие требования к топологическим нормам;
- прямую обработку изображения без преобразования в электричество.

Выводы:

1. Явной альтернативы классическим микроэлектронным технологиям в горизонте 15-20 лет не наблюдается.

2. Все альтернативы классическим микроэлектронным технологиям, которые существуют на сегодняшний день, могут рассматриваться в лучшем случае как гетерогенные системы, то есть как дополнение к существующим. Фотонный вычислитель может более эффективно выполнять тензорные вычисления, квантовый вычислитель – операции по моделированию открытых квантовых систем.

3. Быстро развивающиеся в настоящее время новые технологии позволят в будущем расширить возможности классических микроэлектронных технологий для отдельных применений.

4. Главным направлением развития вычислений в обозримом будущем останется микроэлектроника, которая на сегодняшний день продолжает развиваться. Этот тренд сохранится как минимум еще на ближайшие 20 лет и будет менять наш мир и наше представление о нем, благодаря достижениям, связанным с

возможностью интегральных схем. Эффективность вычислительных возможностей будет расти и применяться не только в линейных вычислениях, но и в области распознавания образов, при анализе объектов технического зрения и обработки больших массивов неоднородных данных.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Горбацевич, А. А. Эффективный интерференционный механизм управления проводимостью элементов молекулярной наноэлектроники / А. А. Горбацевич, Г. Я. Красников, Н. М. Шубин // Российские нанотехнологии. – 2019. – Т. 14, № 9-10. – С. 101-107. – DOI 10.21517/1992-7223-2019-9-10-101-107.

2. Torsten, H. Disentangling Hype from Practicality: On Realistically Achieving Quantum Advantage / Hoefler Torsten, Häner Thomas, Troyer Matthias // Communication of the ACM. – 2023. – Vol. 66, No. 5. – Pp. 82-87.

3. Обоснование технического облика агротехники и стратегических подходов ее проектирования / Г. Я. Красников, О. Н. Дидманидзе, П. В. Сиротин, Е. П. Парлюк // Чтения академика В. Н. Болтинского : Сборник статей, Москва, 25–26 января 2023 года. – Москва: ООО «Сам Полиграфист», 2023. – С. 10-32.

Об авторе:

Красников Геннадий Яковлевич, Президент Российской академии наук (119991, Российская Федерация, Москва, Ленинский проспект, 14), доктор технических наук, профессор, академик РАН.

About the author:

Gennadij J. Krasnikov, President of the Russian Academy of Sciences (119991, Russian Federation, Moscow, Leninsky Prospekt, 14), D.Sc. (Engineering), professor, Academician of the Russian Academy of Sciences.