ДАЙДЖЕСТ №6

ДЕКАБРЬ 2022

ЗНТЦ

news

zntc@zntc.ru | zntc.ru | +7 (499) 720 69 73

Основные темы конференции IEDM в 2022 году

«Новый Ara» для векторных вычислений: высокоэффективный вектор RISC-V V 1.0 с открытым исходным кодом

Ориентация на C-диапазон с помощью сверхвысоковольтных транзисторов типа HEMT

К 2027 году рынок подложек для компаундных полупроводников составит 2,4 миллиарда долларов

Зеленоградский нанотехнологический центр

ДАЙДЖЕСТ №6, декабрь 2022

Уважаемые коллеги!

Совсем скоро 2022 год подойдёт к концу. Перед вами последний выпуск дайджеста в этом году. За полгода реализации проекта мы подготовили и опубликовали шесть выпусков, в которых собрали для вас материалы о новых направлениях и тенденциях развития электроники в мире, анализы рынка и многое другое.

Развитие архитектуры RISC-V в нашей стране особенно актуально при действующих санкциях, действие которых связано как с процессами производства, так и с процессами проектирования, поэтому один из материалов посвящён новому высокоэффективному вектору RISC-V V 1.0 с открытым исходным кодом.

Кроме этого, узнаем, какие темы специалисты обсуждали в начале декабря на конференции IEDM: 2D-транзисторы, память nvRAM, универсальные транзисторы типа HEMT на GaN-на-алмазе и многое другое. Снова расскажем об одном из мировых трендов – использовании GaN для силовой электроники, в частности о его использовании в сверхвысоковольтных транзисторах типа HEMT. А также посмотрим перспективы развития рынка подложек для компаундных полупроводников.

Надеемся, что информация окажется полезной для Вас и Ваших коллег.

Следующий выпуск выйдет уже в 2023 году. Желаем спокойного и продуктивного завершения этого рабочего года и поздравляем с наступающими праздниками!

Приятного чтения!

СОДЕРЖАНИЕ

01

Ключевые слова: 2D-материалы, память, силовая электроника, высокоскоростные устройства, сенсорика

Основные темы конференции IEDM в 2022 году

На конференции IEDM, которая состоялась 3–7 декабря в США, обсуждались разные темы, среди основных: 2D-транзисторы на основе MoS2 от TSMC, самая маленькая и энергоэффективная автономная память nvRAM от Samsung, сверхвысоковольтные устройства с суперпереходом на основе SiC и универсальные транзисторы типа HEMT на GaN-на-алмазе.

«Сейчас полупроводниковая промышленность играет в мире более важную роль, чем когда-либо, и это делает разработки, представленные в этом году на конференции, достаточно значимыми, потому что они в конечном итоге приведут к созданию продуктов, которые сделают нашу жизнь лучше», — заявил один из организаторов мероприятия. «Среди тенденций, которые прослеживались в докладах, можно отметить растущий интерес к использованию систем 2D-материалов для продвинутых, чрезвычайно масштабных устройств».

В этой статье освещаются некоторые из статей и докладов, которые были представлены в области КМОП, памяти, силовых устройств, обработки изображений, высокочастотной электроники и других.

Масштабирование КМОП: 2D-устройства

Почти идеальное подпороговое колебание с 2D MoS2: однослойные дихалькогениды переходных металлов (ДПМ) — это так называемые 2D материалы, потому что они ультратонкие, толщиной всего в один слой атомов. Поскольку транзисторы состоят из нескольких слоёв материалов, использование однослойных ДПМ потенциально может привести к созданию устройств меньшего размера. Однако ключевым недостатком является то, что на них довольно сложно наносить диэлектрические слои или изоляторы без точечных отверстий. Это затрудняет их внедрение в слои материалов, образующих затвор транзистора. Команда под руководством TSMC рассказала, как они интегрировали диэлектрики на основе гафния (образованные путём осаждения атомных слоев) с монослойным ДПМ-материалом MoS_2 для создания nFET с верхним затвором с толщиной диэлектрика 3,4 нм и электрически эквивалентной толщиной оксида ~1 нм (Рис.1). Значение, называемое подпороговым колебанием, является ключевым для МОП-транзисторов, потому что чем оно выше, тем меньший ток протекает через устройство, когда оно выключено. Созданные в рамках исследования устройства имели почти идеальное подпороговое колебание <70 мВ/дек.

Первое 2D устройство на основе GAA: кремниевые нанолистовые транзисторы считаются наиболее перспективными для использования в устройствах следующего поколения, поскольку обеспечивают

улучшенный электростатический контроль, относительно высокий управляющий ток и возможность производства устройств различной ширины. В настоящее время масштабирование длины затвора и высокий электростатический контроль достигаются за счёт утончения кремниевого канала, но в будущем это можно будет делать с помощью монослойных ДПМ. Хотя кремниевые нанолисты, интегрированные в монослойный ДПМ в качестве материала канала, являются перспективными, производительность таких устройств, так и процессы их изготовления ещё предстоит изучить.

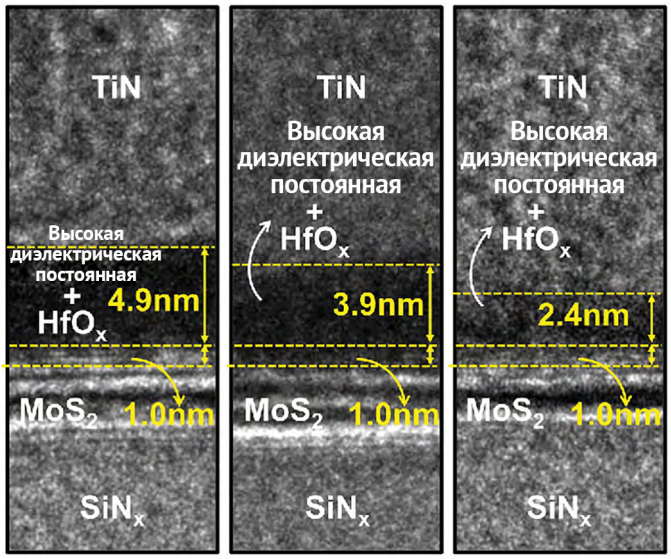


Рисунок 1. Изображения транзистора nFET в просвечивающем электронном микроскопе (ПЭМ)

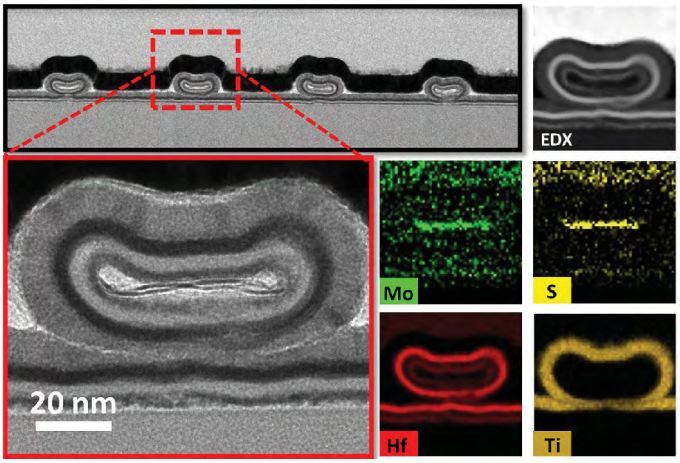


Рисунок 2. Два крупных изображения представляют собой поперечное сечение ПЭМ монослойного нанолистового устройства MoS_2 с блоком затворов вокруг канала. Меньшие изображения показывают соответствующее распределение элементов методом энергодисперсионной рентгеновской спектроскопии.

Специалисты TSMC описали возможный интеграционный поток для монослойных 2D-устройств и решения критических проблем в многослойных 2D-листовых архитектурах. Они использовали эти идеи для создания первого в мире однослойного полевого транзистора с нанолистами MoS_2 в конфигурации с затвором по всему периметру (GAA) (Рис. 2).

Контакты p-типа с низким сопротивлением для 2D-материалов: изготовление металлических контактов с низким сопротивлением для 2D-материалов является проблемой для их использования в КМОП-транзисторах. Был достигнут прогресс с контактами n-типа для использования с полевым транзистором n-типа (nFET), но контакты p-типа с низким сопротивлением для транзистора p-типа (pFET) более сложны в производстве из-за электротермодинамических условий, возникающих между металлами p-типа и 2D-материалами. Это приводит к возникновению барьера Шоттки: чем больше высота барьера Шоттки, тем больше сопротивление протеканию тока. Специалисты компании TSMC провели симуляционные исследования и компьютерное моделирование, чтобы изучить свойства различных материалов для использования в качестве контактов p-типа с 2D-материалом WSe_2 (Рис. 3).

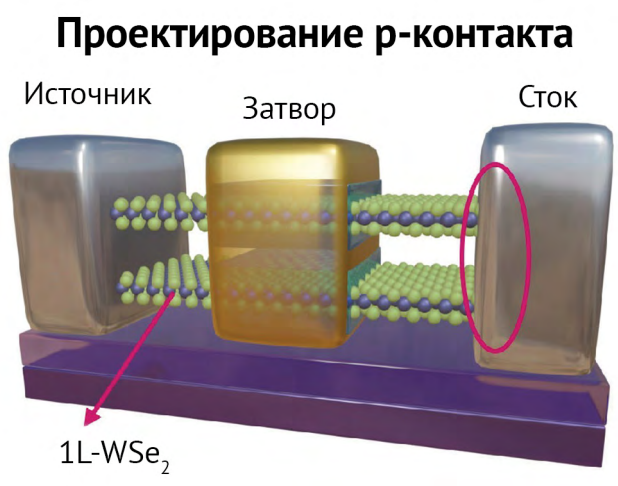


Рисунок 3. Схема однослойного нанолистового транзистора на основе WSe.

В результате этих исследований они определили две стратегии к созданию безбарьерных (т.е. омических) контактов p-типа с низким сопротивлением: металлические контакты Ван-дер-Ваальса с использованием 1T-TiS_2, металлические контакты с использованием различных материалов, например, $Co_3Sn_2S_2$ показал теоретическое контактное сопротивление всего 20 Ом·мкм.

Технологии памяти

Самая маленькая в мире и самая энергоэффективная MRAM: в настоящее время энергонезависимая память с произвольным доступом (nvRAM) используется только на узкоспециализированных рынках, но к ней растёт интерес как к решению с низкой утечкой рабочей памяти (например, для кэш-памяти) для устройств, которые полагаются на массивный сбор и анализ данных, такие как AIoT (слияние концепции Интернета вещей и ИИ) и периферийных устройства для ИИ-вычислений. Снижение общего энергопотребления имеет решающее значение в этих устройствах. Компания Samsung представила автономную энергонезависимую память nvRAM (Рис. 4), основанную на 28-нм технологии встроенной MRAM. Она продемонстрировала лучшую в своем классе энергию записи (25 пДж/бит), а также требования к активной мощности всего 14 мВт (чтение)/27 мВт (запись) при скорости передачи данных 54 МБ/с. Данный тип памяти также имеет наименьшие размеры корпуса (30 мм2 при 16 Мб) и практически неограниченный срок службы (>1E14 циклов). Ключевой частью архитектуры устройства является магнитный туннельный переход; его масштабирование до технологии 14-нм FinFET привело к снижению времени чтения в 2,6 раза.

Магнитный туннельный переход ячеек

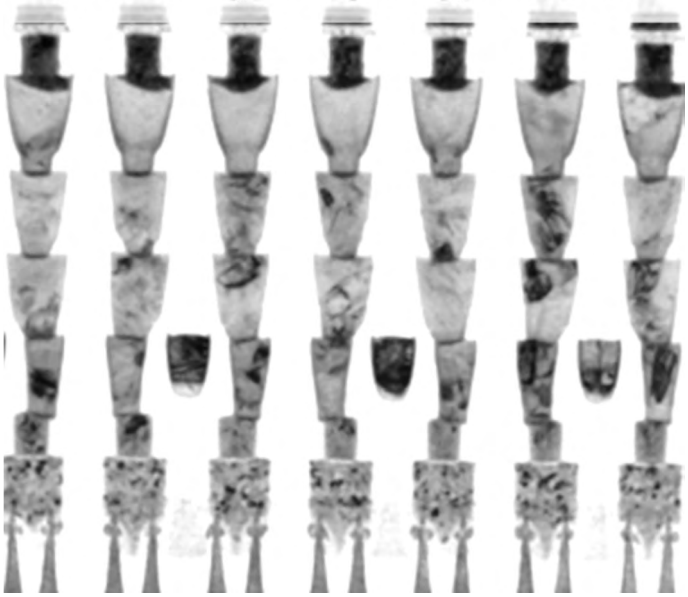


Рисунок 4. Поперечное сечение массива битовых ячеек eMRAM, встроенного в 14-нм логическую платформу в ПЭМ.

Превосходные ферроэлектрические характеристики: по мере того, как электронные системы становятся всё более сложными, промышленность ищет ОЗУ с быстрым временем доступа (<10 нс), высокой выносливостью (> 1014 циклов) и достаточно хорошим сохранением данных в переходных условиях питания, например, при переключении системы в «спящий» режим. Активно идут исследования структуры «один транзистор — один конденсатор» (1T-1K) на основе сегнетоэлектрических материалов (сегнетоэлектрические материалы имеют поляризацию (значения «0» и «1»), которую можно изменить при приложении электрического поля). В частности, специалисты активно изучают цирконат гафния (HZO), так как он совместим с КМОП процессами и пригоден для масштабирования ниже 10 нм. Компания IMEC значительно улучшила сегнетоэлектрические характеристики HZO. Они изготовили как двухслойные, так и трёхслойные конденсаторы HZO, используя затравочный слой TiO_2 толщиной 1 нм и/или верхний слой Nb_2O_5 толщиной 2 нм, а также верхний и нижний электроды из TiN (Рис. 5). В зависимости от используемых химических прекурсоров трёхслойные устройства продемонстрировали либо значительно повышение выносливость до 1011 циклов, либо рекордно высокое сохранение данных (т.е. «остаточную поляризацию») = 66,5 мкКл/см2 после 3x106 циклов при 3 МВ/см.

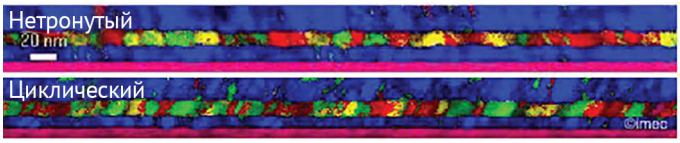


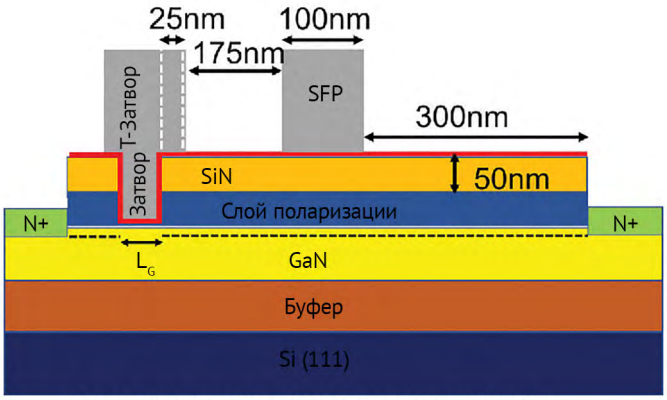
Рисунок 5. Верхнее изображение показывает количество фазы, или возможности поляризации, присутствующие как в чистом, так и в 1010-цикловом однослойном устройстве.

Силовые устройства

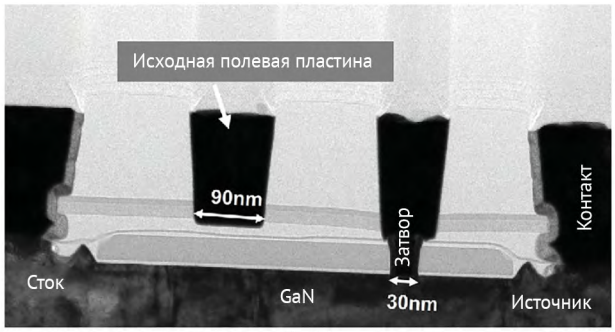
Устройства со сверхвысоким напряжением на основе SiC: полупроводники, такие как силовые МОП-транзисторы, БТИЗ (биполярный транзистор с изолированным затвором) и переключатели/диоды на основе SiC, являются основой приводов электродвигателей и систем преобразования энергии, используемых в морских, железнодорожных, энергетических и других крупномасштабных устройствах. Данные системы ограничены низкими частотами переключения из-за потерь при переключении и диэлектрических потерь на электропроводность. Если бы они работали на более высоких частотах, то можно было бы разработать более компактные, более эффективные, более высоковольтные и менее дорогие системы.

Исследователи General Electric сообщили о сверхпереходных МОП-транзисторах и диодах на основе SiC, образованных с помощью технологии имплантации ионов очень высокой энергии, в результате чего будут созданы устройства с удельным сопротивлением во включенном состоянии ниже однополярного или теоретического предела SiC, что приведёт к меньшим потерям. Была продемонстрирована технология создания PiN-диодов с суперпереходом на SiC на 2 кВ и диоды Шоттки с суперпереходом на SiC на 3,8 кВ. Устройства представляют вариант будущей реализации переключателей 3-20 кВ для силовой электроники.

Рекорд Fmax для 40В на основе GaN-на-Si: Поскольку требования к плотности мощности серверов, графических платформ и других высокопроизводительных систем продолжают расти, необходимы новые варианты повышения эффективности и плотности, а также поддержки более высокой скорости передачи данных. GaN перспективен, так как может работать при более высоких напряжениях и частотах, чем у Si, и с меньшими потерями. В прошлом году компания Intel представила высокомасштабируемый, высокопроизводительный N-МОП-транзистор на основе GaN-на-Si (Рис. 6).



Транзистор на основе GaN в электронном режиме (исходная полевая пластина)



Транзистор на основе GaN в электронном режиме (исходная полевая пластина)

Рис. 6. Два изображения: (сверху) структура транзистора с полевым покрытием и (снизу) микрофотография через ПЭМ.

В этом году исследователи компании рассказали, как они масштабировали технологию и повысили её производительность за счёт интеграции полевой пластины субмикронной длины для управления электрическим полем внутри устройства. Они достигли рабочего напряжения 40 В и Fmax 680 ГГц (Ft = 130 ГГц), что является рекордом для 300-мм устройства GaN-на-Si.

Высокоскоростные устройства

Рекорд скорости для комбинированной технологии изготовления ИС на биполярных и КМОП-транзисторах (BiCMOS): исследователи Global-Foundries обсудили биполярные транзисторы на основе SiGe с гетеропереходом (биполярный гетеротранзистор, БГТ), интегрированные в 45-нм техпроцесс BiCMOS. Они достигли самого высокого значения Fmax (частоты колебаний), когда-либо зарегистрированного для БГТ в любой технологии кремний-на-изоляторе. В тестовой схеме кольцевого генератора время задержки распространения сигнала на каскад составила всего 1,76 пс. Измерения силовой ячейки каскода (т. е. измерения усилителя) показали усиление >18 дБ на частоте 100 ГГц. Эта технология монолитно объединяет БГТ на SiGe, высокоскоростные КМОП и пассивные устройства с малыми потерями на одной платформе и перспективна для высокочастотных устройств в таких областях, как например, связь 5G/6G, спутниковая связь, автомобильные радары и многое другое.

Рекорд скорости для транзисторов типа HEMT с квантовыми ямами на InGaAs: развивающиеся технологии беспроводной связи миллиметрового диапазона, такие как 6G, потребуют более высоких рабочих частот (~ 300 ГГц со скоростью передачи данных, приближающихся к 0,1 Тбит/с), чем это возможно сегодня. Чтобы достичь такой производительности, исследователи изучают устройства, изготовленные из InGaAs и других материалов групп III-V, которые работают быстрее, чем устройства на Si. Группа исследователей из Кёнбукского национального университета (Корея) представила транзисторы на InGaAs с длиной затвора 20 нм, Fmax = 1,1 ТГц и Ft = 0,75, а также описала наилучшее соотношение Fmax и Ft в любой транзисторной технологии и самом высоком Ft в любом полевом транзисторе. Такая производительность связана с индуцированного стоком понижением барьера (т.е. паразитным эффектом, который потенциально может привести к преждевременному включению транзистора) до 60 мВ/В.

Достижения в области визуализации и сенсорного обнаружения

На пути к сверхминиатюрным пикселям: команда STMicroelectronics рассказала об инновационном трёхуровневом КМОП-датчике изображения на задней панели. Он имеет двухуровневую последовательную интеграцию слоёв фотозатвора и пикселей, гибридно связанных со слоем логической схемы, без проблем с выравниванием. Последовательное наложение слоёв фотозатвора и пиксельного транзистора привело к созданию пикселя без мерцания с высоким динамическим диапазоном (106 дБ) и шагом 1,4 мкм. Тонкоплёночная КНД-архитектура пиксельного транзистора и имеет преимущества с точки зрения масштабируемости и производительности, что позволит создать сверхминиатюрные пиксели, которые можно будет интегрировать в сложные архитектуры систем-на-кристалле.

Оптический датчик на основе Ge: в настоящее время у специалистов вызывает интерес использование Ge для фотодетектирования, и в статье Корейского передового института науки и технологий описывается его использование в новом приборе для спектроскопии. Средний инфракрасный диапазон светового спектра (2–15 мкм) всё чаще используется для оптического сенсорного обнаружения для, например, мониторинга окружающей среды. Благодаря колебательному возбуждению их межмолекулярных связей технологии среднего инфракрасного диапазона стали эффективными методами количественного анализа. Однако платформы кремниевой фотоники, основанные на технологии «кремний-на-изоляторе», не могут охватить весь спектр среднего инфракрасного диапазона. Вместо этого исследователи построили газовый датчик Ge-на-изоляторе (Ge-OI) в среднем инфракрасном диапазоне, который имеет превосходные характеристики (низкие потери 1,88 дБ/см), высокое оптическое ограничение (чувствительность и предел обнаружения (0,0885%/ч/млн и 8,5 ч/млн соответственно).

Растущее значение управления температурным режимом

Повышение температуры снижает производительность и надёжность электронных устройств, а эффективное управление температурным режимом становится как никогда важным по мере того, как устройства уменьшаются в размерах.

Универсальные алмазные плёнки рассеивают тепло: устройства на основе GaN имеют большой потенциал в качестве твердотельных усилителей мощности для высокочастотных устройств, таких как 5G и выше, учитывая их более высокую удельную мощность на любой рабочей частоте по сравнению с другими материалами. Но высокая плотность выходной мощности приводит к значительному самонагреву устройства, что серьёзно ограничивает ВЧ-производительность и надёжность, особенно для сверхмасштабируемых устройств. Отвод тепла на уровне корпуса не всегда эффективен. Алмаз является отличным проводником тепла, и учёные из Стэнфорда использовали поликристаллические алмазные плёнки в качестве теплораспределителей, чтобы облегчить удаление горячих точек на уровне устройства. Новый метод затравки с помощью полимера позволил им полностью окружить транзистор типа HEMT на основе GaN изотропным тонкоплёночным алмазом, выращенным методом химического осаждения из паровой фазы, обеспечивая полное покрытие боковых стенок и сводя к минимуму общее тепловое сопротивление подложки. Они использовали три различных метода определения характеристик для контроля температуры каналов устройств и обнаружили, что транзистор типа HEMT на основе GaN с 500-нм круговым алмазом имеет более низкие пиковые и средние температуры канала по сравнению с контрольными устройствами без алмаза. Пиковые температуры были на 98 ± 19°C ниже при мощности постоянного тока 9,5 Вт/мм. Они также имели более равномерный температурный профиль и меньше горячих точек вдоль электрода затвора.

ВЧ-устройства на основе GaN и InP нагреваются на 30–70 % больше, чем предполагалось: Самонагрев в ВЧ-устройствах вызывает особую озабоченность, учитывая существенные требования к мощности ВЧ-цепей. Теплопередача в таких устройствах всегда была диффузионной по своей природе и осуществлялась за счёт проводимости, когда электроны или фононы проходят через объёмный материал и передают тепловую энергию другим частицам, с которыми они сталкиваются. Диффузию можно эффективно анализировать с помощью ПО для моделирования.

Но по мере того, как устройств а уменьшаются, размеры их элементов становятся эквивалентными средним расстояниям, которые проходят электроны и фононы, прежде чем они достигнут другой частицы, и возможностей для рассеивания тепла становится меньше. Обычные компьютерные модели не могут это учитывать, поэтому по мере уменьшения размеров устройств оценки уровня самонагрева транзисторов становятся всё более неточными. Команда IMEC разработала новую структуру теплового моделирования, которая учитывает этот фактор. На основе метода Монте-Карло (метод статистического усреднения) они смоделировали тепловые потоки в ВЧ-устройствах на основе GaN и InPh и обнаружили, что пиковые температуры на уровне транзисторов в ВЧ-устройствах повышаются на 30–70 % больше, чем предполагалось. Эксперименты также были проведены.

Другие темы

Встроенные фильтры объёмной акустической волны и ВЧ-переключатели: Акустическая ВЧ-фильтрация необходима для интерфейсных модулей смартфонов для повышения чувствительности приёмника, снижения энергопотребления передатчика и работы с растущим числом частотных диапазонов (например, Wi-Fi, GPS, LTE, 5G). В современных смартфонах используется около 100 ВЧ-акустических фильтров, при этом фильтры объёмных акустических волн (OAB), созданные с помощью МЭМС процессов, преобладают в высокочастотном диапазоне. Однако по мере увеличения числа частотных диапазонов становится всё труднее вместить >100 фильтров размером с микросхему в компактный модуль. Кроме того, на производительность и стоимость системы сильно влияют повышенные потери из-за разводки проводов на многочисленных печатных платах. Потенциальное решение состоит в том, чтобы монолитно интегрировать OAB-фильтры с внешними схемами вертикально на одном кристалле вместо того, чтобы соединять их рядом через межсоединения ввода-вывода. Для этого исследователи из Сингапурского агентства A*Star разработали процесс монолитной 3D-интеграции на уровне 200-мм пластины. В процессе объединяются фильтры OAB на основе ScAlN с ВЧ-переключателями в КНД-структуре. Встроенный переключаемый фильтр имел отличные характеристики: работая на частоте 2,5 ГГц, он достиг вносимых потерь <3,6 дБ и полосы пропускания >120 МГц на компактной площади <1 мм2.

Понимание улавливания заряда в тонкоплёночных транзисторах на основе IGZO: В последние годы IGZO (индий-галлий-цинковый оксид) стал центром интенсивных исследований для использования его на завершающих операциях обработки полупроводниковых пластин или для DRAM. Было понятно, что устройства на основе IGZO могут соответствовать требованиям к производительности, но не хватало глубокого понимания фундаментальных физических механизмов, которые ограничивают надёжность при существующей толщине подзатворного диэлектрика и условиях эксплуатации. Здесь важно иметь в виду характеристику межфазных ловушек или механизмов деградации в материале. Команда IMEC разработала новый метод световой спектроскопии для изучения улавливания заряда в тонкоплёночных транзисторах на основе IGZO. Это позволило им смоделировать вольт-амперные характеристики (ВАХ) тонкоплёночного транзистора на основе IGZO при различных температурах и условиях освещения, что дало понимание механизмов деградации и выявило сложное взаимодействие между подзатворным диэлектрикам и IGZO-каналом. Этот метод можно будет использовать и с другими материалами с широкой запрещенной зоной, такими как GaN и SiC.

Источник

International Electron Devices Meeting 2022 Preview

Semiconductor Digest November/December 2022

02

Ключевые слова: RISC-V, ISA, вектор, эффективность

«Новый Ara» для векторных вычислений: высокоэффективный вектор RISC-V V 1.0 с открытым исходным кодом

Векторные архитектуры набирают обороты для высокоэффективной обработки задач с параллелизмом данных, управляемых всеми основными стандартами промышленной архитектуры ISA (RISC-V, Arm, Intel) и поддерживаемых основными чипами, например, Fujitsu A64FX на базе Arm SVE, на котором основан суперкомпьютер Fugaku, входящий в ТОП-500. Расширение RISC-V V недавно получило статус 1.0-Frozen.

Здесь мы представляем его первую реализацию с открытым исходным кодом, обсуждаем влияние новой спецификации на микроархитектуру проектирования на основе дорожек и даём представление об ориентированном на производительность проектировании связанных скалярно-векторных процессоров. Наша система обеспечивает более высокие характеристики мощности, производительности и площади, чем современные векторные устройства, использующие более старые версии RVV: площадь увеличена на 15 %, пропускная способность – на 6 %, а использование модулей с плавающей запятой (FPU) > 98,5 % на важных ядрах.

ВВЕДЕНИЕ

В ТОП-500 самых мощных суперкомпьютеров мира входят две группы устройств: системы, использующие производительность от ускорителей, и системы на базе многоядерных процессоров с мощными векторными блоками. В настоящее время лидером является Fujitsu Fugaku [1], [2] который находится в исследовательской лаборатории RIKEN в Японии. Fujitsu Fugaku имеет 159 000 узлов, каждый из которых оснащён 48-ядерным процессором Fujitsu A64FX, работающим на частоте 2,2 ГГц и поддерживающим архитектуру ISA Armv8.2-A с масштабируемым векторным расширением, в котором используется 512-битные векторы.

Sunway TaihuLight [3] также основан на процессорах с векторным расширением. Он имеет 26 000 процессоров Sunway SW26010, каждый из которых работает на частоте 1,45 ГГц с 4 кластерами по 64 вычислительных ядра с 256-битным векторным блоком в каждом.

Векторные процессоры находят применение не только в суперкомпьютерах. Последняя версия ISA от Arm v9 также включает в себя обновленное векторное расширение SVE2 [4] и, как ожидается, станет широко использоваться в смартфонах, а позже в микроконтроллерах, процессорах обработки данных в реальном масштабе времени и прикладные процессорах.

Вышеупомянутые архитектуры ISA являются закрытыми, поэтому их микроархитектуры остаются полностью непрозрачными и не дают возможности реализации с открытым исходным кодом. Это серьёзное препятствие для открытых инноваций. За последние несколько лет RISC-V хорошо зарекомендовал себя как современная и общедоступная альтернатива закрытым ISA, что привело к волне общедоступных реализаций и позволило обсудить новые пользовательские расширения ISA, а также их влияние на микроархитектуру и её потенциальную стандартизацию.

В этой работе мы сосредоточимся на векторном расширении RISC-V (RVV), впервые предложенном в 2015 году [5]. В течение последних шести лет его интенсивно обсуждали, дорабатывали. Сейчас версия v1.0 заморожена и открыта для всех [6]. Утверждение расширения является важно для сообщества RISC-V, поскольку и аппаратное, и программное обеспечение будут опираться на стабильную и стандартизированную, но все же открытую, архитектуру ISA.

Векторные команды работают с векторами переменного размера, длина и размер элементов которых могут быть установлены во время выполнения.

Таблица 1. ОБЗОР ВЕКТОРНЫХ ПРОЦЕССОРОВ RISC-V

	Название	Версия RVV	Объект	XLEN (бит)	Поддержка Float	VLEN (бит)	Разделённый ФВР (дорожки)	Открытый код
	Эта работа	1.0	ASIC	64	да	4096a	да	Да
[8]	SiFive X280	1.0	ASIC	64	Да	512	?	Нет
[9]	SiFive P270	1.0rc	ASIC	64	Да	256	?	Не
[10]	Andes NX27V	1.0	ASIC	64	Да	512	?	Не
[11]	Atrevido 220	1.0	ASIC	64	Да	128-4096	Да	Не
[12]	Vicuna	0.10	FGPA	32	Нет	128-4096	Нет	Да
[13]	Arrow	0.9	FGPA	32	Нет	?	Дас	Нет
[14]	Johns et al.	0.8	FGPA	32	Нет	32	Нет	Нет
[15]	Vitruvius	0.7.1	ASIC	64	Да	16384	Да	Нет
[16]	XuanTie 910	0.7.1	ASIC	64	Да	128b	Да	Нетd
[17]	RISC-V2	?	ASIC	?	Нет	256	Нет	Дае
[7]	Ara	0.5	ASIC	64	Да	4096b	Да	Да
[18]	Hwacha	Не указано	ASIC	64	да	515b	Да	Да

a - Параметрический VLEN. В этой работе мы выбрали 4096
b - VLEN на дорожку.
c – ФВР разделён горизонтально.
d – Векторный модуль закрытый.
e – Скалярное ядро закрытое.

Архитектура использует параллелизм устройств за счёт длинных, глубоких конвейерных шин данных и одиночным потоком команды и несколькими данными (архитектура SIDM, или один поток команд и множество потоков данных) в каждом функциональном блоке. Более того, одна векторная команда запускает вычисление всех элементов вектора. Первоначально это было введено в суперкомпьютерах Cray и иногда называется векторной обработкой в стиле Cray. Среди преимуществ этого подхода можно выделить: меньший размер кода, возможность избежать многократной выборки и декодирования одних и тех же команд, сокращение количества передач кэша команд и повышение эффективности системы. Кроме того, в этом случае улучшается переносимость кода. Сегодня векторные процессоры, подобные Cray, получили новое распространение благодаря гонке за энергоэффективными решениями и необходимости выполнять высокопараллельные задачи. Более того, благодаря своей гибкости и простой модели программирования они стали реальной альтернативой графическим процессорам (GPU - Graphics Processing Units) при работе с длинными векторами. Обратите внимание, что как Cray-, так и SIMD- подобные расширения часто называют векторными, например, Intel Advanced Vector Extension (AVX) и расширения Arm NEON. Однако эти команды с операндами фиксированного размера не являются векторными расширениями в соответствии с приведенным выше определением.

Эта работа представляет первую реализацию с открытым исходным кодом, включая аппаратное и программное обеспечение, RVV 1.0 ISA:

1) В ней описывается конструкция векторного модуля RVV 1.0 как сильно связанного блока расширения для процессора CVA6 RV64GC с открытым исходным кодом. Также представлены новый интерфейс и функции когерентности / непротиворечивости аппаратной памяти.
2) В ней представлено сравнение с векторным блоком RVV 0.5 [7] и анализ влияния, которое новая ISA RVV оказывает на микроархитектуры, использующие дорожки с разделённым файлом векторного регистра.
3) В ней оценивается показатели мощности, производительности и площадь (МПП) устройства, произведённого с использованием технологии полностью обеднённого кремния на изоляторе (FD-SOI) GLOBALFOUNDRIES 22FDX, и доказывается, что данная реализация делает их достаточно конкурентоспособными по сравнению с векторным устройством RVV 0,5.
4) В ней анализируется влияние, которое скалярный процессор оказывает на конечную достигнутую пропускную способность вектора, особенно в случае средних/коротких векторов.

II. СВЯЗАННЫЕ ИССЛЕДОВАНИЯ

Новый векторный модуль RVV 1.0 вдохновлен Ara [7], векторным модулем, реализующим RVV v0.5. Ara — это высокоэффективный векторный сопроцессор, разработанный в 2019 году, который работает в паре с процессором прикладного класса CVA6 (ранее Ariane) [19] и совместим с одной из первых версий RVV [20]. Ara достиг пикового использования в 97% с 256 × 256 матричным умножением с 16 дорожками, пропускной способностью 33 DP-GFLOPS и энергоэффективностью 41 DP-GFLOPS/Вт на частоте более 1 ГГц в типичных условиях технологии 22 нм.

В таблице I мы представляем обзор процессоров RISC-V, которые в настоящее время реализуют векторное расширение. Примечательно, что большинство из них ограничено очень короткой длиной векторного регистра (VLEN), что позволяет избежать многих проблем, особенно тех, которые появились в RVV v1.0 и рассматриваются в этой работе. Что касается производительности, заявлено, что SiFive P270 достигает 5,75 CoreMark/МГц, 3,25 DMIPS/МГц и 4,6 на SPECint 2006 [9]. XuanTie 910 получает 7,1 CoreMark/МГц и 6,11/ГГц на SPECint 2006. Хотя детальное сравнение с другими ядрами, кроме Ara, невозможно, поскольку они либо не соответствуют RVV (Hwacha), либо не (полностью) выпущены, мы можем сравнить доступное использование модулей с FPU во время вычислительных ядер: SiFive X280 и Vicuna заявляют > 90%, а Hwacha получил > 95%. Наша новая архитектура достигает >98% использования и >35 DP-GFLOPS/Вт.

III. ЭВОЛЮЦИЯ RISC-V V

Расширение RISC-V V позволяет обрабатывать несколько данных с помощью одной команды, следуя вычислительной парадигме исходного векторного процессора Cray. Это расширение вводит 32 регистра, организованных в файл векторного регистра (ФВР), где каждый регистр хранит набор элементов данных одного типа (например, FP32). Типичные векторные операции выполняются поэлементно над двумя векторами, т. е. над элементами с одинаковым индексом.

Кроме того, поддерживаются предварительные вычисления, предотвращающие обработку некоторых элементов на основе булевого вектора маски.

Первоначально расширение RVV было предложено в 2015 году [5] и получило несколько обновлений, представленными Крсте Асановичем и Роджером Эспаса [20]–[22]. После 2018 года этот вариант стал поддерживаться официально. Следуя обозначениям, использованным в [7], мы будем ссылаться на последнюю неофициальную спецификацию (2018 г.) как v0.5. Текущая версия спецификации — v1.0, и это замороженная спецификация для публичного рассмотрения. Вместе с V она также описывает различные другие расширения, например, предназначенные для встроенных процессоров: Zve. На протяжении всей этой работы мы будем ссылаться только на основное расширение для прикладных процессоров. Даже если основная концепция расширения RVV оставалась неизменной с течением времени, произошли заметные изменения: 1) организация файла векторного регистра, 2) кодирование инструкций и 3) организация регистров маски. Мы обсудим эти основные изменения в следующих подразделах.

A. Файл векторного регистра (VRF, ФВР)

1) Состояние ФВР: ФРВ является наиболее важной частью конструкции векторного процессора. Он содержит векторные элементы, и его расположение сильно влияет на выбор проектного решения. Когда длина вектора достаточно велика, она обычно реализуется с помощью статических запоминающих устройств с произвольным доступом (SRAM), фактически создавая новый уровень в иерархии памяти.

v0.5: Состояние регистрового файла сохранилось как глобально, так и локально. Пользователь динамически задавал, сколько регистров было включено, а аппаратное обеспечение вычисляло максимальную длину вектора, разделяя пространство байтов файла регистров между всеми включенными регистрами. Затем каждый регистр можно было индивидуально запрограммировать для хранения разных типов данных.

v1.0: Состояние регистрового файла является только глобальным. Файл векторного регистра состоит из 32 векторных регистров с битами VLEN, где VLEN — параметр, зависящий от реализации, и указывает количество битов в одном векторном регистре.

Можно настроить параметр LMUL для изменения детализации файла регистра, например, установка LMUL на 2 означает, что файл регистра будет состоять из 16 2 × VLEN векторных регистров. Более того, регистровый файл не зависит от типа данных хранимых элементов.

2) Объём разбиения: исходная версия чётко не ограничивала структуру байтов векторного файлового регистра. Позже был добавлен параметр объём разбиения (SLEN), чтобы дополнительно указать, как реализации могут организовать расположение байтов в файле внутреннего векторного регистра. Этот параметр стал удобным, особенно в версии 0.9.

v0.9: SLEN ⩽ VLEN: каждый векторный регистр разделен на секции VLEN/SLEN с SLEN битами. Последовательные элементы вектора отображаются в последовательные разделы, возвращаясь к первому разделу, пока векторный регистр не заполнится [23].

v1.0: SLEN = VLEN: ФВР рассматривается как непрерывный объект, а последовательные байты элементов хранятся в его последовательных байтах.

Б. Кодирование команд

v0.5: Поскольку тип данных элементов вектора был указан в управляющем регистре для каждого векторного регистра, может использоваться полиморфное кодирование команд, например, vadd будет использоваться для добавления двух векторных регистров, независимо от их типа данных.

v1.0: кодирование является мономорфным, и существуют разные команды для разных типов данных, т. е. целочисленных, с фиксированной запятой, с плавающей запятой. Таким образом, ISA имеет больше инструкций, являясь одним из самых длинных расширений во всей среде RISC-V.

B. Макет регистра маски

Биты маски используются для поддержки предикации, способа, которым векторные процессоры выполняют условный код. На каждый элемент вектора приходится один бит маски, и ядро выполняет команду для элемента i только в том случае, если i-й бит маски имеет определенное значение.

v0.5: только один векторный регистр (v1) может содержать вектор маски. Каждый элемент этого вектора может содержать один бит маски в своем наименьшем значащем бите (LSB).

v1.0: Каждый регистр ВРФ может быть регистром маски, а биты маски последовательно упаковываются один бит за другим, начиная с наименьшего значащего бита ФВР.

IV. RISC-V V И ДОРОЖКИ

В этом разделе мы обсудим влияние расширения RVV на микроархитектуру. Мы будем рассматривать Ara как пример конструкции, настроенной на RVV 0.5, даже если обсуждение ею не ограничится. Далее мы будем называть её Vector Unit 0.5 (VU0.5), а нашу новую архитектуру — Vector Unit 1.0 (VU1.0).

VU1.0 — это гибкая архитектура с параметрической VLEN, разработанная для достижения высокой производительности и эффективности в широком диапазоне длин векторов. Например, при VLEN = 4096 устройство может обрабатывать векторы размером до 4 КиБ, при LMUL = 8 — с ФВР размером 16 КиБ. Стремление к большим длинам векторов имеет много преимуществ: работа с векторами, которые не соответствуют ФВР, требует разбиения данных, что увеличивает объём требуемых ресурсов и требует увеличения пропускной способности шин данных в памяти и дополнительному расходу энергии на декодирование и запуск дополнительных векторных операций.

A. ФВР и дорожки

В версии VU0.5 ФВР был реализован путём разделения на фрагменты, по одному на дорожку. VU1.0 поддерживает ту же организацию ФВР на основе дорожек. В этом разделе мы обсудим альтернативный способ реализации ФВР с монолитной архитектурой, которая усложнит маршрутизацию к/от каждого банка и области межсоединений, добавляя дополнительную зависимость от количества дорожек. В целом, площадь ФВР на портах банков памяти (Axbar) пропорциональна количеству как ведущих, так и ведомых устройств, так как требует использование демультиплексоров для ведущих устройств и арбитров для ведомых.

Более подробно, в организации на основе дорожек каждая полоса содержит секцию ФВР с 8 банками 1RW SRAM. Все ведущие устройства дорожки (Mlane) подключаются к банку памяти (B), поэтому общая площадь межсоединений равна площади межсоединений одной дорожки, умноженной на количество дорожек (ℓ),

$$A_{xbar}^{split} \propto M_{lane} \times B_{lane} \times \ell = M_{lane} \times 8 \times \ell, (1)$$

в то время как монолитный ФВР соединит каждый банк памяти с ведущим устройством каждой дорожки,

$$A_{xbar}^{mono} \propto (M_{lane} \times \ell) \times B_{lot} = M_{lane} \times 8 \times \ell^2, (2)$$

Квадратичная зависимость от количества дорожек легко ограничит потенциальное масштабирование векторного процессора с монолитным ВРФ. Кроме того, разделённый ВРФ дает дополнительную свободу в размещении макросов при планировании векторного блока и улучшает маршрутизацию, поскольку межсоединение является локальным для полосы.

B. Порядок байтов

Во время операций с векторной памятью векторный процессор сопоставляет байты из памяти с байтами в своём файле векторных регистров. Следуя RVV 1.0, память и порядок ФВР должны быть одинаковыми, т. е. i-й байт вектора в памяти хранится в i-м байте ФВР. Это условие не может выполняться в случае разделения файла векторных регистров, поскольку последующие элементы должны быть сопоставлены с последовательными дорожками, чтобы лучше использовать параллелизм на уровне данных (DLP - Data Level Parallelism) и не усложнять операции с данными разной ширины. Поскольку ширину элемента можно изменить, а сопоставление между элементами и дорожками остается постоянным, то отображение между байтами и дорожками — нет. В зависимости от ширины элемента один и тот же байт отображается на разные дорожки.

Как следствие, процессор должен отслеживать ширину элемента каждого векторного регистра, чтобы иметь возможность восстанавливать его содержимое, и каждый модуль, обращающийся ко всему векторному регистру, должен иметь возможность переназначать его элементы.

C. Перемешивание

Мапирование реализуется с помощью перемешивания (байты в ФВР) и обратного перемешивания (байты из ФВР), которые преобразуются в уровень мультиплексоров байтов, по одному на каждый выходной байт.

Если N дорожек работают на шине данных 8 B, а устройства, имеющие доступ ко всему ФВР, собирают данные с каждой дорожки параллельно для поддержания требований к пропускной способности, N × 8 байтов перемешиваются в каждом цикле с использованием N × 8 мультиплексоров. Поскольку модуль RVV должен поддерживать четыре различных размера элементов (8 бит, 16 бит, 32 бит и 64 бит), каждый мультиплексор имеет четыре входных байта.

D. Проблемы внедрения RVV 1.0

Некоторые изменения, внесенные в RVV 1.0, упрощают интерфейс для программистов, но усложняют конструкцию векторной вычислительной машины, разделённой на дорожки.

1) Устройство маскировки данных (MASKU): Когда векторная операция маскируется, дорожка не должна обновлять байты маскированного элемента в итоговом векторном регистре. Для этого информация о замаскированных индексах должна быть доступна на каждой дорожке. Благодаря новой схеме регистра маски каждый векторный регистр может использоваться и считываться как регистр маски, а из-за новой схемы вектора маски дорожке могут потребоваться биты маски, хранящиеся в другой дорожке. Поскольку данные перемешиваются в каждом регистре ФВР, нам нужен модуль (Mask Unit), способный выбирать и перемешивать данные, зная предыдущую кодировку, использованную для этого регистра, а затем расширять и пересылать маски на правильные дорожки. Введение ещё одного модуля, который обращается ко всем дорожкам, приводит к большей сложности маршрутизации, особенно при увеличении количества дорожек, как уже было замечено в [7].

2) Перемешивание: ограничение порядка байтов ФВР, устанавливающего SLEN = VLEN в архитектурах с дорожками, приводит к специфическим проблемам при исполнении определенных команд. Следуя спецификациям, архитектура также должна поддерживать политику tail-undisturbed, т. е. элементы после последнего активного элемента не должны изменяться. Когда команда записывает векторный регистр vd, который был закодирован с помощью EEWoldvd где EEWnewvd ≠ EEWoldvd, и старое содержимое регистра перезаписывается не полностью, предыдущие данные повреждаются, поскольку байтовое отображение vd более не является уникальным.

Чтобы не повредить хвостовые элементы, VU1.0 должен перемешивать выходной регистр, используя EEWoldvd, и отменять перемешивание, используя EWnewvd. Эта операция выполняется модулем SLDU, поскольку он может получить доступ ко всем дорожкам и уже имеет необходимую логику для выполнения этой операции (которая представляет собой vslide с нулевым шагом и разными EEW для исходного и целевого регистров). Мы назвали эту операцию повторное перемешивание.

Проблема усугубляется тем фактом, что программа может изменять длину вектора и ширину элемента во время выполнения, поэтому невозможно узнать, сколько байтов необходимо повторно перемешать, если только длина вектора и ширина элемента не отслеживаются динамически для каждого векторного регистра. В ином случае архитектура всегда должна повторно перемешать весь реестр. В нашей архитектуре векторный блок вводит операцию повторного перемешивания, как только внешний интерфейс декодирует команду, которая записывает в векторный регистр изменение закодированного EEW. Повторное перемешивание выполняется перед конфликтующей командой и не вводится, если команда записывает весь векторный регистр. Как правило, повторное перемешивание вредит числу команд, выполняемым за цикл (IPC), если задержку перемешивания нельзя скрыть и если эта операция вызывает структурные риски при следовании slide-командам. Эта проблема затрагивает команды, которые имеют одни и те же входные и выходные регистры. Без этапа переименования в архитектуре невозможно разделить входные и выходные регистры, а простой операции повторного перемешивания недостаточно для сохранения хвостовых элементов: при повторном перемешивании в меньший EEW байты, принадлежащие одному и тому же элементу, помещаются в разных дорожках и, следовательно, не могут быть извлечены в пределах одной дорожки с помощью операции сужения.

Использование tail-agnostic политики также создает некоторые проблемы. Хвостовые байты должны быть либо оставлены без изменений, либо перезаписаны «1». Байты, оставленные без изменений, повреждаются, так как информация об их отображении также теряется, а запись всех хвостовых элементов в «1» имеет значительные накладные расходы, поскольку эти дополнительные записи ухудшают число команд, выполняемым за цикл, и, возможно, вызывают новые конфликты банков памяти в ФВР.

Повторное перемешивание является дорогостоящей операцией, поскольку конфликтующая команда всегда имеет зависимость операции «чтения после записи» (RAW) от повторного перемешивания: стоимость выше, если пропускная способность модуля SLDU ниже, чем производительность вычислительного модуля, поскольку цепочка не может двигаться на полной скорости. Компилятор может решить эту проблему путем кластеризации файла векторных регистров, максимально избегая изменения EEW в одном и том же регистре. С аппаратной стороны этап переименования также может помочь в определении приоритетов модификации отображения регистров назначения на физические регистры с тем же EEW.

V. АРХИТЕКТУРА

Наше устройство поддерживает подавляющее большинство RVV 1.0 со следующими исключениями: отсутствие поддержки арифметики с фиксированной точкой, редукций с плавающей запятой и очень специфических команд (округление в сторону нечетного, обратная дробь, обратное вычисление квадратного корня), операции с сегментной памятью (необязательные, т.к. RVV 1.0), gather-compress, скалярные перемещения (мы эмулируем их посредством передачи памяти) и некоторые специальные команды маски (например, vfirst, viota). На рис. 1 показана основная блок-схема системы. Интеграция с процессором CVA6 основана на модуле RVV0.5 [7] со следующими основными улучшениями.

a) Декодирование: благодаря новым спецификациям RISC-V V кодирование векторных инструкций теперь полностью определяет тип данных векторных элементов, с которыми работает команда. Это позволяет переместить большую часть логики декодирования и специфических для вектора регистров управления и состояния (CSR) из CVA6 в векторный блок, что делает CVA6 более независимым от расширения V. В обновлённой архитектуре CVA6 сохраняет только логику предварительного декодирования, необходимую для того, чтобы знать, 1) если векторная команда является векторной командой, чтобы отправить её векторному модулю, когда она достигает модуля табло (scoreboard), 2) если векторная команда является операцией с памятью (необходима для когерентности кэша) и 3) если команде требуется скалярное значение из целочисленных или скалярных регистровых файлов с плавающей запятой.

b) Интерфейс модуля CVA6-Vector: интерфейс между хост-процессором CVA6 и модулем вектора является обобщённым: модуль реализован в виде модульного ускорителя с собственным файлом CSR. При декодировании CVA6 идентифицирует векторные команды, помещает их в очередь диспетчера и отправляет их на ускоритель, как только они перестают быть предположительными.

c) Когерентность памяти: CVA6 и векторный модуль имеют отдельные порты памяти, а CVA6 имеет частный кэш данных L1. В то же время RISC-V ISA требует строго согласования памяти между скалярным и векторным процессорами. В VU0.5 [7] это требование нарушается, и необходим барьер памяти, который выполнит обратную запись и сделает недействительным кэш данных CVA6 между обращениями к областям общей памяти, что значительно увеличит производительность и снизит переносимость кода. В нашей VU1.0 мы расширяем систему с помощью облегченного аппаратного механизма для обеспечения связанности\когерентности. Мы адаптируем кэш данных CVA6 L1 к политике сквозной записи, чтобы основная память, к которой также обращается векторный блок, всегда была актуальной. Когда векторный модуль выполняет сохранение вектора, он делает недействительными соответствующие строки кэша в кэше данных CVA6. Более того, мы выдаем 1) скалярные загрузки только в том случае, если в данный момент не выполняется векторных операций, 2) скалярные сохранения только в том случае, если в процессе отсутствуют загрузки или сохранения векторов и 3) векторные загрузки или сохранения только в случае отсутствия ожидающих сохранения скалярных данных.

d) Устройство маскировки данных: после обновления биты маски не всегда находятся на правильной дорожке. Поэтому мы разрабатываем устройство маскировки данных, который может получить доступ ко всем дорожкам одновременно для извлечения, распаковки и отправки правильных битов маски на соответствующие дорожки.

д) Сокращения: поскольку в нашем проекте есть дорожки, мы реализуем целочисленные сокращения, используя трёхэтапный алгоритм: внутри дорожки, между дорожек и этапы сокращения SIMD. Сокращение внутри дорожки полностью использует локальность данных внутри каждой дорожки, максимально увеличивая эффективность арифметического логического устройства (ALU, АЛУ), сокращая количество всех элементов, уже присутствующих в дорожке.

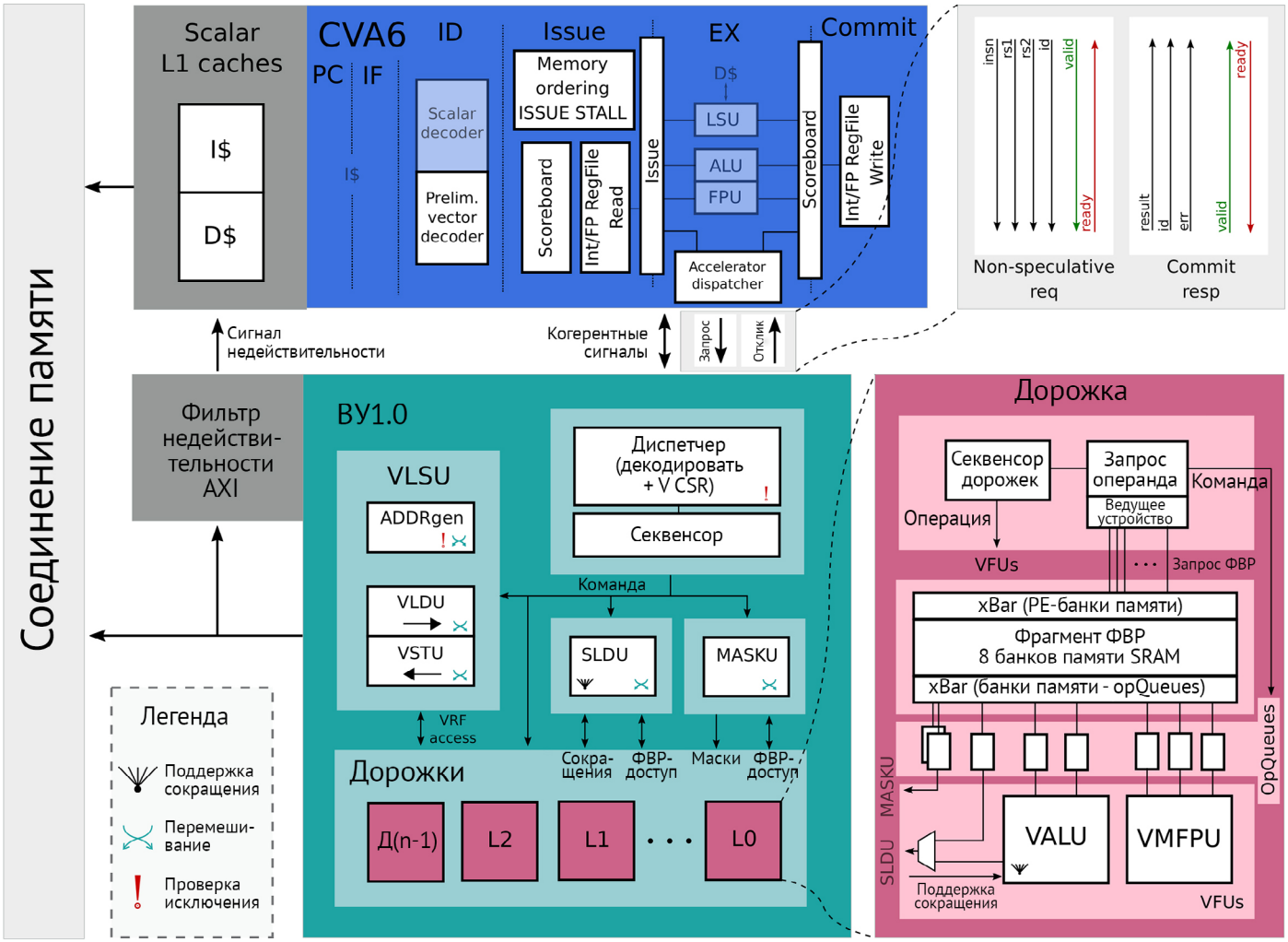


Рисунок 1. Блок-схема верхнего уровня (новой) системы с векторным сопроцессором, отмеченным зеленым, более подробная схема дорожки — пурпурным, а скалярное ядро хоста CVA6 — синим.

Сокращение между дорожками перемещает и уменьшает данные между дорожками с шагами $\log2(\ell)+1$, где ℓ является количеством дорожек, с использованием модуля SLDU; поскольку существует обратная связь по зависимости между слайдом и модулями АЛУ, накладные расходы на задержку связи оплачиваются на каждом этапе.

Наконец, сокращение SIMD уменьшает SIMD слово, если это необходимо; следовательно, его задержка логарифмически зависит от ширины элемента.

VI. РЕЗУЛЬТАТЫ

Одним из основных мотивов разработки векторного процессора является максимальное увеличение пропускной способности за счёт использования встроенного процессора передачи данных. В следующих разделах мы будем использовать понятие «пропускная способность» для обозначения количества полезных результатов вычислений, полученных за такт, например, при добавлении двух векторов длины N пропускная способность равна N/#циклов, где #циклы — количество циклов, требуемого для получения N результатов. В нашем эксперименте мы исследуем, как изменения, внесённые обновлённой спецификацией V, и новые функции системы влияют на пропускную способность. Кроме того, мы размещаем и маршрутизируем наше решение и извлекаем соответствующие метрики мощности, производители и площади.

Технические характеристики

Мы вручную оптимизируем контрольные точки, использованные в [7] (fmatmul, fconv2d с ядром 7 × 7 × 3), адаптируя их к новым спецификациям и архитектуре, и компилируем с помощью LLVM 13.0.0. Чтобы провести сравнение между двумя системами, мы измеряем количество циклов тех же тестов из [7] с точным циклическим моделированием нашего векторного модуля с использованием Verilator v4.214.

Мы настраиваем контрольные точки на ассемблере, так как код, скомпилированный LLVM, показал более низкую производительность по сравнению с оптимизированным.

Чтобы получить дополнительные сведения о системе, мы проводим тот же эксперимент, измеряя, насколько скалярное ядро ограничивает конечную производительность из-за неидеальной скорости выдачи векторных инструкций векторному модулю, показывая, как доступ к скалярной памяти влияет на конечную пропускную способность.

На рис. 2 показана roofline-модель Ara для переменного количества дорожек и результаты тестирования производительности умножения матриц между квадратными матрицами n × n для нескольких размеров матриц n. Арифметическая интенсивность для этого теста пропорциональна n. Горизонтальные пунктирные линии отмечают вычислительный предел архитектуры для соответствующего количества дорожек. Первоначально Ara достигла почти пиковой производительности на ядрах fmatmul и fconv2d, привязанных к вычислительным ресурсам, демонстрируя высокий уровень использования модулей FPU; это также было показано Hwacha с использованием более 95% [24]. Несмотря на то, что размер ФВР составляет 1/4 размера [7], наша новая архитектура обеспечивает сравнимую или лучшую производительность для длинных векторов как для fmatmul, так и для fconv2d, с пиковым использованием> 98,5% с 2 дорожками на 128 × 128 fmatmul. Поскольку использование почти 100%, увеличение размера ФВР едва ли увеличит производительность.

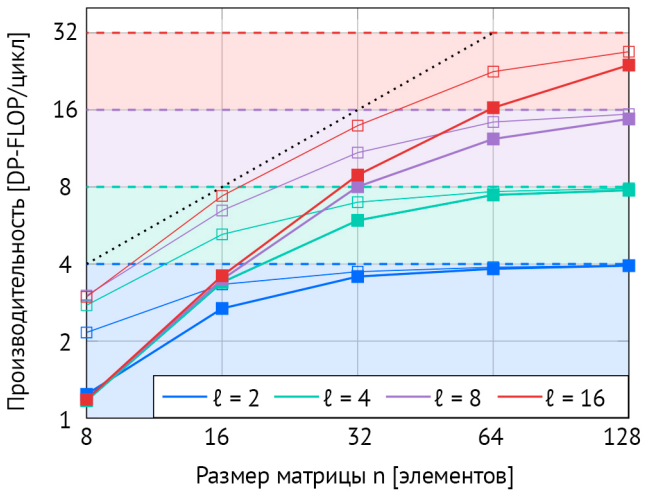


Рисунок 2.Время выполнения ядра умножения матриц размером n × n в нашей системе CVA6+Vector Unit (▪) по сравнению с идеальным диспетчером (□) для нескольких дорожек (ℓ).

На рис. 2 мы показываем производительность реальной системы и производительность, извлечённую с помощью идеального диспетчера. Идеальный диспетчер моделируется путём замены CVA6 предварительно заполненной очередью с соответствующими векторными командами. Система с идеальным диспетчером показывает реальные ограничения производительности векторной архитектуры и отмечает жёсткое ограничение на производительность, достижимую только за счёт оптимизации скалярной части самой системы. Чёрная пунктирная диагональная линия — это предел, заданный скоростью выдачи вычислительных векторных инструкций. В [7] авторы определяют жёсткий предел производительности умножения матриц в системе, особенно для коротких векторов. С RVV v0.5 и их алгоритмом скорость выдачи вычислительных команд для основного ядра составляет одну команду каждые пять циклов. Это связано с наличием команды vins, используемой для перемещения скалярного значения из CVA6 в Ara. В новой спецификации в этом больше нет необходимости, поскольку скалярные операнды можно передавать с помощью векторной команды умножения-накопления. Это улучшает предел скорости выдачи вычислительных команд с 1/5 до 1/4, сдвигая диагональную линию влево.

Чтобы изучить влияние, которое скалярная часть системы оказывает на производительность, мы изменили ширину данных Advanced eXtensible Interface (AXI) порта памяти CVA6, а также ширину строки кэша данных, влияя на накладные расходы в случае обращений к ячейкам памяти, в которых не содержится данных, и временной задержки в результате ошибочных сокращений (при обращении к кэшу) к памяти скалярных данных и, следовательно, влияющая на скорость выдачи векторных инструкций ядра, которое в каждой версии нуждается в новых скалярных элементах, которые пересылаются векторному модулю. На рис. 3 мы суммируем идеальность пропускной способности 16-полосной системы, выполняющей fmatmul на матрице 16 × 16 при изменении параметров памяти скалярного ядра. Увеличение размера строки кэша снижает коэффициент «непопаданий», но если это происходит без увеличения ширины данных AXI, накладные расходы в случае отсутствия нужных данных в кэш-памяти снижаются. Пропускная способность системы, когда ширина строки кэша и данных AXI составляет 512 бит, в 1:54 больше, чем когда они оба установлены на 128 бит, что свидетельствует о важности размера скалярной части памяти системы при повышении производительности. средних/коротких векторов.

a) Короткие векторы: даже если благодаря новой спецификации ограничение скорости выдачи для этого ядра теперь уменьшено, производительность коротких векторов ниже идеальной. Наш ФВР не реализует «полосатую» компоновку ФВР; таким образом, количество действующих банков памяти в каждой дорожке уменьшается, и в системе возникает больше конфликтов между ними. Например, при 16 элементах на 16 дорожках каждый элемент хранится в банке памяти 0 на каждой дорожке, и каждая операция чтения/записи будет нацелена на один и тот же банк памяти. В общем случае, если на банк памяти не будет хотя бы одного элемента (128 элементов на 16 полос, при VLEN=4096), конфликтов банков памяти и связанных с ними зависаний будет больше. «Полосатая» компоновка может смягчить эту проблему, так как также с 16 элементами на 16 дорожках элементы разных регистров занимают разные банки памяти. Это не критическая проблема, так как наша векторная архитектура в первую очередь нацелена на длинные векторы; более того, низкая производительность на коротких векторах уже наблюдалась в Hwacha и Ara, даже при реализации «полосатой» компоновки в их ФВР [7]. Разработчики могут выбрать меньшую длину вектора или выбрать более эффективную невекторную SIMD-архитектуру.

b) Сокращения: В таблице II мы приводим результаты производительности и эффективности при запуске ядра скалярного произведения, изменении числа дорожек, длины векторного байта и ширины элемента. Измеренное количество циклов относится только к фактическому скалярному произведению вычислений, т.е. поэлементному умножению вектор-вектор и последующему сокращению без операций с памятью. Так как умножитель и сумматор нашей единицы принадлежат разным функциональным единицам, произведение и сокращение можно успешно связать в цепочку, чтобы итоговый счёт циклов масштабировался только с количеством элементов в векторе, а не с количеством команд (в данном случае двух). Эффективность рассчитывается для идеальной производительности, которая рассчитывается как $VLB=8\ell + 1 + \log2(\ell)$, где VLB — это длина вектора в байтах, а добавленная единица учитывает цепочку умножения.

1) Более длинные векторы линейно увеличивают время выполнения сокращения внутри дорожки.

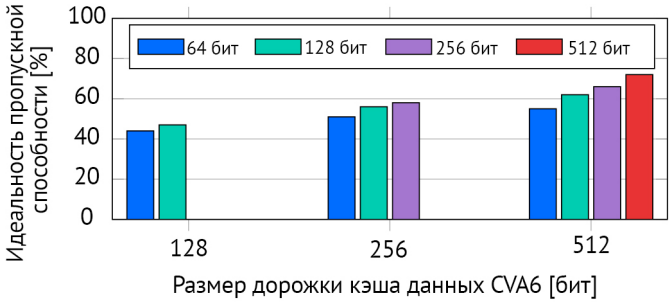


Рисунок 3. Идеальность пропускной способности системы по сравнению с системой с идеальным диспетчером в зависимости от размера строки кэша данных CVA6 и ширины данных AXI.

Чем длиннее вектор, тем больше этот шаг скрывает накладные расходы двух других, способствуя повышению эффективности.

2) Количество дорожек линейно ускоряет сокращение внутри дорожки и логарифмически отрицательно влияет на время, затрачиваемое на сокращение между дорожек. Когда эта фаза недостаточно скрыта (например, короткие векторы), накладные расходы могут быть значительными по сравнению с другим теоретически достижимым максимумом. Чем больше дорожек, тем короче вектор в относительном выражении. Для достижения высокого уровня эффективности конструкция с большим количеством дорожек требует более длинных векторов.

3) Меньшая ширина элементов положительно влияет на пропускную способность, добавляя только логарифмические накладные расходы на этапе сокращения SIMD. Преимущество по сравнению со скалярным ядром имеет решающее значение и может привести к повышению производительности в 380 раз, особенно для элементов с уменьшенной шириной и длинных векторов, где количество скалярных циклов резко возрастает (пик > 24 тыс. циклов): наш векторный блок использует SIMD-подобные вычисления, чтобы сохранить постоянное время выполнения при уменьшении ширины элемента с незначительными накладными расходами на этапе сокращения SIMD.

Для коротких векторов время запуска векторных операций не амортизируется, и общая эффективность падает. Например, нашему векторному модулю требуется около десяти циклов, чтобы получить результаты от сокращения после того, как векторное умножение передано на устройство.

Таблица 2. СЧЁТЧИК ЦИКЛОВ ДЛЯ ОПЕРАЦИИ СНИЖЕНИЯ, С 2/16 ДОРОЖКАМИ, ДЛИНОЙ ВЕКТОРА В БАЙТАХ И РАЗМЕРОМ ЭЛЕМЕНТА ВЕКТОРА. РЕЗУЛЬТАТЫ РАЗЛИЧНЫХ РАЗМЕРОВ ЭЛЕМЕНТОВ РАЗДЕЛЯЮТСЯ СЛЭШЕМ: 8-БИТНЫЕ ЭЛЕМЕНТЫ (СЛЕВА) И 64-БИТНЫЕ ЭЛЕМЕНТЫ (СПРАВА).

	Количество циклов (#)			**Эффективность (% от идеального)**		
	64 B	512 B	4096 B	64 B	512 B	4096 B
2 Дорожки	25 / 23	55 / 51	279 / 275	24% / 26%	62% / 67%	92% / 94%
16 Дорожек	33 / 32	36 / 32	64 / 60	17% / 17%	25% / 28%	58% / 62%

Таблица 3. СРАВНЕНИЕ ФИЗИЧЕСКОЙ РЕАЛИЗАЦИИ МЕЖДУ VU0.5 И VU1.0

	Система VU0.5	**Система VU1.0**	**Обновление показателей**
Размер ФВР [КиБ]	64	16	-75%
Площадь кристалла [мм2]	0.98	0.81	-15%
Площадь ячейки [мм2]	0.43	0.49	+14%
Область макросов памяти [мм2]	-	0.15	-
Частота в худшем случае [МГц]	925	920	-0.5%
Частота TT [ГГц]	1.25	1.34	+7.2%
Производительность [DP-GFLOPS]	9.8	10.4	+6.1%
Мощность при частоте TT [мВт]	259	280	-
Эффективность [DP-GFLOPS/Вт]	37.8	37.1	-1.9%

B. Физическая реализация

Чтобы оценить влияние наших архитектурных модификаций на показатели мощности, производительности и площади системы, мы выполняем синтез, размещение и разводку нашего улучшенного решения с 4 дорожками (CVA6, его кэши и новый векторный модуль) с VLEN = 4096 (16 КиБ ФВР), ориентируясь на технологию GLOBALFOUNDRIES 22FDX FD-SOI. Скалярный кэш команд и кэш данных имеют ширину строки 128 и 256 бит соответственно. Мы используем Synopsys DC Compiler 2021.06 для синтеза топографической информации и Synopsys IC Compiler II 2021.06 для физической реализации. После проектирования топологии мы измеряем занимаемую площадь и рабочую частоту, наконец, соответствующие результаты мощности с помощью моделирования на основе деятельности с обратного аннотацирования SDF в типичных условиях, разработанных с использованием Mentor QuestaSim 2021.2 и Synopsys PrimeTime 2020.09, при выполнении 128 × 128 fmatmul.

Решение размещается и трассируется как макрос 0:81 мм 1:00 мм. Чтобы улучшения существующих характеристик мы разрабатываем процесс, в котором используется модульность конструкции. Наш векторный модуль состоит из параметрического ряда дорожек, которые содержат большую часть логики обработки системы. Все дорожки идентичны, а их синтез требует автоматической пересинхронизации конвейерных ступеней модуля обработки чисел с плавающей точкой. Мы используем иерархический синтез и внутренний поток, в котором дорожки разработаны как пользовательские макросы и синтезированы один раз, чтобы значительно сократить время обработки.

На Рисунке 4 и Рисунке 5 мы представляем физическую схему всей системы и отдельной дорожки соответственно. В верхней части матрицы у нас есть интерфейс AXI, вокруг которого расположены коммутаторы каналов AXI, CVA6 и часть модуля векторной загрузки/хранения (VLSU).

Дорожки расположены с двух сторон, чтобы облегчить маршрутизацию всех устройств с перекрестными дорожками, которые к ним обращаются, таких как устройство маскировки данных (MASKU), модуль слайдов (SLDU), VLSU и основной секвенсор. В области дорожки доминирует модуль векторного умножителя/модуля с плавающей запятой (VMFPU), модуль, который содержит умножитель FPU и умножитель SIMD.

В таблице III показаны параметры и показатели качества нашей реализации по отношению к VU0.5. Поскольку, как показано в разделе VI-A, VU1.0 обеспечивает конкурентоспособную пропускную способность, несмотря на использование ФВР 4 × меньшего размера, чем у VU0.5, мы получаем уменьшение общей площади более чем на 15% от размера кристалла без ущерба для производительности. VU1.0 достигает частоты 920 МГц в наихудших условиях (SS, 0,72 В, 125 °C), практически такой же, как и VU0.5. VU1.0 достигает частоты 1,34 ГГц в типичных условиях (TT, 0,80 В, 25 °C), что на 7,2 % быстрее, чем указано в [7] для VU0.5, благодаря расширенной стратегии иерархической реализации. Это означает, что пиковая пропускная способность fmatmul на 6,1% выше, чем у VU0,5 (10,4 DP-GFLOPS).

VU1.0 потребляет 280 мВт при работе с 128 × 128 fmatmul, что приводит к энергоэффективности 37,1 DPGFLOPS/Вт. Это всего на 1,9% ниже, чем у VU0.5, при этом поддерживается гораздо более полная ISA, включая сокращения и поддержку когерентности памяти. В [18] измеренная эффективность для сопоставимого напряжения (0,8 В) или частоты (1,34 ГГц) ниже, чем 33 DP-GFLOPS/Вт для реальной системы, работающей под управлением Hwacha 4. Эффективность VU1.0 также намного выше заявленной для Hwacha 4.5 [25], хотя прямое сравнение невозможно, поскольку они сообщают только об измерениях мощности полностью изготовленной системы.

На рис. 6 показаны графики мощности двух вариантов запуска fmatmul, демонстрирующие, как векторная архитектура может выдерживать остановку скалярного ядра. Модуль вектора и CVA6 конкурируют за порт AXI к памяти L2, и во время второй итерации CVA6 останавливается из-за промаха кэша данных L1, который не может быть обслужен, поскольку продолжается загрузка вектора. В течение интервала, в течение которого CVA6 не может пересылать новые векторные команды, векторный модуль не простаивает до тех пор, пока не обработает все элементы вектора в своем ФВР, и общее использование остается на пике большую часть времени.

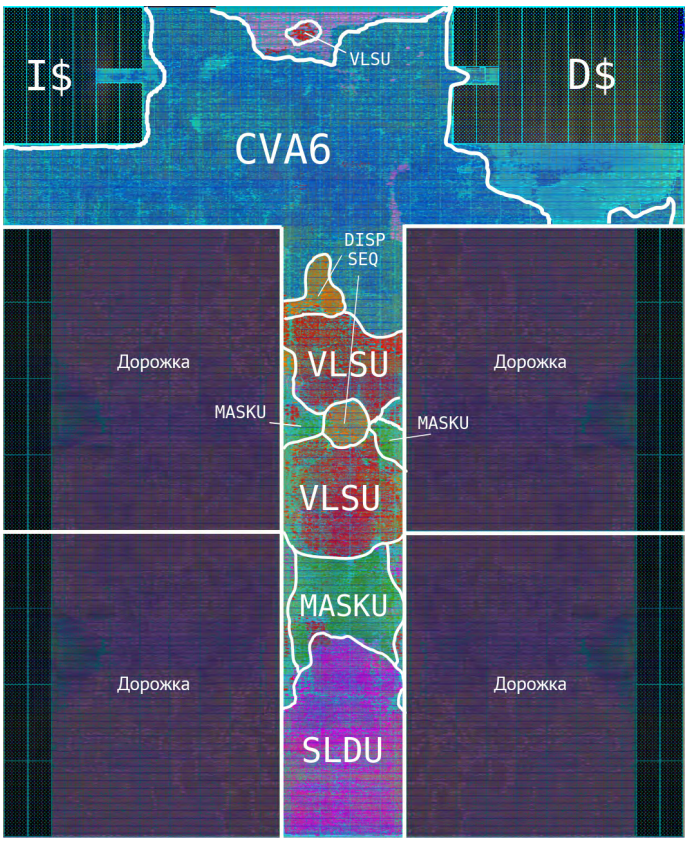


Рисунок 4. Физическая реализация полной системы. Дорожка реализуется и заключается в макрос, а затем помещается на матрицу. Вход и выход системы находятся в верхней части кристалла (интерфейс AXI).

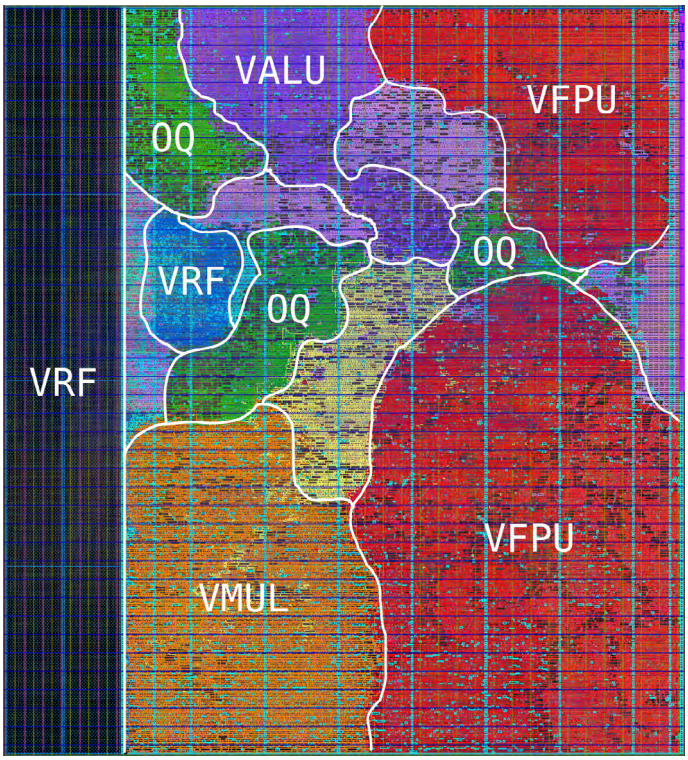


Рисунок 5. Физическая реализация дорожки. Модули без метки: секвенсор дорожек, реквестеры операндов (рядом с ФВР), VDIV и управляющая логика для VMUL и VFPU (посередине).

Средняя разбивка мощности системы с запущенным fmatmul показана на рис. 7. Более 80% мощности потребляется дорожками, на которых происходят вычисления. Как и ожидалось, VMFPU использует большую часть энергии дорожки и вместе с ФВР и очередями операндов занимает почти 90% от общей пропускной способности дорожки. Это показывает, как VU1.0 использует регулярный шаблон выполнения векторных машин, что приводит к уменьшению и упрощению логики контроллера. Для сравнения, CVA6 и его скалярные кэши потребляют менее 12% от общего бюджета мощности, несмотря на значительную площадь, занимаемую модулем табло (scoreboard) и логикой диспетчеризации команд [26].

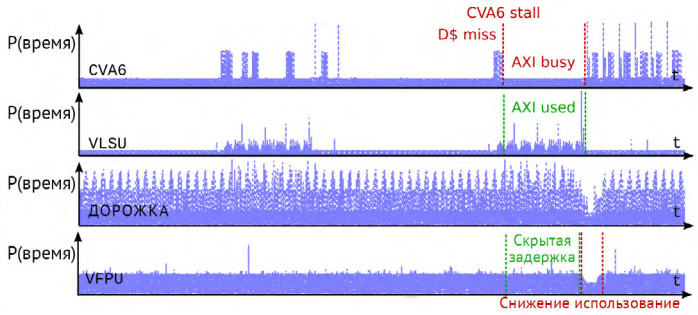


Рисунок 6. Моделирование мощности на основе времени — две итерации fmatmul. VU1.0 выполняет две векторные загрузки, задержка которых скрыта за счёт вычислений в VFPU (при использовании около 97%). К концу временного промежутка CVA6 временно останавливается из-за неудачного обращения в кэш L1D, который не может быть обслужен системой верхней памяти, которая уже используется VLSU VU1.0. Чтобы максимизировать видимость, каждый подграфик масштабируется до максимального энергопотребления.

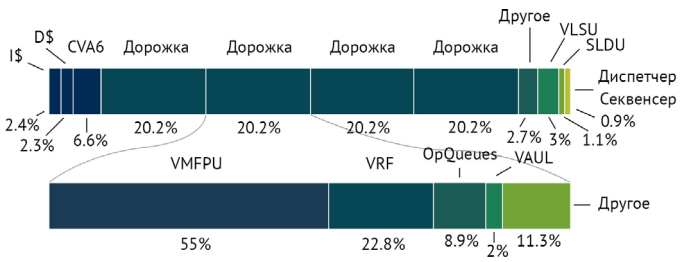


Рисунок 7. Распределение мощности системы. Общее среднее энергопотребление составляет 280 мВт на частоте 1,34 ГГц при выполнении fmatmul.

VII. ВЫВОДЫ

В этой статье мы представляем первый векторный процессор с открытым исходным кодом, совместимый с ядром RVV 1.0. Мы сравниваем нашу конструкцию с блоком RVV 0.5 и обсуждаем влияние, которое обновление спецификации оказывает на архитектуры с разделением дорожек ФВР, на новые добавленные функции. Мы показываем конкурентоспособную пропускную способность и результаты мощности, производительности и площади, используя передовую технологию GLOBALFOUNDRIES 22FDX. Система работает на частоте до 1,34 ГГц с пиковым использованием FPU > 98% на критически важных ядрах Matmul. Мы предоставляем информацию о производительности смешанной скалярно-векторной системы для коротких векторов и анализ нового механизма сокращения, который приводит к ускорению до 380 по сравнению со скалярным ядром. Исходный код опубликован на https://github.com/pulp-platform/ara.

БЛАГОДАРНОСТЬ

Эта работа была поддержана ETH Future Computing Laboratory (EFCL), финансируемой за счёт Huawei Technologies.

СПИСОК ЛИТЕРАТУРЫ

[1] M. Sato, Y. Ishikawa et al., “Co-design for A64FX manycore processor and fugaku,” in Proc. IEEE SC’20, 2020.

[2] J. Dongarra, “Report on the Fujitsu Fugaku system,” University of Tennessee-Knoxville Innovative Computing Laboratory, Tech. Rep. ICLUT-20-06, 2020.

[3] J. Gao et al., “Sunway supercomputer architecture towards exascale computing: analysis and practice,” Science China Information Sciences, vol. 64, no. 4, 2021.

[4] Arm, Arm A64 Instruction Set Architecture - Armv9, for Armv9-A architecture profile, 2021. [Online]. Available: https://developer.arm.com/documentation/ddi0602/2021-12

[5] C. Schmidt et al., RISC-V Vector Extension Proposal, 2015. [Online]. Available: https://riscv.org/wp-content/uploads/2015/06/riscv-vector-workshop-june2015.pdf

[6] K. Asanovic et al., Vector Extension 1.0, 2021. [Online]. Available: https://github.com/riscv/riscv-v-spec/releases/tag/v1.0

[7] M. Cavalcante et al., "Ara: A 1-GHz+ scalable and energy-efficient RISC-V vector processor with multiprecision floating-point support in 22-nm FD-SOI," IEEE TVLSI, vol. 28, no. 2, pp. 530–543, 2020.

[8] SiFive, "SiFive intelligence x280," accessed: 2021-11-20. [Online]. Available: https://www.sifive.com/cores/intelligence-x280

[9] ——, "SiFive performance P270," accessed: 2021-11-20. [Online]. Available: https://www.sifive.com/cores/performance-p270

[10] M. Demler, "Andes plots RISC-V vector heading, NX27V CPU supports up to 512-bit operations," May 2020, accessed: 2021-11-20. [Online]. Available: andestech.com/wp-content/uploads/Andes-Plots-RISC-V-Vector-Heading.pdf

[11] R. Espasa, Introducing SemiDynamics High Bandwidth RISC-V IP Cores, 2021. [Online]. Available: https://www.european-processor-initiative.eu/wp-content/uploads/2021/03/202012.RISCV-SUMMIT.pdf

[12] M. Platzer and P. Puschner, "Vicuna: A timing-predictable RISC-V vector coprocessor for scalable parallel computation," in Proc. ECRTS'21), 2021.

[13] I. A. Assir et al., "Arrow: A RISC-V vector accelerator for machine learning inference," arXiv preprint arXiv:2107.07169, 2021.

[14] M. Johns and T. J. Kazmierski, "A minimal RISC-V vector processor for embedded systems," in Proc. IEEE FDL, 2020.

[15] F. Minervini and O. P. Perez, Vitruvius: And Area-Efficient RISC-V Decoupled Vector Accelerator for High Performance Computing, 2021.

[16] C. Chen, X. Xiang et al., "Xuantie-910: A commercial multi-core 12-stage pipeline out-of-order 64-bit high performance RISC-V processor with vector extension: Industrial product," in Proc. ACM/IEEE ISCA,

2020, pp. 52–64.

[17] K. Patsidis, C. Nicopoulos, G. C. Sirakoulis, and G. Dimitrakopoulos, "RISC-V2: A scalable RISC-V vector processor," in Proc. IEEE ISCAS, 2020.

[18] C. Schmidt, J. Wright, Z. Wang, E. Chang, A. Ou, W. Bae, S. Huang, V. Milovanovi´c, A. Flynn, B. Richards et al., "An eight-core 1.44-GHz RISC-V vector processor in 16-nm FinFET," IEEE Journal of Solid-State Circuits, vol. 57, no. 1, pp. 140–152, 2021.

[19] F. Zaruba et al., "The cost of application-class processing: Energy and performance analysis of a linux-ready 1.7-GHz 64-bit RISC-V core in 22-nm FDSOI technology," IEEE TVLSI, vol. 27, no. 11, 2019.

[20] K. Asanovic et al., RISC-V Vector Extension Proposal, 2018. [Online]. Available: https://riscv.org/wp-content/uploads/2018/05/15.20-15.55-18.05.06.VEXT-bcn-v1.pdf

[21] Krste Asanovic, Vector Extension Proposal v0.2, 2016. [Online]. Available: https://riscv.org/wp-content/uploads/2016/12/Wed0930-RISC-V-Vectors-Asanovic-UC-Berkeley-SiFive.pdf

[22] K. Asanovic et al., The RISC-V Vector ISA, 2017. [Online]. Available: https://riscv.org/wp-content/uploads/2017/12/Wed-1330-RISCVRogerEspasaVEXT-v4.pdf

[23] ——, RISC-V Vector Extension Proposal, 2020. [Online]. Available: https://github.com/riscv/riscv-v-spec/releases/tag/0.9

[24] C. Schmidt, A. Ou, and K. Asanovic, Hwacha V4: Decoupled Data Parallel Custom Extension, 2018.

[25] A. Gonzalez et al., "A 16mm2 106.1 GOPS/W heterogeneous RISCV multi-core multi-accelerator SoC in low-power 22nm FinFET," in ESSCIRC 2021, 2021.

[26] F. Zaruba, F. Schuiki, T. Hoefler, and L. Benini, "Snitch: A 10 kGE pseudo dual-issue processor for area and energy efficient execution of floating-point intensive workloads," Feb. 2020, arXiv:2002.10143v1 [cs.AR].

Мнение эксперта

В настоящее время микроархитектуры, использующие векторные инструкции для работы с данными, входят в сегмент потребительских устройств. Все крупнейшие разработчики архитектур набора команд поддерживают или разрабатывают векторные расширения для своих продуктов. Использование векторных операций позволяет существенно повысить производительность устройств при анализе и обработке данных, что особенно актуально при разработке систем искусственного интеллекта и машинного обучения, а также в других сферах, работающих с большими данными, а высокий потенциал масштабируемости позволяет упростить и ускорить разработку специализированного программного обеспечения.

Ведущий инженер- конструктор, Сергей Сорока

Ключевые слова: транзисторы типа HEMT, GaN, высокая частота, удельная мощность

Ориентация на С-диапазонс помощью сверхвысоковольтных транзисторов типа HEMT

Транзисторы типа HEMT на GaN с минимальной выходной ёмкостью и подавлением гармоник обеспечивают рекордную мощность и эффективность

За последнее десятилетие система GaN-на-SiC стала доминирующей полупроводниковой технологией, которая позволяет обеспечить очень высокую мощность в гигагерцовом диапазоне. Её более известный конкурент, кремниевые LDMOS (металло-оксидные полупроводники с поверхностной диффузией), более популярен и является более дешёвым аналогом для питания большинства систем в СВЧ-диапазоне и более низких УВЧ-диапазонах. Тем не менее, он уступает GaN: у GaN широкая запрещённая зона, он обладает более высокой эффективностью, что сокращает эксплуатационные расходы; и у него более высокая мощность на площадь кристалла, что позволяет проектировать более легкие системы меньшего размера.

Завидное положение, которое занимает GaN на рынке, в первую очередь связано с его способностью поддерживать высокий КПД даже при высоких напряжениях и, следовательно, больших уровнях мощности. Напротив, в полупроводниках типа LDMOS важен баланс эффективности, высокой удельной мощности и производительностью на высоких частотах. Поэтому обычно в конструкции устройства разработчикам приходится идти на компромисс в пользу одной характеристики.

Необходимо понимать физические принципы этих ограничений и точно определять критические параметры устройства, которые требуют тщательной настройки, чтобы оптимизировать устройства на GaN. Масштабирование устройства не может быть бесплатно.

Что касается более высоких частот, то преобладание GaN основано на его способности обеспечивать очень высокую мощность и эффективность. Тем не менее, существуют общепринятые правила масштабирования, подразумевающие, что работа на более высоких частотах должна идти параллельно со снижением напряжения питания. Поэтому в своих продуктовых линейках производители указывают устройства на 65 В с частотой до 2 ГГц, в то время как устройства, способные работать на частотах 12 ГГц и 18 ГГц, похоже, обозначают последний рубеж для транзисторов на 50 В и 40 В соответственно. Кроме того, устройства рассчитаны на работу с напряжением 28 В, если вообще рассчитаны.

Прошлым летом компания Integra Technologies (США) выпустила корпусный транзистор типа HEMT на GaN на 100 В. Этот продукт раздвигает границы напряжения питания и, в конечном счёте, выходной мощности на устройство. Их транзистор типа HEMT разработан для радаров авионики L-диапазона (от 1,030 ГГц до 1,090 ГГц).

Существуют устройства, требующие более высокой мощности в диапазоне кВт, например, в диапазоне С (от 4 ГГц до 8 ГГц) и даже в диапазоне X (от 8 ГГц до 12 ГГц).

Поэтому в этих случаях имеет смысл использовать GaN, пробуя сочетать его высокочастотные характеристики и высокую мощность.

Контроль заряда

К сожалению, когда разработчики переходят на более высокие напряжения, это даёт им мало пользы. На самом деле, почти каждый критический параметр производительности, по-видимому, ухудшается при увеличении питания — есть проблемы, связанные с частотой переходов и снижением надёжности, например. Все эти проблемы возникают из-за растущего электрического поля, и это является основным фактором, который следует учитывать при попытке сохранить производительность при работе с более высокими напряжениями.

Ещё одним существенным недостатком сильного электрического поля является то, что оно может привести к так называемым «эффектам короткого канала». Когда это происходит, затвор, который в идеальных условиях отвечает исключительно за управление электронами в канале, с трудом выполняет свою работу. Эффекты короткого канала, как правило, связаны с высокочастотными технологиями, потому что их появление обычно приводит к «недостаточному масштабированию» электрических параметров (они масштабируются меньше, чем предполагают правила). Говоря в общих чертах, увеличение частоты или напряжения приводит к аналогичным проблемам, связанным с полем, если не будут приняты соответствующие меры.

Одной из таких мер является внедрение полевых пластин, концепция, используемая в большинстве GaN-технологий вплоть до диапазона Ka. Роль полевой пластины состоит в том, чтобы управлять электрическим полем, распространяя его от затвора к электроду стока. При её добавлении происходит уменьшение максимального электрического поля вблизи контакта затвора и нивелирование эффектов короткого канала. Ещё одним преимуществом является то, что устройство может реализовать более высокое напряжение пробоя при заданном сопротивлении во включенном состоянии. Поскольку более высокое сопротивление приводит к увеличению потерь, крайне важно при достижении максимальной эффективности поддерживать его на максимально низком уровне.

Наша команда разрабатывает устройства, которые должны соответствовать этим критериям. Измерения показывают, что напряжение пробоя для наших 100-вольтовых транзисторах типа HEMT на GaN превышает 350 В (Рис. 1).

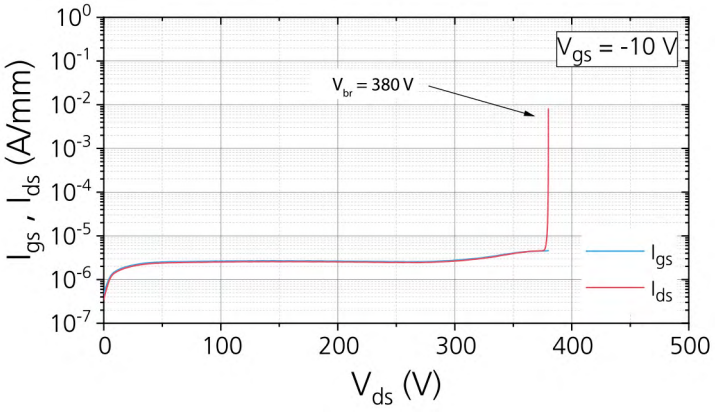


Рисунок 1. Характеристики пробоя сток-источник для устройства с общей шириной затвора 1,2 мм (6 x 200 мкм). Для измерения источник прибора заземляют, а потенциал затвора устанавливают на -10 В. Затем потенциал стока быстро повышают до тех пор, пока регистрируемый ток затвора и стока не превысит значение соответствия 1 мА/мм, что считается как пробой.

Между тем, сопротивление во включенном состоянии составляет всего 4,2 Ом/мм, что всего на одну треть выше, чем у GaN-процессов с напряжением 50 В.

Гармоники

На первый взгляд может показаться довольно консервативным иметь такое высокое напряжение пробоя для устройства на 100 В. Однако есть веская причина – гармоническое завершение. Если вы немного углубитесь в проектирование высокоэффективных усилителей, вы, скорее всего, столкнетесь с такой техникой управления гармоническими импедансами.

В конечном счёте, эффективность идеального усилителя ограничивается его классом работы. В то время как класс A и класс B, вероятно, знакомы большинству из нас, некоторые конструкторы выбирают класс E, класс F или класс F-1, который также известен как обратный класс F.

Три последних класса имеют теоретическую эффективность 100%, в то время как класс B и класс A ограничены 78,5% и 50% соответственно. На практике все эти цифры невозможно получить из-за потерь в реальных устройствах. Однако транзисторы типа HEMT на GaN продемонстрировали эффективность до 90%. Не секрет, что гармоническое завершение обходится дорого. Придание устройству определенных импедансов на его гармонических частотах — в идеале короткое замыкание или разомкнутая цепь — изменяет форму сигнала во временной области на выходе устройства.

Для классов E и F-1 могут быть формы сигналов с величиной напряжения, в три с половиной раза превышающей напряжение питания. Такое высокое напряжение возникает из-за высокого импеданса второй гармоники (при разомкнутой цепи или близко), что требуется для режимов работы. Поэтому достаточно высокое напряжение пробоя является необходимым условием для достижения наивысшей эффективности.

На более высоких частотах концепция завершения гармоник постепенно перестаёт работать. Это происходит из-за того, что реальные устройства имеют конечную выходную ёмкость и собственный источник тока. При повышении частоты сопротивление выходной ёмкости уменьшается, формирует путь ответвлённого тока к земле. Если теперь разработчик попытается заставить устройство работать в режиме класса E или класса F-1, выходная ёмкость «обойдёт» разомкнутую цепь на второй гармонике. На определенной частоте это приводит к короткому замыканию для второй гармоники и всех высших гармоник. Такой набор укороченных гармонических импедансов приводит к состоянию класса B, ограничивая максимальную теоретическую эффективность от 100 до 78,5%.

Помимо этого препятствия на пути к достижению высокой эффективности на высоких частотах, существует ещё одна проблема, связанная с высокими напряжениями. В данном случае частью проблемы являются полевые пластины. Больше всего проблем вызывают так называемые полевые пластины с оконечной нагрузкой — на них может приходиться большая часть выходной ёмкости устройства. Если просто исключить эту форму полевой пластины из макета устройства, такой шаг будет противоречить цели надлежащего контроля поля. Лучше искать компромисс.

Мы верим, что возможно больше. Нет причин думать, что C-диапазон является окончательным пределом для 100-вольтовых транзисторов типа HEMT на GaN. Высокие уровни производительности наших устройств на этих частотах означают, что ещё есть возможности для продвижения технологии в направлении X-диапазона.

Повышение эффективности

Ранее мы продемонстрировали эффективность преобразования энергии постоянного тока в ВЧ энергию более 77% на частоте 1,0 ГГц, что на тот момент было рекордом для 100-вольтовых транзисторов типа HEMT на GaN. Однако мы обнаружили ряд слабых мест, когда сравнили это устройство с нашими базовыми устройствами на 50 В.

Недостатки включали максимальную эффективность, которая отставала от базовых устройств примерно на 7%. Мы также наблюдали явный спад эффективности при измерении 100-вольтовых устройств на более высоких частотах.

Чтобы устранить эти недостатки, мы начали работу сначала. Это включало переработку эпитаксиальных слоёв и внутренних характеристик устройства с целью ужесточения электростатики. Наши результаты подтвердились, когда мы извлекли вызванное стоком понижение барьера — меру, которая количественно определяет паразитный контроль над электронами с помощью контактов стока, причём чем ниже значение, тем лучше. Этот ключевой показатель был снижен в пять раз, что дало значение ниже 1 мВ/В. Данный результат показывает, что при высоких напряжениях затвор управляет электронами без каких-либо помех со стороны стока.

Дальнейшие измерения подчеркнули возможности нашего нового устройства. Удалось выявить рекордную эффективность преобразования энергии постоянного тока в ВЧ энергию для нашего 100-вольтового транзистора типа HEMT на GaN на частоте 2,0 ГГц, составившую 84,7%— это увеличение почти на 8%, несмотря на удвоение частоты. Максимальная удельная мощность составила 15,5 Вт/мм (Рис. 2).

Изменив внутреннюю компоновку устройства, мы максимально уменьшили выходную ёмкость без ухудшения электростатических характеристик. Это должно позволить транзисторам эффективно работать на гораздо более высоких частотах, чем раньше.

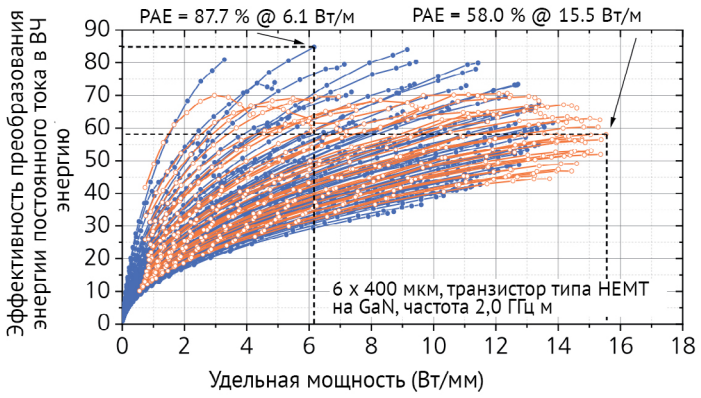


Рисунок 2. Измерения импульсной нагрузки транзистора типа HEMT на GaN на 100 В с общей шириной затвора 2,4 мм (6 x 400 мкм) на частоте 2,0 ГГц. Ширина импульса составляет 100 мкс при коэффициенте заполнения 10%. Здесь показаны два условия. Оранжевый — для нагрузки на основной частоте без прерывания гармоник, а синий — для основной нагрузки с импедансами второй гармоники на входе и выходе, настроенными на максимальную эффективность.

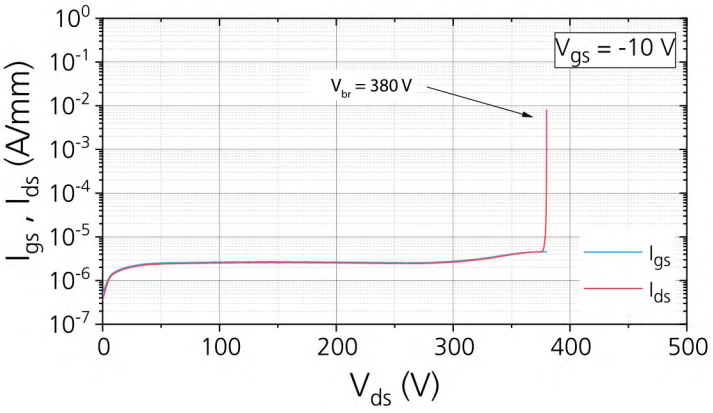


Рисунок 3. Измерения импульсной нагрузки транзистора типа HEMT на GaN на 100 В с общей шириной затвора 0,9 мм (6x150 мкм) на частоте 7,2 ГГц. Ширина импульса составляет 100 мкс при коэффициенте заполнения 10%. Ни одна гармоника не была прекращена во время измерения.

При уменьшении выходной ёмкости на 40% теоретически можно предположить, что устройства будут подвержены гармоническим искажениям на более высоких частотах.

Лабораторные данные подтверждают эту точку зрения. Измерения на верхней частоте С-диапазона 7,2 ГГц, используемой для отслеживания космических аппаратов, дают очень впечатляющие результаты. Согласно фундаментальным измерениям нагрузочная удельная мощность лишь немного ниже значения для 2,0 ГГц и достигает впечатляющих 13,8 Вт/мм. Между тем, эффективность преобразования энергии постоянного тока в ВЧ энергию достигает максимума в 57,2%.

Тем не менее, оставался вопрос, повысит ли эффективность оконечное гармоническое сопротивление. Из-за ряда ограничений мы не могли одновременно настраивать вход и выход, как это было сделано для измерения на частоте 2,0 ГГц. Поэтому мы настроили только вторую выходную гармонику и выбрали основную нагрузку, обеспечивающую баланс эффективности и удельной мощности. Затем мы оптимизировали фазу второй гармоники, пока не нашли настройку, обеспечивающую максимальную эффективность (Рис. 4, где показаны результаты развертки мощности).

Уменьшение выходной ёмкости окупилось. Эффективность преобразования энергии постоянного тока в ВЧ энергию возросла до 66,0%, что примерно на 9% больше по сравнению с согласованием нагрузки без подавления второй гармоники. Это значение также устанавливает новый эталон эффективности 100-вольтовых транзисторов типа HEMT на GaN в С-диапазоне.

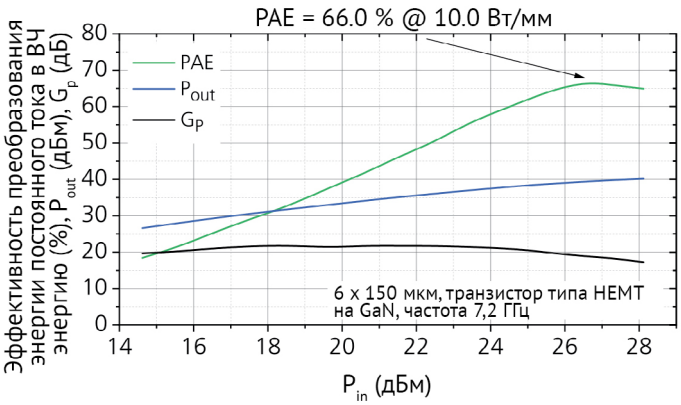


Рисунок 4. Измерения импульсной нагрузки транзистора типа HEMT на GaN на 100 В с общей шириной затвора 0,9 мм (6 x 150 мкм) на частоте 7,2 ГГц. Ширина импульса составляет 100 мкс при коэффициенте заполнения 10%. Для этого измерения был выбран «компромисс» основного импеданса, который дал хорошие результаты сочетания эффективности и удельной мощности. Дополнительно вторая гармоника на выходе устройства была установлена на значение, которое показало наибольшую эффективность преобразования энергии постоянного тока в ВЧ энергию.

Не менее впечатляющими являются соответствующая удельная мощность в 10 Вт/мм и коэффициент усиления в 18,5 дБ, которые достигаются при максимальной эффективности. Важно отметить, что эти показатели производительности реализуются одновременно, что отличает эти 100-вольтовые устройства от уже доступных GaN-технологий. В то время как эффективность преобразования энергии постоянного тока в ВЧ энергию наших новых транзисторов типа HEMT сравнима с лучшими технологиями 40 В и 50 В, удельная мощность и коэффициент усиления наших 100 В устройств — это большой шаг вперед по сравнению с возможностями сегодняшних доступных транзисторов типа HEMT на GaN.

Мы верим, что возможно больше. Нет причин думать, что С-диапазон является окончательным пределом для 100-вольтовых транзисторов типа HEMT на GaN. Высокие уровни производительности наших устройств на этих частотах означают, что ещё есть возможности для продвижения технологии в направлении X-диапазона. Мы не знаем, где именно ограничения частоты этих высоковольтных устройств, но мы готовы и хотим это выяснить.

Источник

Targeting the C-band with ultra-high-voltage HEMTs

Ключевые слова: рынок, компаундные полупроводники, новые материалы

К 2027 году рынок подложек для компаундных полупроводников составит 2,4 миллиарда долларов

Согласно новому отчёту Yole Intelligence о состоянии отрасли за 2022 год, рынок подложек для компаундных полупроводников к 2027 году достигнет 2,4 миллиардов долларов США, а среднегодовой темп роста составит 16% в период с 2021 по 2027 год.

Компаундные полупроводники использовались во многих устройствах на протяжении десятилетий; а совсем недавно им на замену пришли SiC и GaN в энергетике, GaN и GaAs в высокочастотных (ВЧ) устройствах, GaAs и InP в фотонике, а также LED и microLED в дисплеях. В результате ожидается рост рынков подложек и эпитаксиальных пластин.

Один из представителей Yole Intelligence заявил: «Компания Wolfspeed является ведущим поставщиком SiC подложек и эпитаксиальных пластин для силовых устройств на SiC и ВЧ на GaN. Подложки большего размера являются стратегическим ресурсом в производстве устройств следующего поколения, а открытие заводов по производству 8-дюймовых пластин и расширение производственных мощностей говорит о том, что представители отрасли имеют достаточно амбициозные цели на ближайшее десятилетие».

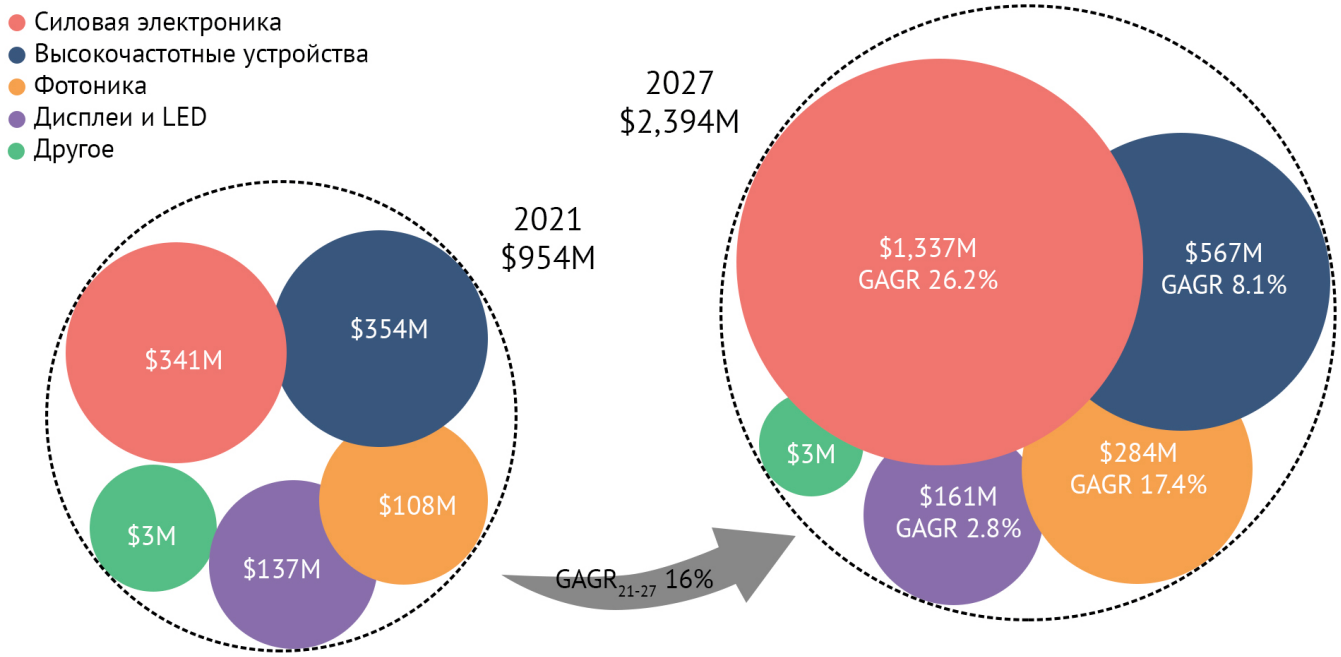
Компания Coherent сейчас является ведущей среди производителей фотонных устройств, а также поставщиков подложек из SiC для силовых и ВЧ-устройств. Кроме того, она работает с компанией SEDI над производством ВЧ-устройств на GaN и запустила производство силовых устройств на SiC вместе с GE. Оба игрока укрепляют свою конкурентоспособность среди производителей устройств.

AXT, Sumitomo Electric, Freiberger и SICC являются ведущими поставщиками подложек из GaAs, InP и полуизолирующих подложек из SiC. В их случае увеличение дохода будет связано с использованием новых материалов для компаундных полупроводников: например, подложки из GaAs и InP для ВЧ, фотонных и microLED устройств. Кроме того, производители полуизолирующих подложек из SiC начинают использовать SiC n-типа, поскольку этот рынок имеет более высокие темпы роста.

По мнению экспертов Yole Intelligence, разная динамика рынка эпитаксиальных подложек для компаундных полупроводников играет на руку поставщикам. Компания IQE участвует в различных категориях рынках (например, ВЧ GaAs и GaN), так как среднегодовые темпы роста фотоники InP и GaAs отражаются двузначными числами и представляют рынки разные как по объёму, так и по масштабу.

Прогноз развития рынка подложек для компаундных полупроводников 2021-2027 (в млн $)

Источник: отчёт компании Yole Intelligence, 2022



Рынок microLED – быстро развивающийся, и ожидается, что он будет удваиваться каждый год в ближайшие пять лет. VPEC удалось стать крупнейшим поставщиком эпитаксиальных подложек на ВЧ GaAs на открытом рынке, и компания продолжает увеличивать свой интерес к фотонике.

Первый переломный момент у компаундных полупроводников на GaAs и InP случился в 1990-х годах с появлением LED, усилителей мощности для мобильных телефонов, телекоммуникаций и передачи данных.

По мере того как рынок предъявляет требования к 5G, быстрым зарядным устройствам для смартфонов, компаундные полупроводники будут расти как в объёме, так и в стоимости.

Заглядывая в будущее, можно сказать, что следующий переломный момент настанет, когда, наряду с появлением новых подложек и новых устройств, в компаундных полупроводниках начнут использоваться новые материалы для производства устройств для электромобилей, устройств с более высоким напряжением, датчиков в различных системах, дисплеях microLED, а также при переходе от 5G к 6G.

Источник

$2.4 billion CS substrate market by 2027

Compound semiconductor, Issue IX 2022