DOI: 10.22184/1992-4178.2022.215.4.76.88

Цифровые ИС: перспективы развития схем памяти и их применение в системах ИИ

И. Черепанов[1], М. Макушин[2]

УДК 621.37 | ВАК 05.27.01

Сектор схем памяти динамично развивается. Появляются новые типы приборов, совершенствуются уже устоявшиеся. Происходит освоение подходов 3D-интеграции, в том числе создания приборов, сочетающих этажерки памяти и слои логических приборов, процессорных ядер и т. п. Это, в свою очередь, существенно расширяет возможности конечных электронных систем, открывает новые области их использования. Соответственно, возникают новые возможности ускорения процесса цифровизации.

Основой процесса цифровизации являются аппаратное и программное обеспечение. Аппаратное обеспечение представлено интегральными схемами и другими полупроводниковыми приборами. К основным тенденциям развития цифровых ИС, помимо масштабирования и гетерогенной интеграции, можно отнести рост их доли в структуре стоимости конечных электронных систем. Что касается схем памяти, то здесь есть еще одна тенденция – рост емкости памяти в конечных электронных системах. Например, за несколько последних лет суммарная емкость ИС ЗУ во флагманских моделях смартфонов более чем удвоилась. Расширяется использование ИС ЗУ и в таких применениях, как искусственный интеллект (ИИ), машинное / глубокое обучение, вычисления в памяти, краевые вычисления и т. п. При всем этом, несмотря на успехи в освоении перспективных типов памяти, на этом рынке в среднесрочной перспективе продолжат доминировать ДОЗУ и флеш-память NAND-типа.

ПРОГНОЗ РАЗВИТИЯ ПЕРСПЕКТИВНЫХ СХЕМ ПАМЯТИ

В августе 2021 года исследовательские фирмы Objective Analysis и Coughlin Associates опубликовали прогноз развития рынка перспективных типов схем памяти (ОЗУ на эффекте изменения фазового состояния – PCRAM, магнитная память на эффекте переключения спинового момента электрона – spin-transfer torque MRAM, STT-MRAM, резистивные ОЗУ – ReRAM). В нем прогнозируется, что к 2031 году емкость рынка таких приборов составит 44 млрд долл. Утверждается, что они начнут вытеснять существующие типы ИС ЗУ, а именно флеш-память NOR-типа, СОЗУ и ДОЗУ (как в секторе автономных схем памяти, так и в сегменте приборов, встраиваемых в микроконтроллеры, ASIC), и даже компьютерные процессоры. Их общий прогноз развития технологий памяти представлен на рис. 1.

Рост доходов от продаж новых типов ИС ЗУ будет стимулировать потребность в новых инструментальных средствах для поддержки различных процессов и использования разнообразных материалов. Это придаст импульс развитию рынка капитального оборудования. Например, предполагается, что увеличится общий доход от оборудования для производства схем MRAM. При этом доходы от продаж автономных MRAM и STT-MRAM вырастут примерно до 1,7 млрд долл., то есть более чем в 42 два раза по сравнению с доходами от продаж автономных MRAM в 2020 году. Встраиваемые схемы MRAM наряду со встраиваемыми схемами резистивных ОЗУ (ReRAM) будут все активнее продвигаться в качестве альтернативы СОЗУ и флеш-памяти NOR-типа в системах-на-кристалле (SoC).

По оценкам, объем продаж схем 3D XPoint* к 2031 году может превысить 20 млрд долл., что будет способствовать

[1] АО «ЦНИИ «Электроника», главный специалист.

[2] АО «ЦНИИ «Электроника», главный специалист, mmackushin@gmail.com.

* 3D Xpoint (3D Crosspoint) – «трехмерный координатный переключатель». Технология памяти на основе эффекта изменения фазового состояния. Бестранзисторная схема памяти, в которой пара «селектор – ячейка памяти» располагается в точке пересечения перпендикулярных проводников. Запись бита происходит при изменении агрегатного состояния вещества при подаче на селектор напряжения определенной величины.

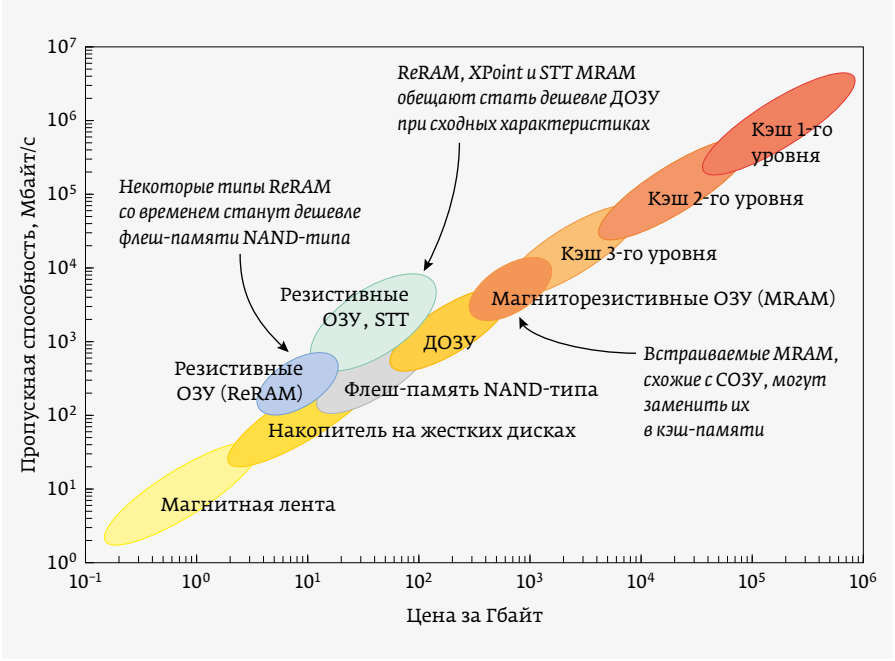


Рис. 1. Прогноз развития технологий памяти. STT (Spin-transfer torque) – память на эффекте переключения спинового момента электрона

увеличению использования PCRAM. В настоящее время ИС семейства Optane от Intel является единственной коммерчески доступной памятью типа 3D Xpoint.

При создании перспективных типов схем памяти главным препятствием для их широкого внедрения всегда была рентабельность. Непосредственные возможности развития рынков MRAM и ReRAM – сектор встраиваемых решений для SoC, где традиционные встраиваемые решения – флеш-память NOR-типа и СОЗУ – столкнулись с ограничениями в плане дальнейшего применения. SoC на основе MRAM уже поставляются на рынок, а SoC на основе ReRAM, как предполагается, появятся в ближайшем будущем.

Привлекательность и перспективы внедрения STT MRAM обусловлены совместимостью процессов MRAM и STT-RAM со стандартными КМОП-процессами. Соответственно, эти ИС ЗУ могут формироваться непосредственно поверх КМОП-логических пластин. Это делает MRAM жизнеспособной флеш-памятью NOR-типа и СОЗУ, поскольку флеш-память не имеет такой же совместимости с обычными КМОП-процессами, а экономия энергии энергонезависимых и более простых MRAM и STT MRAM по сравнению с СОЗУ весьма значительна (рис. 2) [1].

Тем не менее, в ближайшие годы на рынке схем памяти будут доминировать хорошо известные схемы памяти – ДОЗУ и флеш-память NAND-типа.

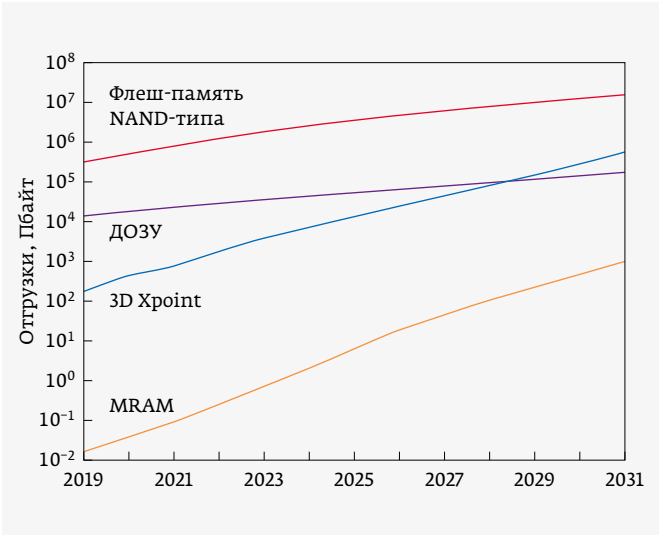


Рис. 2. Отгрузки автономных ИС ЗУ по емкости памяти в период 2019–2031 годов. Пбайт = 10^{15} байт

НЕКОТОРЫЕ АСПЕКТЫ РАЗВИТИЯ 3D-ФЛЕШ-ПАМЯТИ NAND-ТИПА

Ведущие производители схем флеш-памяти NAND-типа уже несколько лет производят трехмерные приборы, наращивая объемы их выпуска. Сейчас наиболее сложными являются 192-уровневые приборы, производство которых в основном осуществляют Samsung и SK Hynix. Однако разработкой данной технологии занимаются не только они, о чем свидетельствуют соответствующие доклады на ISSCC-2022 (табл. 1) [2]. Есть сообщения о разработке подобных приборов рядом китайских фирм. Так, в августе 2018 года корпорация YMTC представила прорывную архитектуру Xtacking, позволяющую создавать приборы с числом уровней более 200 (при этом поверх флеш-памяти можно размещать слой логики – с использованием гибридных соединений). В сентябре 2020 года другая китайская фирма – IC League – описала технологию гетерогенной интеграции флеш-памяти и логики на кристалле (HITOC), предназначенную для создания ИС для систем искусственного интеллекта [3]. Капиталовложения ведущих производителей флеш-памяти в производственные мощности и НИОКР на 2021–2022 годы и дальнейшую перспективу во многом связаны с разработкой 3D-приборов и структур с более чем 200 уровнями памяти. Переход с планарных на 3D-архитектуры флеш-памяти

Таблица 1. Темы докладов по 3D-флеш-памяти NAND-типа на ISSCC-2022

Фирмы/организации, представившие доклад	Тема доклада
Western Digital (Милпитас, шт. Калифорния) и KIOXIA (Йокогама, Япония)	«162-слойная 3D-флеш-память емкостью 1 Тбит на 4-битных ячейках, разделенная на 4 плоские матрицы, с быстродействующим интерфейсом (2,4 Гбит/с) устройств ввода/вывода»
Micron Technology (Сан-Хосе, шт. Калифорния, и Авеццано, Италия)	«3D-флеш-память NAND-типа емкостью 1 Тбит на 4-битных ячейках и 176-уровневой технологии, разделенная на 4 независимые плоские матрицы, для считывания с КМОП-структурой под матрицей»
SK Hynix (Инчхон, Ю. Корея)	«3D-флеш-память NAND-типа емкостью 1 Тбит на 4-битных ячейках и 176 уровнями числовых шин с уменьшенной задержкой считывания и плотностью памяти 14,8 Гбит/мм^2»
Samsung Electronics (Ю. Корея)	«3D-флеш-память NAND-типа 8-го поколения емкостью 1 Тбит на 3-битных ячейках со скоростью записи 164 Мбайт/с и быстродействием интерфейса в 2,4 Гбит/с»
Macronix, Национальный университет Цинхуа, Центральная научно-исследовательская академия (Тайвань), Городской университет Гонконга (Гонконг, КНР)	«512-Гбит 3D-флеш-память NAND для вычислений в памяти, поддерживающая операции сопоставления сходных векторов на приборах с краевым ИИ»

NAND-типа требует инноваций в области материалов и оборудования.

Общие тенденции развития 3D-флеш-памяти NAND-типа

Переход к трехмерности увеличивает емкость и снижает издержки

Переход к 3D-структурам был вызван потребностью существенного увеличения емкости памяти. Этажирование 2D-NAND-подобных слоев (уровней) вело к росту числа этапов технологического процесса, что значительно увеличило издержки производства. Основная идея «истинного» 3D-NAND-подхода заключается в этажировании ячеек в вертикальную строку, что обеспечивает бóльшую емкость на единицу площади. В этой конфигурации ячейки по-прежнему адресуются горизонтальными числовыми шинами. Увеличивая число уровней вместо уменьшения проектных норм, производители флеш-памяти NAND-типа отказались от классического масштабирования. Первые коммерческие продукты 3D-AND появились в 2013 году – у них было 24 слоя числовых шин, а емкость достигала 128 Гбит. В зависимости от поставщика существуют вариации структур – V-NAND, BICS...

От плавающих затворов к ловушкам заряда

Для уменьшения сложности 3D-процесса и увеличения емкости памяти были введены и продолжают вводиться различные инновации. Так, увеличение числа хранимых в ячейке бит данных до четырех является существенным технологическим преимуществом. Такие многоуровневые ячейки используют 16 дискретных уровней заряда в каждом отдельном транзисторе, что обеспечивается достаточно большим окном памяти.

Другим заметным новшеством стала замена ячейки с плавающим затвором ячейкой с ловушкой заряда, что упрощает технологический процесс. Принцип работы обоих типов ячеек относительно схож, но в ячейке с ловушкой заряда улавливающий слой представляет собой изолятор (обычно нитрид кремния), что обеспечивает меньшие электростатические помехи между соседними ячейками. В настоящее время ячейки с ловушками заряда являются основой большинства 3D-NAND-структур.

Направление развития – увеличение удельной плотности памяти

В целях сохранения действенности маршрутной карты развития технология флеш-памяти NAND-типа

Мощные и надёжные анализаторы спектра

АКИП-4205
АКИП-4212
АКИП-4213

- Диапазон частот: 9 кГц… 1,5 ГГц/ 2,1 ГГц/ 3,2 ГГц/ 5 ГГц/ 7,5 ГГц
  Анализатор спектра реального времени (АКИП-4213):
  полоса анализа 25 МГц, опция - 40 МГц; POI от 7,2 мкс
- Встроенный трекинг генератор, предусилитель
- Средний уровень собственных шумов: -161 дБм/Гц
- Уровень фазовых шумов: до -95 дБн/Гц при отстройке 10 кГц (на fнес=1 ГГц)
- Разрешение ПЧ 1 Гц
- Измерительные функции: измерение мощности в канале и соотношение мощностей в смежных каналах, измерение мощности во временной области, измерение ширины занимаемой полосы частот, соотношение сигнал шум, фильтры ЭМС и квазипиковый детектор, анализ параметров модуляции АМн, ЧМн, ФМн, QAM, АМ, ЧМ; измерение коэффициента стоячей волны (VSWR) и коэффициента затухания.

prist.ru

У точности есть имя!
ПРИСТ

119071, **г. Москва**, 2-й Донской пр., д. 10, стр. 4; тел.: +7 (495) 777-5591; факс: +7 (495) 640-3023
196006, **г. Санкт-Петербург**, ул. Цветочная, д. 18, лит. В, офис 202; тел./факс: +7 (812) 677-7508
620089, **г. Екатеринбург**, ул. Цвиллинга, д. 58, офис 1; тел./факс: +7 (343) 317-3999; ek@prist.ru

некоторые крупные производители недавно объявили об увеличении числа уровней до 500 и более. Предполагается, что число слоев NAND-флеш ИС к 2030 году может дойти до 1000. Увеличение числа уровней приводит к росту сложности технологического процесса, в частности операций осаждения и травления, а также вызывает накопление напряженности внутри слоев. Решая некоторые из этих проблем, производители недавно начали разделять уровни на два или более ярусов, этажируя индивидуально обработанные ярусы друг на друга.

Существуют опасения, что без серьезных инноваций эта эволюция постепенно приведет к снижению экономической эффективности изделий флеш-памяти NAND-типа. Увеличение числа слоев требует инвестиций в перспективные инструменты для нанесения покрытий и травления. А тенденция этажирования нескольких ярусов значительно увеличит число шаблонов, этапов обработки и длительность процесса. Это может привести к замедлению реализации упомянутой маршрутной карты и 1000-уровневые этажерки до 2030 года не появятся.

По мере увеличения числа слоев растет необходимость уменьшения толщины слоя и регулирования высоты этажерки – по причинам формирования рисунка и снижения напряженностей. Одно из предлагаемых решений – вертикальное масштабирование шага

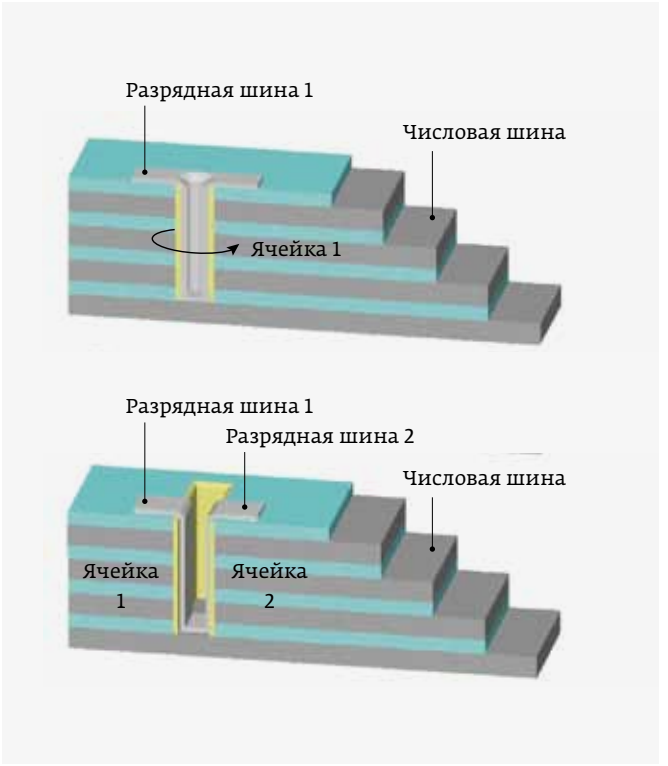


Рис. 3. Сопоставление GAA NAND-ячейки (верхняя) и ячейки с канавкой (нижняя)

элементов, подразумевающее снижение высоты слоев всех используемых в этажерке материалов, включая металлы и оксиды числовых шин.

Вертикальное масштабирование шага

Вертикальное масштабирование шага (по оси *Z*), вероятно, будет дополнено дальнейшим снижением горизонтальных (по осям *X* и *Y*) размеров. Это может потребовать серьезных инноваций в ячейке памяти, остававшейся неизменной на протяжении многих лет разработки 3D-флеш-памяти NAND-типа. Поэтому, в качестве альтернативы современным ячейкам флеш-памяти на транзисторах с круговым затвором (GAA), сейчас изучаются новые материалы и архитектуры ячеек.

Одной из примечательных разработок является подход, напоминающий архитектуру с использованием канавок. Ячейки реализуются на боковых стенках канавки с двумя транзисторами на ее противоположных концах, что значительно увеличивает плотность памяти в битах. С точки зрения эксплуатации эта ячейка с канавкой напоминает плоскую элементарную ячейку, поставленную вертикально (рис. 3).

Ячейка с канавкой, предлагаемая для архитектуры флеш-памяти NAND следующего поколения как альтернативное GAA-ячейкам решение, позволит сократить как площадь самой ячейки, так и шаг в плоскости X-Y со 140 нм до примерно 30 нм. Правда при этом подходе происходит небольшое ухудшение электрических характеристик, например, окна программирования / стирания [4].

Проблемы развития технологии 3D-флеш-памяти с точки зрения поставщиков материалов и оборудования

Для поддержания неуклонного развития сектора 3D-флеш-памяти NAND-типа необходимы специальные производственные материалы и оборудование, позволяющие решать сложные технические задачи. Так, системы травления должны обеспечивать возможность формирования глубоких канальных отверстий от верхней части прибора до подложки. От установок осаждения требуется получение высококачественных, бездефектных тонких пленок толщиной в несколько нанометров. Для мониторинга технологических процессов и поддержания высоких уровней выхода годных нужны высокоточные и инструментальные средства метрологии / контроля. Большая часть усилий в области НИОКР также направлена на поиск новых решений в области материалов. Например, в настоящее время интенсивно исследуются новые материалы жестких масок с высокой селективностью (избирательностью), такие как карбиды, легированные металлами. Для формирования контактных линий (токопроводящих дорожек) требуется

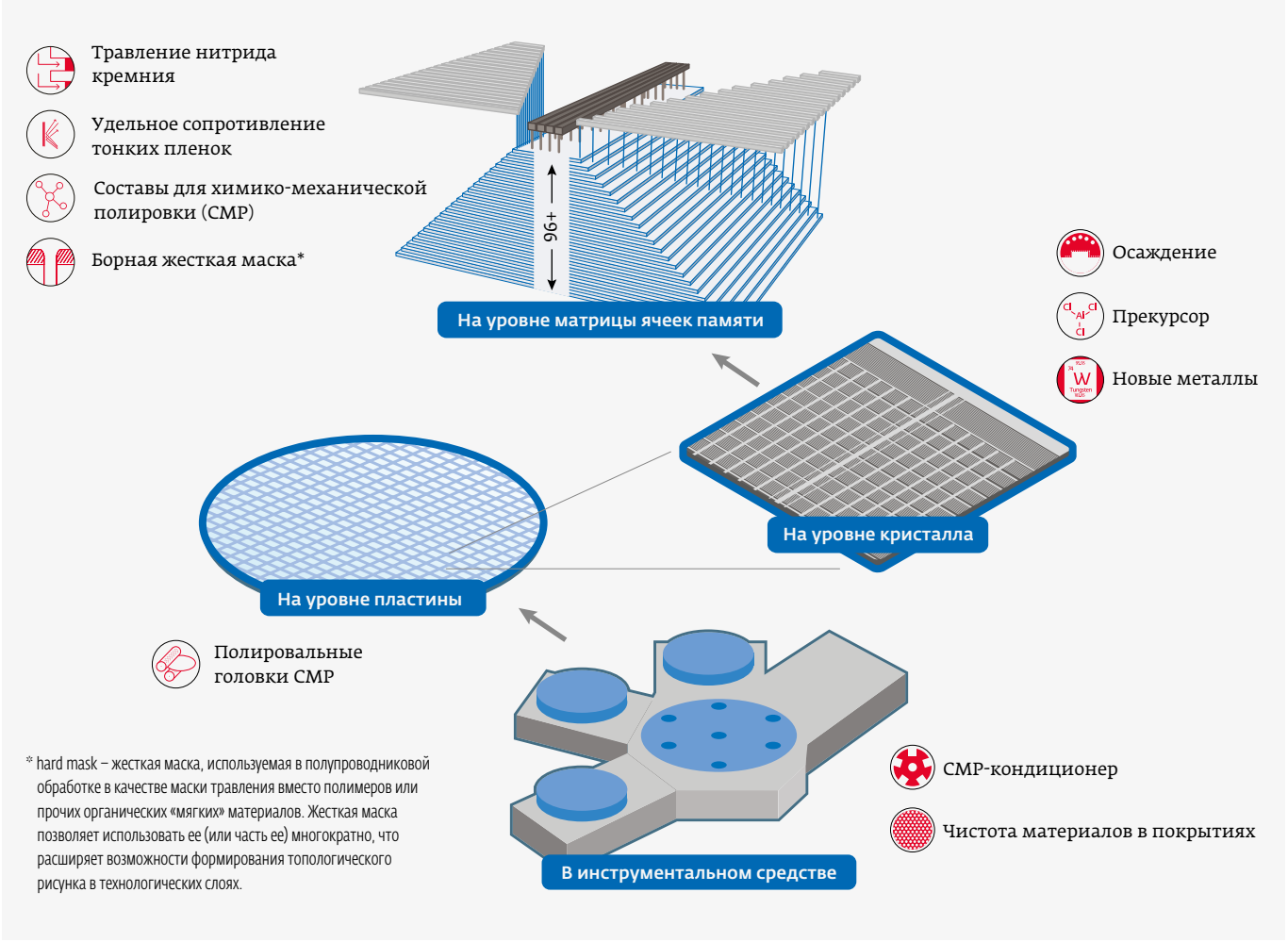


Рис. 4. Решение проблем интеграции и материалов 3D-флеш-памяти NAND-типа при увеличении числа слоев до 192 и более

использование новых металлов, необходимы альтернативные прекурсоры (материалы-предшественники) для осаждения вольфрама (W), новые материалы каналов (позволяющие избежать ухудшения подвижности носителей заряда) и многое другое. Специалисты фирмы Entegris предлагают рассматривать решение проблем интеграции и материалов 3D-флеш-памяти как комплексную задачу от уровня пластины до уровня технологического оборудования (рис. 4).

С точки зрения поставщика материалов самыми большими возможностями и проблемами являются окна процесса* и высокая чувствительность ячеек памяти к присутствию металлических загрязнений (источником которых могут быть прекурсоры и контейнеры / сосуды). Особенность окон процесса – обеспечение избирательного жидкостного травления SiN (а также травления выемок для металлических затворов, ALD**-слоев ячеек, ALD металлических затворов, специальных операций, связанных с жесткой маской для травления отверстий с высоким аспектным отношением) – имеет решающее значение при высоких аспектных отношениях и массивных матричных структурах.

Существуют и другие важные факторы, способствующие развитию технологий 3D-флеш-памяти NAND-типа. Например, это сухое травление отверстия канала с высоким аспектным отношением (классические химические методы сухого травления исчерпали свои возможности), снижение общей проводимости при увеличении длины канального слоя, избирательное удаление SiN,

* process window – окно процесса, набор значений параметров процесса, позволяющий изготавливать ИС и работающий при желаемых спецификациях. Например, окно процесса литографии обычно определяется как набор пунктов (фокус, экспозиция и т. д.) для контроля разбросов критических размеров в пределах 10%.

** ALD (atomic layer deposition) – атомарно-слоевое осаждение слоев.

Таблица 2. Доклады различных фирм/организаций по ДОЗУ на ISSCC-2022

Фирма/организация	Название доклада
SK Hynix (Ю. Корея)	«192-Гбит, 12-слойное 896-Гбит/с HBM3 ДОЗУ со схемой автокалибровки и оптимизацией топологии на основе машинного обучения»
Samsung Electronics, (Ю. Корея)	«16-Гбит, 27-Гбит/с/вывод GDDR6 ДОЗУ с объединенными шиной MUX TX, оптимизированными WCK операциями и альтернативной шиной данных»
	«16-Гбит, 9,5-Гбит/с/вывод LPDDR5X синхронное ДОЗУ с маломощными схемами динамического масштабирования напряжения/частоты и усилителями считывания со смещением, калиброванные по смещению, реализованное по 10-нм ДОЗУ-процессу 4-го поколения»
Университет Корё, Инчхонский национальный университет (оба Ю. Корея)	«0,385-пДж/бит, 10-Гбит/с двухкодовый приемопередатчик с выравниванием задержки, корректирующим кодом и калибровкой рассогласований для интерфейсов HBM»
Сеульский национальный университет (Ю. Корея)	«78,8-фДж/бит/мм, 12,0-Гбит/с/шина – емкостно-управляемый внутрикристальный канал связи более 5,6 мм с использованием FFE-комбинированной методики с принудительным смещением по заземлению для глобальной шины в рамках 65-нм КМОП-процесса»
Samsung Electronics, SK Hynix, Корейский институт авиакосмических исследований, Корейский институт передовых технологий (оба Ю. Корея)	«Сеть распределения тактовых импульсов, подавляющая дрожание, вызванное помехами питания, для мобильных LPDDR5 ДОЗУ с адаптивным фильтром 2-го порядка»

необходимость оптимизации или замены материалов-предшественников (прекурсоров), травителях, тонких пленок, оборудования и т. п. [5].

НЕКОТОРЫЕ ПЕРСПЕКТИВЫ РАЗВИТИЯ ДОЗУ

По мнению многих отраслевых экспертов, дальнейшее развитие ДОЗУ будет связано с развитием таких перспективных направлений, как технологии DDR5, 3D ДОЗУ и использование ДОЗУ в высокопроизводительной памяти HBM3 (high-bandwidth memory, 3rd generation). Это подтверждается названиями докладов, представленных в феврале 2022 года на Международной конференции по твердотельным ИС (International Solid-State Circuit Conference) (табл. 2) [2].

Развитие технологии DDR5

Развитие центров обработки данных (ЦОД) обуславливает ужесточение требований к используемым схемам памяти. Необходимо повышение емкости и пропускной способности. Одним из вариантов решения этой проблемы является технология DDR5, пятое поколение технологии ввода/вывода данных через интерфейс ДОЗУ с удвоенной скоростью (double data rate, DDR) ДОЗУ. Одной из фирм-разработчиков, активно продвигающих технологию DDR5, является корпорация Rambus.

Она уже закладывает основу для внедрения экосистемы поддержки DDR5, которое планируется не ранее, чем через год. В начале 2021 года ее заказчикам были представлены опытные образцы своих DDR5 регистрирующих синхрогенераторов (RCD) 2-го поколения с быстродействием 5600 млн передач данных в секунду (MT/s) (рис. 5). В DDR5 RCD наряду с буферами данных (DB) будут использоваться в DDR5 RDIMM и DDR5 DIMM с уменьшенной нагрузкой (Load Reduced DIMM, LRDIMM) для обеспечения более высокой пропускной способности, производительности и емкости по сравнению с небуферизованными DIMM. RDIMMS и LRDIMMs применяются для снижения нагрузки на центральный процессор и улучшения целостности сигнала* командных/адресных шин. Роль RCD – функционирование в качестве одного из основных кристаллов ИС плоскости управления,

* Signal Integrity – целостность сигналов, наличие достаточных для безошибочной передачи качественных характеристик электрического сигнала.

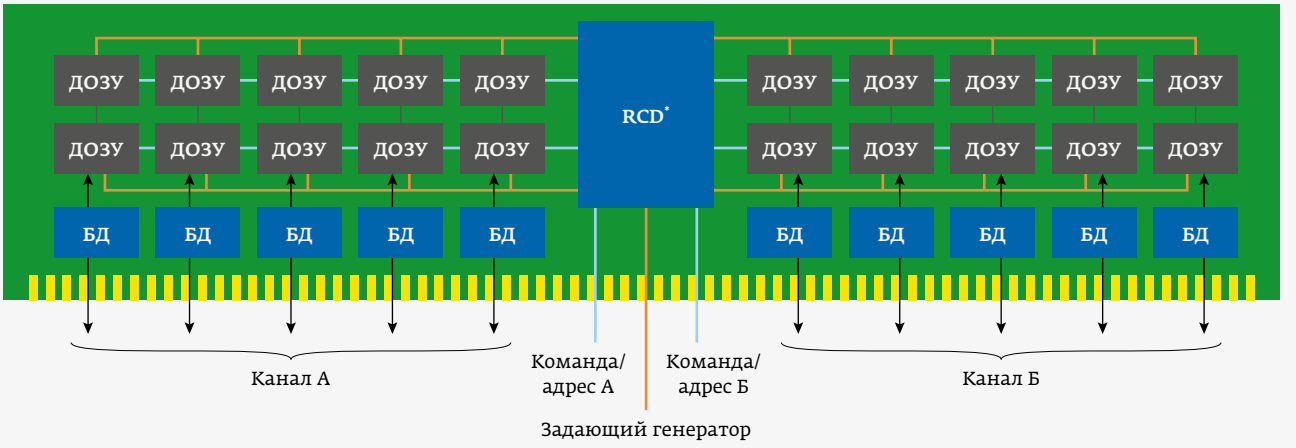
ИНТЕЛЛЕКТ. КАЧЕСТВО.

АО «МИКРОВОЛНОВЫЕ СИСТЕМЫ»
Москва, Щелковское шоссе, д.5, стр.1
Тел. (499) 644-21-03, (499) 644-25-62
(многоканальный)
Факс +7(499) 644-19-70
E-mail: mwsystems@mwsystems.ru
www.mwsystems.ru

- СОВРЕМЕННОЕ ПРОИЗВОДСТВО И ТЕХНОЛОГИИ
- ОПТИМАЛЬНОЕ СООТНОШЕНИЕ ЦЕНА/КАЧЕСТВО

- ПОЛНЫЙ СПЕКТР УСЛУГ ПО ПРОЕКТИРОВАНИЮ И ПРОИЗВОДСТВУ МОНОЛИТНЫХ И ГИБРИДНЫХ ИНТЕГРАЛЬНЫХ СХЕМ, ТВЕРДОТЕЛЬНЫХ МОДУЛЕЙ, МНОГОФУНКЦИОНАЛЬНЫХ СВЧ-УСТРОЙСТВ И БЛОКОВ РЭА (0,3 - 22 ГГц)

АКЦИОНЕРНОЕ ОБЩЕСТВО «МИКРОВОЛНОВЫЕ СИСТЕМЫ»



* RCD (Registering Clock Driver) – регистрирующий формирователь синхронизирующих импульсов, синхронизатор и распределитель тактового сигнала. БД – буфер данных.

Рис. 5. Отличие DDR5 DIMM от DDR4 DIMM – большая степень параллелизма и увеличенная скорость передачи данных, что обеспечивает большие емкость и пропускную способность

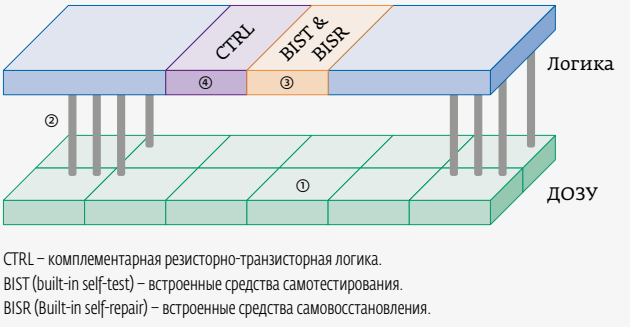
который распределяет сигналы команд / адресов и синхронизирует кристаллы ДОЗУ в DIMM. Один RCD корпорации Rambus может поддерживать DDR5 LRDIMM совместно с десятком буферов данных на модуль. Это позволяет снизить нагрузку на шину данных и использовать в модуле ДОЗУ большей емкости – без увеличения времени ожидания.

Отличительной особенностью памяти DDR5 является то, что в модули DIMM встроено больше интеллектуальных функций. Это позволяет вдвое увеличить скорость передачи данных и в четыре раза увеличить пропускную способность по сравнению с модулями DIMM DDR4. Рабочие характеристики новейшего DDR5 RCD корпорации Rambus обеспечивают увеличение скорости передачи данных на 17% по сравнению с DDR5 RCD первого поколения (4 800 MT / s) при меньших времени ожидания и потребляемой мощности. Кроме того, оптимизированы параметры синхронизации [6].

3D не только для флеш-памяти NAND-типа

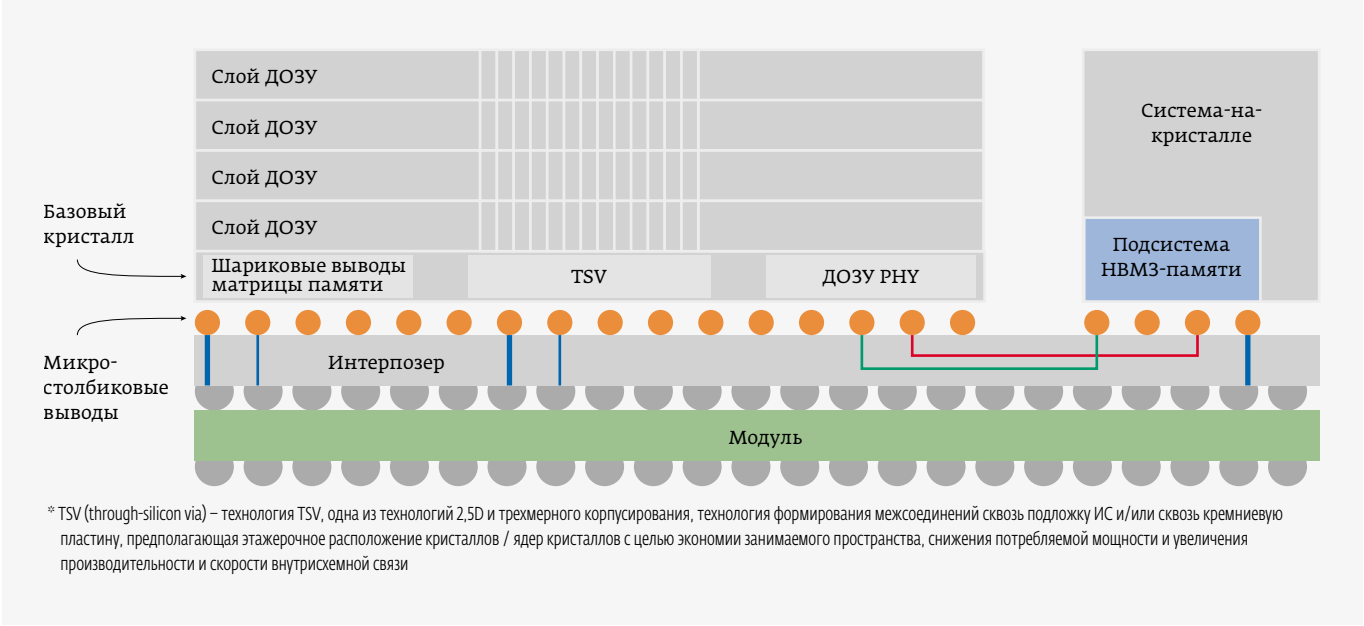
Создание 3D-ДОЗУ считается возможным уже сейчас, но неясно, как это произойдет. Один из способов их создания предложила фирма Xi'an UniICSemiconductors, использующая технологию 3D-интеграции для формирования встраиваемых ДОЗУ. Предполагается, что эти приборы преодолеют проблему «стены памяти», присущей принстонской архитектуре. Эта проблема – разрыв в производительности между процессором и памятью, который становится все больше, на передачу данных тратится больше времени, чем на обработку данных, для передачи используется больше энергии, чем для реальных вычислений.

Созданное специалистами Xi'an UniICSemiconductors этажированное встраиваемое ДОЗУ (stacked embedded DRAM (SeDRAM) использует 3D гибридный процесс соединения, при этом логика помещается поверх SeDRAM (рис. 6). Некоторые из преимуществ этого подхода включают в себя межсоединения межслойного уровня в одном кристалле ИС – от SoC до ДОЗУ, гибкий интерфейс «логика-память» и вертикальные межсоединения. Конструкция не требует быстродействующей шины данных и дополнительных крупных приборов физического уровня (PHY) в SoC и ДОЗУ. Энергопотребление, необходимое



CTRL – комплементарная резисторно-транзисторная логика.
BIST (built-in self-test) – встроенные средства самотестирования.
BISR (Built-in self-repair) – встроенные средства самовосстановления.

Рис. 6. Этажированное встраиваемое ДОЗУ (SeDRAM) использует 3D-гибридное присоединение, при этом логика размещается поверх SeDRAM



* TSV (through-silicon via) – технология TSV, одна из технологий 2,5D и трехмерного корпусирования, технология формирования межсоединений сквозь подложку ИС и/или сквозь кремниевую пластину, предполагающая этажерочное расположение кристаллов / ядер кристаллов с целью экономии занимаемого пространства, снижения потребляемой мощности и увеличения производительности и скорости внутрисхемной связи

Рис. 7. 2,5D/3D-архитектура системы с НВМ3-памятью

для передачи данных, снижается по сравнению с другими вариантами интеграции ИС ЗУ. Конструкция системы намного проще, а затраты намного ниже.

Некоторые применения SeDRAM включают в себя кэш последнего уровня, который сочетает в себе логическую пластину центрального процессора и пластину SEDRAM. Это делает его недорогим решением для кэша с очень высокой плотностью или как часть «сэндвич-решения» для системы в одном кристалле с верхним слоем для вычислений, средним слоем для кэша данных и нижним слоем для хранения данных. Это многообещающее решение для чувствительных к потребляемой мощности приложений, таких как Интернет вещей, позволяющее достичь низкого уровня энергопотребления и в то же время очень малых форм-факторов [7].

Перспективная память с высокой пропускной способностью

Хотя реализации НВМ3 могут показаться достаточно простыми, это обманчивое впечатление. Поскольку эти ЗУ часто используются в критически важных приложениях, обеспечение их работы должным образом требует дополнительных усилий. Для оценки того, насколько хорошо конечная система в целом будет работать с системами/подсистемами НВМ3, пользователи обычно используют среду тестирования и моделирования, предоставляемую поставщиком.

Специалисты корпорации Rambus отмечают, что с точки зрения общей эффективности системы, одной из проблем НВМ всегда была реализация на физическом уровне – из-за малой площади. В перспективных конструкциях, использующих центральный процессор или графический процессор, а также четыре или больше НВМ ДОЗУ (занимающих относительно мало места) (рис. 7), основными вопросами при реализации на физическом уровне также являются потребляемая мощность, целостность сигнала, производственная надежность.

В целях получения максимальной производительности интерпозера и конструкции корпуса/модуля, когда скорости передачи данных доходят до 3,2 и 3,6 Гбит/с, многие фирмы-производители пытаются добиться хорошей целостности сигнала именно за счет интерпозеров. Ситуация осложняется тем, что у каждого кремниевого завода существуют свои нормы проектирования интерпозеров: у одних сложнее, у других – проще. С помощью НВМ3 можно увеличить число слоев памяти и расширить возможности интерпозера – толщину диэлектрика и т. д. Это немного облегчает ряд проблем. Но даже на приборах предыдущего поколения многие клиенты не могли понять: как эта штука работает со скоростью 3,2 Гбит/с? [8].

Рассмотрим в качестве примера развитие технологии НВМ3 корпорации SK Hynix. В июне прошлого года она представила 12-слойное НВМ3 ДОЗУ с максимальной скоростью передачи данных в 5,2 Гбит/с на вывод и пропускной способностью 665 Гбайт/с. К октябрю ее производительность была улучшена на 23%, а на ISSCC-2022 представители корпорации описывали систему пропускной способностью 896 Гбайт/с, то есть производительность выросла еще на 10% (табл. 3). В целом, емкость 12-слойного НВМ3 ДОЗУ составляет 196 Гбит (24 Гбайт).

Таблица 3. Память с высокой пропускной способностью корпорации SK Hynix разных поколений

Параметры	HBM2	HBM2e	HBM3
Число устройств ввода/вывода (интерфейс шины)	1024	1024	1024
Максимальная скорость передачи данных/пропускная способность	256 Гбайт/с	460,8 Гбайт/с	5,2 Гбит/с/665 Гбайт/с (июнь 2021) 6,4 Гбит/с/819 Гбайт/с (октябрь 2021) **7,0 Гбит/с/896 Гбайт/с (февраль 2022)**
Максимальное число слоев ДОЗУ	8	8	12
Максимальная емкость	8 Гбайт	16 Гбайт	24 Гбайт
Напряжение	1,2 В	1,2 В	ТВА*

* TBA (to be advised/to be announced) – рассматривается, будет объявлено дополнительно о новой продукции, еще не выведенной на рынок.

Это было достигнуто за счет автоматической калибровки TSV* и оптимизации топологии с помощью машинного обучения [9].

Вопросы применения ИС ЗУ в модулях / системах с ИИ

На ISSCC-2022 китайская корпорация Alibaba описала свое устройство для рекомендательных систем, созданное методом гибридного соединения логики и ДОЗУ, осуществляющее вычисления в непосредственной близости к памяти. Его основные характеристики – быстродействие 184 запроса в секунду (queries-per-second, QPS) на ватт и удельная емкость памяти 64 Мбит/мм^2. То есть речь идет о высокопроизводительных распределенных системах, предназначенных для хранения и обработки данных в оперативной памяти в реальном масштабе времени (in-memory computing) с использованием ИИ, производительность которых на порядки (по утверждениям специалистов Alibaba до 1 000 раз) выше, чем системы с дисковыми накопителями. Такие системы ускоряют обработку больших объемов данных и, по мере расширения использования технологий больших данных, приобретают все большую популярность. Здесь снова возникает проблема «стены памяти» (memory wall) – ограничения производительности вычислительной системы в целом пропускной способностью памяти, из-за чего процессоры не могут работать на полную мощность. Преодоление этой проблемы имеет решающее значение для вычислений ИИ, учитывая быстрое ужесточение требований к вычислительным возможностям моделей ИИ (табл. 4).

* TSV (through-silicon via) – технология TSV, одна из технологий 2,5D- и 3D-корпусирования, технология формирования межсоединений сквозь подложку ИС и/или сквозь кремниевую пластину, предполагающая этажерочное расположение кристаллов/ядер кристаллов с целью экономии занимаемого пространства, снижения потребляемой мощности и увеличения производительности и скорости внутрисхемной связи

В приборе корпорации Alibaba для подключения многобанковой ДОЗУ непосредственно к логическим устройствам ИИ-процессоров используется гибридное соединение. Обычно, размер кристаллов доступных на рынке ДОЗУ довольно мал, меньше 50 мм^2, отчасти из-за необходимости обеспечения высокого выхода годных и ограничений стандарта JEDEC. Размеры же 3D-кристалла «Логика-на-ДОЗУ» от Alibaba существенно больше – 602,22 мм^2. При разработке этой системы важной задачей было проектирование логики и соответствующей ДОЗУ как полноценной системы с несколькими банками ДОЗУ, напрямую подключающимися к многоядерной логике под ними.

В дальнейших планах Alibaba – расширение ее концепции 3D – «Логика-на-ДОЗУ» до крупноформатного кристалла масштаба пластины – наподобие изделия Wafer-Scale-Engine (CS-2) корпорации Cerebra. Но в CS-2 используются только СОЗУ и его емкость – 40 Гбайт. Если же специалистам Alibaba удастся создать аналогичное изделие на основе ДОЗУ, то емкость его памяти превысит 1 Тбайт, или увеличится, по крайней мере, в 25 раз.

Превосходство разработки Alibaba над обычными системами на основе связки центрального процессора (ЦП) и ДОЗУ иллюстрируется табл. 5, 6. По сравнению

Таблица 4. Проблема «стены памяти» в эпоху искусственного интеллекта

Параметр	Фактор роста за два года
Увеличение вычислительных потребностей модели ИИ	750
Рост производительности аппаратного обеспечения	3,1
Ускорение пропускной способности системы памяти	1,4

Radel

САМОЕ ПОСЕЩАЕМОЕ ОТРАСЛЕВОЕ МЕРОПРИЯТИЕ СЕВЕРО-ЗАПАДА РОССИИ!*

21-23 СЕНТЯБРЯ 2022 САНКТ-ПЕТЕРБУРГ

XXI МЕЖДУНАРОДНАЯ ВЫСТАВКА

РАДИОЭЛЕКТРОНИКА & ПРИБОРОСТРОЕНИЕ

НАПОЛНЯЙТЕ КЛИЕНТСКУЮ БАЗУ - ОСНОВУ ВАШЕЙ ЭКОСИСТЕМЫ!

*Выставку 2021 года посетили более 7 700 специалистов

FarEXPO IFE

www.radelexpo.ru
(812) 718-35-37

Таблица 5. Сопоставление систем на основе ЦП-ДОЗУ и «Логики-на-ДОЗУ» по пропускной способности, энергоэффективности и удельной эффективности площади

Параметры	ЦП-ДОЗУ	«Логика-на-ДОЗУ»	
		Стандартное значение	Пиковое значение
Пропускная способность, QPS	41	401	512
Энергоэффективность, QPS/Вт	0,58	184,11	235,08
Эффективность использования площади, QPS/мм^2	0,0095	6,27	8,0

Таблица 6. Сопоставление систем на основе ЦП-ДОЗУ и «Логики-на-ДОЗУ» по проектным нормам, частоте, площади и потребляемой мощности

Параметры	ЦП*-ДОЗУ	«Логика-на-ДОЗУ»
Проектные нормы, нм	14	55
Частота, ГГц	2,2	0,3
Площадь, мм^2	4294	64 (4 Гбит)
Потребляемая мощность**	70,17	2,178

* Центральный процессор – Intel Xeon Cold 5220, 2,2 ГГц, тестирование с использованием Pytorch.

** Для ЦП измерена с помощью PyRAPL.

с системой ЦП-ДОЗУ разработка Alibaba обладает пропускной способностью, большей в 9,78 раза, энергоэффективность лучше в 317,43 раза, а эффективность использования площади – в 660 раз. При этом пропускная способность и емкость памяти могут быть улучшены за счет увеличения количества блоков гибридной связи или использования более совершенных технологических процессов (для обслуживания более сложных рекомендательных моделей). Что самое интересное, что эти результаты были достигнуты при использовании относительно зрелого 55-нм технологического процесса (для логики), а само сопоставление осуществлялось с 14-нм процессором Xeon Gold от Intel.

Кроме того, результаты разработчиков Alibaba существенно лучше, чем в случае вертикального кэша (V-Cache), наращивающего емкость памяти ЦП Ryzen корпорации AMD, также использующего гибридное соединение. Кроме того, у AMD шаг вертикальных соединений 9 мкм, а у китайцев – 3 мкм, а в некоторых случаях даже 1 мкм [3].

❊ ❊ ❊

Анализ развития сектора схем памяти показывает, что, несмотря на существенные успехи перспективных ИС ЗУ, доминирующими приборами до 2030 года и какое-то время после него будут оставаться ДОЗУ и флеш-память NAND-типа. Освоение 3D-подходов, начавшееся в сегменте флеш-памяти, уверенно охватывает и сегмент ДОЗУ. Многие перспективные типы ИС трехмерны по своей природе. 3D-интеграция позволяет увеличивать удельную емкость памяти, а также широко использовать, помимо минимальных новейших, и более зрелые проектные нормы, что способствует снижению сложности разработки, проектирования и производства. Все более широкое применение получают ИС ЗУ в системах с искусственным интеллектом.

ЛИТЕРАТУРА

1. **Hilson G.** Emerging Memories Look to Displace NOR, SRAM // EE Times. 08.30.2021.
2. ISSCC 2022 Advance Program 2-17-2022.
3. **Zvi Or-Bach.** China May Win in AI Computing // EE Times. 03.04.2022.
4. **Maarten Rosmeulen and Jan van Houdt.** How 3D NAND flash works, what lies ahead in its density roadmap // EDN. March 8. 2022.
5. **Bertolazzi S.** Focus on 3D NAND Manufacturing Challenges – An interview with Entegris // i-Micronews. April 29. 2021.
6. **Hilson G.** DDR5 Ecosystem Ramps Up // EE Times. 11.01.2021.
7. **Hilson G.** IMW Highlights 3D Architectures, In-Memory Computing // EE Times. 06.08.2021.
8. **Ann Steffora Mutschler.** HBM3: Big Impact On Chip Design // Semiconductor Engineering. October 14th. 2021.
9. SK hynix to discuss 24GB 896GB/s HBM3 memory, Samsung to showcase 27Gbps GDDR6 memory at ISSCC 2022 // VideoCardz.com. 18th Jan. 2022.

20-я МЕЖДУНАРОДНАЯ ВЫСТАВКА ЭЛЕКТРОНИКИ

ChipEXPO-2022

КОМПОНЕНТЫ | ОБОРУДОВАНИЕ | ТЕХНОЛОГИИ

ВЫСТАВКА ПРОЙДЕТ

13-15.09

В ТЕХНОПАРКЕ ИННОВАЦИОННОГО ЦЕНТРА

СКОЛКОВО

ТЕМАТИЧЕСКИЕ ЭКСПОЗИЦИИ:

- ☑ Предприятия радиоэлектронной промышленности России
- ☑ Поставщики электронных компонентов
- ☑ Участники конкурса "Золотой Чип"
- ☑ Новинки производителей электроники
- ☑ Стартапы в электроники (стенд Инновационного центра Сколково)
- ☑ Дизайн-центры электроники

ОФИЦИАЛЬНАЯ ПОДДЕРЖКА:

Sk Сколково

МИНПРОМТОРГ РОССИИ

Ростех

ОРГАНИЗАТОРЫ:

ЗАО «ЧипЭКСПО», 111141, Москва, Зеленый пр-т, д.2
Тел.: +7 (495) 221-50-15, E-mail: info@chipexpo.ru
http://www.chipexpo.ru