ФЕДЕРАЛЬНОЕ БЮДЖЕТНОЕ ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

МИКРОПРОЦЕССОРНЫЕ СИСТЕМЫ
ЛЕКЦИЯ №9

«Подсистема памяти микропроцессорной системы» (продолжение)

Лектор:
доцент каф. ЭАФУ ФТИ
Горюнов А.Г.

Томск 2012 г.

План лекции

9.1 Оперативные запоминающие устройства (ОЗУ);

9.1.1 Статические ОЗУ;

9.1.2 Динамические ОЗУ;

9.2 Буферная память;

9.3 Кэш-память.

9.1 Оперативные запоминающие устройства (ОЗУ)

Операти́вная па́мять (также оперативное запоминающее устройство, ОЗУ) — в информатике — память, часть системы памяти ЭВМ, в которую процессор может обратиться за одну операцию (jump, move и т. п.).

Предназначена для временного хранения данных и команд, необходимых процессору для выполнения им операций.

Оперативная память передает процессору данные непосредственно, либо через кеш-память.

Каждая ячейка оперативной памяти имеет свой индивидуальный адрес.

ОЗУ может изготавливаться как отдельный блок или входить в конструкцию однокристальной ЭВМ или микроконтроллера.

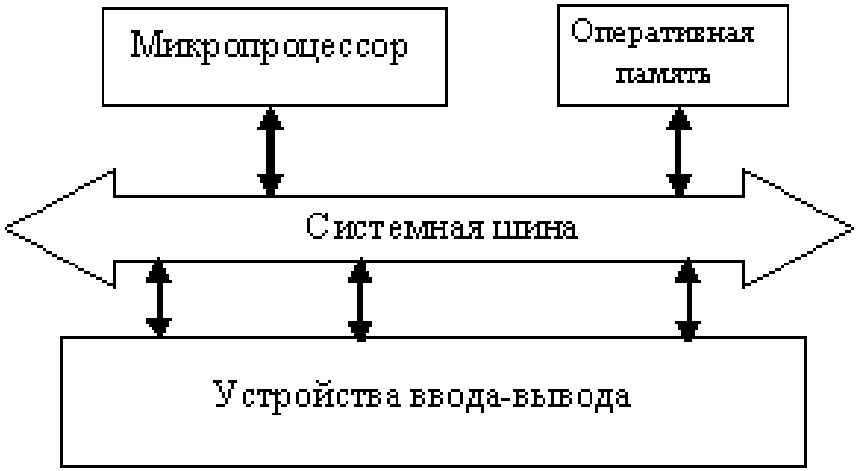


Рисунок 9.1 – Простейшая схема взаимодействия оперативной памяти с ЦП

На сегодня наибольшее распространение имеют два вида ОЗУ:

- **SRAM (Static RAM)**
- **DRAM (Dynamic RAM)**

9.1.1 Статическое ОЗУ

Статическая оперативная память с произвольным доступом *(SRAM — Static Random Access Memory)* — полупроводниковая оперативная память, в которой каждый двоичный разряд хранится в схеме с положительной обратной связью, позволяющей поддерживать состояние сигнала без постоянной перезаписи.

Тем не менее, сохранять данные без перезаписи SRAM может только, пока есть питание, т.е. SRAM остается энергозависимым типом памяти.

Двоичная SRAM

Типичная ячейка статической двоичной памяти (двоичный триггер) на КМОП-технологии состоит из двух перекрестно (кольцом) включенных инверторов и ключевых транзисторов для обеспечения доступа к ячейке.

Часто для увеличения плотности упаковки элементов на кристалле в качестве нагрузки применяют поликремниевые резисторы.

Недостатком такого решения является рост статического энергопотребления.

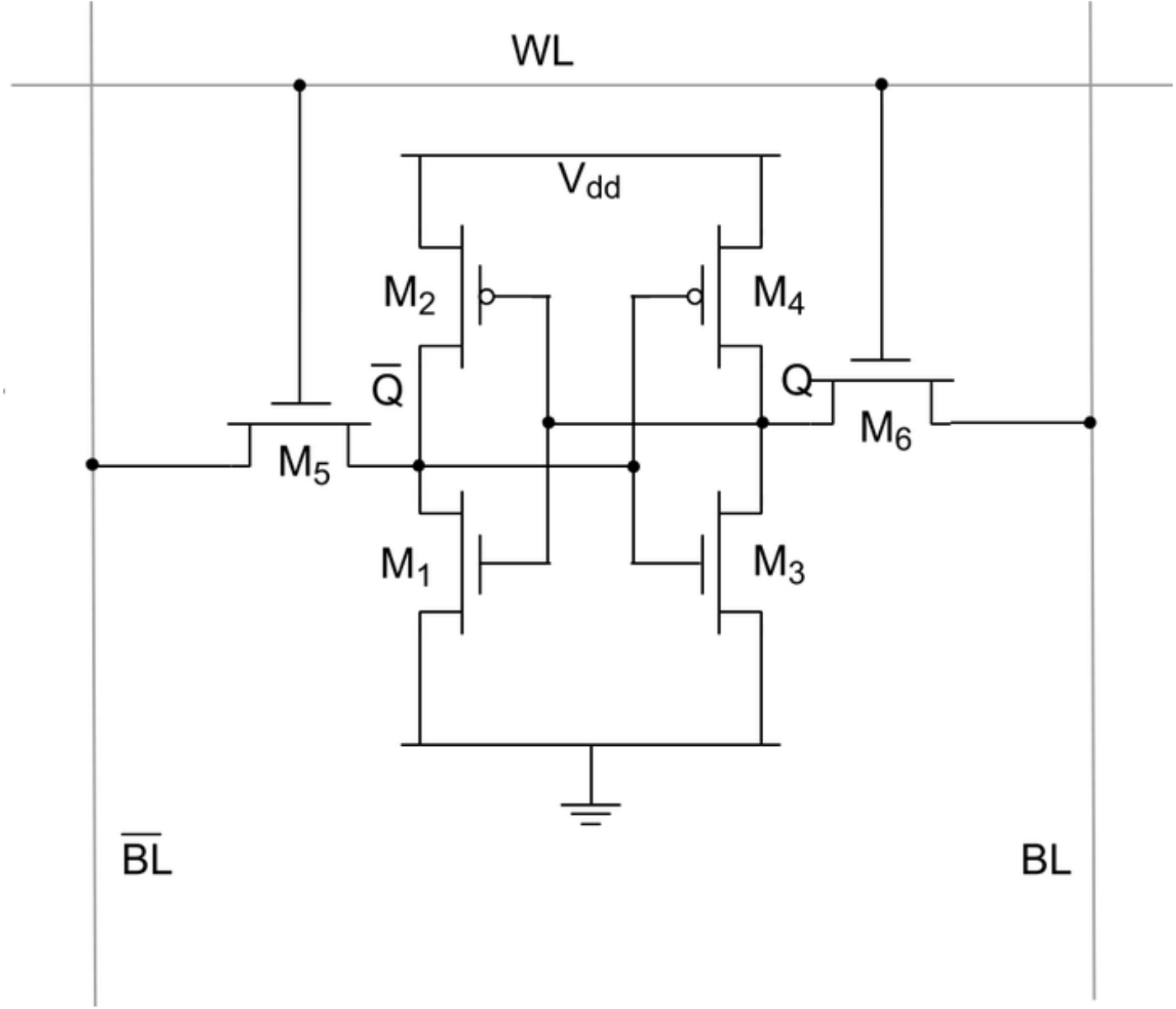


Рисунок 9.2 – Типичная ячейка
статической двоичной памяти

Линия WL (Word Line) управляет двумя транзисторами доступа. Линии !BL и BL (Bit Line) – битовые линии, используются и для записи данных и для чтения данных.

Запись. При подаче «0» на линию !BL или BL параллельно включенные транзисторные пары (M5 и M1) и (M6 и M3) образуют логические схемы 2ИЛИ, последующая подача «1» на линию WL открывает транзистор M5 или M6, что приводит к соответствующему переключению триггера.

Чтение. При подаче «1» на линию WL открываются транзисторы M5 и M6, уровни записанные в триггере выставляются на линии !BL и BL, и попадают на схемы чтения.

Для выбора ячеек (WL) используются дешифратор адреса.

Logic block diagram

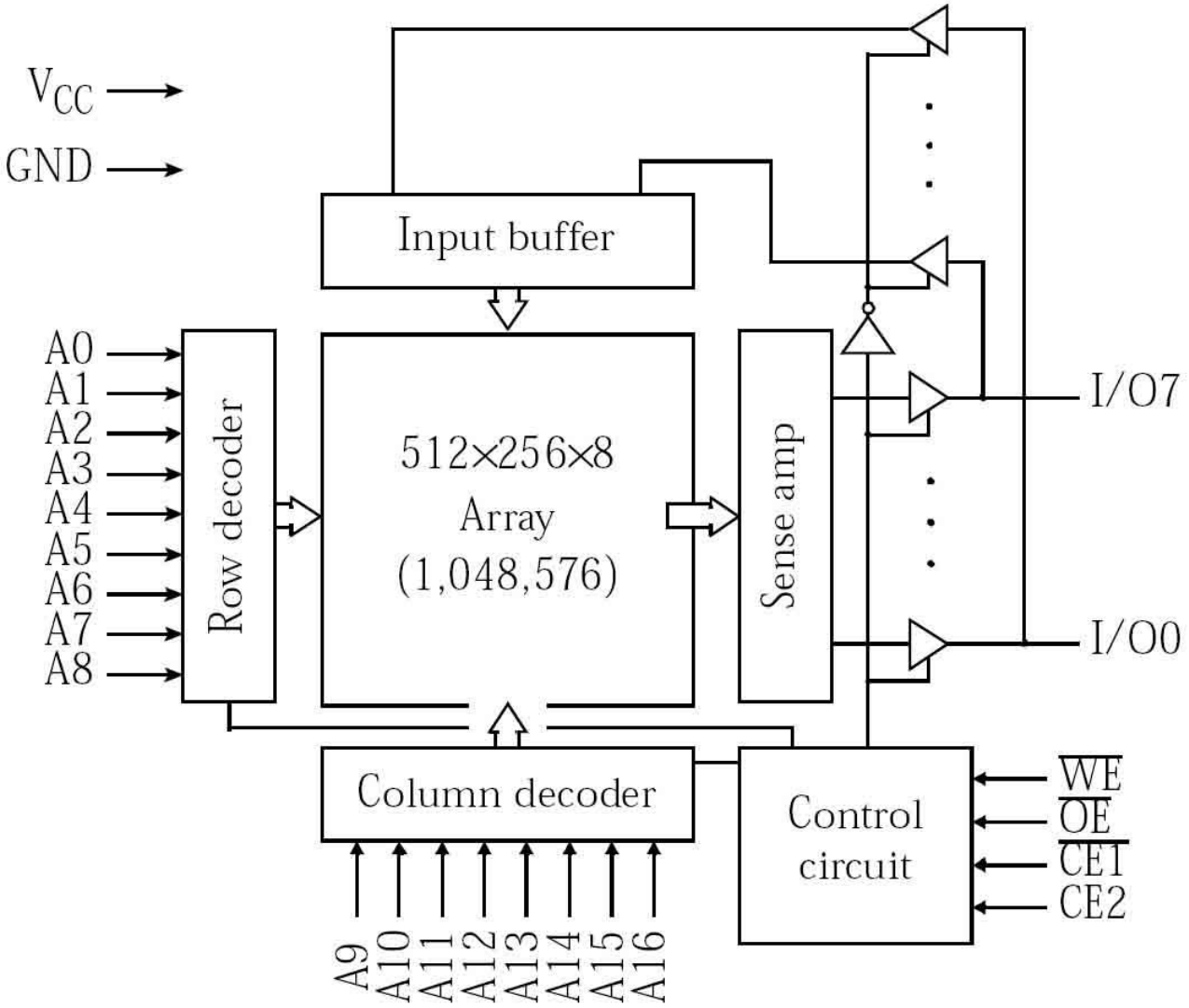


Рисунок 9.3 – Дешифратор адреса

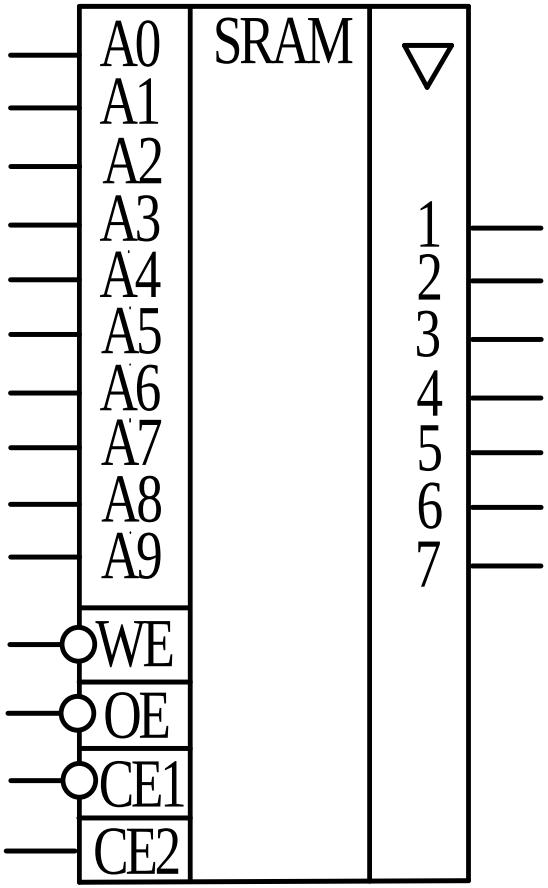
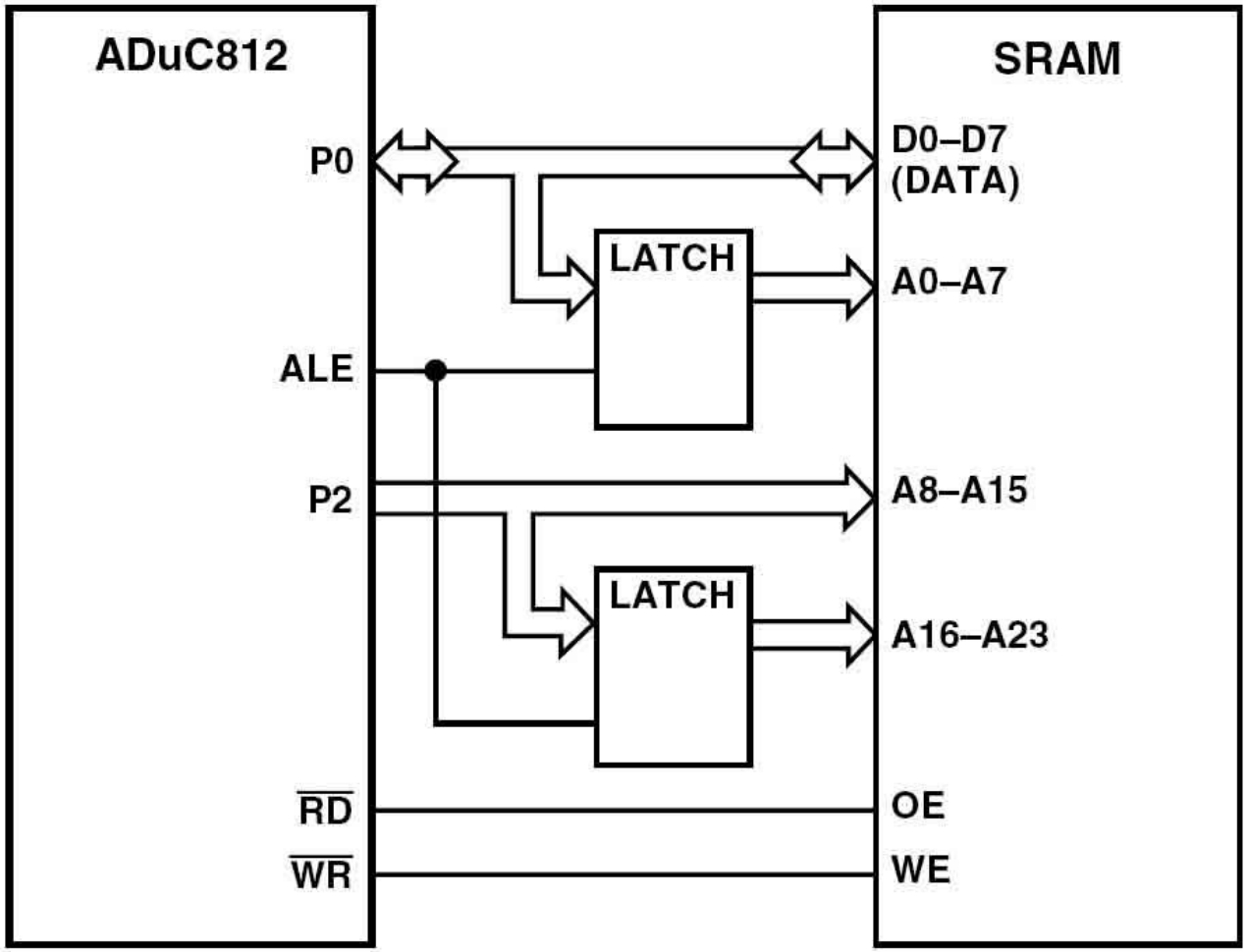


Рисунок 9.4 – Условно-графическое обозначение микросхем SRAM



External Data Memory Interface (16 M Bytes Address Space)

Рисунок 9.5 – Пример подключения SRAM к микроконтроллеру

Преимущества:

- Быстрый доступ. SRAM — это действительно память произвольного доступа, доступ к любой ячейке памяти в любой момент занимает одно и то же время.

- Простая схемотехника — SRAM не требуются сложные контроллеры.

- Возможны очень низкие частоты синхронизации, вплоть до полной остановки синхроимпульсов.

Недостатки:

- Невысокая плотность записи (шесть элементов на бит, вместо двух у DRAM).

- Высокое энергопотребление.

Вследствие чего — дороговизна килобайта памяти.

Тем не менее, высокое энергопотребление не является принципиальной особенностью SRAM, оно обусловлено высокими скоростями обмена с данным видом внутренней памяти процессора.

Энергия потребляется только в момент изменения информации в ячейке SRAM.

Применение

- SRAM применяется в микроконтроллерах и ПЛИС, в которых объем ОЗУ невелик (единицы килобайт), зато нужны низкое энергопотребление (за счет отсутствия сложного контроллера динамической памяти), предсказываемое с точностью до такта время работы подпрограмм и отладка прямо на устройстве.

- В устройствах с большим объемом ОЗУ рабочая память выполняется как DRAM.

- SRAM’ом же делают регистры и кэш-память.

9.1.2 Динамические ОЗУ

Динамическая память — **Dynamic RAM** — получила свое название от принципа действия ее запоминающих ячеек, которые выполнены в виде конденсаторов, образованных элементами полупроводниковых микросхем.

При отсутствии обращения к ячейке со временем за счет токов утечки конденсатор разряжается и информация теряется, поэтому такая память требует периодической подзарядки конденсаторов (обращения к каждой ячейке) — память может работать только в динамическом режиме.

Этим она принципиально отличается от статической памяти, реализуемой на триггерных ячейках и хранящей информацию без обращений к ней сколь угодно долго (при включенном питании).

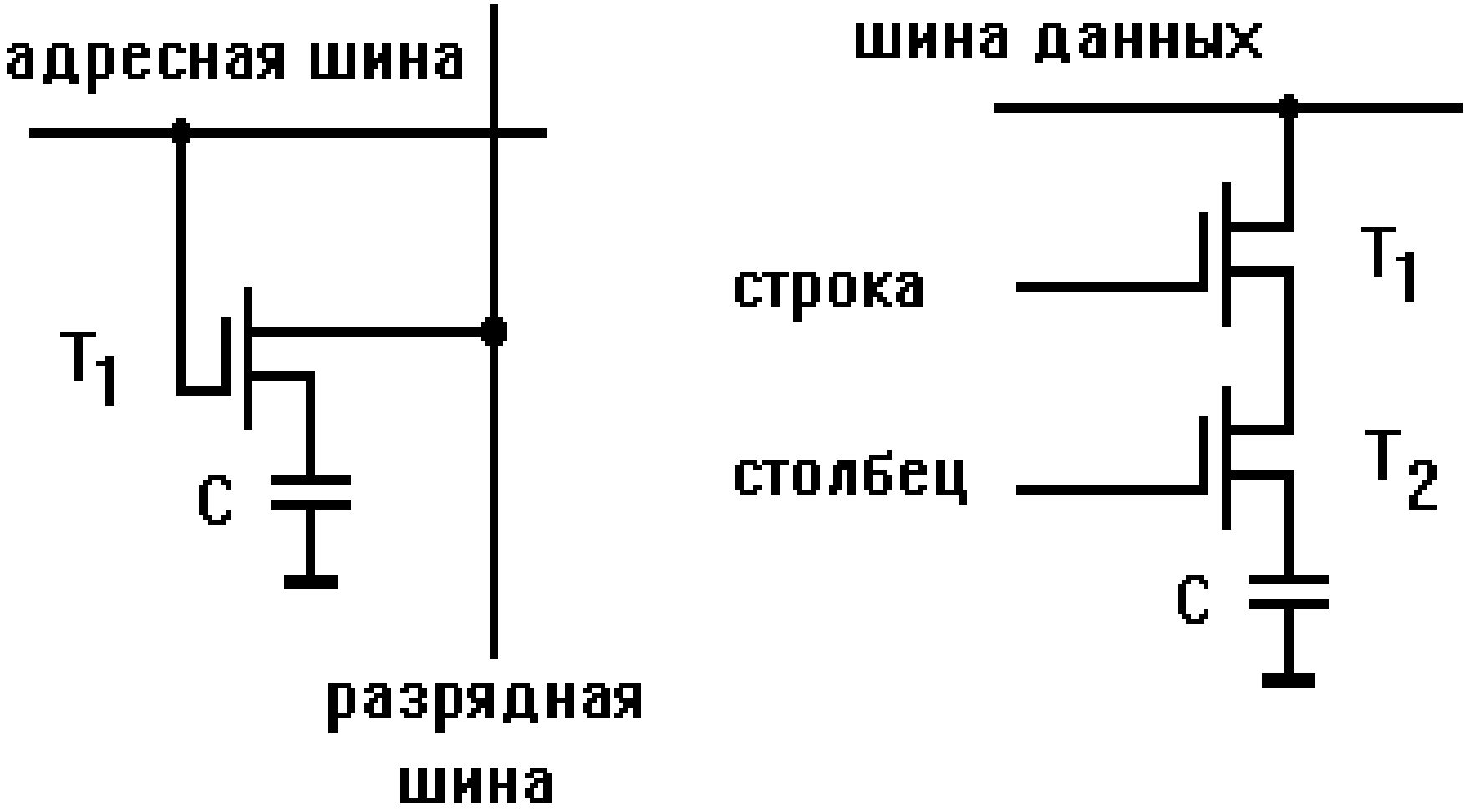


Рисунок 9.6 – Запоминающая ячейка динамического ОЗУ

Запоминающие ячейки микросхем DRAM организованы в виде двумерной матрицы.

Адреса строки и столбца передаются по мультиплексированной шине адреса **MA** (Multiplexed Address) и стробируются по спаду импульсов **RAS#** (Row Access Strobe) – строка и **CAS#** (Column Access Strobe) – столбец.

Совокупность ячеек DRAM образуют условный «прямоугольник», состоящий из определённого количества строк и столбцов.

Один такой «прямоугольник» называется страницей, а совокупность страниц называется банком.

Весь набор ячеек условно делится на несколько областей.

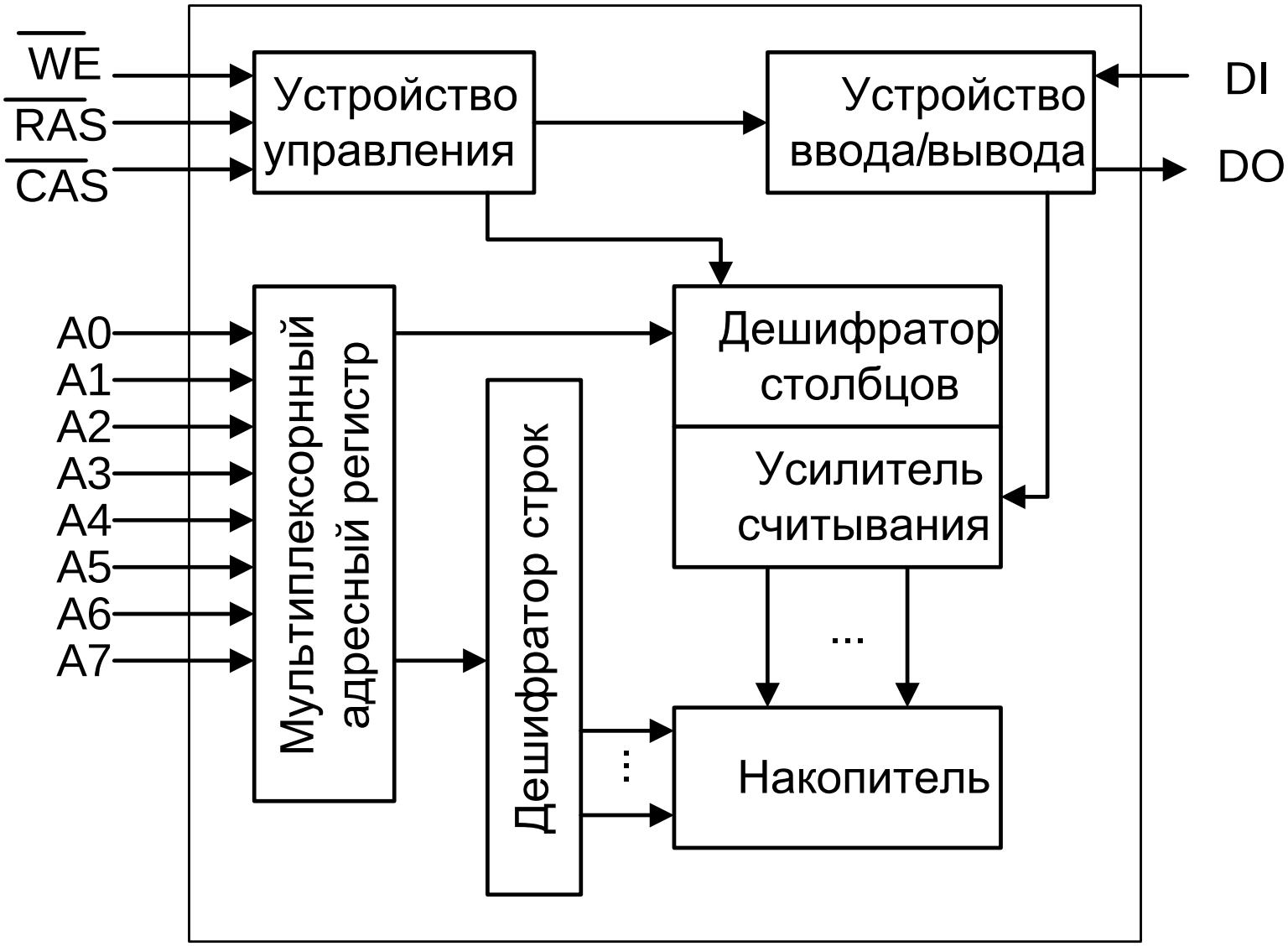


Рисунок 9.7 – Типовая структура микросхемы динамической ОЗУ

Основными характеристиками DRAM являются рабочая частота и тайминги.

При обращении к ячейке памяти контроллер памяти задает номер банка, номер страницы в нем, номер строки и номер столбца и на все эти запросы тратится время, помимо этого довольно большой период уходит на открытие и закрытие банка после самой операции.

На каждое действие требуется время, называемое «**таймингом**».

Основными «**таймингами**» DRAM являются:

- задержка между подачей номера строки и номера столбца, называемая **временем полного доступа** (англ. **RAS to CAS delay**);

- задержка между подачей номера столбца и получением содержимого ячейки, называемая **временем рабочего цикла** (англ. **CAS delay**);

- задержка между чтением последней ячейки и подачей номера новой строки (англ. **RAS precharge**).

Тайминги измеряются в наносекундах, и чем меньше величина этих таймингов, тем быстрее работает оперативная память.

Типы DRAM

Перечислить типы DRAM. Студенты должны рассмотреть самостоятельно.

Страничная память

Страничная память (англ. page mode DRAM, PM DRAM) являлась одним из первых типов выпускаемой компьютерной оперативной памяти.

Память такого типа выпускалась в начале 90-х годов, но с ростом производительности центральных процессоров и ресурсоемкости приложений требовалось увеличивать не только объём памяти, но и скорость ее работы.

Быстрая страничная память

Быстрая страничная память (англ. fast page mode DRAM, FPM DRAM) появилась в 1995 году.

Принципиально новых изменений память не претерпела, а увеличение скорости работы достигалось путем повышенной нагрузки на аппаратную часть памяти.

Данный тип памяти в основном применялся для компьютеров с процессорами Intel 80486 или аналогичных процессоров других фирм.

Память могла работать на частотах 25 МГц и 33 МГц с временем полного доступа 70 нс и 60 нс и с временем рабочего цикла 40 нс и 35 нс соответственно.

EDO DRAM — память с усовершенствованным выходом

С появлением процессоров Intel Pentium память FPM DRAM оказалась совершенно неэффективной.

Поэтому следующим шагом стала память с усовершенствованным выходом (англ. extended data out DRAM, EDO DRAM).

Эта память появилась на рынке в 1996 году и стала активно использоваться на компьютерах с процессорами Intel Pentium и выше. Ее производительность оказалась на 10—15 % выше по сравнению с памятью типа FPM DRAM.

Ее рабочая частота была 40 МГц и 50 МГц, соответственно, время полного доступа — 60 нс и 50 нс, а время рабочего цикла — 25 нс и 20 нс.

Эта память содержит регистр-защелку (англ. data latch) выходных данных, что обеспечивает некоторую конвейеризацию работы для повышения производительности при чтении.

SDRAM — синхронная DRAM

В связи с выпуском новых процессоров и постепенным увеличением частоты системной шины, стабильность работы памяти типа EDO DRAM стала заметно падать.

Ей на смену пришла синхронная память (англ. synchronous DRAM, SDRAM).

Новыми особенностями этого типа памяти являлись использование тактового генератора для синхронизации всех сигналов и использование конвейерной обработки информации.

Также память надежно работала на более высоких частотах системной шины (100 МГц и выше).

Если для FPM и EDO памяти указывается время чтения первой ячейки в цепочке (время доступа), то для SDRAM указывается время считывания последующих ячеек.

Цепочка — несколько последовательных ячеек.

На считывание первой ячейки уходит довольно много времени (60-70 нс) независимо от типа памяти, а вот время чтения последующих сильно зависит от типа.

Рабочие частоты этого типа памяти могли равняться 66 МГц, 100 МГц или 133 МГц, время полного доступа — 40 нс и 30 нс, а время рабочего цикла — 10 нс и 7,5 нс.

С этим типом памяти применялась одна интересная технология — Virtual Channel Memory (VCM).

VCM использует архитектуру виртуального канала, позволяющую более гибко и эффективно передавать данные с использованием каналов регистра на чипе.

Данная архитектура интегрирована в SDRAM. VCM, помимо высокой скорости передачи данных, была совместима с существующими SDRAM, что позволяло делать апгрейд системы без значительных затрат и модификаций.

Это решение нашло поддержку у некоторых производителей чипсетов.

Enhanced SDRAM (ESDRAM)

Для преодоления некоторых проблем с задержкой сигнала, присущих стандартной DRAM памяти, было решено встроить небольшое количество SRAM в чип, то есть создать на чипе кэш.

ESDRAM — это по существу SDRAM плюс немного SRAM. При малой задержке и пакетной работе достигается частота до 200 МГц. Как и в случае внешней кэш-памяти, DRAM-кэш предназначен для хранения и выборки наиболее часто используемых данных.

Отсюда и уменьшение времени доступа к данным медленной DRAM.

Одним из таких решений, заслуживающих внимания, являлась ESDRAM от Ramtron International Corporation.

Пакетная EDO RAM

Пакетная память EDO RAM (англ. burst extended data output DRAM, BEDO DRAM) стала дешевой альтернативой памяти типа SDRAM.

Основанная на памяти EDO DRAM, ее ключевой особенностью являлась технология поблочного чтения данных (блок данных читался за один такт), что сделало ее работу быстрее, чем у памяти типа SDRAM.

Однако невозможность работать на частоте системной шины более 66 МГц не позволила данному типу памяти стать популярным.

Video RAM

Специальный тип оперативной памяти Video RAM (VRAM) был разработан на основе памяти типа SDRAM для использования в видеоплатах.

Он позволял обеспечить непрерывный поток данных в процессе обновления изображения, что было необходимо для реализации изображений высокого качества.

На основе памяти типа VRAM, появилась спецификация памяти типа Windows RAM (WRAM), иногда ее ошибочно связывают с операционными системами семейства Windows.

Ее производительность стала на 25 % выше, чем у оригинальной памяти типа SDRAM, благодаря некоторым техническим изменениям.

DDR SDRAM

По сравнению с обычной памятью типа SDRAM, в памяти SDRAM с удвоенной скоростью передачи данных (англ. **double data rate SDRAM, DDR SDRAM или SDRAM II**) была вдвое увеличена пропускная способность.

Первоначально память такого типа применялась в видеоплатах, но позднее появилась поддержка DDR SDRAM со стороны чипсетов.

У всех предыдущих DRAM были разделены линии адреса, данных и управления, которые накладывают ограничения на скорость работы устройств.

Для преодоления этого ограничения в некоторых технологических решениях все сигналы стали выполняться на одной шине.

Двумя из таких решений являются технологии DRDRAM и SLDRAM.

Они получили наибольшую популярность и заслуживают внимания.

Стандарт SLDRAM является открытым и, подобно предыдущей технологии, SLDRAM использует обе перепада тактового сигнала.

Что касается интерфейса, то SLDRAM перенимает протокол, названный SynchLink Interface и стремится работать на частоте 400 МГц.

Память DDR SDRAM работает на частотах в 100, 133, 166 и 200 МГц, ее время полного доступа — 30 нс и 22,5 нс, а время рабочего цикла — 5 нс, 3,75 нс, 3 нс и 2,5 нс.

Так как частота синхронизации лежит в пределах от 100 до 200 МГц, а данные передаются по 2 бита на один синхроимпульс, как по фронту, так и по срезу тактового импульса, то эффективная частота передачи данных лежит в пределах от 200 до 400 МГц.

Такие модули памяти обозначаются DDR200, DDR266, DDR333, DDR400.

Direct RDRAM, или Direct Rambus DRAM

Тип памяти RDRAM является разработкой компании Rambus. Высокое быстродействие этой памяти достигается рядом особенностей, не встречающихся в других типах памяти.

Первоначальная очень высокая стоимость памяти RDRAM привела к тому, что производители мощных компьютеров предпочли менее производительную, зато более дешевую память DDR SDRAM.

Рабочие частоты памяти — 400 МГц, 600 МГц и 800 МГц, время полного доступа — до 30 нс, время рабочего цикла — до 2,5 нс.

DDR2 SDRAM

Конструктивно новый тип оперативной памяти DDR2 SDRAM был выпущен в 2004 году.

Основываясь на технологии DDR SDRAM, этот тип памяти за счёт технических изменений показывает более высокое быстродействие и предназначен для использования на современных компьютерах.

Память может работать с тактовой частотой шины 200, 266, 333, 337, 400, 533, 575 и 600 МГц.

При этом эффективная частота передачи данных соответственно будет 400, 533, 667, 675, 800, 1066, 1150 и 1200 МГц.

Некоторые производители модулей памяти помимо стандартных частот выпускают и образцы, работающие на нестандартных (промежуточных) частотах.

Они предназначены для использования в разогнанных системах, где требуется запас по частоте.

Время полного доступа — 25 нс, 11,25 нс, 9 нс, 7,5 нс и менее.

Время рабочего цикла — от 5 нс до 1,67 нс.

DDR3 SDRAM

Этот тип памяти основан на технологиях DDR2 SDRAM с увеличенной еще вдвое частотой передачи данных по шине памяти. отличается пониженным энергопотреблением по сравнению с предшественниками.

Частота полосы пропускания лежит в пределах от 800 до 2400 МГц (рекорд частоты - более 3 000 МГц), что обеспечивает большую пропускную способность по сравнению со всеми предшественниками.

Рассмотреть SDRAM и DDR.

9.2 Буферная память

Эффективно обмен данными между подсистемами с различным быстродействием реализуется при наличии между ними специальной буферной памяти.

Рисунок 9.8 – Обмен между медленнодействующей и быстродействующей памятью посредством буферной памяти

Данные от подсистемы 1 временно запоминаются в буферной памяти до готовности подсистемы 2 принять их. Емкость буферной памяти должна быть достаточной для хранения тех блоков данных, которые подсистема 1 формирует между считываниями их подсистемой 2.

Отличительной особенностью буферной памяти является запись данных с быстродействием и под управлением подсистемы 1, а считывание – с быстродействием и под управлением подсистемы 2 ("эластичная память").

В общем случае память должна выполнять операции записи и считывания совершенно независимо и даже одновременно, что устраняет необходимость синхронизации подсистем.

Буферная память должна сохранять порядок поступления данных от подсистемы 1, т. е. работать по принципу "первое записанное слово считывается первым" (**First Input First Output – FIFO**).

Таким образом, под буферной памятью типа **FIFO** понимается ЗУПВ, которое автоматически следит за порядком поступления данных и выдает их в том же порядке, допуская выполнение независимых и одновременных операций записи и считывания.

На рисунке 9.9 приведена структурная схема буферной памяти типа FIFO емкостью 64x4.

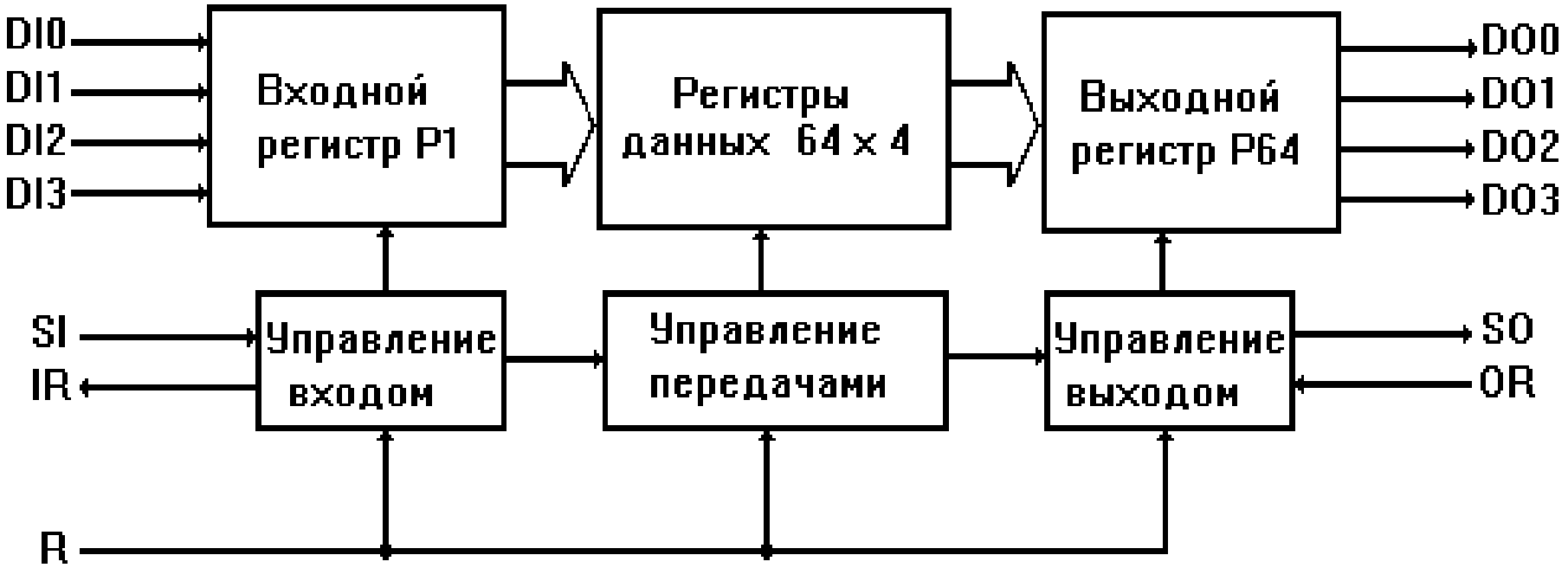


Рисунок 9.9 – Структурная схема буферной памяти типа FIFO емкостью 64x4.

DI – линии входных данных; SI – shift in – запись данных; DO – линии выходных данных; SO – shift out – считывание данных; IR – input ready – готовность ввода данных; OR – output ready – готовность вывода данных; R – сброс содержимого буфера.

На кристалле размещены 64 4-битных регистра с независимыми цепями сдвига, организованных в 4-х последовательных 64-битных регистрах данных, 64-битный управляющий регистр, а также схема управления.

Буферы можно наращивать как по числу слов, так и по их длине.

9.3 Кэш

Кэш или **кеш** (англ. **cache**, от фр. **cacher** — прятать; произносится [kæʃ] — кэш) — **промежуточный буфер с быстрым доступом**, содержащий копию той информации, которая хранится в памяти с менее быстрым доступом, но с наибольшей вероятностью может быть оттуда запрошена.

Доступ к данным в кэше идет быстрее, чем выборка исходных данных из медленной памяти или их перевычисление, за счет чего уменьшается среднее время доступа.

История

Впервые слово «кэш» в компьютерном контексте было использовано в 1967 году во время подготовки статьи для публикации в журнале «IBM Systems Journal».

Статья касалась усовершенствования памяти в разрабатываемой модели 85 из серии IBM System/360. Редактор журнала Лайл Джонсон попросил придумать более описательный термин, нежели «высокоскоростной буфер», но из-за отсутствия идей сам предложил слово «кэш».

Статья была опубликована в начале 1968 года, авторы были премированы IBM, их работа получила распространение и впоследствии была улучшена, а слово «кэш» вскоре стало использоваться в компьютерной литературе как общепринятый термин.

Функционирование КЕШ

Кэш состоит из набора записей. Каждая запись ассоциирована с элементом данных или блоком данных (небольшой части данных), которая является копией элемента данных в основной памяти. Каждая запись имеет идентификатор, определяющий соответствие между элементами данных в кэше и их копиями в основной памяти.

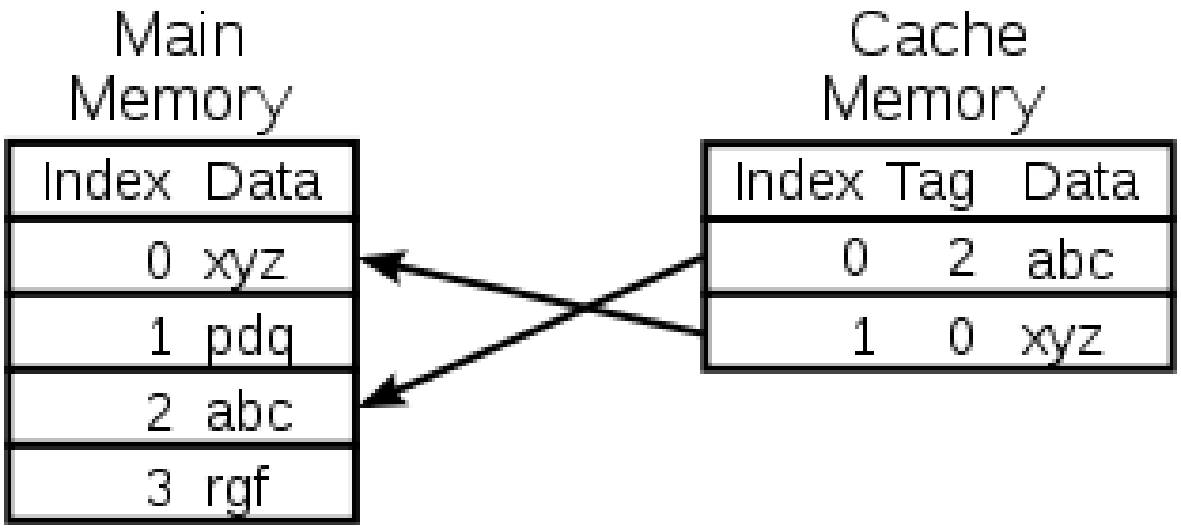


Рисунок 9.10 – Кэширование данных

1. Когда клиент кэша (ЦПУ) обращается к данным, прежде всего исследуется кэш.

2. Если в кэше найдена запись с идентификатором, совпадающим с идентификатором затребованного элемента данных, то используются элементы данных в кэше.

Такой случай называется попаданием в кэш.

3. Если в кэше не найдено записей, содержащих затребованный элемент данных, то он читается из основной памяти в кэш, и становятся доступным для последующих обращений.

Такой случай называется промахом кэша.

4. Процент обращений к кэшу, когда в нем найден результат, называется уровнем попаданий или коэффициентом попаданий в кэш.

5. Если кэш ограничен в объеме, то при промахе может быть принято решение отбросить некоторую запись для освобождения пространства.

6. Для выбора отбрасываемой записи используются разные алгоритмы вытеснения.

7. При модификации элементов данных в кэше выполняется их обновление в основной памяти.

8. Задержка во времени между модификацией данных в кэше и обновлением основной памяти управляется так называемой политикой записи.

9. В кэше с немедленной записью каждое изменение вызывает синхронное обновление данных в основной памяти.

10. В кэше с отложенной записью (или обратной записью) обновление происходит в случае вытеснения элемента данных, периодически или по запросу клиента.

11. Для отслеживания модифицированных элементов данных записи кэша хранят признак модификации (изменённый или «грязный»).

12. Промах в кэше с отложенной записью может потребовать два обращения к основной памяти: первое для записи заменяемых данных из кэша, второе для чтения необходимого элемента данных.

13. В случае, если данные в основной памяти могут быть изменены независимо от кэша, то запись кэша может стать неактуальной.

14. Протоколы взаимодействия между кэшами, которые сохраняют согласованность данных, называют протоколами когерентности кэша.

Кэш центрального процессора

Ряд моделей центральных процессоров (ЦП) обладают собственным кэшем, для того чтобы минимизировать доступ к оперативной памяти (ОЗУ), которая медленнее, чем регистры.

Кэш-память может давать значительный выигрыш в производительности, в случае когда тактовая частота ОЗУ значительно меньше тактовой частоты ЦП.

Тактовая частота для кэш-памяти обычно ненамного меньше частоты ЦП.

Уровни кэша

Кэш центрального процессора разделен на несколько уровней. Для универсальных процессоров — до 3.

Кэш-память уровня N+1 как правило больше по размеру и медленнее по скорости обращения и передаче данных, чем кэш-память уровня N.

Самой быстрой памятью является кэш первого уровня — L1-cache.

По сути, она является неотъемлемой частью процессора, поскольку расположена на одном с ним кристалле и входит в состав функциональных блоков.

L1 кеш состоит из кэша команд и кэша данных. Некоторые процессоры без L1 кэша не могут функционировать.

На других его можно отключить, но тогда значительно падает производительность процессора.

L1 кэш работает на частоте процессора, и, в общем случае, обращение к нему может производиться каждый такт (зачастую является возможным выполнять даже несколько чтений/записей одновременно).

Латентность доступа обычно равна 2–4 тактам ядра. Объём обычно невелик — не более 128 Кбайт.

Вторым по быстродействию является L2-cache — кэш второго уровня.

Обычно он расположен либо на кристалле, как и L1, либо в непосредственной близости от ядра, например, в процессорном картридже (только в слотовых процессорах).

В старых процессорах — набор микросхем на системной плате.

Объем L2 кэша от 128 Кбайт до 1−12 Мбайт.

В современных многоядерных процессорах кэш второго уровня, находясь на том же кристалле, является памятью раздельного пользования — при общем объеме кэша в 8 Мбайт на каждое ядро приходится по 2 Мбайта.

Обычно латентность L2 кэша, расположенного на кристалле ядра, составляет от 8 до 20 тактов ядра.

В отличие от L1 кэша, его отключение может не повлиять на производительность системы.

Однако, в задачах, связанных с многочисленными обращениями к ограниченной области памяти, например, СУБД, производительность может упасть в десятки раз.

Кэш третьего уровня наименее быстродействующий и обычно расположен отдельно от ядра ЦП, но он может быть очень внушительного размера — более 32 Мбайт.

L3 кэш медленнее предыдущих кэшей, но все равно значительно быстрее, чем оперативная память.

В многопроцессорных системах находится в общем пользовании.

Отключение кэша второго и третьего уровней обычно используется в математических задачах, например, при обсчете полигонов, когда объём данных меньше размера кэша.

В этом случае, можно сразу записать все данные в кэш, а затем производить их обработку.