15.Основная память. Проблема когерентности памяти. Время доступа к памяти, пропускная способность, многоканальный доступ, доступ с чередованием.

Основная память - это устройство для хранения информации. Она состоит из оперативного запоминающего устройства (ОЗУ) и постоянного запоминающего устройства (ПЗУ).
Оперативное запоминающее устройство (ОЗУ)
ОЗУ-быстрая, полупроводниковая, энергозависимая память. В ОЗУ хранятся исполняемая в данный момент программа и данные, с которыми она непосредственно работает. Это значит, что когда вы запускаете какую-либо компьютерную программу, находящуюся на диске, она копируется в оперативную память, после чего процессор начинает выполнять команды, изложенные в этой программе. Часть ОЗУ, называемая "видеопамять", содержит данные, соответствующие текущему изображению на экране. При отключении питания содержимое ОЗУ стирается.
ОЗУ - это память, используемая как для чтения, так и для записи информации. При отключении электропитания информация в ОЗУ исчезает (энергозависимость).
Постоянное запоминающее устройство (ПЗУ)
ПЗУ - быстрая, энергонезависимая память. ПЗУ - это память, предназначенная только для чтения. Информация заносится в нее один раз (обычно в заводских условиях) и сохраняется постоянно (при включенном и выключенном компьютере). В ПЗУ хранится информация, присутствие которой постоянно необходимо в компьютере.
Когерентность памяти (англ. memory coherence) — свойство компьютерных систем, содержащих более одного процессора или ядра, имеющих доступ к одной области памяти, заключающееся в том, что изменённая одним ядром/процессором ячейка памяти принимает новое значение для остальных ядер/процессоров.
В однопроцессорных системах (более строго — в одноядерных) работу с памятью выполняет один процессорный узел: только один узел может читать данные из памяти или записывать данные в память. После записи нового значения в ячейку памяти, доступную по какому-либо адресу, при чтении данных из той же ячейки будет получено записанное значение (даже при наличии кэширования).
В многопроцессорных (многоядерных) системах несколько процессорных узлов работают одновременно и могут одновременно (параллельно) обращаться к одной ячейке памяти (для чтения или для записи). Узлы могут одновременно прочитать значение из одной ячейки памяти, могут сохранить прочитанное значение в своих кешах. Как только один из узлов запишет в ячейку новое значение, значения, сохранённые в локальной памяти других узлов, должны помечаться как устаревшие. Необходим механизм уведомления всех узлов о том, что значение, сохранённое в их кешах, устарело; такой механизм называется протоколом когерентности (англ. memory coherence protocol). Если в системе используется подобный протокол, то говорят, что система имеет «когерентную память» (англ. coherent memory).
Точная природа и смысл механизма когерентности определяются моделью консистентности/связанности, реализованной в протоколе. Для составления правильных «параллельных» программ программисты должны знать о том, какая именно модель/способ консистентности/связанности кеш-памяти используется в их системах.
Если протокол когерентности/синхронизации реализован аппаратно, для выяснения применяемой модели консистентности/связанности программисты могут использовать сниффинг (снупинг) шины (англ.), могут читать специальные таблицы-справочники (англ. directory-based). В качестве примера протокола когерентности можно привести протокол MSI (англ. modified, shared, invalid) (англ.) и его разновидности (MESI (англ.), MOSI (англ.), MOESI, MESIF).
Пропускная способность
Как известно, главной характеристикой памяти является ее пропускная способность, то есть

максимальное количество данных, которое можно считать из памяти или записать в память в единицу времени. Именно эта характеристика прямо или косвенно отражается в названии типа памяти. Пропускная способность памяти зависит от ширины шины данных и частоты работы памяти. Ширина шины данных определяет количество бит, передаваемых за один такт, а частота работы памяти — количество тактов в единицу времени. Поэтому для того, чтобы определить пропускную способность памяти, нужно умножить частоту системной шины на ширину шины данных. Память SDRAM имеет 64-битную (8-байтную) шину данных.

К примеру, память DDR400, функционирующая на эффективной частоте 400 МГц, имеет пропускную способность 400 МГц x 8 байт = 3,2 Гбайт/с, а память DDR2-800, функционирующая на эффективной частоте 800 МГц, — 800 МГц x 8 байт = 6,4 Гбайт/с.

Кроме того, необходимо учесть, что синхронная память DDR и DDR2 может функционировать в двухканальном режиме и в этом случае максимальная пропускная способность удваивается.

Тайминги памяти

Кроме максимальной пропускной способности, память характеризуется латентностью. Причем во многих случаях латентность памяти оказывает большее влияние на производительность всей системы в целом, нежели тактовая частота работы памяти.

Под латентностью принято понимать задержку между поступлением команды и ее реализацией. Латентность памяти определяется ее таймингами, то есть задержками, измеряемыми в количествах тактов, между отдельными командами. Принято различать несколько разных таймингов памяти, соответствующих задержкам между различным командами. Для того чтобы разобраться с разными таймингами, рассмотрим последовательность команд при чтении или записи данных в память.

время доступа - Время, требующееся программе или устройству для получения порции данных и подготовки этой порции для обработки компьютером. Чипы DRAM (динамической памяти произвольного доступа) для персональных компьютеров имеют время доступа от 50 до 150 наносекунд (одна миллиардная секунды). Статическая память произвольного доступа (SRAM) имеет время доступа ниже 10 наносекунд. В идеале, время доступа память должно быть достаточно быстрым, чтобы не отставать от CPU (центрального процессора). Если это не так, Центральный Процессор будет тратить впустую некоторое число тактовых циклов, что делает его медленнее.

Обратите однако, внимание, что обещаемое время доступа может не соответствовать истинному, потому что большинство чипов памяти, особенно чипы Динамической Оперативной Памяти (DRAM), требует паузы между противоположными видами доступа (чтение после записи, запись после чтения). Это одна из причин, по которой статическая память (SRAM) намного быстрее динамической (DRAM), даже если говорят, что время доступа у них одинаковое; SRAM не требует регенерации, так что нет никакой паузы между доступами. Более важная мера скорости чипов - время цикла, т.е. время, через которое можно сделать следующее обращение к памяти после предыдущего противоположного обращения (запись после чтения или чтение после записи).

Понятие время доступа часто используется при описании скорости дисковых приводов. Время доступа диска измеряется в миллисекундах (одна тысячная секунды), обозначается как ms. Быстрые жесткие диски для персональных компьютеров имеют время доступа приблизительно от 9 до 15 миллисекунд. Обратите внимание, что это - приблизительно в 200 раз медленнее, чем средняя скорость DRAM.

Время доступа для жесткого диска включает в себя время, необходимое головке чтения/записи для установки на нужный сектор диска (оно называется временем позиционирования). Это - среднее время, так как оно зависит от того, как далеко головка находится от требуемых данных.

Многоканальный режим (англ. Multi-channel architecture) — режим работы оперативной памяти (RAM) и её взаимодействия с материнской платой, процессором и другими

компонентами компьютера, при котором может быть увеличена скорость передачи данных между ними за счёт использования сразу нескольких каналов для доступа к объединённому банку памяти (это можно проиллюстрировать на примере ёмкостей, через горлышко одной из которых жидкость будет выливаться дольше, чем из двух других с такими же общим суммарным объёмом и горлышками, но с большей пропускной способностью — двумя горлышками). Таким образом, система при использовании, например, двух модулей памяти в двухканальном режиме может работать быстрее, чем при использовании одного модуля, равного их суммарному объёму.
Двухканальный режим — режим параллельной работы двух каналов памяти. Наиболее популярный режим для бытовых настольных компьютеров и для ряда ноутбуков. Позволяет увеличить пропускную способность до 2 раз по сравнению с одноканальным режимом.
Не следует путать термин Двухканальный режим с двойной скоростью передачи данных (DDR), в котором обмен данными происходит дважды во время одного тика DRAM. Эти две технологии являются независимыми друг от друга.
Трёхканальный режим — режим работы оперативной памяти компьютера (RAM), при котором осуществляется параллельная работа трёх каналов памяти. То есть параллельно работают 3 модуля или три пары модулей. Теоретически дает прирост пропускной способности в размере около 3 раз по сравнению с одноканальным режимом (1,5 по сравнению с более популярным двухканальным).
Четырёхканальный режим — режим работы оперативной памяти компьютера (RAM), при котором осуществляется параллельная работа четырех каналов памяти. То есть параллельно работают 4 модуля или четыре пары модулей. Теоретически дает прирост пропускной способности в размере около 4 раз по сравнению с одноканальным режимом (двух раз по сравнению с двухканальным). Поддерживается на платформах LGA 2011, LGA 2011v3, TR4, SP3.
https://studfiles.net/preview/2495755/
тут про память с чередованием
https://www.ixbt.com/mainboard/ram-faq-2006.shtml
тут про доступ с чередованием банков. Что подразумевается здесь — хз.
Про проблему когерентности тут есть https://studfiles.net/preview/4339737/

18. Система с общей глобально-адресуемой памятью, однородный и неоднородный доступ к памяти.

Однородный доступ памяти (UMA) - архитектура совместно используемой памяти, используемая в параллельных компьютерах. Все процессоры в модели UMA разделяют физическую память однородно. В архитектуре UMA время доступа к местоположению памяти независимо, к которым процессор обращается с просьбой или какая микросхема памяти содержит переданные данные. Однородные архитектуры ЭВМ доступа памяти часто противопоставляются архитектуре неоднородного доступа памяти (NUMA). В архитектуре UMA каждый процессор может использовать частный тайник. Периферия также разделена некоторым способом. Модель UMA подходит для общей цели и приложений режима разделения времени многочисленных пользователей. Это может использоваться, чтобы ускорить выполнение единственной большой программы в важных приложениях времени.

Неоднородный доступ памяти (NUMA) - дизайн машинной памяти, используемый в мультиобработке, где время доступа памяти зависит от местоположения памяти относительно процессора. Под NUMA процессор может получить доступ к своей собственной местной памяти быстрее, чем нелокальная память (память, местная к другому процессору или памяти, разделенной между процессорами). Выгода NUMA ограничена особой рабочей нагрузкой, особенно на серверах, где данные часто связываются сильно с определенными задачами или пользователями.

http://ru.knowledgr.com/00130407/%D0%9D%D0%B5%D0%BE%D0%B4%D0%BD%D0%BE%D1%80%D0%BE%D0%B4%D0%BD%D1%8B%D0%B9%D0%94%D0%BE%D1%81%D1%82%D1%83%D0%BF%D0%9F%D0%B0%D0%BC%D1%8F%D1%82%D0%B8

тут

Системы с общей (разделяемой) оперативной памятью образуют современный класс ВС — многопроцессорных супер-ЭВМ. Одинаковый доступ всех процессоров к программам и данным представляет широкие возможности организации параллельного вычислительного процесса (параллельных вычислений). Отсутствуют потери реальной производительности на межпроцессорный (между задачами, процессами и т.д.) обмен данными.

19. Распределенная память, проблема отсутствия глобального адресного пространства в такой системе. Способы организации систем с распределенной памятью.

В вычислительных системах с распределенной памятью оперативная память имеется у каждого процессора. Процессор имеет доступ только к своей памяти. В этом случае отпадает необходимость в шине или переключателе. Нет и конфликтов по доступу к памяти, так как каждый процессор работает только со своей собственной памятью. Нет присущих системам с разделяемой памятью ограничений на число процессоров, нет, разумеется, и проблемы с кэшкогерентностью. Но, с другой стороны, возникают проблемы с организацией обмена данными между процессорами. Обычно такой обмен осуществляется при помощи обмена сообщениямипосылками, содержащими данные. Для формирования такой посылки требуется время, для получения и считывания полученных данных тоже требуется определенное время. Эти дополнительные затраты времени - плата за все те преимущества, о которых шла речь.

Программирование для систем с распределенной памятью - более сложная задача. Оно требует разбиения исходной вычислительной задачи на подзадачи, выполнение которых может быть разнесено на разные процессоры.

Одним из наиболее известных компьютеров такого типа является вычислительная система CM-5 фирмы Thinking Machines. Она состоит из процессорных элементов, построенных на основе микропроцессора SPARC и соединенных сетью со специальной топологией типа "дерево". У каждого процессорного элемента имеется локальная память объемом 32 мегабайта. Пропускная способность шины больше у корня дерева и меньше у ее ветвей.

http://lawbooks.news/arhitektura-kompyutera_971/raspredelennaya-obschaya-pamyat-60732.html

Сегодня к распределенным вычислительным системам относят: вычислительные кластеры, SMP – симметричные мультипроцессоры, DSM – системы с распределенной разделяемой памятью, MPP – массово-параллельные системы и мультикомпьютеры. Данная классификация основывается на функциональных возможностях с точки зрения конечного пользователя. Также в литературе [6-11] встречаются подходы к построению структурно-функциональной систематизации распределенных вычислительных систем. Классификация Флинна различает следующие параллельные архитектуры: SIMD, MIMD, MISD и MSIMD. [1,12,13]

http://technology.snauka.ru/2015/04/6452

Там про классификацию