10 вопрос – Характеристики системы памяти. Иерархия ЗУ

В любой ВМ, вне зависимости от ее архитектуры, программы и данные хранятся в памяти. Функции памяти обеспечиваются запоминающими устройствами (ЗУ), предназначенными для фиксации, хранения и выдачи информации в процессе работы ВМ. Процесс фиксации информации в ЗУ называется *записью,* процесс вы-дачи информации — *чтением* или *считыванием,* а совместно их определяют как *процессы обращения* к ЗУ.

Характеристики систем памяти

Перечень основных характеристик, которые необходимо учитывать, рассматривая конкретный вид ЗУ, включает в себя:

- место расположения;
- емкость;
- единицу пересылки;
- метод доступа;
- быстродействие;
- физический тип;
- физические особенности;
- стоимость.

По месту расположения ЗУ разделяют на процессорные, внутренние и внешние. Наиболее скоростные виды памяти (регистры, кэш-память первого уровня) обычно размещают на общем кристалле с центральным процессором, а регистры общего назначения вообще считаются частью ЦП. Вторую группу (внутреннюю память) образуют ЗУ, расположенные на системной плате. К внутренней памяти относят основную память, а также кэш-память второго и последующих уровней (кэш-па-мять второго уровня может также размещаться на кристалле процессора). Медленные ЗУ большой емкости (магнитные и оптические диски, магнитные ленты) называют внешней памятью, поскольку к ядру ВМ они подключаются аналогично устройствам ввода/вывода.

Емкость ЗУ характеризуют числом битов либо байтов, которое может храниться в запоминающем устройстве. На практике применяются более крупные единицы, а для их обозначения к словам «бит» или «байт» добавляют приставки: кило, мега, гига, тера, пета, экза (kilo, mega, giga, tera, peta, exa). Стандартно эти пристав ки означают умножение основной единицы измерений на 10^3, 10^6, 10^9, 10^{12}, 10^{15} и 10^{18} соответственно. В вычислительной технике, ориентированной на двоичную систему счисления, они соответствуют значениям достаточно близким к стандарт тным, но представляющим собой целую степень числа 2, то есть 2^{10}, 2^{20},2^{30},2^{40}, 2^{50}, 2^{60}. Во избежание разночтений, в последнее время ведущие международные организации по стандартизации, например IEEE (Institute of Electrical and Electronics Engineers), предлагают ввести новые обозначения, добавив к основной приставке слово binary (бинарный): kilobinary, megabinary, gigabinary, terabinary, petabinary; exabinary. В результате вместо термина «килобайт» предлагается термин «кибибайт», вместо «мегабайт» — «мебибайт» и т. д. Для обозначения новых единиц предлагаются сокращения: Ki, Mi, Gi, Ti, Pi и Ei.

Важной характеристикой ЗУ является *единица пересылки.* Для основной памяти (ОП) единица пересылки определяется шириной шины данных, то есть количеством битов, передаваемых по линиям шины параллельно. Обычно единица пересылки равна длине слова, но не обязательно. Применительно к внешней памяти данные часто передаются единицами, превышающими размер слова, и такие единицы называются *блоками.*

При оценке быстродействия необходимо учитывать применяемый в данном типе *ЗУ метод доступа* к данным. Различают четыре основных метода доступа:

Последовательный доступ. ЗУ с последовательным доступом ориентировано на хранение информации в виде последовательности блоков данных, называемых записями. Для доступа к нужному элементу (слову или байту) необходимо прочитать все предшествующие ему данные. Время доступа зависит от положения требуемой записи в последовательности записей на носителе информации и позиции элемента внутри данной записи. Примером может служить ЗУ на магнитной ленте.

Прямой доступ. Каждая запись имеет уникальный адрес, отражающий ее физическое размещение на носителе информации. Обращение осуществляется как адресный доступ к началу записи, с последующим последовательным доступом к определенной единице информации внутри записи. В результате время доступа к определенной позиции является величиной переменной. Такой режим характерен для магнитных дисков.

Произвольный доступ. Каждая ячейка памяти имеет уникальный физический адрес. Обращение к любой ячейке занимает одно и то же время и может производиться в произвольной очередности. Примером могут служить запоминающие устройства основной памяти.

Ассоциативный доступ. Этот вид доступа позволяет выполнять поиск ячеек содержащих такую информацию, в которой значение отдельных битов совпадает с состоянием одноименных битов в заданном образце. Сравнение осуществляется параллельно для всех ячеек памяти, независимо от ее емкости. По ассоциативному принципу построены некоторые блоки кэш-памяти.

Быстродействие ЗУ является одним из важнейших его показателей. Для количественной оценки быстродействия обычно используют три параметра:

Время доступа ($Г_д$). Для памяти с произвольным доступом оно соответствует интервалу времени от момента поступления адреса до момента, когда данные заносятся в память или становятся доступными. В ЗУ с подвижным носителем информации — это время, затрачиваемое на установку головки записи/считывания (или носителя) в нужную позицию.

Длительность цикла памяти или *период обращения* ($Г_ц$). Понятие применяется к памяти с произвольным доступом, для которой оно означает минимальное время между двумя последовательными обращениями к памяти. Период обращения включает в себя время доступа плюс некоторое дополнительное время. Дополнительное время может требоваться для затухания сигналов на линиях, а в некоторых типах ЗУ, где считывание информации приводит к ее разрушению, — для восстановления считанной информации.

Скорость передачи. Это скорость, с которой данные могут передаваться в память или из нее. Для памяти с произвольным доступом она равна $\mathbf{1/Т_ц}$. Для других видов памяти скорость передачи определяется соотношением:

$$T_N = T_A + N/R$$

где T_N — среднее время считывания или записи N битов; T_A — среднее время доступа; R — скорость пересылки в битах в секунду.

Говоря о *физическом типе* запоминающего устройства, необходимо упомянуть три наиболее распространенных технологии ЗУ — это полупроводниковая память, память с магнитным носителем информации, используемая в магнитных дисках и лентах, и память с оптическим носителем — оптические диски.

В зависимости от примененной технологии следует учитывать и ряд *физических особенностей* ЗУ, например энергозависимость. В энергозависимой памяти информация может быть искажена или потеряна при отключении источника пищания. В энергонезависимых ЗУ

записанная информация сохраняется и при отключении питающего напряжения. Магнитная и оптическая память — энергонезависимы. Полупроводниковая память может быть как энергозависимой, так и нет, в зависимости от ее типа. Помимо энергозависимости нужно учитывать, приводит ли считывание информации к ее разрушению.

Стоимость ЗУ принято оценивать отношением общей стоимости ЗУ к его емкости в битах, то есть стоимостью хранения одного бита информации.

Иерархия запоминающих устройств

Память часто называют «узким местом» фон-неймановских ВМ из-за ее серьезного отставания по быстродействию от процессоров, причем разрыв этот неуклонно увеличивается. Так, если производительность процессоров ежегодно возрастает вдвое примерно каждые полтора года, то для микросхем памятиприрост быстродействия превышает 9% в год (удвоение за 10 лет), что выражается в увеличении разрыва в быстродействии между *процессором* и *памятью* приблизительно на 50% в год.

Если проанализировать используемые в настоящее время типы ЗУ, выявляется следующая закономерность:

- чем меньше время доступа, тем выше стоимость хранения бита;
- чем больше емкость, тем ниже стоимость хранения бита, но больше время доступа.

При создании системы памяти постоянно приходится решать задачу обеспечения требуемой ёмкости и высокого быстродействия за приемлемую цену. Наиболее распространенным подходом здесь является построение системы памяти ВМ по иерархическому принципу. *Иерархическая память* состоит из ЗУ различный типов (**рис. 5.1**), которые, в зависимости от характеристик, относят к определенному уровню иерархии. Более высокий уровень меньше по емкости, быстрее и имеет большую стоимость в пересчете на бит, чем более низкий уровень. Уровни иерархии взаимосвязаны: все данные на одном уровне могут быть также найдены на более низком уровне, и все данные на этом более низком уровне *могут быть* найдены на следующем нижележащем уровне и т. д.

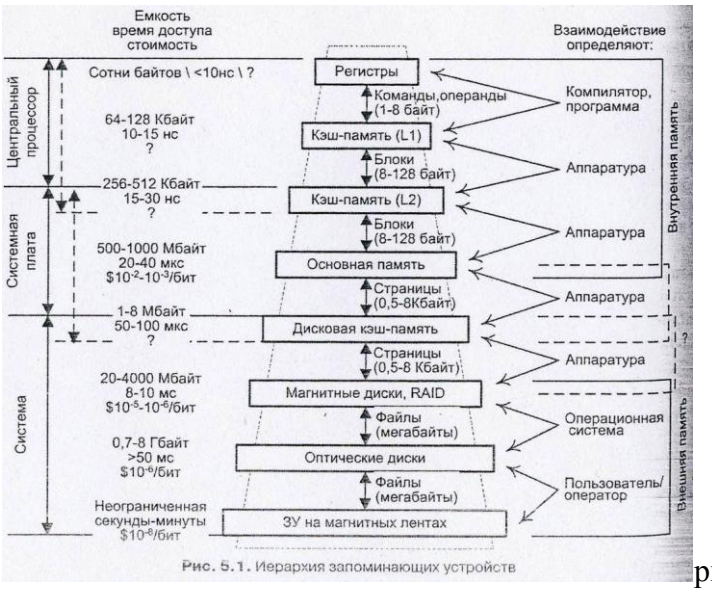

рис.5.1

Четыре верхних уровня иерархии образуют *внутреннюю память* ВМ, а все нижние уровни – это *внешняя* или *вторичная* память. По мере движения вниз по иерархической структуре:

1.Уменьшается соотношение «стоимость/бит».
2.Возрастает емкость.
3.Растет время доступа.
4.Уменьшается частота обращения к памяти со стороны центрального процессора.

Если память организована в соответствии с пунктами 1-3, а характер размещения в ней данных и команд удовлетворяет пункту 4, иерархическая организация Введет к уменьшению общей стоимости при заданном уровне производительности.

Справедливость этого утверждения вытекает из принципа *локальности по обращению*. Если рассмотреть процесс выполнения большинства программ, то можно заметить, что с очень высокой вероятностью адрес очередной команды программы либо следует непосредственно за адресом, по которому была считана текущая команда, либо расположен вблизи него. Такое расположение адресов называется *пространственной локальностью программы*. Обрабатываемые данные, как правило, структурированы, и такие структуры обычно хранятся в последовательных ячейках памяти. Данная особенность программ называется *пространственной локальностью данных*. Кроме того, программы содержат множество небольших циклов и подпрограмм. Это означает, что небольшие наборы команд могут многократно повторяться в течение некоторого интервала времени, то есть имеет 1 место *временная локальность*. Все три вида локальности объединяет понятие *локальностъ по обращению*. Принцип локальности часто облекают в численную форму и представляют в виде так называемого правила «90/10»: 90% времени работы программы связано с доступом к 10% адресного пространства этой программы.

Из свойства локальности вытекает, что программу разумно представить в виде последовательно обрабатываемых фрагментов — компактных групп команд и данных. Помещая такие фрагменты в более быструю память, можно существенно снизить общие задержки на обращение, поскольку команды и данные, будучи один раз переданы из медленного ЗУ в быстрое, затем могут использоваться многократно, и среднее время доступа к ним в этом случае определяется уже более быстрым ЗУ. вЭто позволяет хранить большие программы и массивы данных на медленных, емких, но дешевых ЗУ, а в процессе обработки активно использовать сравнительно небольшую быструю память, увеличение емкости которой сопряжено с высокими затратами.

На каждом уровне иерархии информация разбивается на *блоки,* выступающие в качестве наименьшей информационной единицы, пересылаемой между двумя соседними уровнями иерархии. Размер блоков может быть фиксированным либо переменным. При фиксированном размере блока емкость памяти обычно кратна его размеру. Размер блоков на каждом уровне иерархии чаще всего различен и увеличивается от верхних уровней к нижним.

При доступе к командам и данным, например, для их считывания, сначала производится поиск в памяти верхнего уровня. Факт обнаружения нужной информации называют *попаданием* (hit), в противном случае говорят о *промахе* (miss). При промахе производится поиск в ЗУ следующего более низкого уровня, где также возможны попадание или промах. После обнаружения необходимой информации выполнятся последовательная пересылка блока, содержащего искомую информацию, с нижних уровней на верхние. Следует отметить, что независимо от числа уровней иерархии пересылка информации может осуществляться только междй двумя соседними уровнями.

При оценке эффективности подобной организации памяти обычно используют следующие характеристики:

коэффициент попаданий (hit rate) — отношение числа обращений к памяти, при которых произошло попадание, к общему числу обращений к ЗУ данного уровня иерархии;

коэффициент промахов (miss rate) — отношение числа обращений к памяти, при которых имел место промах, к общему числу обращений к ЗУ данного уровня иерархии;

время обращения при попадании (hit time) — время, необходимое для поисюм нужной информации в памяти верхнего уровня (включая выяснение, является ли обращение попаданием), плюс время на фактическое считывание данных;

потери на промах (miss penalty) — время, требуемое для замены блока в памятиЭ более высокого уровня на блок с нужными данными, расположенный в ЗУ слеш дующего (более низкого) уровня. Потери на промах включают в себя: *время доступа* (access time) — время обращения к первому слову блока при промахе и *время пересылки* (transfer time) — дополнительное время для пересылки оставшихся слов блока. Время доступа обусловлено задержкой памяти более низкого уровня, в то время как время пересылки связано с полосой пропускания канала между ЗУ двух смежных уровней.

Описание некоторого уровня иерархии ЗУ предполагает конкретизацию четырех моментов:

размещения блока — допустимого места расположения блока на примыкающем сверху уровне иерархии;

идентификации блока — способа нахождения блока на примыкающем сверху уровне;

замещения блока — выбора блока, заменяемого при промахе с целью освобождения места для нового блока;

согласования копий (стратегии записи) — обеспечения согласованности копищ. одних и тех же блоков, расположенных на разных уровнях, при записи новой информации в копию, находящуюся на более высоком уровне. Самый быстрый, но и минимальный по емкости тип памяти — это внутренние регистры ЦП, которые иногда объединяют понятием *сверхоперативное запоминающее устройство* — СОЗУ. Как правило, количество регистров невелико, хотя в архитектурах с сокращенным набором команд их число может доходить до нескольких сотен. Основная память (ОП), значительно большей емкости, располагается несколькими уровнями ниже. Между регистрами ЦП и *основной* памяттью часто размещают кэш-память которая по емкости ощутимо проигрывает ОП, но существенно превосходит последнюю по быстродействию, уступая в то же время СОЗУ. В большинстве современных ВМ имеется несколько уровней кэш-памяти, которые обозначаютбуквой L и номером уровня кэш-памяти. На рис. 5.1 показаны два таких уровня. В последних разработках все чаще появляется также третий уровень кэш-памяти (L3), причем разработчики ВМ говорят о целесообразности введения и четвертого уровня — L4. Каждый последующий уровень кэш-памяти имеет большую емкость, но одновременно и меньшее быстродействие по сравнению с предыдущим. Как бы то ни было, по «скорости» любой уровень кэш-памяти превосходит основную память. Все виды внутренней памяти реализуются на основе полупроводниковых технологий и в основном,являются энергозависимыми.

Долговременное хранение больших объемов информации (программ и данных) обеспечивается внешними ЗУ, среди которых наиболее распространены запоминающие устройства на базе магнитных и оптических дисков, а также магнитоленточные ЗУ.

Наконец, еще один уровень иерархии может быть добавлен между основной памятью и дисками. Этот уровень носит название дисковой кэш-памяти и реализуется в виде самостоятельного ЗУ, включаемого в состав магнитного диска. Дисковая кэш-память существенно улучшает производительность при обмене информацией между дисками и основной памятью.

Иерархия может быть дополнена и другими видами памяти. Так, некоторые модели ВМ фирмы IBM включают в себя так называемую расширенную память (expanded storage), выполненную на основе полупроводниковой технологии, но имеющую меньшее быстродействие и стоимость по сравнению с ОП. Строго говоря, этот вид памяти не входит в иерархию, а представляет собой ответвление от нее, поскольку данные могут передаваться только между расширенной и основной памятью, но не допускается обмен между расширенной и внешней памятью.