**Тема 5. Управление памятью**

**Лекция 17 «Функции ОС по управлению памятью. Логическая организация памяти. Физическая организация памяти».**

Главная задача компьютерной системы – выполнять программы. Программы вместе с данными к которым они имеют доступ в процессе выполнения должны находиться в оперативной памяти. Операционной системе приходится решать задачу распределения памяти между пользовательскими процессами и компонентами ОС. Эта деятельность называется управлением памятью. Память является важнейшим ресурсом, требующим тщательного управления. Часть ОС, которая отвечает за управление памятью, называется менеджером памяти.

**Функции ОС по управлению памятью:**

* отслеживание свободной и занятой памяти;
* выделение памяти процессам и освобождение памяти по завершении процессов;
* вытеснение кодов и данных процессов из оперативной памяти на диск (полное или частичное), когда размеры основной памяти недостаточны для размещения в ней всех процессов, и возвращение их в оперативную память, когда в ней освобождается место;
* настройка адресов программы на конкретную область физической памяти;
* динамическое распределение памяти, т.е. выполнение запросов приложений на выделение им дополнительной памяти во время их выполнения. После того как приложение перестает нуждаться в дополнительной памяти, оно может возвратить ее системе;
* дефрагментация памяти;
* выделение памяти для создаваемых ОС служебных информационных структур, таких как описатели процессов и потоков, различные таблицы распределения ресурсов, буферы, используемые процессами для обмена данными, синхронизирующие объекты и т.п.;
* защита памяти, которая состоит в том, чтобы не позволить выполняемому процессу записывать или читать данные из памяти, назначенной другому процессу.

**Физическая организация памяти компьютера.**

Запоминающие устройства компьютера разделяют, как минимум, на два уровня:

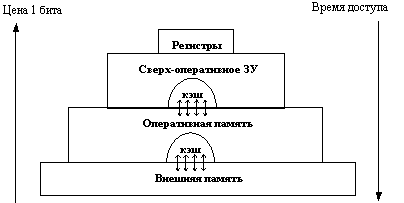
1. **Основная память (внутренняя, оперативная, физическая)** представляет собой упорядоченный массив однобайтовых ячеек, каждая из которых имеет свой уникальный адрес (номер). Процессор извлекает команду из основной памяти, декодирует и выполняет ее. Для выполнения команды могут потребоваться обращения еще к нескольким ячейкам основной памяти. Обычно основная память изготавливается с применением полупроводниковых технологий и теряет свое содержимое при отключении питания.
2. **Вторичную память (внешняя, диски, flash накопители и др.)** можно рассматривать как одномерное линейное адресное пространство, состоящее из последовательности байтов. В отличие от оперативной памяти, она является энергонезависимой, имеет существенно большую емкость и используется в качестве расширения основной памяти.

Рис. 1. Иерархия памяти

Многоуровневую схему используют следующим образом. Информация, которая находится в памяти верхнего уровня, обычно хранится также на уровнях с большими номерами. Если процессор не обнаруживает нужную информацию на i-м уровне, он начинает искать ее на следующих уровнях. Когда нужная информация найдена, она переносится в более быстрые уровни.

Адреса в основной памяти, характеризующие реальное расположение данных в физической памяти называются физическими адресами. Набор физических адресов, с которым работает программа называют физическим адресным пространством.

**Логическая память.**

Большинство программ представляет собой набор модулей, созданных независимо друг от друга. Иногда все модули, входящие в состав процесса, располагаются в памяти один за другим, образуя линейное пространство адресов. Однако чаще модули помещаются в разные области памяти и используются по-разному.

Схема управления памятью, поддерживающая этот взгляд пользователя на то, как хранятся программы и данные, называется сегментацией.

Сегмент – область памяти определенного назначения, внутри которой поддерживается линейная адресация. Сегменты содержат процедуры, массивы, стек или скалярные величины, но обычно не содержат информацию смешанного типа.

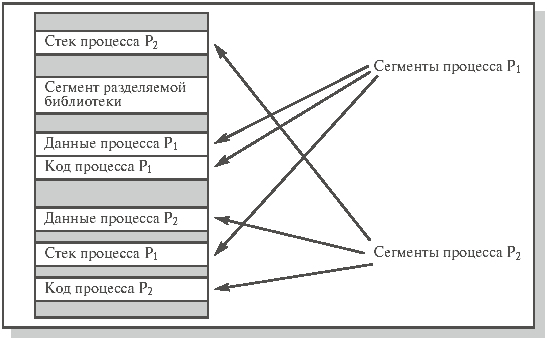


Рис. 2. Расположение сегментов процессов в памяти компьютера

Большинство современных ОС поддерживают сегментную организацию памяти. В некоторых архитектурах (Intel, например) сегментация поддерживается оборудованием.

Адреса, к которым обращается процесс, таким образом, отличаются от адресов, реально существующих в оперативной памяти. В каждом конкретном случае используемые программой адреса могут быть представлены различными способами. Например, адреса в исходных текстах обычно символические. Компилятор связывает эти символические адреса с перемещаемыми адресами (такими, как n байт от начала модуля). Подобный адрес, сгенерированный программой, обычно называют логическим (в системах с виртуальной памятью он часто называется виртуальным) адресом. Совокупность всех логических адресов называется логическим (виртуальным) адресным пространством.

Связывание логического адреса, порожденного оператором программы, с физическим должно быть осуществлено до начала выполнения оператора или в момент его выполнения.

Логические и физические адресные пространства ни по организации, ни по размеру не соответствуют друг другу. Максимальный размер логического адресного пространства обычно определяется разрядностью процессора и в современных системах значительно превышает размер физического адресного пространства. Следовательно, процессор и ОС должны быть способны отобразить ссылки в коде программы в реальные физические адреса, соответствующие текущему расположению программы в основной памяти. Такое отображение адресов называют **трансляцией** (привязкой) адреса или **связыванием адресов**.

Привязка инструкций и данных к памяти может быть сделана на следующих шагах.

**Этап компиляции** (Compile time). Когда на стадии компиляции известно точное место размещения процесса в памяти, тогда непосредственно генерируются физические адреса. При изменении стартового адреса программы необходимо перекомпилировать ее код.

**Этап загрузки** (Load time). Если информация о размещении программы на стадии компиляции отсутствует, компилятор генерирует перемещаемый код. В этом случае окончательное связывание откладывается до момента загрузки. Если стартовый адрес меняется, нужно всего лишь перезагрузить код с учетом измененной величины.

**Этап выполнения** (Execution time). Если процесс может быть перемещен во время выполнения из одной области памяти в другую, связывание откладывается до стадии выполнения. Здесь желательно наличие специализированного оборудования, например, регистров перемещения. Их значение прибавляется к каждому адресу, сгенерированному процессом. Большинство современных ОС осуществляет трансляцию адресов на этапе выполнения.

**Простейшие схемы управления памятью.**

Первые ОС применяли очень простые методы управления памятью. Вначале каждый процесс пользователя должен был полностью поместиться в основной памяти, занимать непрерывную область памяти, а система принимала к обслуживанию дополнительные пользовательские процессы до тех пор, пока все они одновременно помещались в основной памяти.

**1. Схема с фиксированными разделами.** Разбиение памяти на несколько разделов фиксированной величины. Поступающие процессы помещаются в тот или иной раздел. Подсистема управления памятью оценивает размер поступившего процесса, выбирает подходящий для него раздел, осуществляет загрузку процесса в этот раздел и настройку адресов.

**2. Один процесс в памяти.** Частный случай схемы с фиксированными разделами. В памяти размещается один пользовательский процесс. Остается определить, где располагается пользовательская программа по отношению к ОС – в верхней части памяти, в нижней или в средней.

**3. Оверлейная структура.** Так как размер логического адресного пространства процесса может быть больше, чем размер выделенного ему раздела, используется техника, называемая оверлей (overlay) или организация структуры с перекрытием. Основная идея – держать в памяти только те инструкции программы, которые нужны в данный момент.

**4. Динамическое распределение. Свопинг**– перемещению процессов из главной памяти на диск и обратно целиком. Частичная выгрузка процессов на диск осуществляется в системах со страничной организацией (paging).

**5. Схема с переменными разделами.** В этом случае вначале вся память свободна и не разделена заранее на разделы. Вновь поступающей задаче выделяется строго необходимое количество памяти, не более. После выгрузки процесса память временно освобождается. По истечении некоторого времени память представляет собой переменное число разделов разного размера.