

Лекции по Операционным системам

Сверстал: Кузякин Никита Александрович

По лекциям ИТМО

Плейлист с лекциями — [тут](#)

СОДЕРЖАНИЕ

I Основы архитектуры ПК и операционных систем	4
1 Архитектура компьютерных систем	4
2 Обзор элементов компьютерных систем	7
2.1 Процессор	7
3 Общие сведения об операционных системах	9
3.1 Функции OS.....	9
3.2 Оператор ЭВМ	9
3.3 Пакетная обработка	9
3.4 Многозадачность	9
3.5 Разделение времени	10
4 Основные задачи OS	11
4.1 Управление процессами.....	11
4.2 Виртуальная память	12
4.3 Безопасность	12
4.4 Диспетчеризация и планирование ресурсов	13
5 Современные архитектурные концепции OS	14
5.1 Архитектура ядер	14
5.2 Многопоточность	15
5.3 SMP и ASMP	15
5.4 Виртуализация	16
6 Основные понятия надежности операционной системы	17
6.1 Надежность и отказоустойчивость	17
6.2 Сбои	17
7 Общая архитектура UNIX / Linux.....	19
8 Общая архитектура Windows	20
9 Средства для отладки Linux	22
9.1 Стандартные средства.....	22
9.2 /proc	23
9.3 Трассировщики	23
9.4 perf	23
9.5 System tap	25

9.6 Kernel debugger	25
10 Средства для отладки Windows	27
10.1 Встроенные средства	27
10.2 SysInternals.....	27
10.3 Отладчик ядра WinDbg и KD	28
II Процессы	28
11 Основы процессов	28
11.1 Вычисления	28
11.2 Процесс. Характеристики процесса	28
11.3 Состояние процессов и разделение ресурсов	29
12 Пейджинг и свопинг	33
12.1 Paging Swapping.....	33
12.2 Дополнительные состояния процессора	34
13 Управление процессами	35
13.1 Управляющие таблицы	35
13.2 Образ процесса	35
13.3 Функции OS.....	37
13.4 Процессы SVR4	38
14 Потоки	39
14.1 Понятие потока	39
14.2 Связь потоков и процессов	39
14.3 Состояния потока	40
14.4 Варианты реализации	40
14.5 Закон Амдала	41
15 Параллельные вычисления. Блокировки	42
15.1 Параллельность программ	42
15.2 Функции OS поддержки параллельности	42
15.3 Проблемы	42
15.4 Функции OS поддержки параллельности	43
15.5 Взаимодействие процессов / потоков	43
16 Примитивы синхронизации OS	44
16.1 Семафоры, мьютексы	44
17 Процессы и потоки в Linux	47

17.1	Создание процесса со стороны пользователя	48
17.2	Завершение процесса	50
18	Примитивы синхронизации в Linux	50
18.1	Spinlock	50
18.2	Semaphore.....	52
18.3	Mutex	52
19	Процессы и потоки в Windows	54
19.1	Типы процессов	54
19.2	Структура процессов Windows	55
19.3	Состояние процессов Windows	56
19.4	Создание процессов	57
19.5	Завершение процессов	58
20	Примитивы синхронизации в Windows	59
20.1	Объекты диспетчера	59
20.2	События, мьютексы, семафоры	59
20.3	Spinlocks	60
21	Полезные утилиты	61

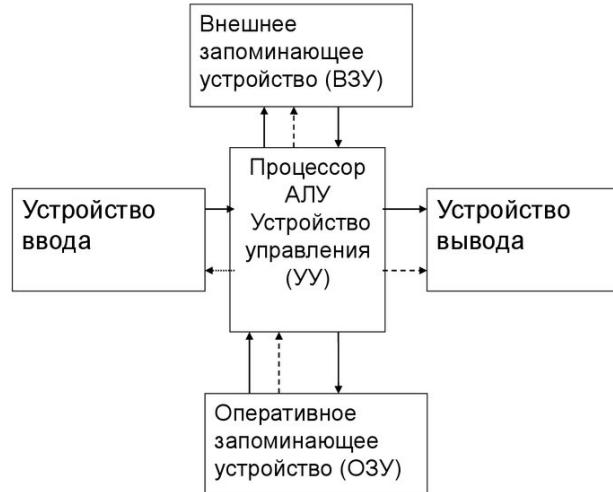
Часть I

Основы архитектуры ПК и операционных систем

1 Архитектура компьютерных систем

Первоначальными двумя архитектурами компьютерных систем являются Гарвардская и Неймановская архитектуры.

Архитектура по Фон Нейману



6

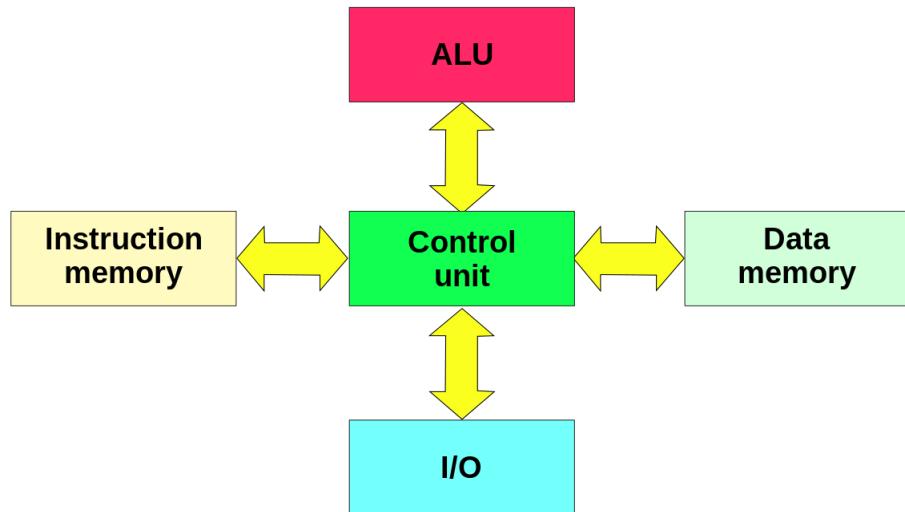


Рисунок 1 – Гарвардская архитектура ЭВМ

Любая вычислительная машины состоит из управляющего устройства (организует вычисления) и арифметико - логического устройства (производит вычисление арифметических операций), а также различных видов памяти.

В архитектуре фон Неймана предполагается, что есть единое управляющее устройство, память при этом общая (и данная, и программа в одно блоке).

Принципы архитектуры фон Неймана:

- Принцип однородности памяти — команды и данные хранятся в одной и той же памяти (внешне неразличимы).
- Принцип адресности — память состоит из пронумерованных ячеек, процессору доступна любая ячейка.
- Принцип программного управления — вычисления представлены в виде программы, состоящей из последовательности команд.
- Принцип двоичного кодирования — вся информация, как данные, так и команды, кодируются двоичными цифрами 0 и 1.

UMA / NUMA

В архитектуре **UMA** подразумевается, что все устройства являются одноранговыми. Те у любого устройства в системе равные права на доступ к памяти и системные характеристики обращения к ней.

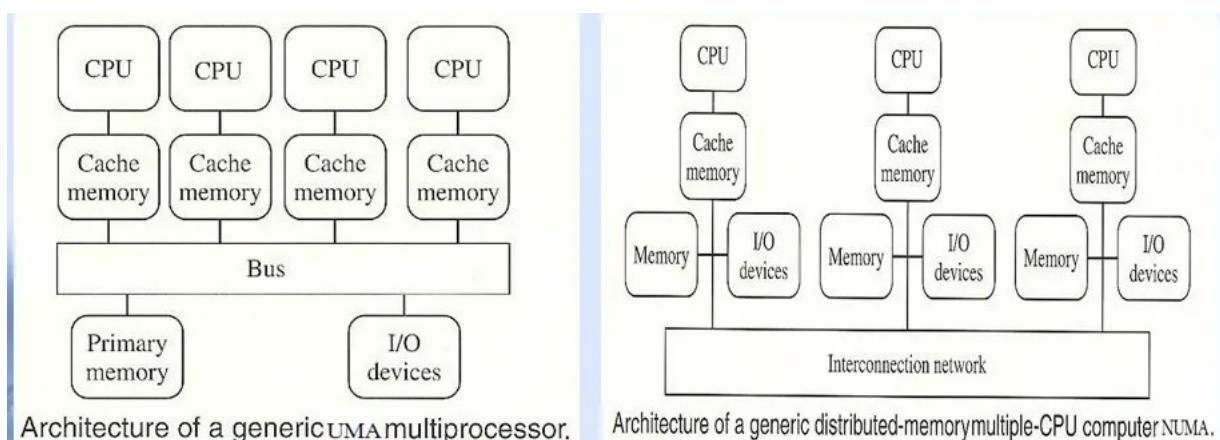


Рисунок 2 – Гарвардская архитектура ЭВМ

Минусом данной архитектуры является, то что тяжело организовать доступ к памяти для большого числа процессоров.

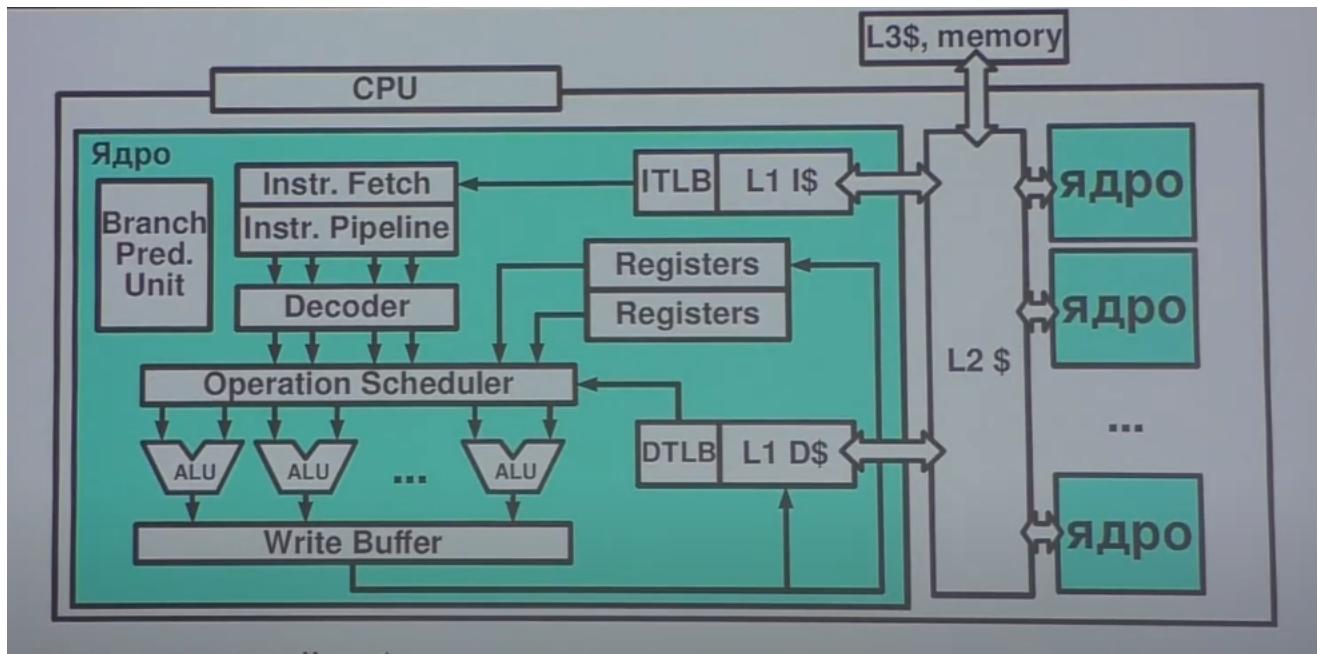
В архитектуре **NUMA** у нас есть память, которая находится ближе к какому-то процессору и память, которая доступна через коммутатор (передает данные через порты).

Адресное пространство для данной архитектуры является общим.

Огромным плюсом является, что можно заменять ее части прямо во время работы, что сильно повышает надежность системы.

2 Обзор элементов компьютерных систем

2.1 Процессор



Составляющие:

1. Арифметико-логическое устройство (АЛУ), выполняющее действия над операндами.
2. Буфер ассоциативной трансляции (TLB) — хранит информацию, есть ли такие-то данные в данном кэше.
3. Кэш процессора, используемый микропроцессором компьютера для уменьшения среднего времени доступа к компьютерной памяти. Делится на L1 i и L1 d. Один из них хранит набор инструкций для работы с кэшем, другой данные.
4. Регистры для хранения данных, адресов и служебной информации.
5. Декодер команд.
6. Буфер для записи — хранит данные, пока буфер не освободится для записи.
7. Branch Pred. Unit — предполагает куда будут записаны данные, по какому адресу (последовательно или с каким-то отступом).
8. Instr. Pipeline — это метод реализации параллелизма на уровне команд в пределах одного процессора.

Важно помнить, что процессор выполняет команды последовательно. Пока один компонент выполняет одно действие, другой выполняет другое (они не останавливаются пока одни данные пройдут от начала до конца).

Определение 1. Виртуальная память — это подход к управлению памятью компьютером, который скрывает физическую память (в различных формах, таких как: оперативная память, ПЗУ или жесткие диски) за единым интерфейсом, позволяя создавать программы, которые работают с ними как с единым непрерывным массивом памяти с произвольным доступом.

	Объем	Тд	*	Тип	Управл.
CPU	100-1000 б.	<1нс	1с	Регистр	компилятор
L1 Cache	32-128Кб	1-4нс	2с	Ассоц.	аппаратура
L2-L3 Cache	0.5-32Мб	8-20нс	19с	Ассоц.	аппаратура
Основная память	0.5Гб-4Тб	60-200нс	50-300с	Адресная	программно
SSD	128Гб-1Тб/drive	25-250мкс	5д	Блочн.	программно
Жесткие диски	0.5Тб-4Тб/drive	5-20мс	4м	Блочн.	программно
Магнитные ленты	1-6Тб/к	1-240с	200л	Последов.	программно

Управляется компилятором — означает, что именно компилятор определяет, как именно ваша программа будет взаимодействовать с данным блоком памяти, те что в какие регистры запишется и тд.

3 Общие сведения об операционных системах

3.1 Функции OS

- Разработка программ.
- Выполнение программ.
- Доступ к устройствам ввода / вывода.
- Контролируемый доступ к файлам.
- Доступ к системе и системным ресурсам.
- Обнаружение и обработка ошибок.
- Учет пользования и диспетчеризация ресурсов.
- Предоставление ключевых интерфейсов (ISA — набор команд, ABI — бинарный интерфейс приложения, API — интерфейс прикладных программ).

3.2 Оператор ЭВМ

Что должен делать оператор?

1. получить программу с данными от программиста.
2. подготовить программу к загрузке.
3. загрузить программу и компилятор.
4. запустить программу на вычисление.
5. распечатку с результатом передать программисту.

Минусами оператора ЭВМ является: наличие расписания машинного времени и долгое время подготовки к работе.

3.3 Пакетная обработка

В следствие минусов оператора ЭВМ появилась пакетная обработка.

Появился первый Системный монитор, который включал в себя: обработчик прерываний, драйверы устройств, планировщик заданий, интерпретатор командного языка и было отведено пространство под пользовательские программы и данные.

3.4 Многозадачность

Одним из главных минусов первых ЭВМ было то, что вовремя вывода, ввода или работы других устройств процессор простоявал. В следствие этого появилась концепция многозадачности.

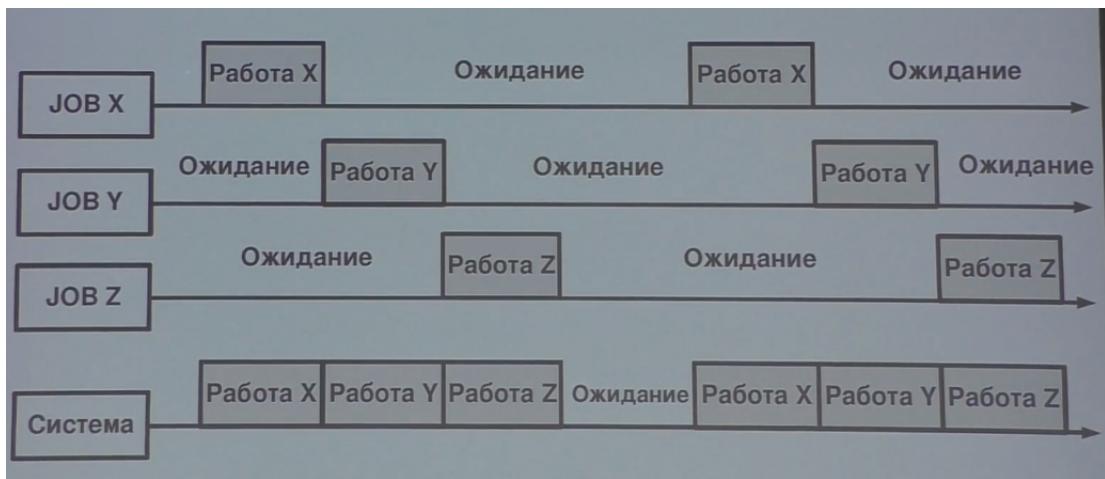


Рисунок 3 – Схема многозадачности первых ЭВМ

3.5 Разделение времени

Следующим нововведением в ЭВМ стало исключение оператора и добавление пользователей. Каждому пользователю выдавалось часть времени процессора с использованием квантового времени. В следствие этого появились проблемы разделения ресурсов и защита одних программ от других.

4 Основные задачи OS

4.1 Управление процессами

Определение 2. Процесс (с точки зрения обывателя) — экземпляр программы во время ее исполнения.

Определение 3. Процесс (с точки зрения OS) — единица потребления ресурсов OS, в которой существует последовательность действий, текущее состояние и набор связанных ресурсов.

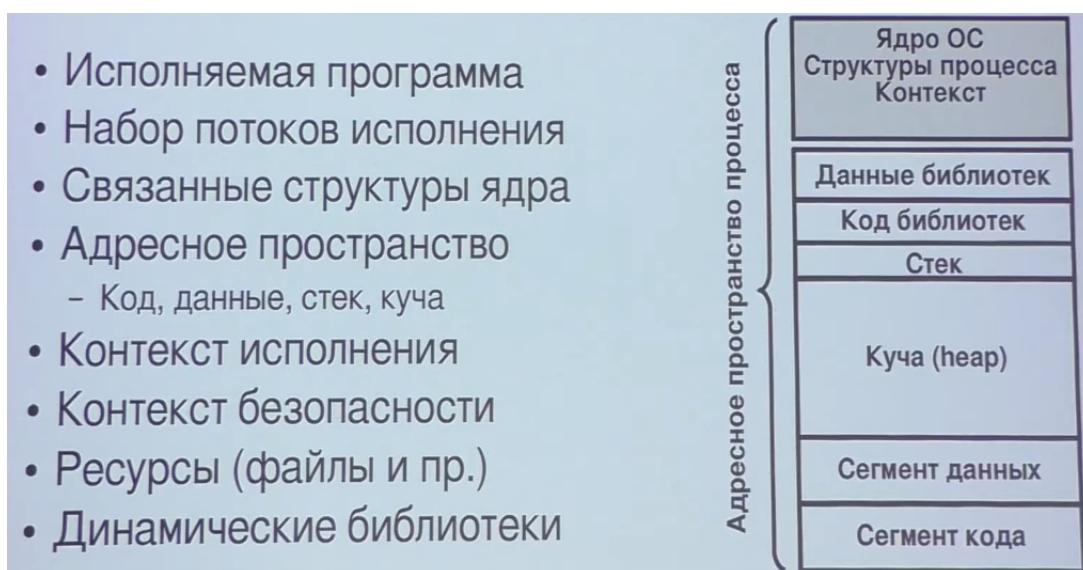


Рисунок 4 – Структура процесса

Для того, чтобы создать процесс, необходимо создать все части адресного пространства представленного на рисунке 4.

Процесс создается не так быстро, поэтому для вычислений на процессоре можно просто создать поток (по сути он будет представлять набор регистров) и с помощью него провести вычисления. Это все и является контекстом.

Когда создается процесс, ядро OS должно построить для ресурсов, которое он будет потреблять систему (описание ресурсов) (в линуксе task structure).

Проблемы современных процессов:

- Защита памяти процессов — недетерминированное поведение процесса, к примеру обращение не к своей памяти, может нарушить другие процессы.
- Взаимные блокировки — есть два процесса, один из них захватил один ресурс, другой другой, и они пытаются также добавить к себе захваченный другим процессом ресурс. (deadlock, livelock, starvation)

- Проблема синхронизации — тк у нас может быть несколько процессов, а адресное пространство для них одно.
- Взаимное исключение доступа ресурсов.

4.2 Виртуальная память

Для решения проблемы с единым адресным пространством была придумана виртуальная память.

Управление памятью:

- Изоляция процессов.
- Управление выделением и освобождением памяти (аллокаторы и менинг памяти).
- Поддержка модулей (модульности) — динамическая загрузка и выгрузка модулей.
- Защита и контроль доступа — права на сегменты памяти.
- Долговременное хранение — запись информации на диск.
- Страницочный обмен.

Определение 4. Виртуальная память — отдельное виртуальное адресное пространство для каждого процесса и ядра.

Также виртуальная память подразумевает, что некоторые страницы нельзя выгружать из памяти, к примеру если они используются в большом количестве процессов.

4.3 Безопасность

Также важный аспект OS это то, на сколько она безопасна, на сколько она обеспечивает безопасность данных.

Самым важным аспектом безопасности является протокол работы с информацией.

Что должна обеспечивать OS:

1. Безопасность доступа к системе — защита от несанкционированного доступа.
2. Конфиденциальность — невозможность неавторизованного доступа к данным.
3. Целостность данных — защита данных от неавторизованного и нецелостного изменения.
4. Аутентификация и авторизация.

4.4 Диспетчеризация и планирование ресурсов

Что важно учесть при планирование ресурсов (с точки зрения OS):

- Равноправие — пользователи, процессы и тд должны получать ресурсы равноправно. (Интересный факт: в UNIX приоритет процесса развернутого окна на 15 пунктов выше других).
- Дифференциация отклика — в некоторых задачах нужно понизить время отклика, к примеру в задачах выполняющихся в реальном времени.
- Общесистемная эффективность.
- Планировщик процессов, дисков и тд.

5 Современные архитектурные концепции OS

5.1 Архитектура ядер

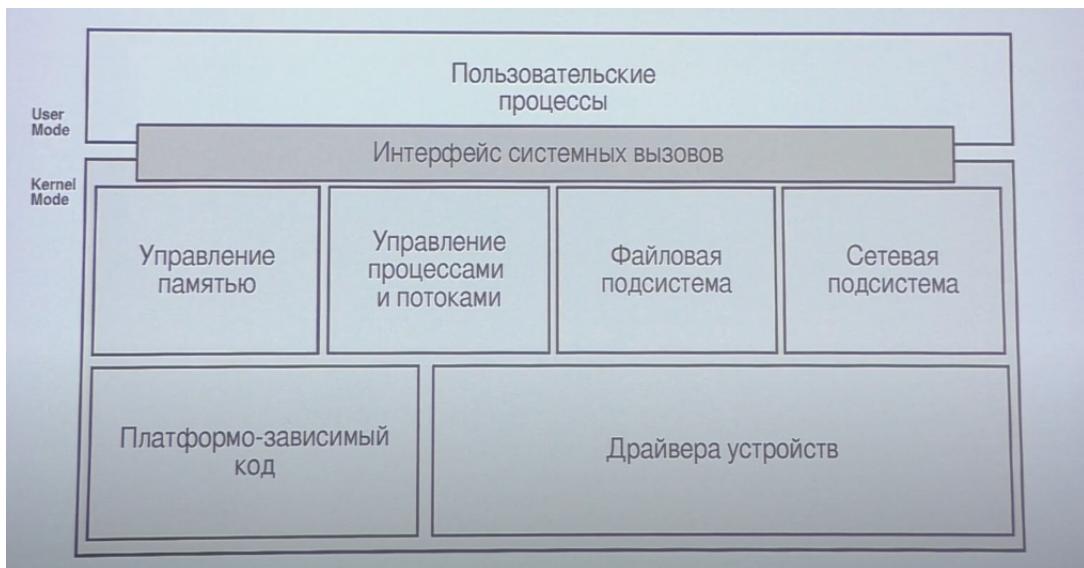


Рисунок 5 – Схема ядра ОС

Управление памятью, процессами и потоками, файловая подсистема и сетевая подсистема работают на основе драйверов и платформо - зависимого кода.

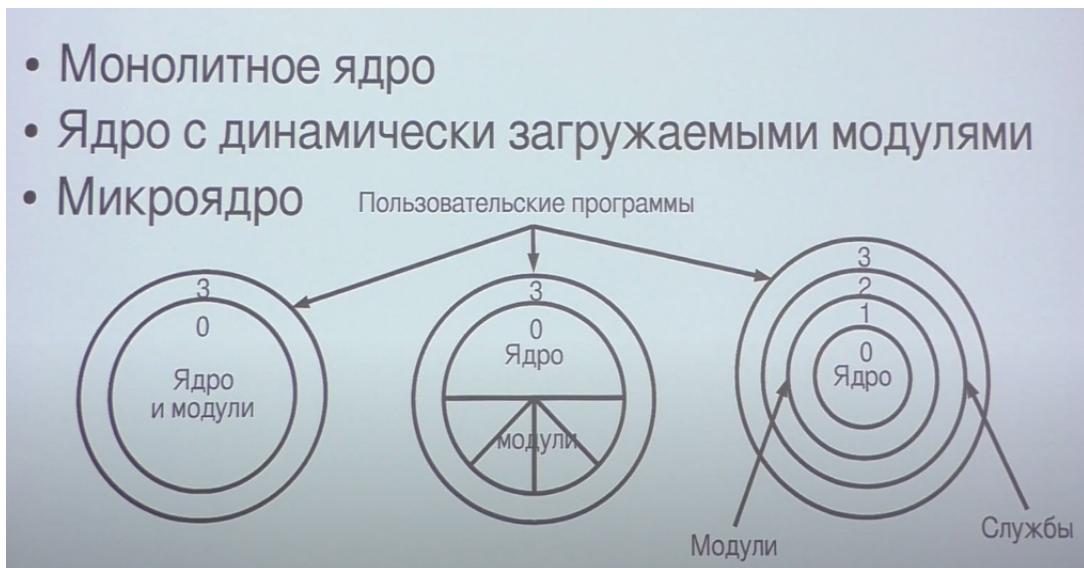


Рисунок 6 – Виды архитектур ядер ОС

Монолитное ядро — подразумевает, что для изменения чего-то в ядре придется перекомпилировать OS. Подходит для систем где набор устройств

определен и не будет изменяться. 0 уровень — ядро и встроенные в него модули, 3 уровень — пользовательские программы. 1 и 2 не используются.

Ядро с динамически загружаемыми модулями имеет возможность загрузить модули во время выполнения операционной системы.

Микроядро — концепция, в которой само ядро занимается базовыми задачами: диспетчеризация процессов и выделение памяти. 1 и 2 уровни занимают остальные задачи, реализованные в виде сервисов. Пользовательские приложения работают на 3 уровне. Из-за частого переключения контекстов это работает очень медленно.

5.2 Многопоточность

Из-за сложности создания процесса была придумана концепция реализации внутри процесса потоков. Thread — нить / поток.

Библиотека порождающая потоки на UNIX системах Posix Threads.

Существует множество концепций реализации потоков, они будут рассмотрены в следующих параграфах.

5.3 SMP и ASMP

Symmetric multiprocessing — процессы равны, процесс выполняется на нескольких процессорах одновременно. Это дает следующие плюсы: простота разработки и производительность, более высокая надежность (при отказе выполнить процесс, его могут выполнить другие), масштабируемость приложений, динамическое добавление ресурсов процессора.

Asymmetric multiprocessing — в системе с асимметричной многопроцессорностью не все процессоры играют одинаковую роль. Например, система может использовать (либо на аппаратном, либо на уровне операционной системы) только один процессор для выполнения кода операционной системы, или поручать только одному процессору выполнение операций ввода-вывода. В других AMP-системах все процессоры могут выполнять код операционной системы и операции ввода-вывода, так что с этой стороны они ведут себя как симметричная многопроцессорная система, но определенная периферийная аппаратура может быть подсоединенна только к одному процессору, так что со стороны работы с этой аппаратурой система предстаёт асимметричной. Более дешевая альтернатива в системах, которые поддерживали SMP.

Многопоточность ! = Многопроцессорность

5.4 Виртуализация

Виртуальные машины (интерпретаторы) — по сути программы, которые работают под выполнением другой программы. Как примеры: JS в браузере, python, JAVA VM. Это позволяет поднять уровень абстракции.

Определение 5. Интерпретация — построчный анализ, обработка и выполнение исходного кода программы или запроса, в отличие от компиляции, где весь текст программы, перед запуском анализируется и транслируется в машинный или байт-код без её выполнения.

Контейнеры приложений — позволяет писать приложения один раз и запускать их где угодно. Разработчики могут создавать и развертывать приложения быстрее и безопаснее, чем при традиционном подходе к написанию кода — когда он разрабатывается в определенной вычислительной среде, а его перенос в новое место, например из тестовой среды в продуктивную, часто приводит к ошибкам выполнения кода.

Определение 6. Контейнер приложения — экземпляр исполняемого программного обеспечения (ПО), который объединяет двоичный код приложения вместе со всеми связанными файлами конфигурации, библиотеками, зависимостями и средой выполнения.

Смысл и главное преимущество технологии в том, что контейнер абстрагирует приложение от операционной системы хоста, то есть остается автономным, благодаря чему становится легко переносимым — способным работать на любой платформе.

Примеры: Docker, Solaris containers, Linux containers.

Аппаратурная виртуализация — виртуализация с поддержкой специальной процессорной архитектуры. В отличие от программной виртуализации с помощью данной техники возможно использование изолированных "гостевых" операционных систем.

Примеры: Virtual BOX, KVM.

Облачные технологии — по сути облачная виртуализация, главным плюсом является, что в случае сбоя одной физической системы, данные иммигрируют на другую систему и продолжат выполняться. Данные технологии построены на базе аппаратурной виртуализации.

6 Основные понятия надежности операционной системы

6.1 Надежность и отказоустойчивость

Отказоустойчивость — способность системы продолжать работу при аппаратных или программных ошибках.

Для обеспечения отказоустойчивости нужно:

- Избыточность аппаратуры(двойное, тройное резервирование).
- Аппаратная "горячая"замена компонентов.
- Программная поддержка OS выведения компонентов из системы и их подключения.
- Организация уровней хранения RAID в дисковой подсистеме.

Надежность — вероятность бесперебойной работы системы до времени t , при условии ее корректной работы в $t = 0$.

Среднее время наработка на отказ MTTF = $\int_0^x R(t) dt$, включает в себя время на перезагрузку, ремонта или замены неисправного компонента, установки (переустановки) OS или ПО.

Коэффициент доступности — процент времени, когда система или служба доступна для запросов пользователей.

Простой (downtime) — время, в течение которого система недоступна

Безотказная работа — когда система находится в продуктивной работе.

6.2 Сбои

Какие бывают отказы:

- Ошибочное состояние аппаратуры или ПО в результате сбоя компонентов.
- Ошибки оператора.
- Физические помехи окружающей среды.
- Ошибки проектирования, программирования, структуры данных и тд.
- Могут быть: постоянные, временные (однократные или периодические).

Методы резервирования:

- Физическая избыточность (компонентов, серверов).
- Временная избыточность (повтор вычислений).
- Информационная избыточность (ECC, RAID).

Методы повышения отказоустойчивости:

- Изоляция процессов
- Разрешение блокировок при параллелизме
- Виртуализация
- Точки восстановления и откаты

7 Общая архитектура UNIX / Linux

В UNIX появилась ключевая концепция: все есть файл или процесс. Также появился принцип: одна программа — одна функция. Также использовалась концепция минимизации ядра, реализация на С и унификация файлов.

SINGL UNIX Specification — общее название для семейства стандартов, которым должна удовлетворять операционная система, чтобы называться "UNIX".

UNIX системы — AIX, MAC OS X, Solaris.

UNIX like системы — Linux, Free BSD, Open Solaris.

Подробно архитектуру ядра Linux можно посмотреть по ссылке:

<https://makelinux.github.io/kernel/map/>

Основные подсистемы Linux:

- Процессы и планировщик задач — создает, управляет и планирует процессы.
- Виртуальная память — выделяет виртуальную память для процессов и управляет ею.
- Физическая память — управляет пулом кадров страниц и выделяет страницы для виртуальной памяти.
- Файловая система — представляет глобальное иерархическое пространство имен для файлов и функции для работы с файлами.
- Драйверы символьных устройств — управление устройствами, которые требуют от ядра отправки и получения данных по одному байту.
- Драйверы блочных устройств — управление устройствами, которые читают и записывают данные блоками.
- Сетевые протоколы (TCP/IP) — поддержка пользовательского интерфейса сокетов для набора протоколов.
- Драйверы сетевых устройств.
- Ловушки и отказы — обработка генерируемых прерываний.
- Прерывания — обработка прерываний от периферийных устройств.
- Сигналы и IPC — управление межпроцессорным взаимодействием.

8 Общая архитектура Windows

В первых версиях Windows была надстройкой над операционной системой DOS. В данный момент у Windows есть несколько линеек: для мобильных устройств, для ПК и серверная.

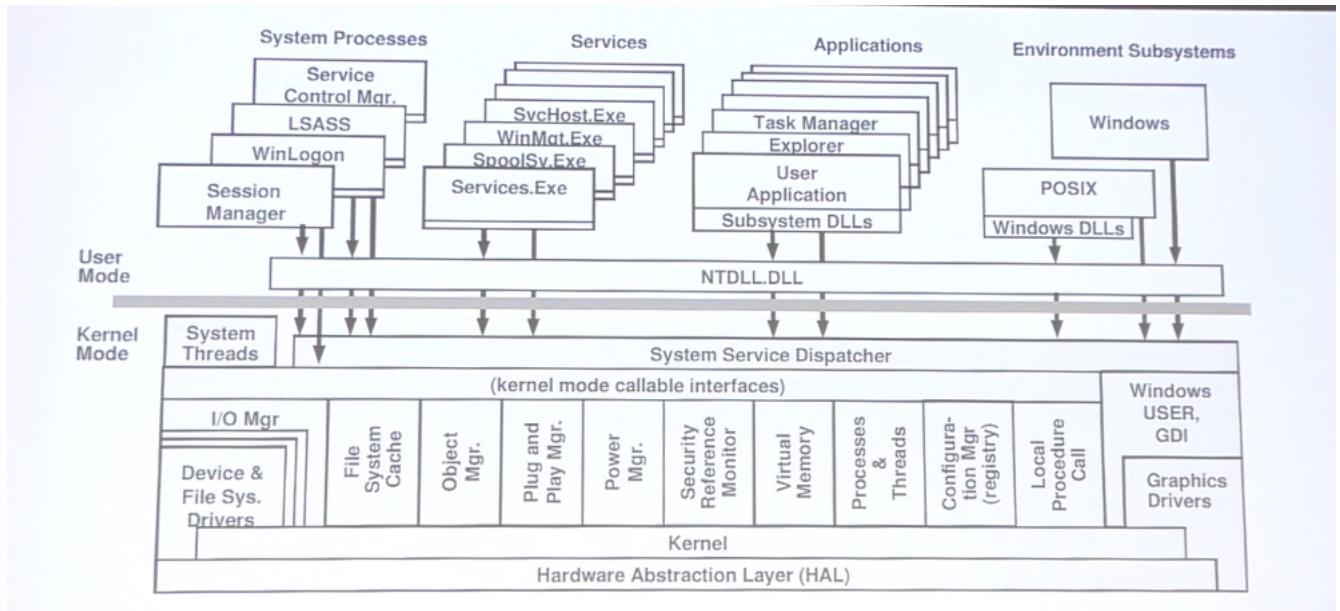


Рисунок 7 – Архитектура Windows

Как не странно архитектура Windows очень похожа на архитектуру UNIX.

Интерфейс системных вызовов — NTDLL.dll

Также подсистемы Windows похожи на подсистемы в Linux.

Plug and play Mgr. — система позволяющая легко ставить драйверы, без указания портов и тд.

Одним из отличий является графический интерфейс интегрированный в ядро.

WinAPI — это библиотеки динамической компоновки (DLL), которые являются частью Windows операционной системы. Они используются для выполнения задач, когда сложно написать эквивалентные процедуры. Например, Windows предоставляет функцию с именем FlashWindowEx, которая позволяет сделать заголовок строки приложения чередующимся между светлыми и темными оттенками. Позволяет легко написать приложение, которое будет совместимо с любой версией Windows.

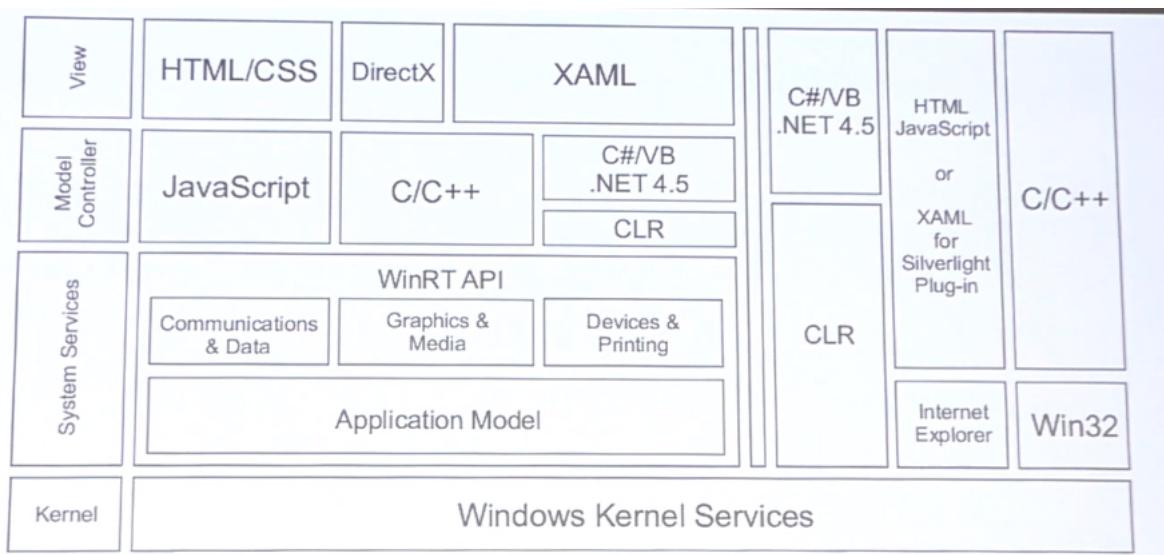


Рисунок 8 – Схема WinAPI

Другие важные компоненты Windows:

- Гипервизор Hyper-V — запуск гостевых операционных систем.
- Firmware — содержимое энергонезависимой памяти любого цифрового вычислительного устройства — микрокалькулятора, сотового телефона, GPS-навигатора и т. д., в которой содержится его программа.
- Terminal Servers.
- Объекты — все вещи в системе сделаны в виде объектов.
- Реестр.
- Оснастки — специальная вспомогательная программа для администрирования выделенного пула задач.

9 Средства для отладки Linux

9.1 Стандартные средства

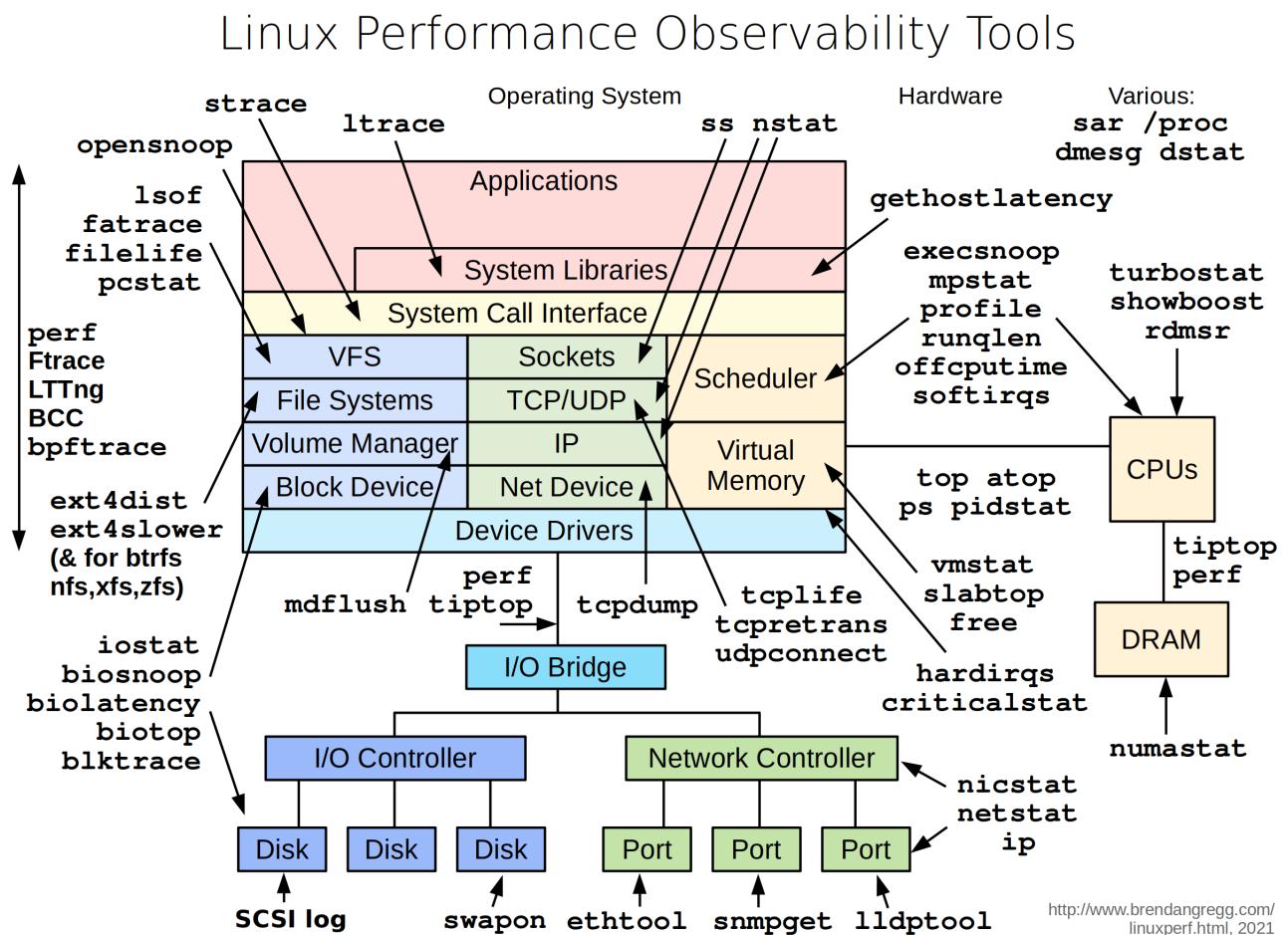


Рисунок 9 – Утилиты для отладки линукс

Ядро операционной системы накапливает большое количество различных счетчиков. И все представленные на рисунке 9 утилиты, по сути просто дают доступ к этим счетчикам, те никаких дополнительных вычислений не производится.

Стандартные средства наблюдения за счетчиками

sar — утилита, которая позволяет посмотреть информацию о счетчиках любой подсистемы Linux.

Подробно ознакомиться можно здесь: <https://greendail.ru/node/monitoring-proizvoditelnosti-linux-na-primere-sar>

- Процессор: ps, top, tiptop, turbostat, rdmsr, numastat, uptime
- Виртуальная память: vmstat, slabtop, pidstat, free
- Дисковая подсистема: iostat, iotop, blktrace
- Сеть: netstat, tcpdump, iptraf, ethtool, nicstat, ip
- Интерактивные (типа top) или с указанием количества запуска и интервала (типа sar)
- Некоторые работают только с правами root!

Рисунок 10 – Другие встроенные утилиты

Утилиты обычно двух типов: интерактивные (можно изменять параметры системы) и статичные (просто предоставляют информацию).

9.2 /proc

/proc — виртуальная файловая система, которая содержащая файлы статистики и управляющая модулями ядра. По сути вся информация представлена в виде файловой системы. К примеру, информация о процессоре будет лежать в каталоге `/proc/cpuinfo`.

9.3 Трассировщики

- Трассировка системных вызовов: strace
- Трассировка вызовов библиотек: ltrace
- Трассировка lock -ф: bpftrace

Также одно из средств отладки, с помощью которого легко увидеть логи системных вызовов.

9.4 perf

Профилировщики — собирает системную информацию, которую вы указали.

Основное предназначение профилировщиков — это взять ваше готовое приложение и посмотреть, что находится в ядре во время его запуска.

Суть в том, что perf может собрать весь стэк трейс запущенной программы.

Естественно, запущенный perf будет вносить задержку в работу всей системы.

Но у нас есть флаг -F #, где # — частота сэмплирования, измеряемая в Гц.

К примеру perf record df -h запишет данные любой команды Perf, которую вы хотите сохранить для использования в будущем.

```
• usage: perf [-version] [-help] [OPTIONS] COMMAND [ARGS]
• The most commonly used perf commands are:
  • bench      General framework for benchmark suites
  • c2c        Shared Data C2C/HITM Analyzer.
  • config     Get and set variables in a configuration file.
  • data       Data file related processing
  • diff       Read perf.data files and display the differential profile
  • evlist    List the event names in a perf.data file
  • ftrace     simple wrapper for kernel's ftrace functionality
  • kallsyms  Searches running kernel for symbols
  • kmem      Tool to trace/measure kernel memory properties
  • list      List all symbolic event types
  • lock      Analyze lock events
  • mem       Profile memory accesses
  • record    Run a command and record its profile into perf.data
  • report    Read perf.data and display the profile
  • sched     Tool to trace/measure scheduler properties (latencies)
  • script    Read perf.data and display trace output
  • stat      Gather performance counter statistics on command
  • timechart Tool to visualize total system behavior
  • top       System profiling tool.
  • probe     Define new dynamic tracepoints
  • trace     strace inspired tool
```

Рисунок 11 – Базовые команды в perf

Одно из удобных визуальных представлений, что сохраняет профилировщик — FlameGraph.

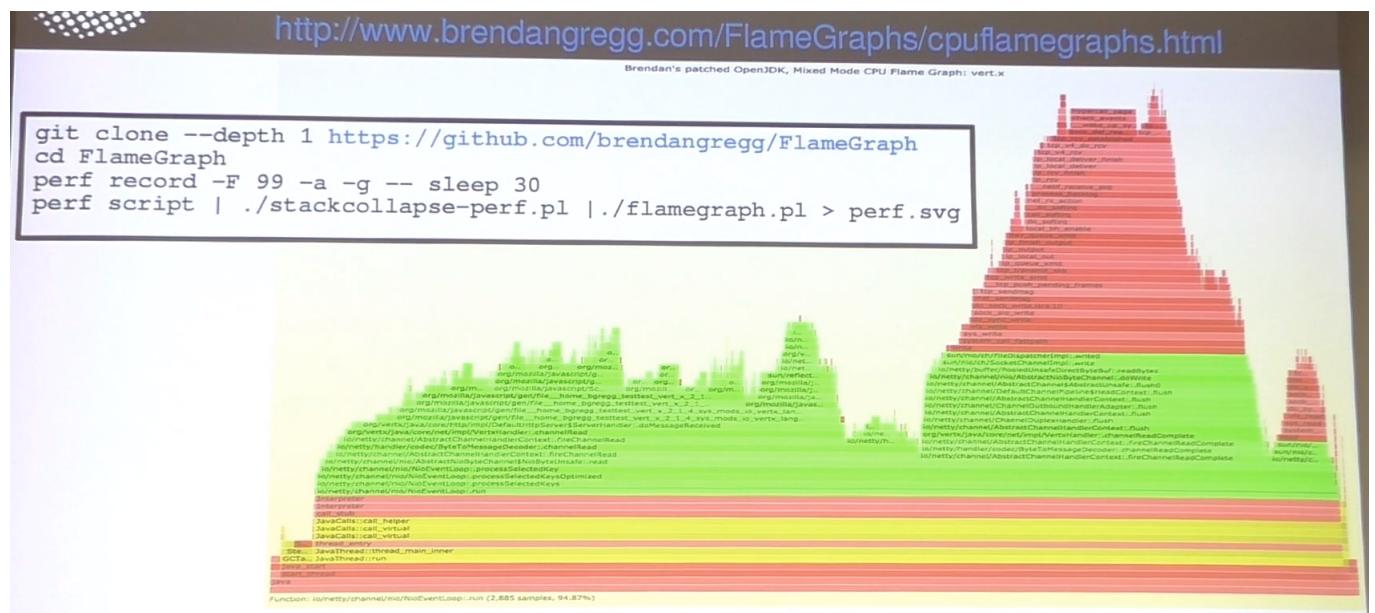


Рисунок 12 – FlameGraph

9.5 System tap

Еще одно средство сбора информации о подсистеме ядра или пользователя, при этом имеет минимальное воздействие на систему. SystemTap по сути имеет скриптовый синтаксис.

Основная идея SystemTap состоит в том, чтобы обозначить события и назначить для них обработчики.

Во время выполнения скрипта, SystemTap занимается мониторингом событий и, как только произойдет событие, ядро системы выполнит обработчик. Событиями могут быть начало или конец сессии SystemTap, срабатывание таймера и другие.

Обработчиком является последовательность скриптовых операторов, которые будут выполнены после срабатывания события. Обычно обработчики извлекают информацию из контекста события или выводят информацию на экран.

Сессия SystemTap начинается тогда, когда мы выполняем скрипт. В это время происходит следующая последовательность действий:

1. Сначала SystemTap проверяет библиотеку «тапсетов» на наличие использованных в скрипте;
2. Потом SystemTap транслирует скрипт в Си (язык программирования) и запускает системный компилятор, чтобы создать модуль ядра из скрипта;
3. SystemTap загружает модуль и активирует все события в скрипте;
4. Как только происходит событие выполняется обработчик данного события;
5. Когда все события выполнены, модуль выгружается и сессия завершается;

9.6 Kernel debugger

Существует два режима у отладчика ядра: локальный отладчик (предустановлен в системе) и удаленный (предоставляет информацию об системе, находящейся на другом компьютере).

Чтобы включить компиляцию kdb, вы должны сначала включить kgdb. Параметры компиляции тестов kgdb описаны в главе kgdb test suite

<https://docs.kernel.org/dev-tools/kgdb.html>.

Kdb - это упрощенный интерфейс в стиле оболочки, который можно использовать на системной консоли с клавиатурой или последовательной консолью. Вы можете использовать его для проверки памяти, регистров, списков

процессов, dmesg и даже установки точек останова для остановки в определенном месте. Kdb не является отладчиком исходного кода, хотя вы можете устанавливать точки останова и выполнять некоторые базовые элементы управления запуском ядра. Kdb в основном предназначена для проведения некоторого анализа, чтобы помочь в разработке или диагностике проблем ядра.

10 Средства для отладки Windows

10.1 Встроенные средства

Windows SDK — включает в себя: отладчик, множество утилит, поддерживающих сборку приложений.

DTrace on Windows — по сути System tap был скопирован с Dtrace для Windows, тк что по сути у них похожий функционал.

Администрирование — Disk Cleanup, Performance Monitor (очень удобное средство, в котором удобно можно задать параметры), Resource Monitor, Registry Editor Services, System Configuration и тд. (чтобы попасть туда нужно перейти по такому пути: Control Panel -> System and Security -> Administrative Tools).

Task manager — ну, тут не нужно лишних слов.

Также ссылка на сторонние программы: <https://habr.com/ru/company/ua-hosting/blog/280578/>

10.2 SysInternals

Множество скриптов и программ для управления, диагностики, устранения неполадок и мониторинга всей среды Microsoft Windows. Автор Марк Руссинович, в настоящее время сотрудник Microsoft (Соавтор книги Windows Internals).

Top SysInternals utils:

- PsList and PsKill – просмотр и остановка процессов (в том числе и удаленно)
- Process Explorer - просмотр ресурсов процесса, замена Task Manager
- Process Monitor – просмотр связанных с процессом ресурсов реестра
- Autoruns - поиск автозапускаемых программ
- Contig - дефрагментирует конкретный файл
- PSFile – позволяет показать открытые файлы, в том числе и удаленно
- MoveFile - перемещает заблокированные файлы во время перезагрузки. • Sync - синхронизация файловой системы
- TCPview - информация о открытых сетевых соединениях
- SDelete - удалить файлы и папки без возможности восстановления

Тут очень много средств для отлавливания вирусов.

10.3 Отладчик ядра WinDbg и KD

По умолчанию Windows не загружается в режиме отладчика, для этого есть специальные команды.

livekd (SysInternals) — позволяет перейти в режим отладки без перезагрузки системы.

Для использования WinDbg необходимо получить символьную информацию (тк в ядре информация хранится в виде 1 и 0, то нужно создать папку для хранения символов переменных, **ДЛЯ КАЖДОЙ СБОРКИ ВИНДЫ НУЖНА СВОЯ СИМВОЛИЧЕСКАЯ ИНФОРМАЦИЯ**).

Пример вызова отладчика:

```
set _NT_SYMBOL_PATH=srv*c:\symbols*http://msdl.microsoft.com/download/symbols
C:\Program Files (x86)\SysinternalsSuite>livekd64.exe -w -k "C:\Program Files (x86)\Windows Kits\10\Debuggers\x64\windbg.exe"
```

Рисунок 13 – FlameGraph

Часть II

Процессы

11 Основы процессов

11.1 Вычисления

По сути наша система компьютерная состоит из ресурсов: процессор, память устройства ввода-вывода. Приложения, которые мы пишем, решают практическую задачу: Входные данные → Обработка → Выходные данные. OS по сути находится между оборудованием и приложением. По сути OS представляет абстракции ресурсов и предоставляет их пользователям.

11.2 Процесс. Характеристики процесса

Каждая из подсистем рассматривает процесс в своем ключе.

И так, повторим, процесс это:

- Выполняемая программа;

- Экземпляр программы, выполняющийся на компьютере;
- Сущность, которая может быть назначена процессору и выполнена на нем;
- Единица активности, характеризуемая выполнением последовательных команд, текущим состоянием и связанным с ней множеством системных ресурсов;

Характеристики процесса в момент времени:

- Уникальный идентификатор;
- Состояние (выполнение, очередь, ожидание);
- Приоритет по отношению к другим процессам;
- Счетчик команд;
- Указатели на область памяти процесса;
- Контекст процесса (регистры, user / kernel);
- Статус ввода-вывода;
- Счетчики системных ресурсов;
- Права доступа процесса;
- и тд;

11.3 Состояние процессов и разделение ресурсов

Рассмотрим ситуацию, где есть два процесса с одинаковым приоритетом и числом доступных ресурсов. Вспомним, что в таком случае у нас для этих процессов будет выделено одинаковое количество процессорного времени.

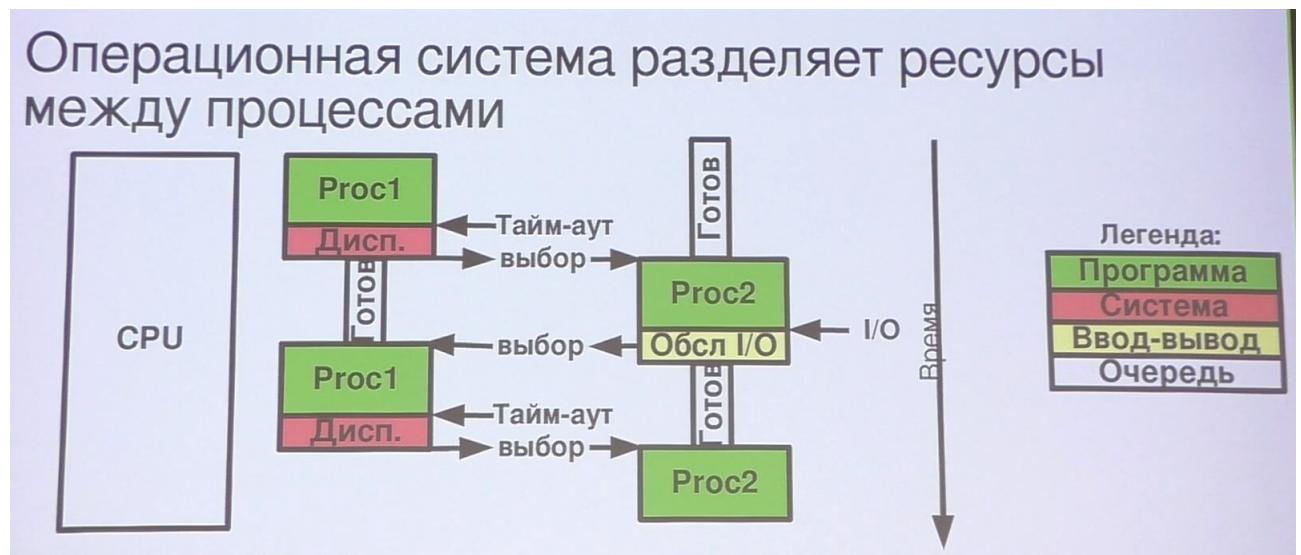


Рисунок 14

Как происходит распределение ресурсов процессора между процессами?

Определение 7. Квант — единица времени, периодичность, с которой система проверяет, не выполняется ли процесс слишком долго.

Раз в квант система проверяет, закончилось ли процессорное время у процесса, если да, то диспетчер берет другой, а этот снова кладет в очередь.

Это все делается для того, чтобы примерно одинаковые процессы выполнялись одинаковое количество времени.

Рассмотрим состояния процесса

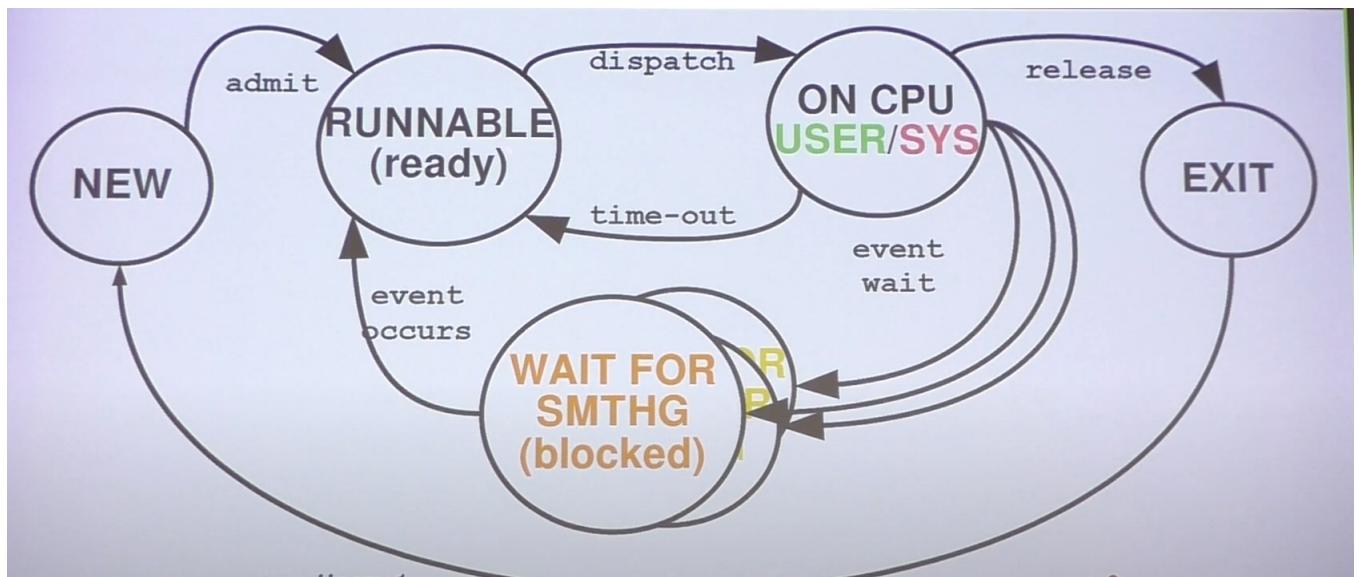


Рисунок 15 – Цикл работы процесса

- Ready — готов к выполнению, это значит процессор может взять этот процессор на исполнение. Есть все необходимые ресурсы, которые позволяют ему выполнится. К примеру, страницы памяти, сигнал завершения ввода вывода;
- On CPU — если процессор свободен, то мы можем назначить на него процесс. Процесс может работать в двух режимах user и sys. Во время выполнения процесса эти режимы могут переключаться;
- Выход из состояния On CPU в двух случаях — процесс выполнен или у него закончилось системное время;

New state — процесс создан, но еще не размещен в очереди процессов, готов к исполнению.

Exit state — процесс не может продолжить выполнение. (Структуры процесса еще существуют).

Причины попадания в состояние:

- Нормальное завершение (вызов exit);
- превышение лимитов на время выполнения;
- Недостаток памяти;
- Ошибки границ и защиты памяти;
- Арифметическая ошибка;
- Ошибка ввода-вывода;
- Неправильная или привилегированная инструкция;
- Команда оператора или OS;
- Завершение или запрос родительского процесса;

Runnable state — процесс обладает всеми ресурсами для выполнения, но нет возможности исполниться.

Причины попадания в состояние:

- Низкий приоритет по сравнению с другими процессами;
- Ожидания освобождения CPU;
- Закончился квант времени;

Runnable state — процесс выполняется на процессоре.

Процесс остается в этом состояние если:

- не истек квант времени;
- Ожидание на спин-блокировке;
- В Runnable состоянии нет процессов с более высоким приоритетом;
- Обслуживание высокоприоритетных прерываний;
- Нет блокирующих вызовов (ввод вывод, ожидание блокировки);

Wait (blocked) state — ожидания событий OS, освобождения блокировки. Процесс не расходует ресурсов CPU, процесс может находиться в ожидании неопределенно долго. Бывает процесс не убивается, пока он не будет разблокирован (самое простое, перезагрузить OS).

Состояние продолжается пока:

- освободится блокировка;
- Придет сообщение OS о наступление ожидаемого события.

12 Пейджинг и свопинг

12.1 Paging Swapping

Как правило, основной памяти всегда мало, если появляется свободная память программисты сразу пытаются ее заполнить. А также в системе обычно большое количество запущенных процессов.

Давайте поместим блокируемый процесс на диск и освободим основную память для других процессов.

Определение 8. Paging — выгрузка или загрузка неиспользуемых страниц процесса на диск.

Определение 9. Swapping — выгрузка всего процесса, кроме критически важных для ядра структур.

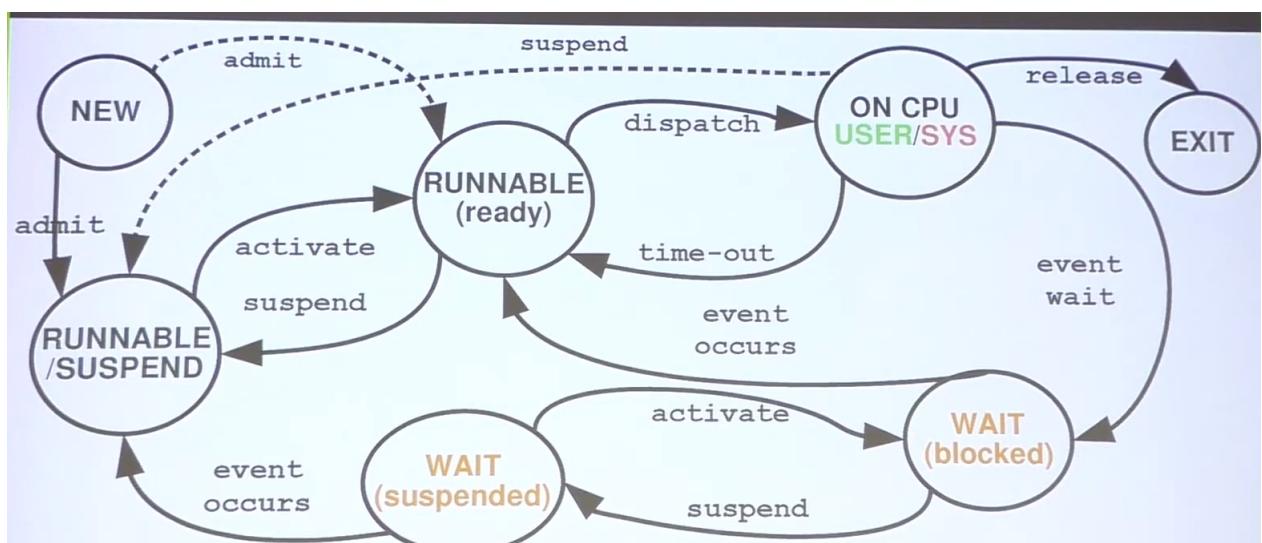


Рисунок 16 – Модель цикла процесса с 7 состояниями

На картинке можно наблюдать цикл состояния процесса с Paging / Swapping.

Изменения в том, что мы разделили состояние Wait и Runnable на две части.

Почему стрелочка из New идет в Runnable Suspend? Тк процесс создается долго, в современных системах, создается минимальный костяк процесса, без мепинга, и отправляется в приостановленное состояние, чтобы можно было быстро отчитаться о создание процесса.

Также интересное замечание, что может создаться такая ситуация, что при большом количестве процессов, может выстроиться очередь на блокировку процессов, тогда вся система начнет глобально тормозить.

12.2 Дополнительные состояния процессора

Wain / Suspend state — процесс приостановлен и выгружен в область подкачки. По событию Suspend процесс выгружается на диск. По событию Active загружается в основную память. (Данные действия повышают нагрузку на дисковую подсистему).

Причины попадания в состояние:

- Длительное ожидание события операционной системы;
- Недостаток памяти (зачем держать в памяти процесс, который не имеет возможности исполниться).

Runnable / Suspend state — процесс готов к выполнению, но выгружен из памяти.

Причины:

- Был неготов, выгружен, но произошло событие, которое позволяет выполниться;
- Desperate memory conditions;
- Команда пользователя;
- Создание процесса в минимальном варианте, без, к примеру создания сегментов памяти.

13 Управление процессами

13.1 Управляющие таблицы

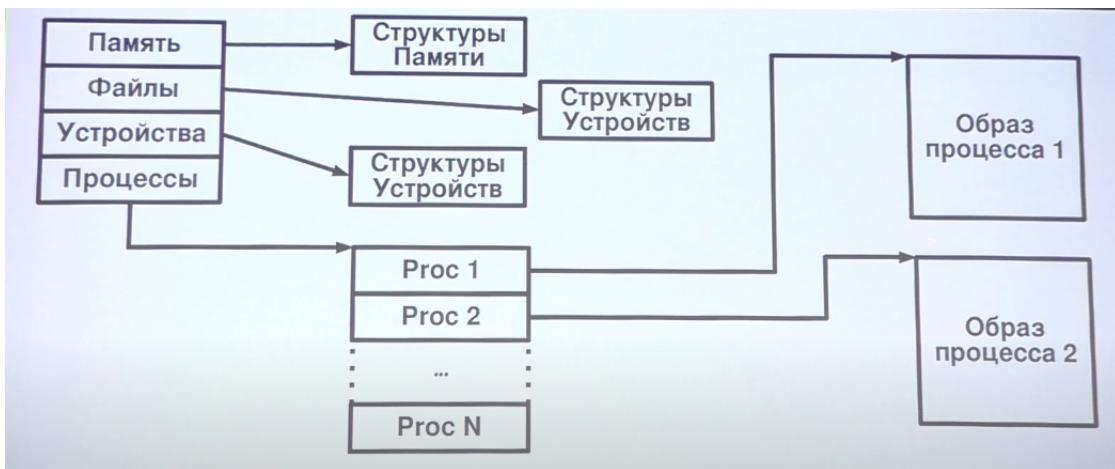


Рисунок 17 – Общая структура управляющей таблицы

13.2 Образ процесса

Допустим у нас есть код написанный на языке С.

```
#include <stdio.h>

int var_int_data[1024]; static char var_char_data[4096];

void foo(int var_inc) {
    int var_a = 10;
    static int var_sa = 10;

    var_a += var_inc;
    var_sa += var_inc; printf("a = %d, sa = %d\n", var_a, var_sa);
}

int main(int argc, char**argv) {
    int var_i;

    for (var_i = 0; var_i < 10; ++var_i)
        foo(var_i);
}
```

Давайте попробуем понять, где что находится?

```

serge@ra:/tmp$ gcc -o a_prog -g a.c
serge@ra:/tmp$ gdb a_prog
GNU gdb (Ubuntu 9.1-0ubuntu1) 9.1
[...]
(gdb) break foo
Breakpoint 1 at 0x1149: file a.c, line 7
(gdb) run
Starting program: /tmp/a_prog
Breakpoint 1, foo (var_inc=32767) at a.c:7
7 {
(gdb) step 2
10     var_a += var_inc;
(gdb) print &var_char_data
[...] &var_int_data, &var_a, &var_sa, &var_inc
$4 = (char (*)[4096]) 0x555555558040 <var_char_data>
$3 = (int (*)[1024]) 0x555555559040 <var_int_data>
$5 = (int *) 0x7fffffffde5c #var_a
$6 = (int *) 0x555555558010 <var_sa>
$7 = (int *) 0x7fffffffde4c #var_inc

```

Address	Kbytes	RSS	Dirty	Mode	Mapping
...					
0000555555555000	4	4	4	r-x--	a_prog
...					
0000555555559000	8	0	0	rw---	[anon]
00007ffff7dc0000	148	148	0	r----	libc-2.31.so
00007ffff7de5000	1504	484	8	r-x--	libc-2.31.so
...					
00007ffff7fab000	12	12	12	rw---	libc-2.31.so
00007ffff7fae000	24	20	20	rw---	[anon]
00007ffff7fc0000	12	0	0	r----	[anon]
00007ffff7fce000	4	4	4	r-x--	[anon]
00007ffff7fcf000	4	4	0	r----	ld-2.31.so
00007ffff7fd0000	140	140	24	r-x--	ld-2.31.so
...					
00007ffff7ffd000	4	4	4	rw---	ld-2.31.so
00007ffff7ffe000	4	4	4	rw---	[anon]
00007ffff7fffe000	132	12	12	rw---	[stack]
ffffffffffff600000	4	0	0	--x--	[anon]
total kB	2368	960	120		

Рисунок 18

Запомним: rw — сегмент данных или куча (anon), а также у нас снизу есть еще и стек. Строки где одни ——, значит, что к этим страницам обратиться нельзя.

Пройдемся дебагером, и увидим что массив чаров у нас ушел в сегмент данных, по адресу8040. А массив интов, ушел в кучу по адресу9040. Важно помнить, что разные компиляторы ведут себя по разному. Также мы видим, статические переменные попали в стек.



Рисунок 19 – Управляющий блок процесса

В управляемом блоке процесса содержится информация о: инициализации процесса, какие регистры ему принадлежат, а также многое другое).

Если смотреть на наш процесс внутри операционной системы, то есть несколько очередей, в которых содержатся ссылки на управляющие блоки процессов. Все эти структуры тесно связаны между собой, иногда переходят из одной очереди в другую.

13.3 Функции OS

Функции OS связанные с процессами:

Управление процессами

- Создание и завершение процессов;
- Планирование и диспетчеризация процессов;
- Переключение процессов;
- Синхронизация и поддержка обмена информацией между процессами;
- Организация управляющих блоков процессов

Управление вводом-выводом

- Управление буферами;
- Выделение процессам каналов и устройств ввода - вывода

Управление памятью

- Выделение адресного пространства процессам;
- Управление страницами и сегментами;
- Пейджинг и Свопинг;

Функции поддержки

- Обработка прерываний;
- Учет пользования ресурсами;
- Текущий контроль системы;

Вспомним, что создание процесса включает в себя:

- Присвоение уникального идентификатора;
- Выделение памяти;
- Инициализация PCB;
- Постановка процесса в очередь ядра;
- Создание потоков ввода-вывода;

- Создание других управляющих структур данных.

13.4 Процессы SVR4

На картинке представлен цикл создания процесса в UNIX.

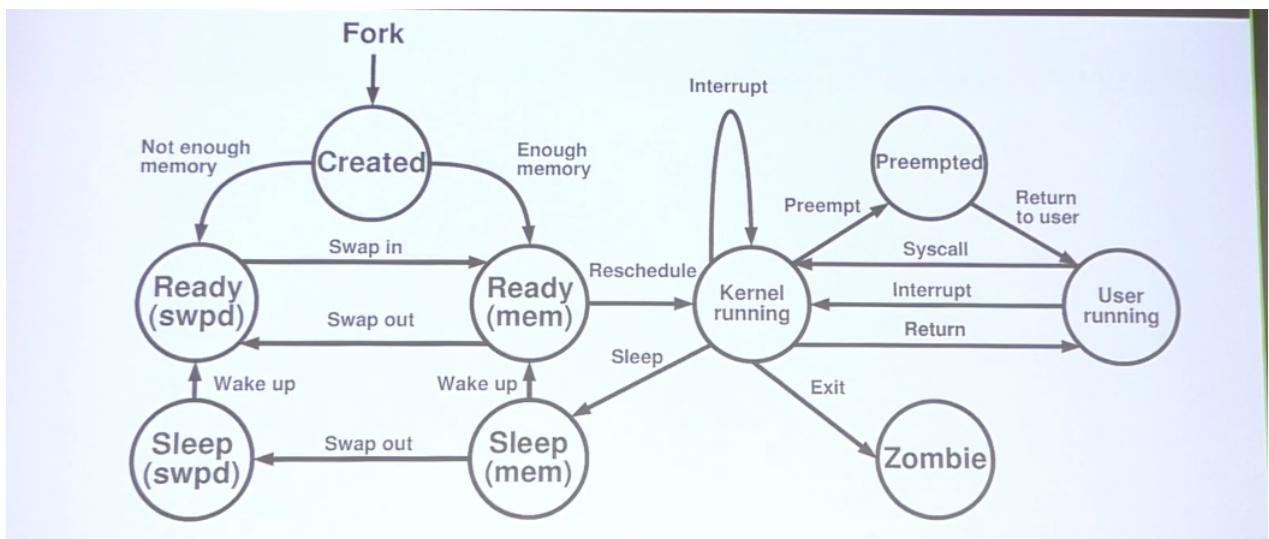


Рисунок 20 – Создание процессов в UNIX

Можем наблюдать явно использование процесса в user и kernel режимах.

14 Потоки

14.1 Понятие потока

Абстрактная модель процесса подразумевает владение ресурсами, а также планирование и диспетчеризацию.

По сути, процесс является единицей группировки ресурсов, а Thread — единица выполнения программного кода.

Thread содержит: Состояние выполнения, Сохранение контекста потока, Стек, Локальные переменные, Доступ к памяти и ресурсам процесса владельца.

14.2 Связь потоков и процессов

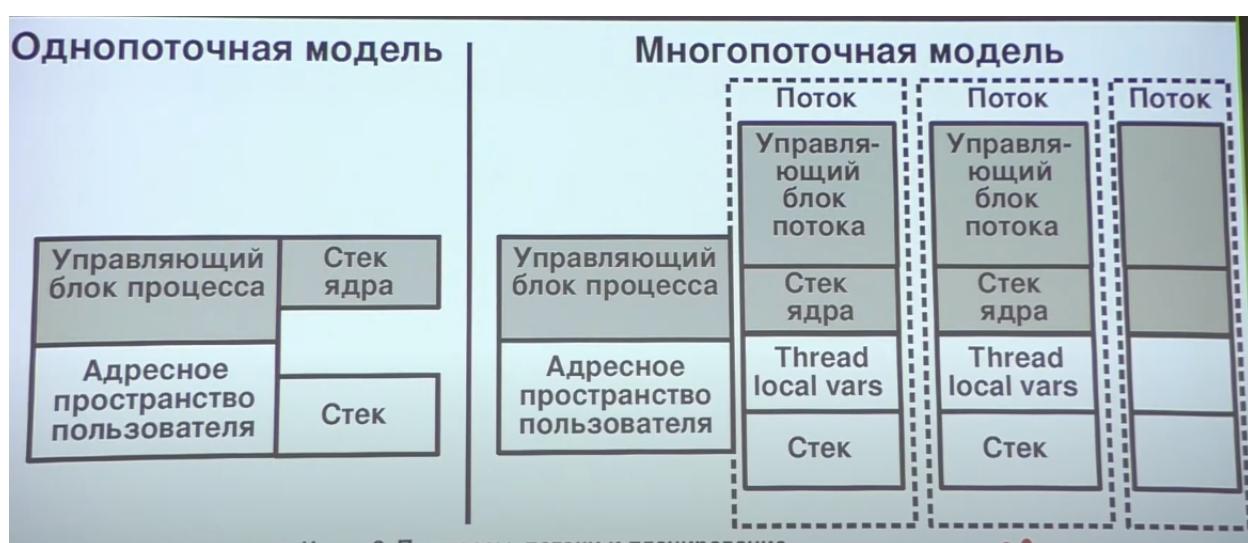


Рисунок 21 – Связь структур ядра и потока

Как мы видим у нас также есть управляемый блок процесса. Важно заметить, что структуры отвечающие за управления у потоков разные, а ресурсы одни.

Главное преимущество потоков в быстродействие, все действия для них выполняются быстрее (Переключение между потоками быстрее, тк переключается только контекст, убиваются они тоже быстрее, но лучше лишний раз не убивать, тк это все равно энергозатратно). Даже в однопоточных моделях есть преимущества (работа в приоритетном и фоновом режиме, асинхронная работа частей программы, модульная структура программы).

14.3 Состояния потока

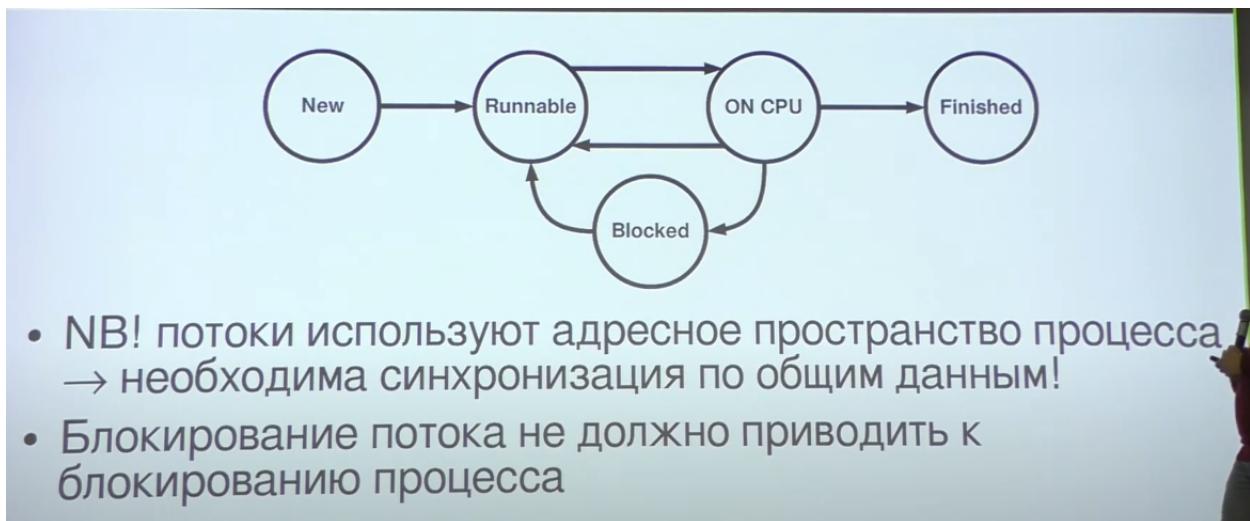


Рисунок 22 – Состояние потоков

Как мы видим, состояния потоков не сильно отличаются от состояний процессов, за исключением того, что нет структур отвечающих за диспетчери-зацию.

14.4 Варианты реализации

User level Thread — реализуется библиотеками или приложениями на сто-роне пользователя.

Kernel level Thread — реализуется ядром.

Основная разница заключается в том, что для User lvl для переключения между потоками не нужно уходить в ядро.

В данный момент чаще всего используется модель KLT, сколько потоков мы породили на уровне пользователя, столько получим и на уровне ядра.

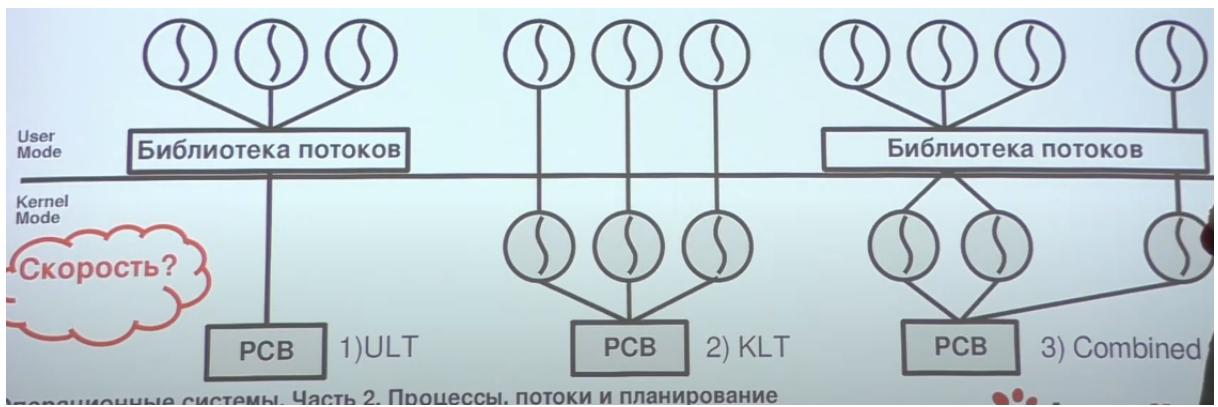


Рисунок 23 – Модели потоков

14.5 Закон Амдала

На сколько можно ускорить программу на N процессах с использованием потоков?

Теорема 1. Закон Амдала — ускорение = время работы на одном процессоре / время работы на N процессах.

$$\text{ускорение} = \frac{T*(1-f)+T*f}{T*(1-f)+\frac{T*f}{N}} = \frac{1}{(1-f)+\frac{f}{N}}$$

Где Т — время работы, f — доля распараллеливания [0,1). Когда f мало, использование параллельного выполнения неэффективно. Когда N стремится к бесконечности, то ускорение ограничено $\frac{1}{1-f}$.

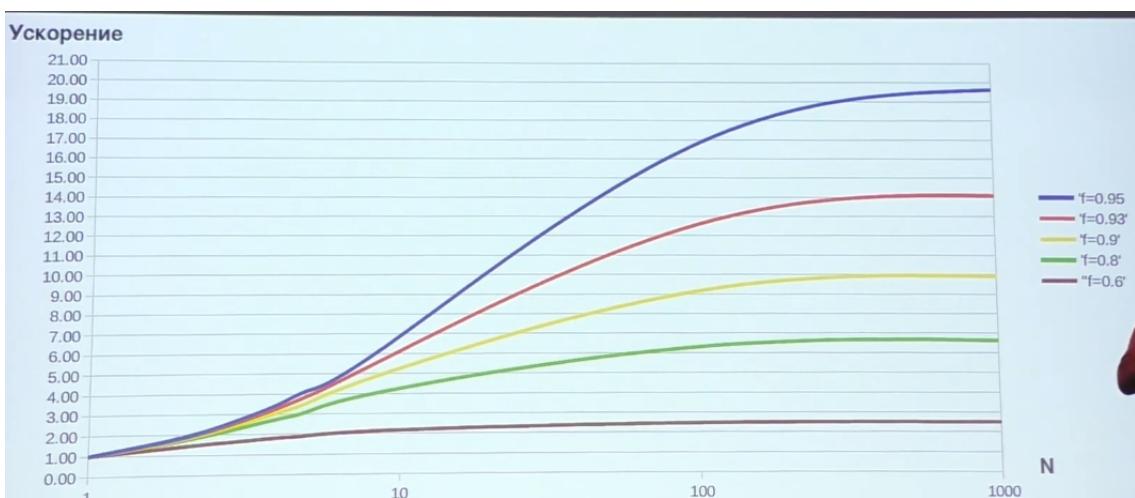
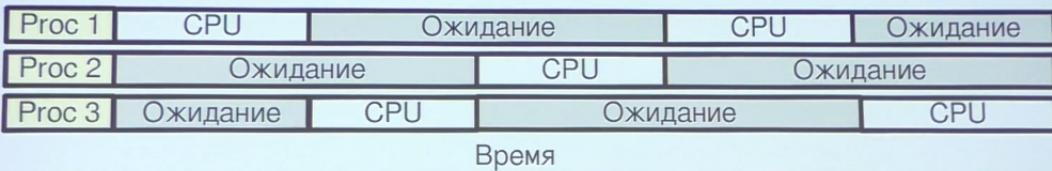


Рисунок 24 – Визуализация закона Амдала

15 Параллельные вычисления. Блокировки

15.1 Параллельность программ

- Однопроцессорные системы — процессы чередуются



- Многопроцессорные — чередуются и перекрываются.

- Конкуренция за общие ресурсы. Голодание.

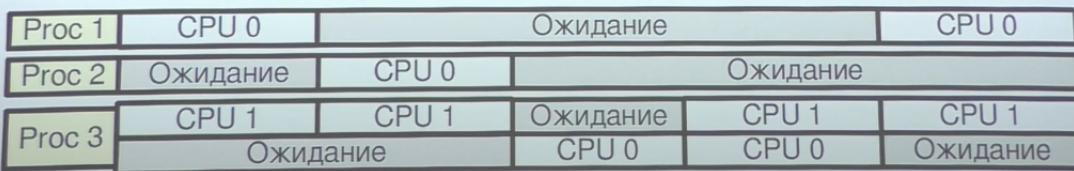


Рисунок 25 – Механизм параллельных вычислений

Голодание часто встречается в ситуации, где блокировкой процесса А владеет низкоприоритетный процесс В. В следствие этого, процесс В отодвигают процессы с более высоким приоритетом, из-за этого процесс А не может выполниться.

15.2 Функции OS поддержки параллельности

Требуемые функции OS:

- Отслеживание ресурсов процесса или потока;
- Распределение и освобождение ресурсов для процесса или потока;
- Защита ресурсов процесса или потока от непреднамеренного воздействия на них других процессов или потоков;
- Независимость результата процесса или потока, от скорости его выполнения;

15.3 Проблемы

- Взаимоисключения (Mutual Exclusion) — процессы или потоки не должны одновременно использовать критический ресурс.
- Взаимоблокировки (Dead Lock) — процессы или потоки не должны взаимозахватывать требуемые ресурсы;

- Голодание (Starvation) — конкуренция за ресурсы не должна порождать невозможность доступа к ресурсу;

15.4 Функции OS поддержки параллельности

Требования OS к взаимным исключением

- Взаимные исключения осуществляются в принудительном порядке (в критическом участке только один процесс / поток);
- Процесс / поток не должен влиять на другие процессы / потоки в некритическом участке;
- Противодействие бесконечному ожиданию доступа к критическому участку;
- Вход в свободный критический участок должен незамедлительно предоставляться;
- Отсутствие предположений о количестве процессов и их скорости;
- Ограничение времени нахождения в критическом участке;

Аппаратной поддержкой взаимных исключений являются атомарные инструкции: TAS, CAS, CMPXCHG и тд.

15.5 Взаимодействие процессов / потоков

Процессы / потоки могут не знать ничего друг о друге, тогда происходит конкуренция за ресурсы, взаимоисключения, взаимоблокировки, голодание.

Процессы / потоки сотрудничают используя общий ресурс, это может вызвать: взаимоисключения, взаимоблокировки, голодание, связь данных.

Процессы / потоки совместимо выполняются, это может вызвать: взаимоблокировки, голодание.

16 Примитивы синхронизации OS

16.1 Семафоры, мьютексы

Определение 10. Семафоры — захват и освобождение множественного ресурса.

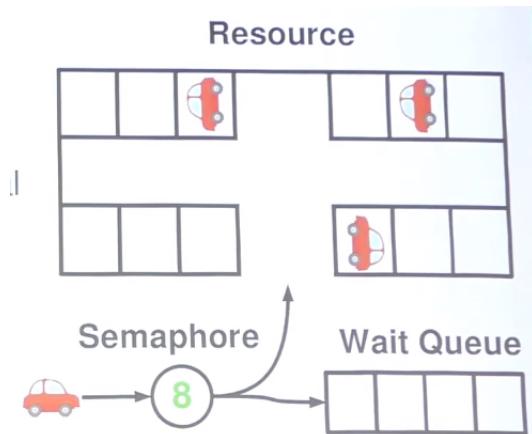


Рисунок 26 – Пример работы семафора

Определение 11. Мьютексы — блокировка или освобождение ресурса единственным процессом или потоком (создание критических секций). Существует большое количество мьютексов: блокирующие, спин, адаптивные и тд.

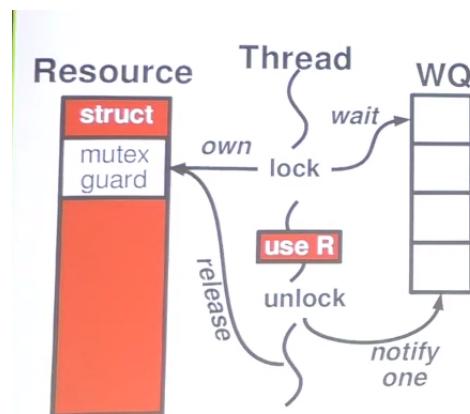


Рисунок 27 – Пример работы мьютекса

Для каждого мьютекса или семафора очередь ожидания своя.

Важно помнить, что захват мьютекса должен быть как можно короче, желательно не больше пары строчек кода.

Если сообщение о разблокировке мьютекса приходит не от того потока, который заблокировал его, то выводится сообщение об ошибке.

Частое решение проблемы Priority Inversion, когда у нас поток с высоким приоритетом не может выполниться, тк мьютекс заблокирован потоком, который также не может пройти дальше из-за низкого приоритета — это наследование приоритета до разблокировки критической секции.

Определение 12. Условные переменные — примитив синхронизации, обеспечивающий блокирование одного или нескольких потоков до момента поступления сигнала от другого потока о выполнении некоторого условия или до истечения максимального промежутка времени ожидания. Условные переменные используются вместе с ассоциированным мьютексом и являются элементом некоторых видов мониторов.

У условных переменных есть 3 состояния: wait, signal, broadcast

Producer / Consumer Problem — проблема загрузки в буфер, когда из него при этом происходит чтение. Для этого существуют два индекса, которые все контролируют.

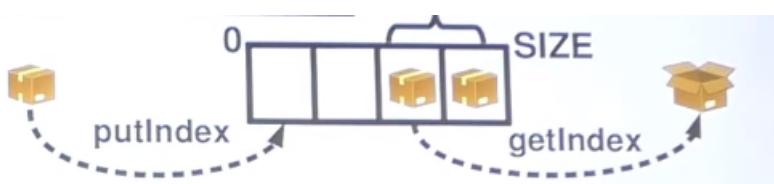


Рисунок 28 – Producer / Consumer Problem

Определение 13. Мониторы — конструкции высокоуровневых языков программирования, которые скрывают низкоуровневые примитивы синхронизации.

Определение 14. Флаги событий — блокировка секций путем проверки условий на флагах. Похожая реализация у event в python.

Multiple reader, single writer locks — когда читатели читают, они захватывают read lock, количество одновременных читателей содержится в rwlock. Когда писателю нужно записать, он устанавливает требование записи (want write) и ожидает на rwlock, который ждет освобождения readlock.

Определение 15. Message passing — является одной из популярных концепций параллельного программирования. Она часто используется при создании сложных распределенных систем с высокой степенью параллелизма. Реализация этой концепции представлена в языках программирования в качестве актёров (actor) или агентов (agent).

- Решение состоит из изолированных компонент, которые работают параллельно (в параллельных потоках из пула потоков). Взаимодействие между компонентами идет через обмен сообщениями по определенному протоколу. Сетевые компоненты могут использовать TCP, UDP, HTTP, и т.д. Локальные взаимодействуют через протокол, определенный конкретный языком его реализации или библиотекой.
- Компонент определяет логику обработки входных сообщений. Последние попадают в очередь (queue) и последовательно достаются из нее для обработки.
- Компонент может быть владельцем некоторых ресурсов и быть их провайдером для других компонент. Ресурсом могут быть: данные в определенном формате в оперативной памяти, аппаратно-программный ресурс или их комбинация.
- Компонент имеет определенное состояние, которое может инкапсулировать ресурс (из предыдущего пункта), или, как в случае машины состояний (state machine), может быть выражен в виде определенного алгоритма обработки сообщений, который переводит его в другое состояние.
- Интерфейс между компонентами:
 - postSync, postAsync — послать сообщение компоненте синхронно или асинхронно.
 - receiveSync, receiveAsync — получить сообщение от компонента синхронно или асинхронно. Асинхронное ожидание состоит в том, что поток как ресурс возвращается системе и может быть использован для другой работы. В этом случае система регистрирует функцию обратного вызова (callback) на определенное событие.
 - tryReceive функции — аналогичные вышеперечисленным, но имеющим определенную задержку для получения данных.
- Концепция тесно связана с понятием асинхронного исполнения.

17 Процессы и потоки в Linux

Каждая задача в ядре Linux описывается структурой task struct, а список задач хранится в виде циклического двусвязного списка (Связный список). Описание структуры task struct можно найти в файле: include -> Linux -> sched.h исходников ядра. task struct ещё называют дескриптором процесса и в нём находится вся важная информация об исполняемом процессе.

<https://elixir.bootlin.com/linux/latest/source/include/linux/sched.h>

Визуально можно представить это данным образом:

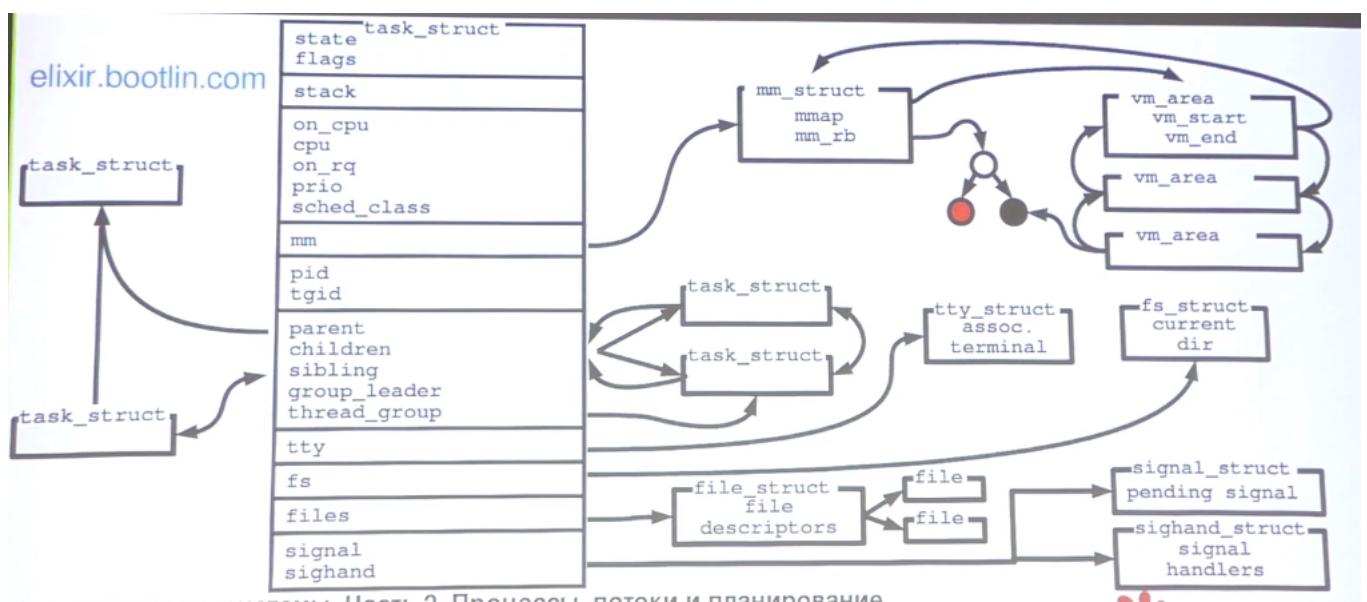


Рисунок 29 – Task-struct

Утилита ps3 показывает дерево процессов.

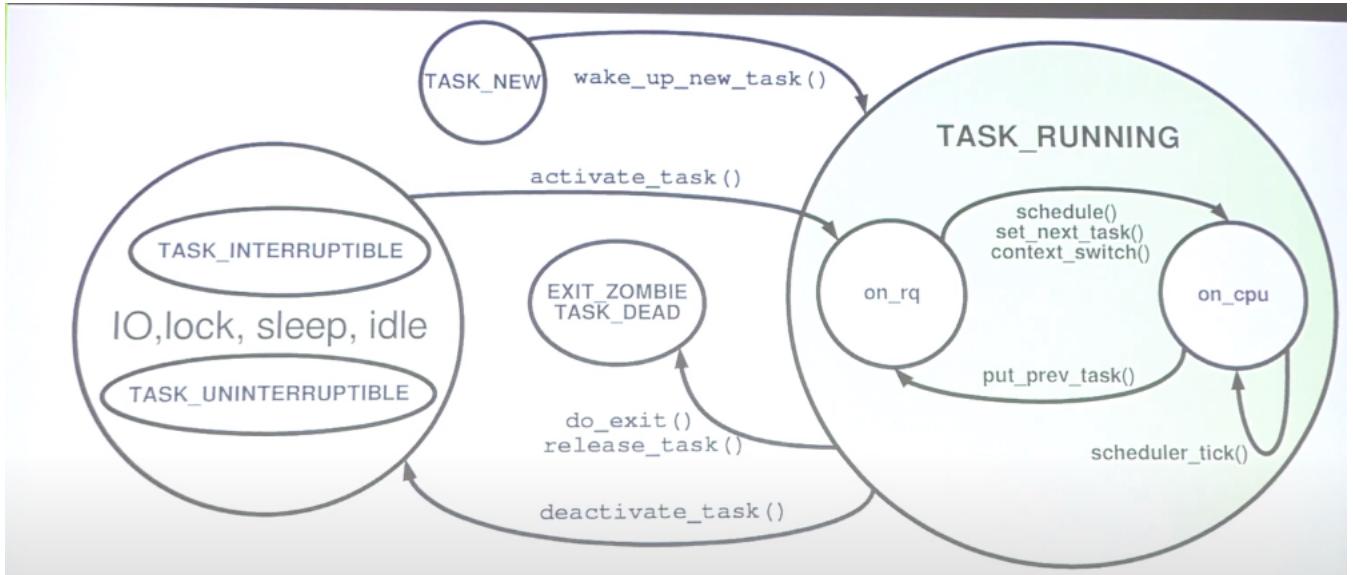


Рисунок 30 – Диаграмма состояний процесса

Помним, что процессы сами себя диспетчеризируют, и к примеру, когда закончился квант времени у процесса, вызывается функция put prev task () и процесс сам уходит в очередь ожидания.

17.1 Создание процесса со стороны пользователя

- fork() — создает процесс, наследуясь от другого процесса, возвращает номер процесса родителю и 0 созданному процессу, -1 в случае ошибки.
- vfork() — создать процесс с копией адресного пространства родителя.
- clone() — создать процесс, управляя копированием выбранных частей процесса.
- execv ()— перекрытие образа процесса.

В Linux нет примитива tread соответствующего пользовательскому потоку. Альтернатива tread это вызов clone.

Пример:

```

clone (CLONE_VM | CLONE_FILES | CLONE_FS | ClonE_High aNd | ClonE_ThReAd |
CLOnE_seT TLs | CLONE_PARENT_SETTID | CLONE_CHILD_CLEARTID CLONE_SYSVSEM, 0);
(NPTL version)

```

Визуализация создания процесса:

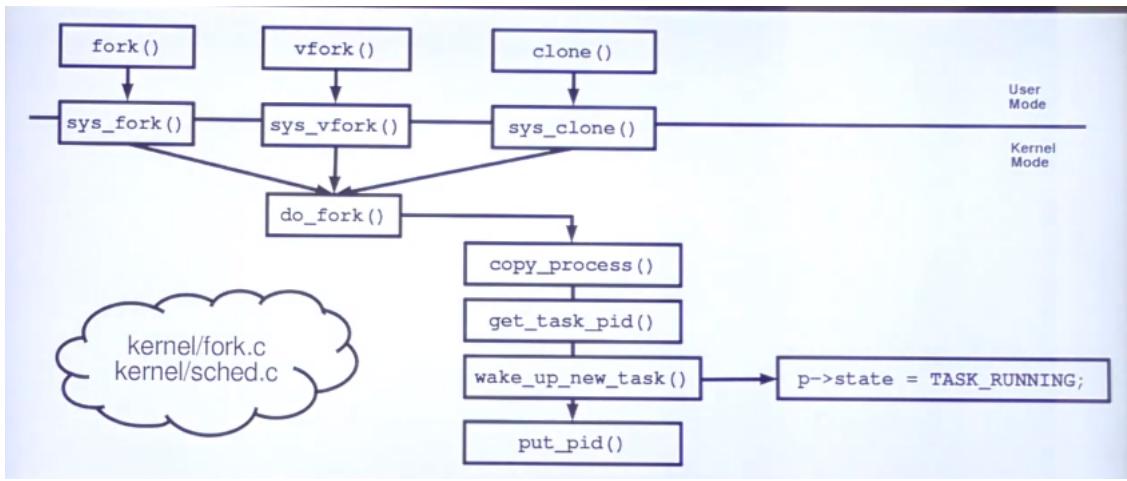


Рисунок 31 – Создание процесса

Функция `copy_process` отвечает за дублирования процесса. Тут не используется стандартное копирование структур, когда мы копируем поля и методы, а используется усовершенствованное.

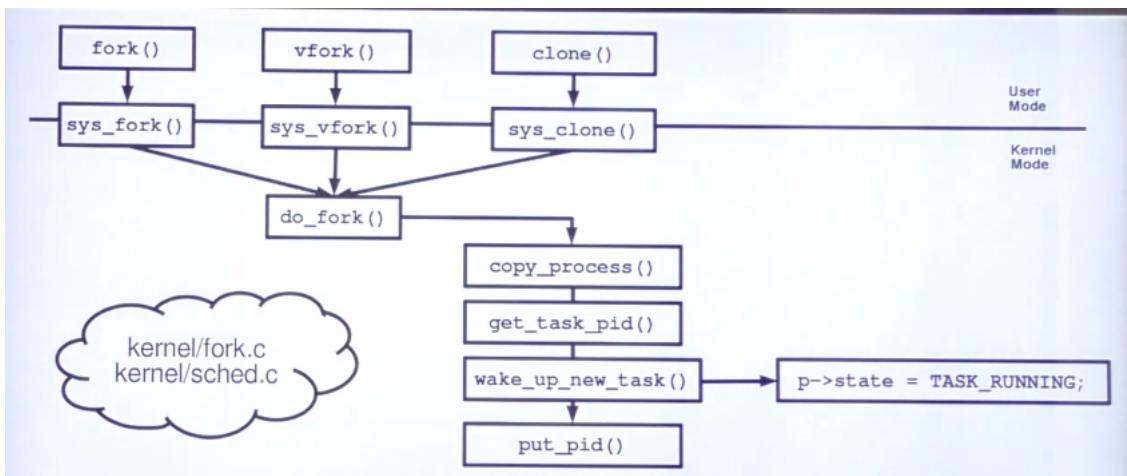


Рисунок 32 – Как выглядит внутри `copy process`

Определение 16. KThread — предназначен для выполнения фоновых операций в ядре, нет своего адресного пространства (используется кэш ядра). По сути является процессом фонового назначения. Особенность заключается в том, что другие потоки не могут остановить его, а только кинуть сообщение об остановке.

Определение 17. Tasklets — придуманы для того, чтобы разнести обработку при вызове прерывания. Обработка разносилась на две части Top Half (быстрая регистрация прерывания) и Bottom half (вызывается по тригеру и обработать прерывания асинхронно). Один tasklet может исполняться только на одном ядре

(процессора).

17.2 Завершение процесса

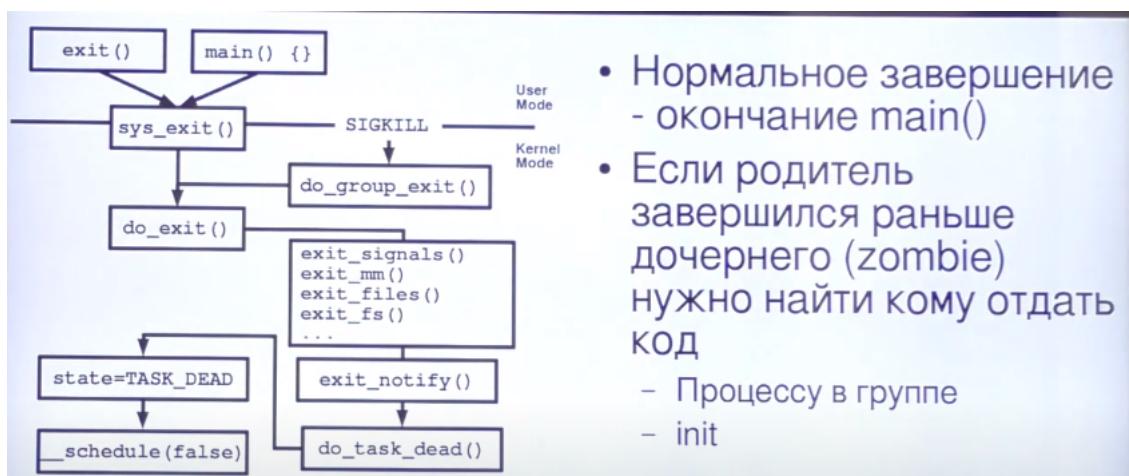


Рисунок 33 – Завершение процесса

18 Примитивы синхронизации в Linux

18.1 Spinlock

Определение 18. **Spinlock** — низкоуровневый примитив синхронизации, применяемый в многопроцессорных системах для реализации взаимного исключения исполнения критических участков кода с использованием цикла активного ожидания. Применяется в случаях, когда ожидание захвата блокировки предполагается недолгим либо если контекст выполнения не позволяет переходить в заблокированное состояние.

Чтобы посмотреть как выглядит спинлок зайдите в `include -> linux -> spinlock types.h`

Спинлок по сути состоит из одного большого Union в C (структуры с единой областью памяти).

В зависимости от архитектуры у вас будет разные спинлоки `raw`, `arch`, `q`.

При раскрытие полей дебага спинлока, нужно пресобрать ядро.

```
include/asm-generic/qspinlock_types.h
typedef struct qspinlock {
    union {
        atomic_t val;
    #ifdef __LITTLE_ENDIAN
        struct {
            u8 locked;
            u8 pending;
        };
        struct {
            u16 locked_pending;
            u16 tail;
        };
    #else
        // reverse order and alignment
    #endif
    };
} arch_spinlock_t;
```

Рисунок 34 – Структура спинлока

Спинлок может запрещать прерывания на момент выполнения, сохранять состояния прерывания, также есть функции проверки блокировки спинлока, функции разблокировки и тд.

В кончном итоге спринлок использует Qspinlok очередь, которого разбивается на несколько для повышения производительности. Очереди определяются полем value. Сохраняется один бит, под второй поток приходящий в очередь из тех же соображений. Очередь начинает заполняться только с третьего потока. Очереди существуют для каждого процессора и для каждого из них по 4 очереди.

18.2 Semaphore

```
include/linux/semaphore.h
struct semaphore {
    raw_spinlock_t lock;
    unsigned int count;
    struct list_head wait_list;
};

#define __SEMAPHORE_INITIALIZER(name, n) \
{ \
    .lock = __RAW_SPIN_LOCK_UNLOCKED((name).lock), \
    .count = n, \
    .wait_list = LIST_HEAD_INIT((name).wait_list), \
}

static inline void sema_init(struct semaphore *sem, int val)
{
    static struct lock_class_key __key;
    *sem = (struct semaphore) __SEMAPHORE_INITIALIZER(*sem, val);
    lockdep_init_map(&sem->lock.dep_map, "semaphore->lock", &__key, 0);
}
```

- void **down**(struct semaphore *sem);
- void **up**(struct semaphore *sem);
- int **down_interruptible**(struct semaphore *sem);
- int **down_killable**(struct semaphore *sem);
- int **down_trylock**(struct semaphore *sem);
- int **down_timeout**(struct semaphore *sem, long jiffies)

Рисунок 35 – Структура семафора

Также на картинке представлены функции для захвата симафора.

При захвате симафора важно отпускать потом его, тк иначе ресурс будет заблокирован, и приложение умрет.

Списки ядра медленнее, чем у спинлока.

18.3 Mutex

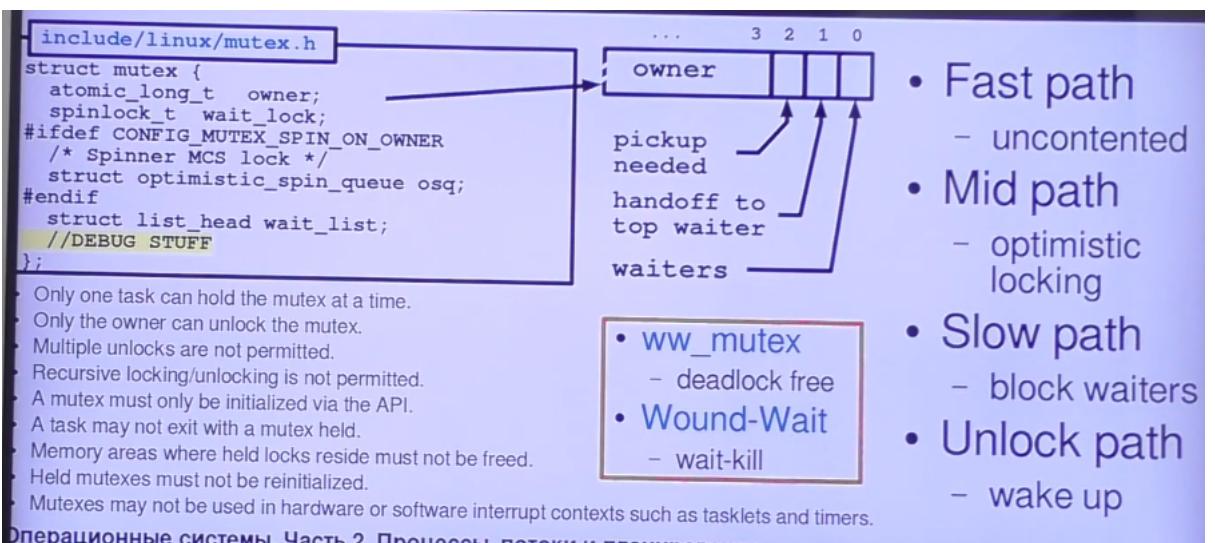


Рисунок 36 – Структура мьютекса

Поле owner указывает на того, кто захватил мьютекс. Есть спинлок, который защищает очередь ожидающих спинлока.

Также на изображение представлен список требований к мьютексу.
Hand off позволяет передавать мьютекс без освобождения.

19 Процессы и потоки в Windows

19.1 Типы процессов

Современные процессы — Universal Windows Platform process (приложения из магазина майкрософт). В них также есть Protected Processes, которые недоступны даже для администратора, чтобы избежать пиратства.

Minimal процессы — нет пользовательского пространства и работают как часть ядра (system process и Memory Compression process)

Pico процессы — ограниченный доступ к системным функциям представляющий набор callback.

Trustlets процессы — процессы для запуска внутри системы визуализации (Virtual trust level)

Jobs процессы — средство группировки процессов.

WOW — 32 битные процессы на 64 разрядной системе.

Обычные потоки (User-Kernel) — базовая реализация posix tread.

Fider потоки — пользовательская реализация потоков, невидимая ядру. (kernel32.dll). Данные потоки сами освобождают ресурсы ядра для след процесса.

Asynchronous Procedure Call (APS) и Deffered Procedure Call (DPC) — концепции, которые позволяют выполняться на уровне прерываний, который выше, чем у пользователя. (используются при программирование на ядро)

Определение 19. IRQL — «уровень запроса прерывания». Механизм программно-аппаратной приоритизации, применяемый для синхронизации в операционных системах.

19.2 Структура процессов Windows

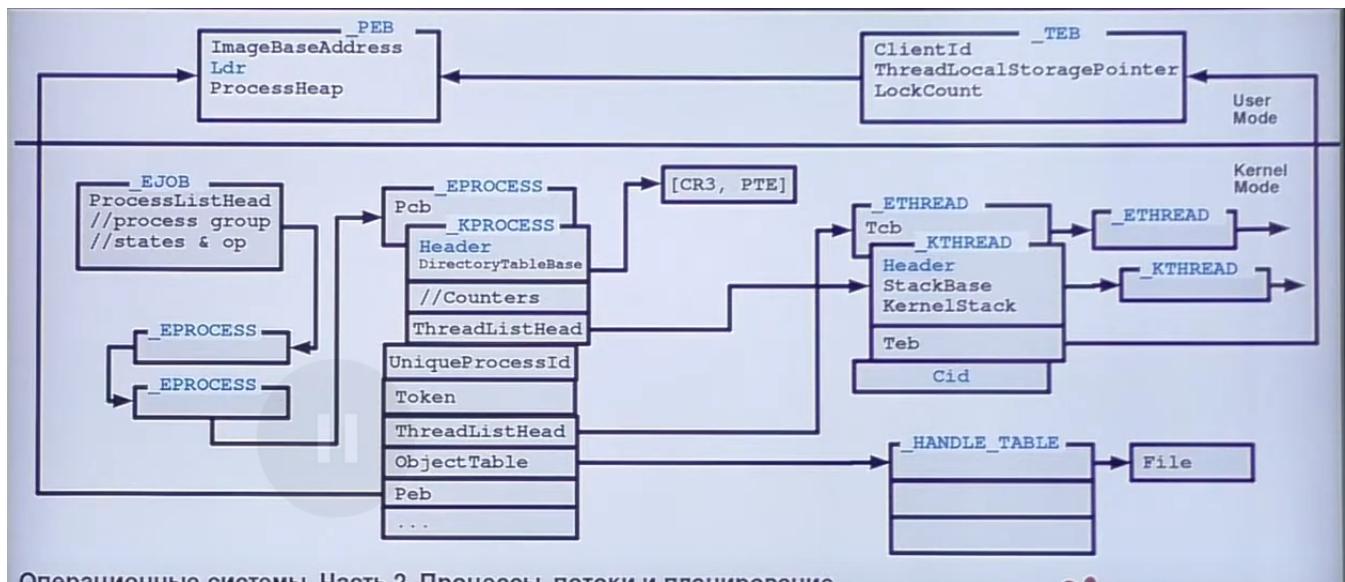


Рисунок 37 – Структура процессов Windows

Главным элементом является структура **Eprocess**, в которой содержится индикатор процесса, токен безопасности, список потоков принадлежащих процессов.

В целом большинство вещей в ядре Windows являются объектами (максимально приближенно к концепции ООП).

В Eprocess встроен **Kprocess** или **kernel process**, который содержит инструкции исполнения на процессоре, также счетчики производительности

PCB — управляющий блок процесса. **TCB** — управляющий блок потоков.

Аналогично Kprocess есть Ktread.

Все это находится на Kernel уровне, а чтобы получить доступ к частям этих структур на пользовательском режиме используют структуры PEB (описание процесса и ссылки на его части) и (ТЕВ, то же самое с потоками).

CSRSS — Процесс исполнения клиент-сервер, также помогает с пользовательского уровня добраться до kernel процессов.

В Windows нет в явном виде parent процессов. Есть процессы и их потоки, которые хранят ссылки на друг друга.

19.3 Состояние процессов Windows

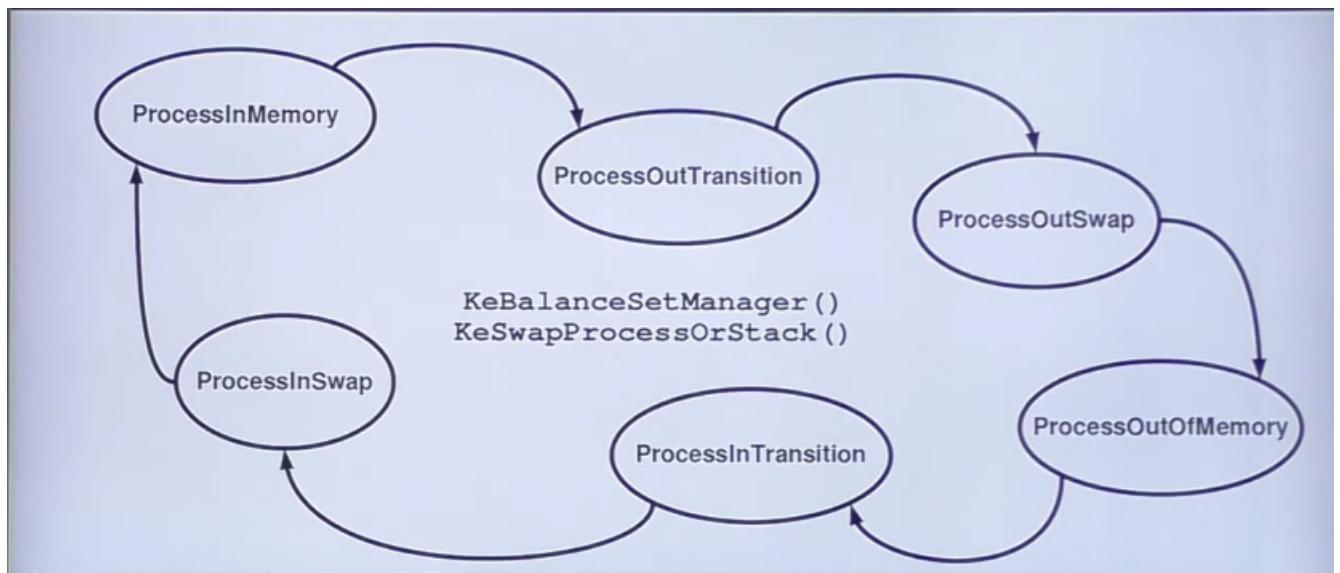


Рисунок 38 – Состояния процессов Windows

KeSwapProcess — функция, которая занимается загрузкой и выгрузкой процессов.

Состояние транзита, когда процесс начал выгружаться в swap область.
(Помечен для выгрузки, очередь на диспетчеризацию загрузки).

ProcessOutSwap — фактическая выгрузка в swap.

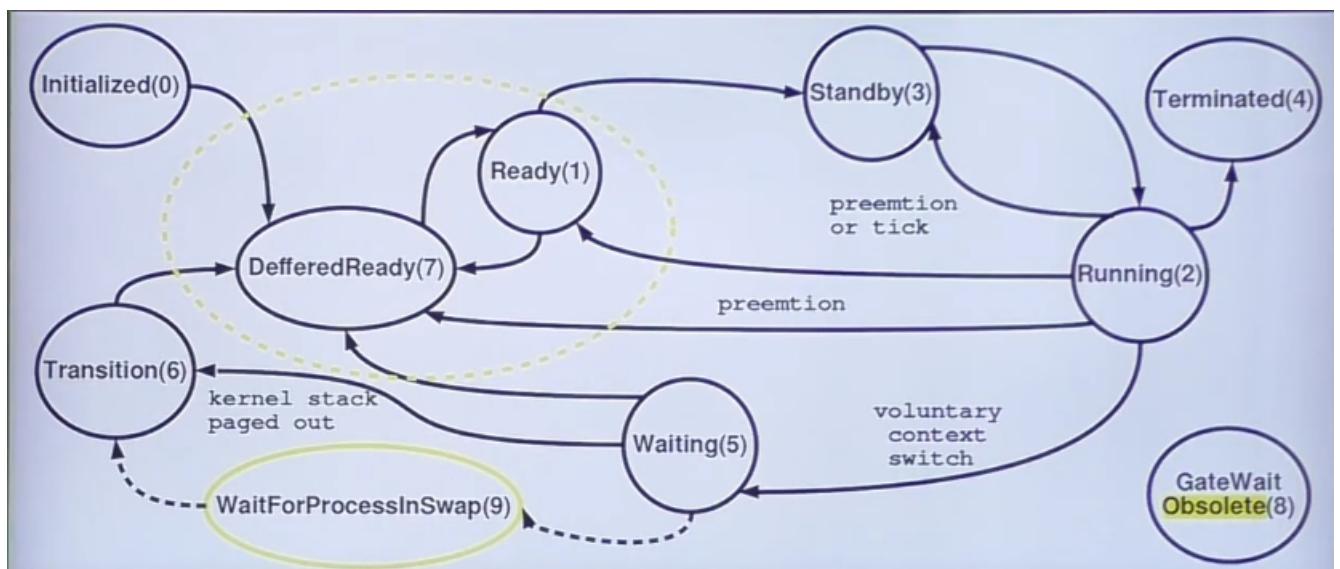


Рисунок 39 – Состояния потоков Windows

- Поля _EPROCESS
 - PCB
 - Защищающие блокировки
 - UniqueProcessID
 - Сылки списка процессов (начало в PsActiveProcessHead)
 - Флаги
 - Времена создания и завершения
 - Информация о квотах
 - Ссылка на сессию
 - Основной токен доступа
 - Ссылка на задание
 - Объекты процесса (Handle Table)
 - Окружение процесса (PEB)
 - Имя фала образа процесса
 - Счетчики производительности
 - Список потоков
 -
 - Win32K структура
 - WOW64 структура
 - Pico Context
 - DirectX process
 - Набор рабочих страниц WorkingSet
 - Секции (сегменты) образа
- Поля _KPROCESS
 - Заголовок диспетчера
 - Ссылка на таблицу страниц
 - Kernel/User/Cycle Times
 - Context Switches
 - Список Thread
 - Аффинити
 - Флаги
 - Базовый приоритет

Рисунок 40 – Поля процессов и потоков Windows

19.4 Создание процессов

NtCreateUserProcess — главная функция, которая осуществляется в ядре вызов создания процесса. Nt.dll — библиотека системных вызовов нижнего уровня, к ней подключаются все библиотеки, которые вы можете использовать как user.

Avapi.dll — библиотека, которая создает процесс с использованием вашего логина или токена безопасности. Она не лезет на прямую к нижнему уровню, а использует процесс SvcHost.exe.

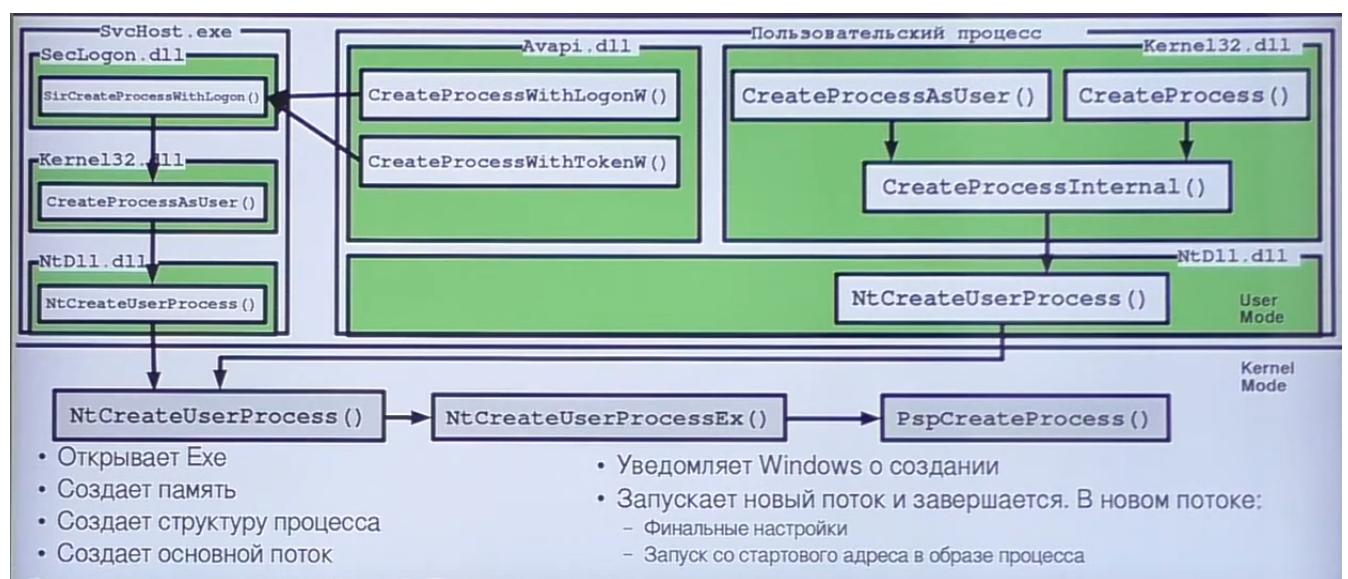


Рисунок 41 – Создание процессов в Windows

19.5 Завершение процессов

Корректное завершение — ExitProcess();

Прекращение процесса другим процессом — TerminateProcess();

Последовательность завершения процесса:

1. Оповещение DLL (если не использован TerminateProcess())
2. Закрытие всех handles и kernel objects
3. Закрываются все активные потоки
4. Код возврата изменяется на указанный
5. Когда все ссылки в процессе равны 0, объект процесса удаляется.

- A1pc - Расширенные локальные вызовы процедур
- Cc - Общий кэш
- Cm - Диспетчер конфигурации
- Dbg - Поддержка отладки ядра
- Dbgk - Отладочная инфраструктура для пользовательского режима
- Em - Диспетчер ошибок
- Etw - Трассировка событий для Windows
- Ex - Исполнительные вспомогательные функции
- FsRtl - Библиотека файловой системы времени выполнения
- Hv - Библиотека Hive
- Hvl - Библиотека гипервизора
- Io - Диспетчер ввода/вывода
- Kd - Отладчик ядра
- Ke - Ядро
- Ki - Внутренние функции ядра
- Kse - Оболочка совместимости ядра
- Lsa - Локальная система безопасности
- Psp - внутренние вспомогательные функции процессов.
- Mm - Диспетчер памяти

Рисунок 42 – Префиксы функций в Windows и их значение

20 Примитивы синхронизации в Windows

20.1 Объекты диспетчера

Любой объект ядра может быть двух типов dispatcher object и control object (который не рекомендуется трогать).

dispatcher object — любой объект ядра, на котором можно ожидать появление события. (dispatcher header — структура, которая управляет знаниями процессора об разных событиях, есть и в ktread и тд). Dispatcher object может находиться в двух состояниях signaled и not signaled, наступило событие или нет.

События ожидания для всех объектов представлена одними и теми же функциями.

Гибкие функции вызова ожидания: KeWaitForSignaled и KeWaitForMultipleObjects ожидают один или несколько объектов.

Объекты, которые можно ожидать: Event, Mutex, Semaphore, Timers, Processes, Thread, Directories.

Также важно отметить, что нет четкого порядка завершения ожидания.

EventObject — может сбрасываться в ручную, автоматически и pulsed. Используется для оповещения о наступление события двух типов: синхронизация(автоматически сбрасывает состояние signality) и нотификация(требует ручного сброса состояния signality)

20.2 События, мьютексы, семафоры

Мьютексы реализованы через основные функции поддержки Mutants (общая структура kMutant).

Структура для обоих примитивов синхронизации одинаковая.

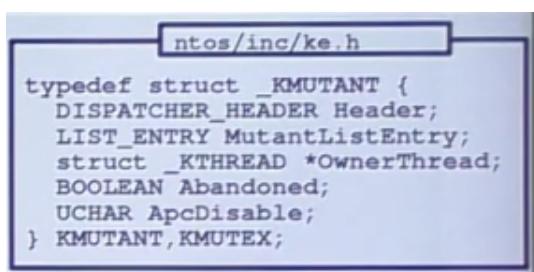


Рисунок 43 – Структура мьютексов и мутантов

Мьютекс должен быть освобожден тем потоком, который его захватил. А мутант любым. APC запрещены в мьютексах, а в мутантах нет. Также мутанты

можно использовать в области ядра и в области пользовавателя.

Оба примитива используют рекурсивный захват.

Мьютексы и семафоры сразу создаются в signaled.

Захват ресурсов происходит функцией ожидания, а освобождение release semaphore.

20.3 Spinlocks

Не отличается от спинлоков в linux. Используется атомарная операция test-and-modify. Один владелец в один момент времени. Как и в linux спинлоки предназначены для быстрой обработки, соответсвенно нормальные уровне Dispatcher блокируют поток на уровне ниже 2, то смысла в спинлоках нет.

21 Полезные утилиты

Linux kernel map — <https://makelinux.github.io/kernel/map/>

Сайт с рекомендациями по отладке Linux —

<https://brendangregg.com/linuxperf.html>

Сайт с исходниками линукса —

<https://elixir.bootlin.com/linux/latest/source/include/linux/sched.h>