

Лекции по Операционным системам

Сверстал: Кузякин Никита Александрович

По лекциям ИТМО

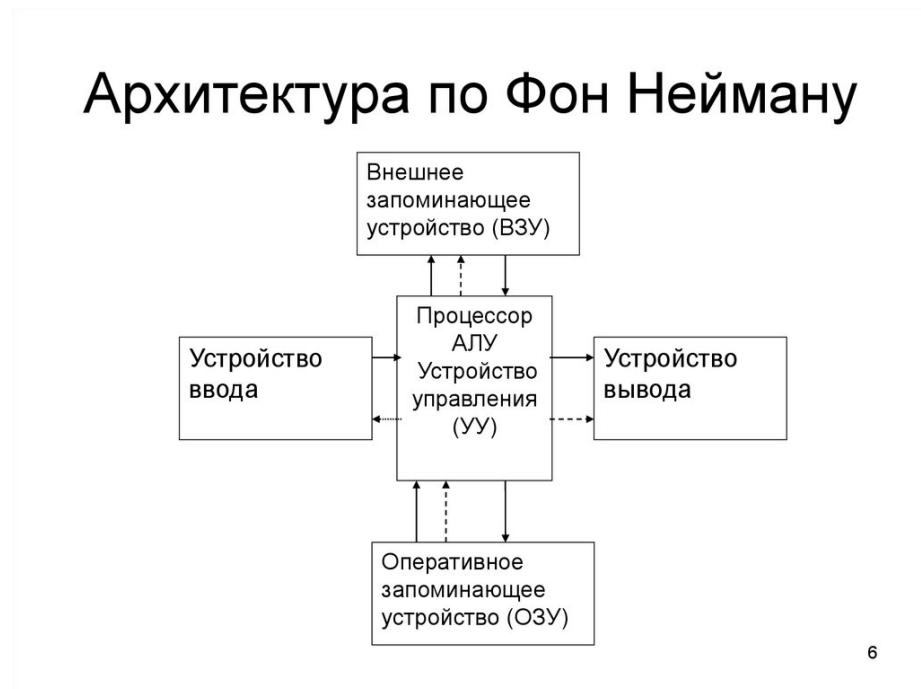
Плейлист с лекциями — [тут](#)

СОДЕРЖАНИЕ

1	Архитектура компьютерных систем	3
2	Обзор элементов компьютерных систем	6
2.1	Процессор	6
3	Общие сведения об операционных системах	8
3.1	Функции OS.....	8
3.2	Оператор ЭВМ	8
3.3	Пакетная обработка	8
3.4	Многозадачность	8
3.5	Разделение времени	9
4	Основные задачи OS	10
4.1	Управление процессами.....	10
4.2	Виртуальная память	11
4.3	Безопасность	11
4.4	Диспетчеризация и планирование ресурсов	12
5	Современные архитектурные концепции OS	13
5.1	Архитектура ядер	13
5.2	Многопоточность	14
5.3	SMP и ASMP	14
5.4	Виртуализация	15
6	Основные понятия надежности операционной системы	16
6.1	Надежность и отказоустойчивость	16
6.2	Сбои	16
7	Общая архитектура UNIX / Linux.....	18
8	Общая архитектура Windows	19
9	Средства для отладки Linux	21
9.1	Стандартные средства.....	21
9.2	/proc	22
9.3	Трассировщики	22
9.4	perf	22
9.5	System tap	24
9.6	Kernel debugger	24
10	Средства для отладки Windows	26
11	Полезные утилиты	26

1 Архитектура компьютерных систем

Первоначальными двумя архитектурами компьютерных систем являются Гарвардская и Неймановская архитектуры.



6

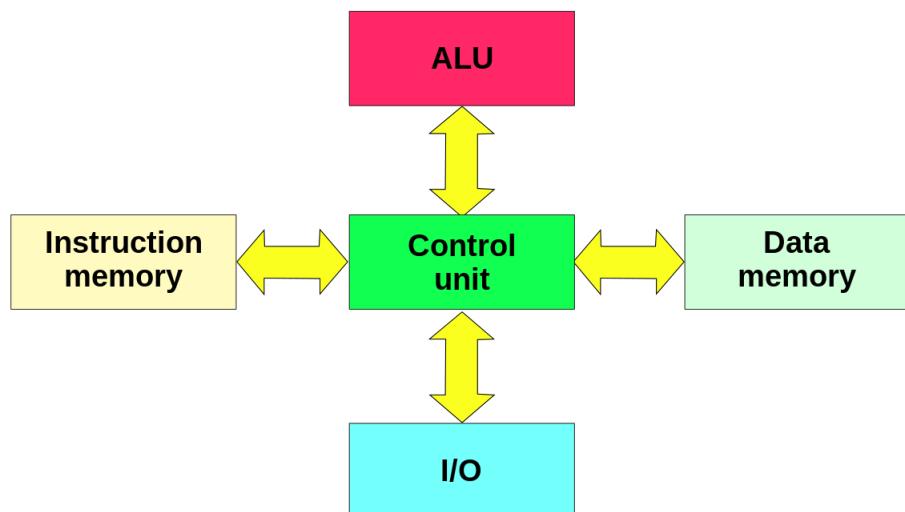


Рисунок 1 – Гарвардская архитектура ЭВМ

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 управляющие 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.

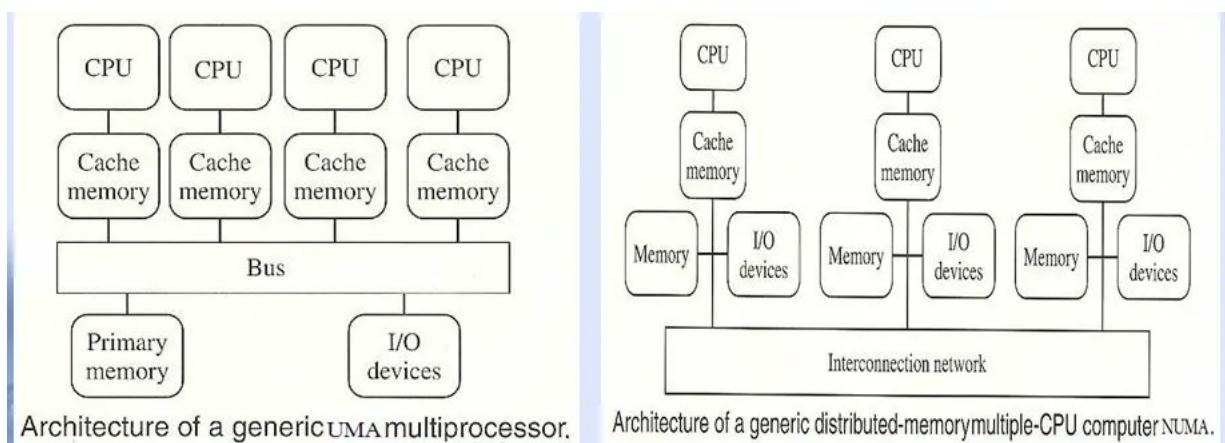


Рисунок 2 – Гарвардская архитектура ЭВМ

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

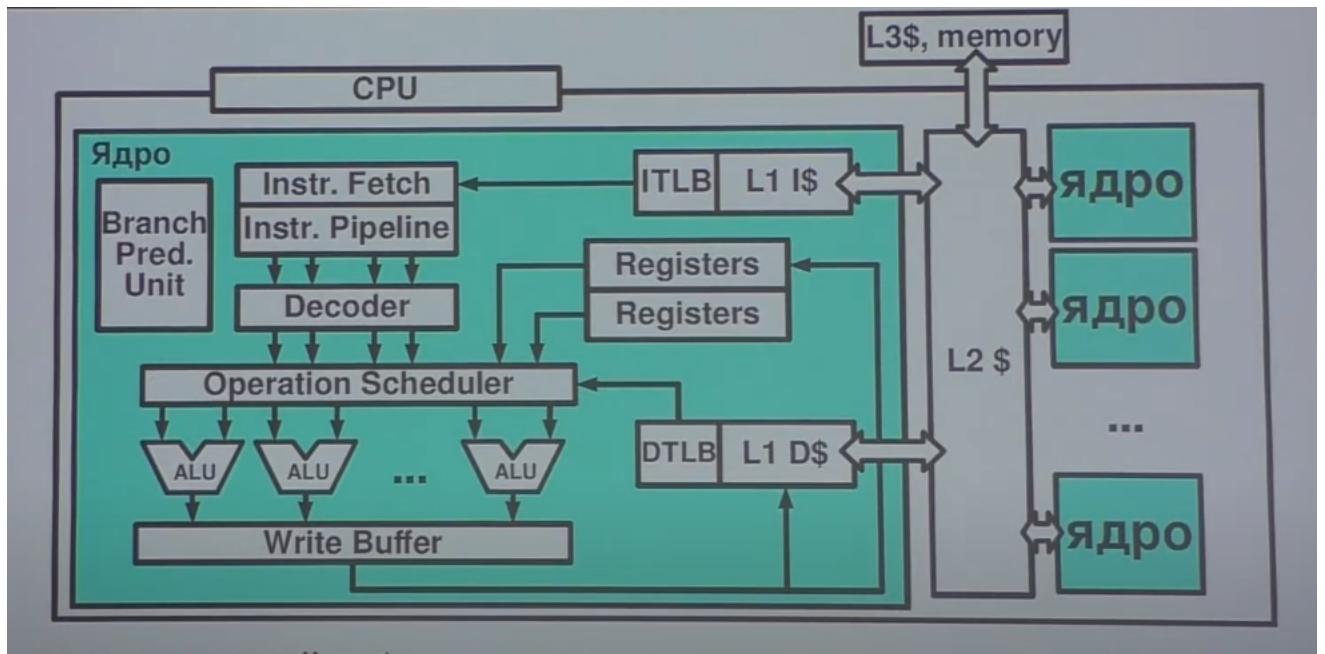
В архитектуре NUMA у нас есть память, которая находится ближе к какому-то процессору и память, которая доступна через коммутатор (передает данные через порты).

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.

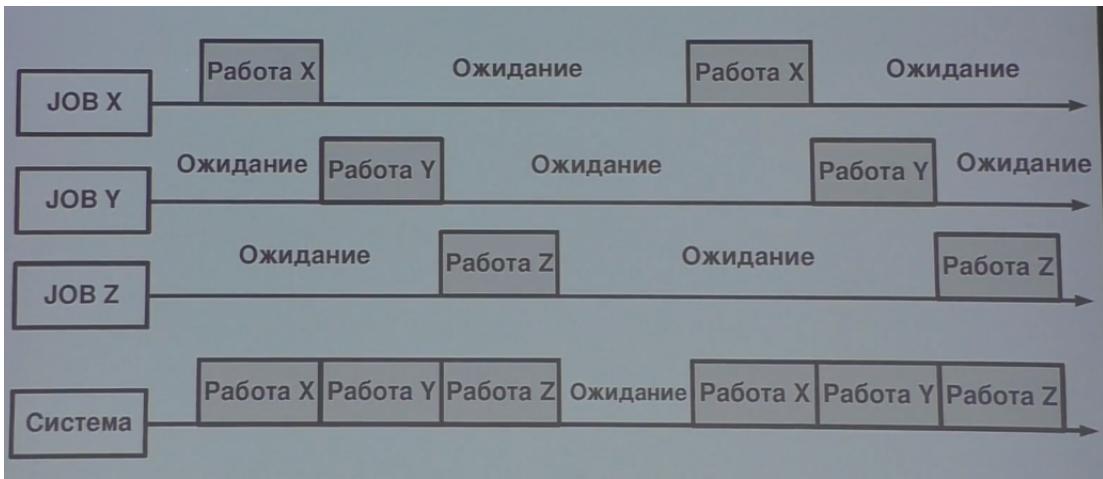


Рисунок 3 – Схема многозадачности первых ЭВМ

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.

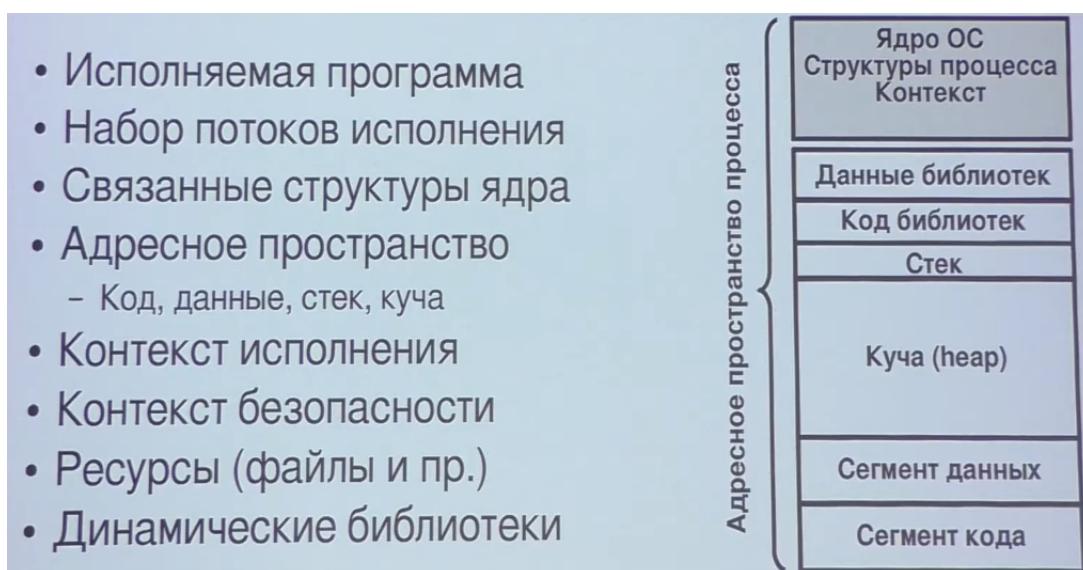


Рисунок 4 – Структура процесса

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

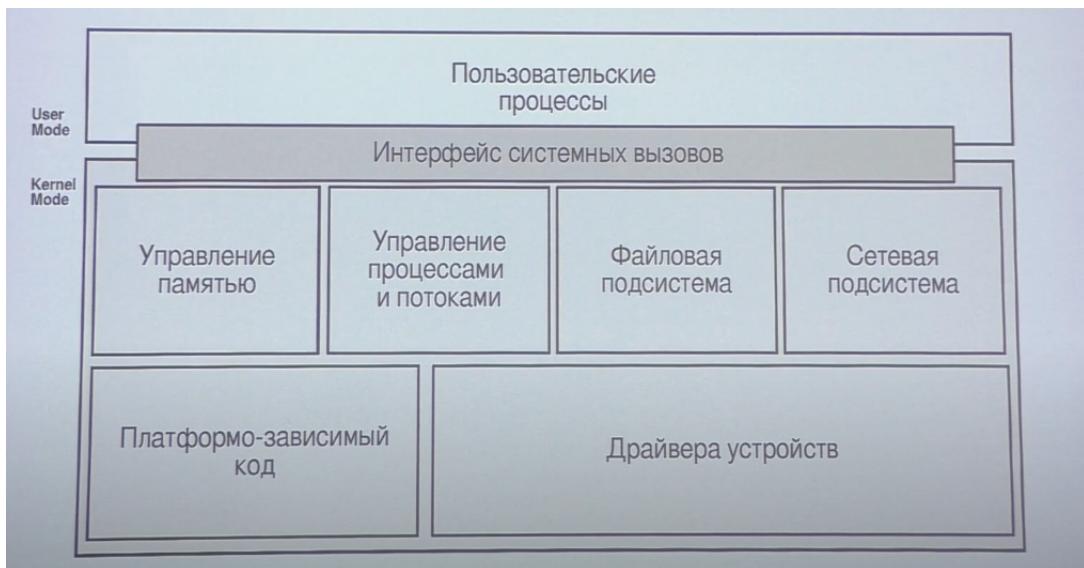


Рисунок 5 – Схема ядра ОС

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.

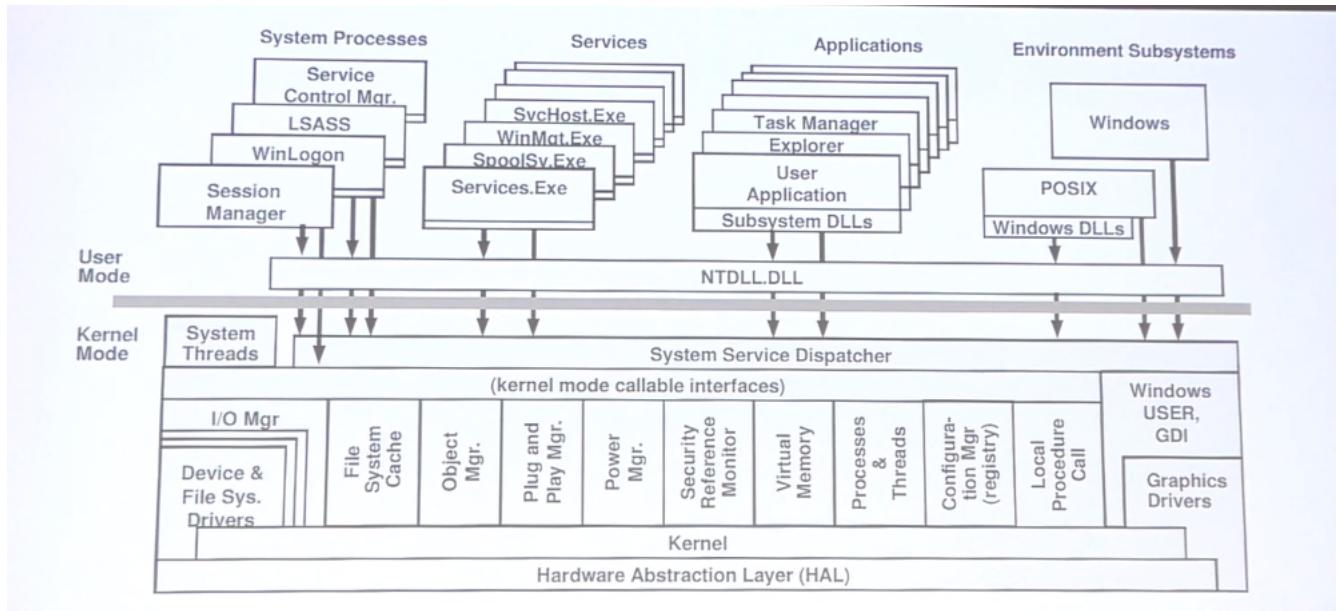


Рисунок 7 – Архитектура Windows

Как не странно архитектура Windows очень похожа на архитектуру UNIX.

Интерфейс системных вызовов — NTDLL.dll

Также подсистемы Windows похожи на подсистемы в Linux.

Plag 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.

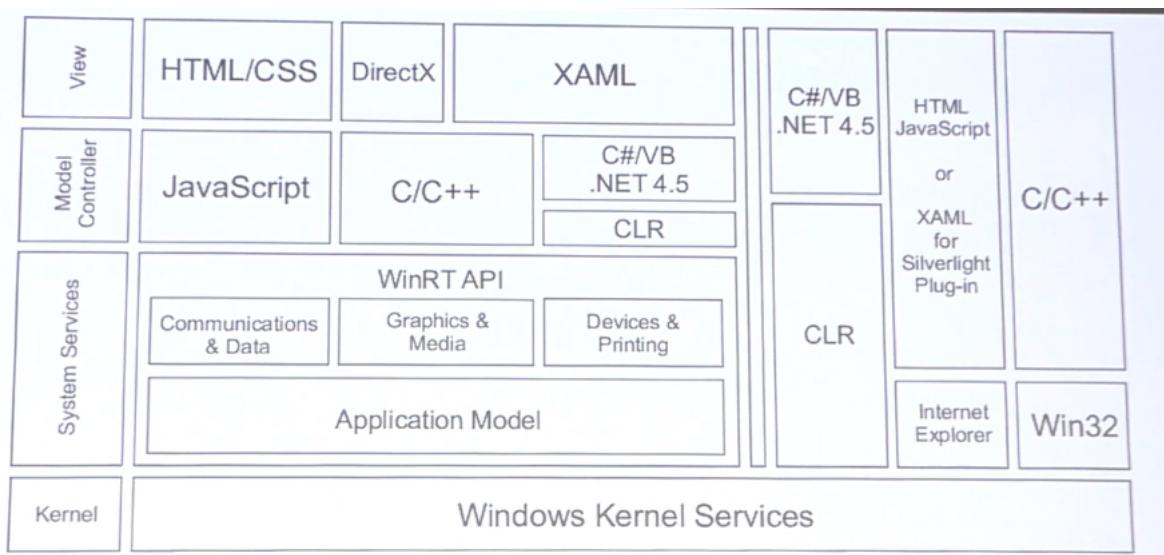


Рисунок 8 – Схема WinAPI

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

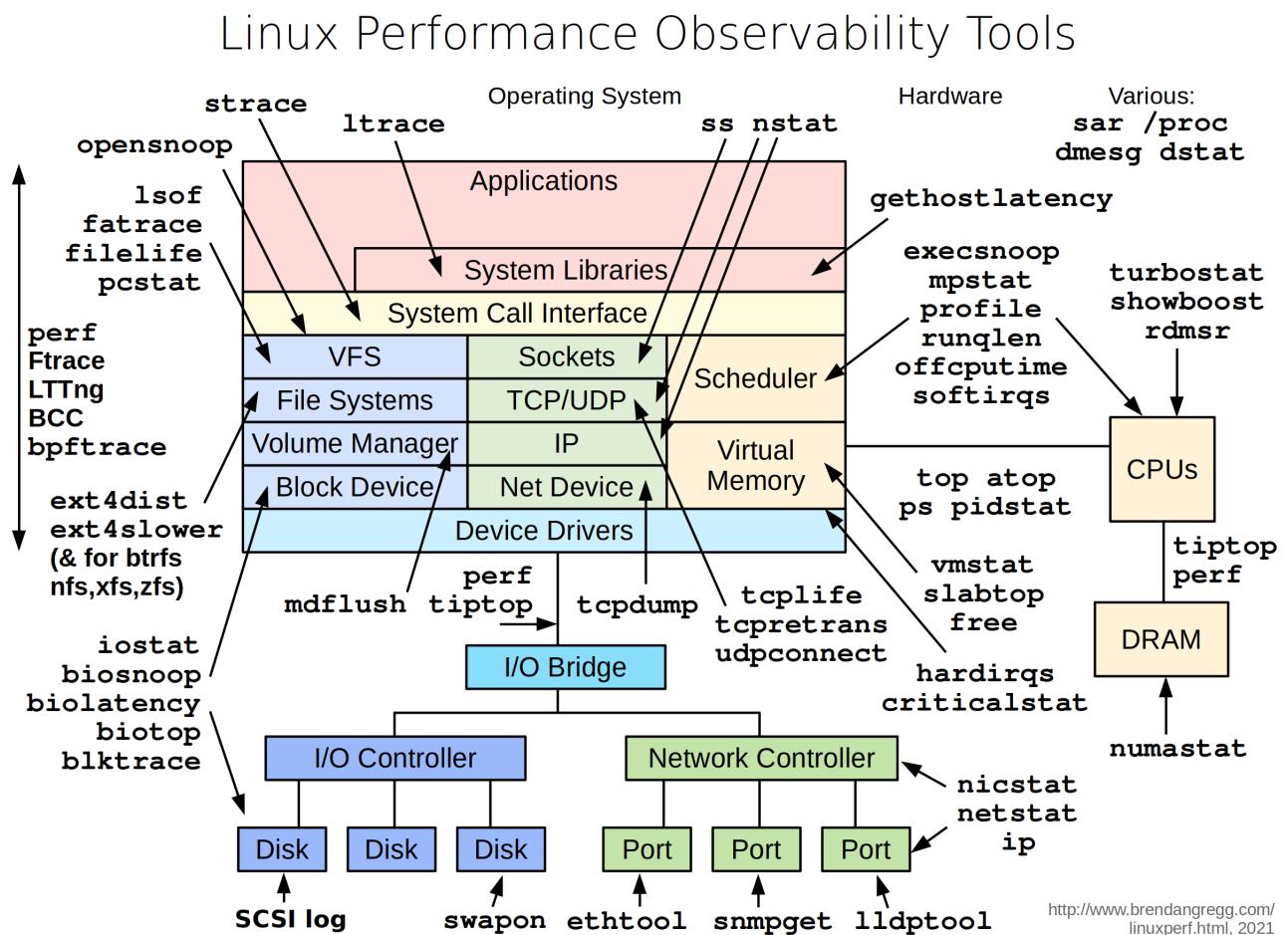


Рисунок 9 – Утилиты для отладки линукс

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.

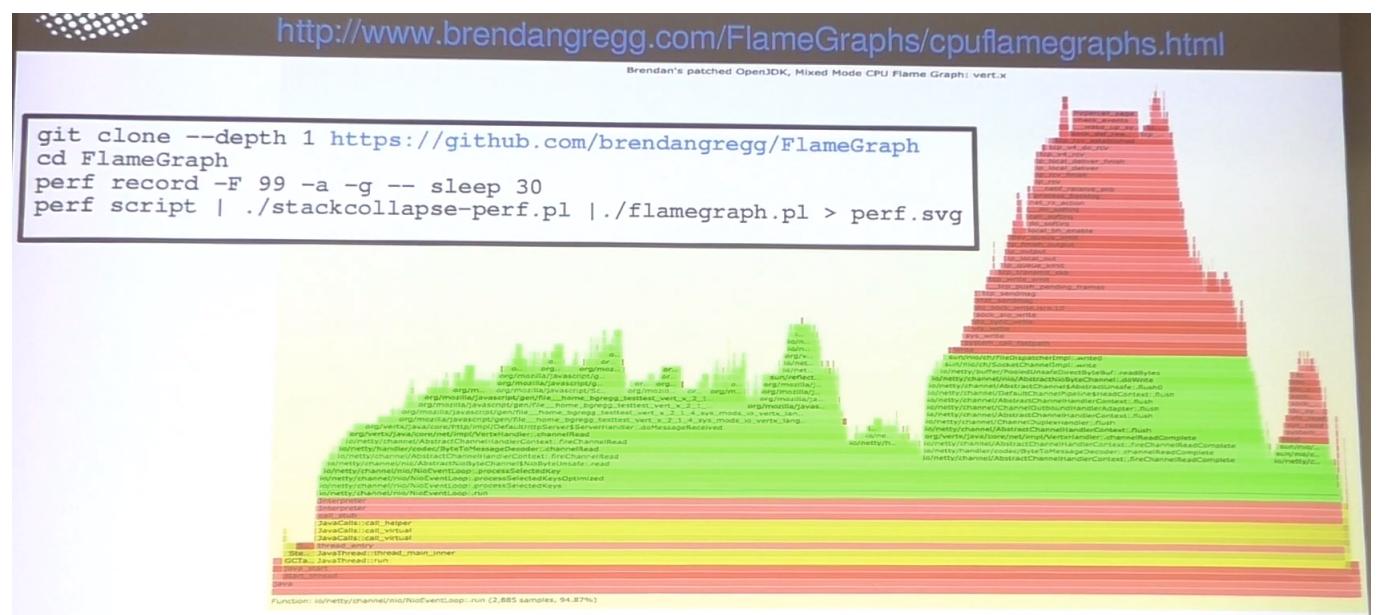


Рисунок 12 – FlameGraph

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/devel/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

11 Полезные утилиты

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

Сайт с рекомендациями по отладке Linux —

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