

# Распределенные системы хранения и обработки данных

Владислав Белогрудов, EMC

[vlad.belogrudov@gmail.com](mailto:vlad.belogrudov@gmail.com)

# Лекция 2

## Среда систем хранения данных

# Содержание лекции

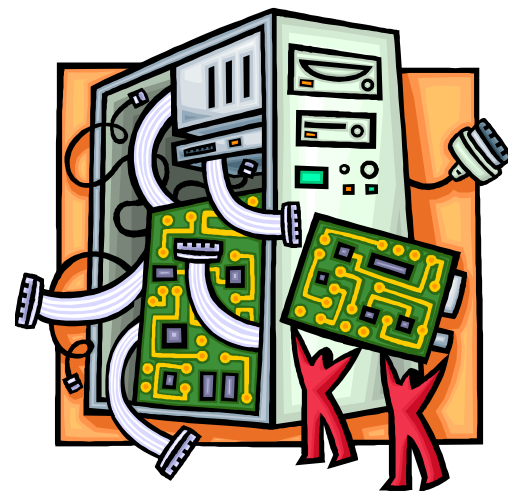
- Основные элементы
- Компоненты жесткого диска и его производительность
- Накопители SSD
- Виртуализация

# Компоненты среды хранения данных

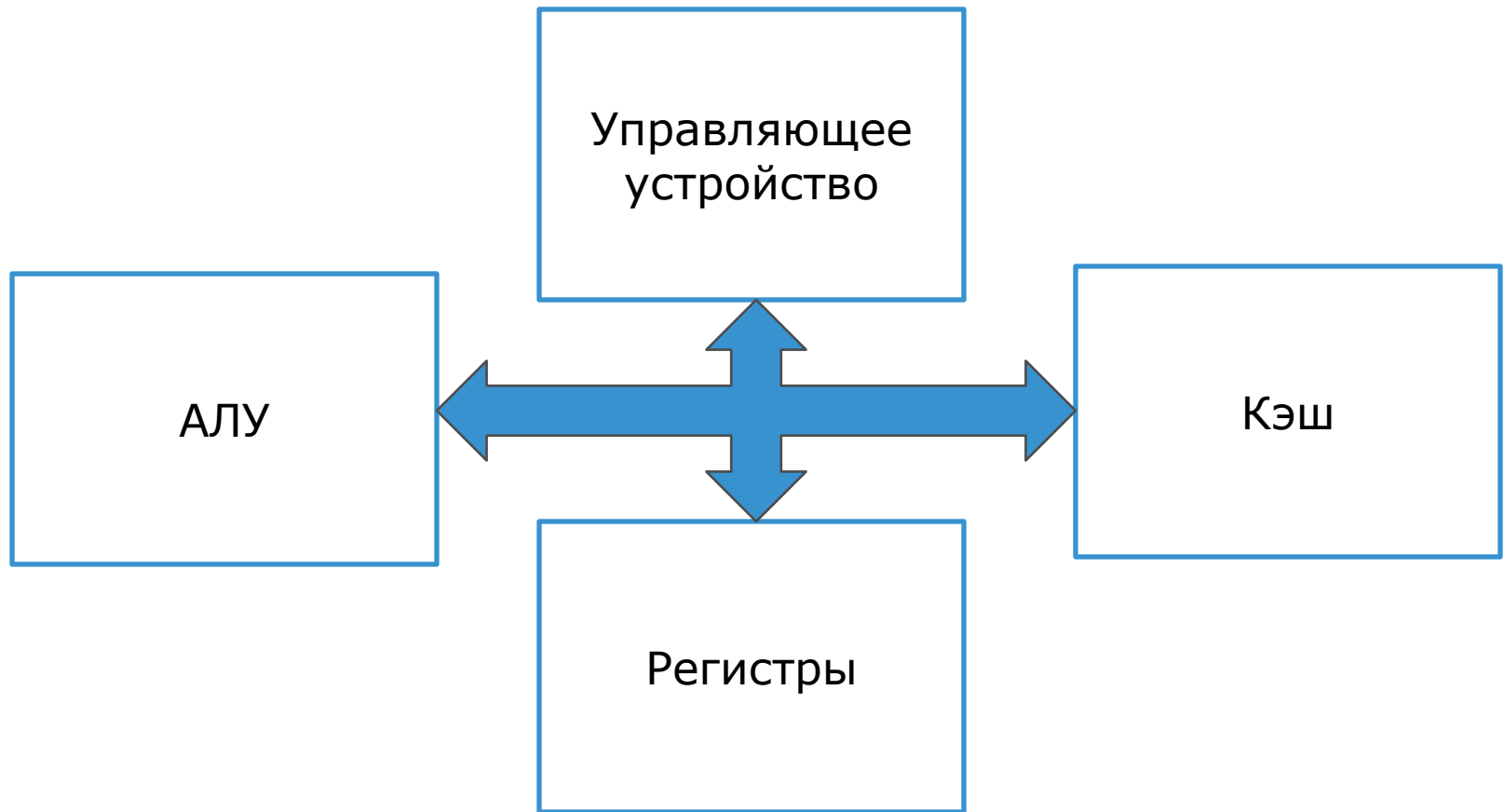


# Хост

- Лэптоп, РС, группа серверов, которые запускают приложение
- Три физических компонента
  - центральный процессор
  - устройства хранения (память, диски)
  - Устройства ввода/вывода



# Центральный процессор



# Запоминающие устройства

- RAM
  - Энергозависимая память, хранит данные и программы во время их исполнения
- ROM
  - Энергонезависимая память, для запуска системы, тестирования компонентов

# Execute in place (XIP)

- Программа выполняется с диска
- ОЗУ используется для изменяемых данных
- Меньше памяти
- Быстрая загрузка ОС
- Пререквизиты
  - СХД имеет тот же интерфейс, что и основная память
  - NOR flash для побайтного доступа (медленно и дорого)

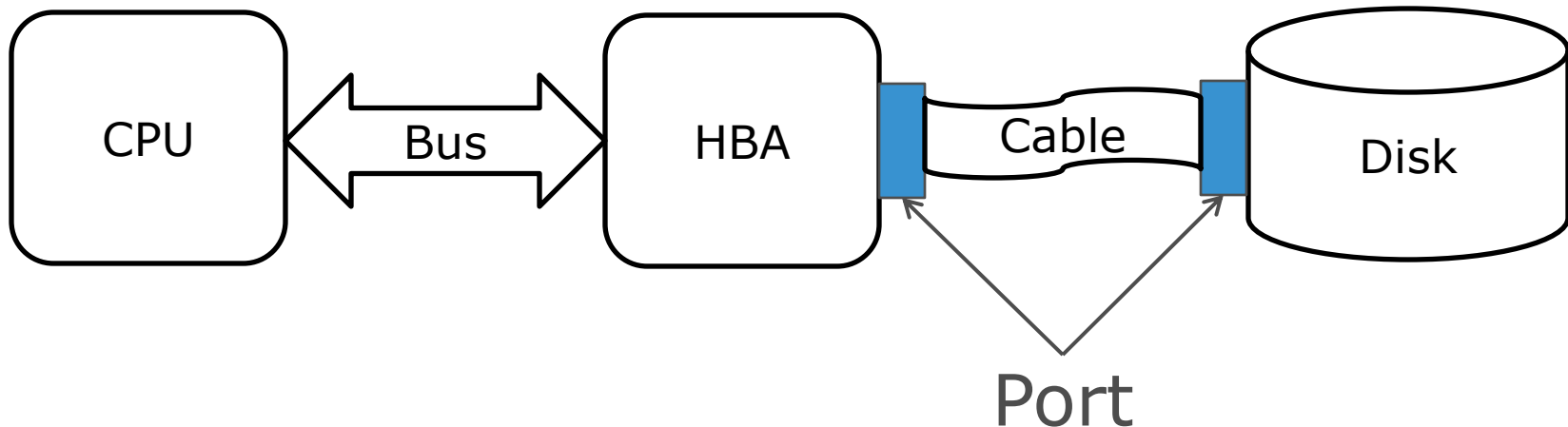


# Устройства ввода-вывода

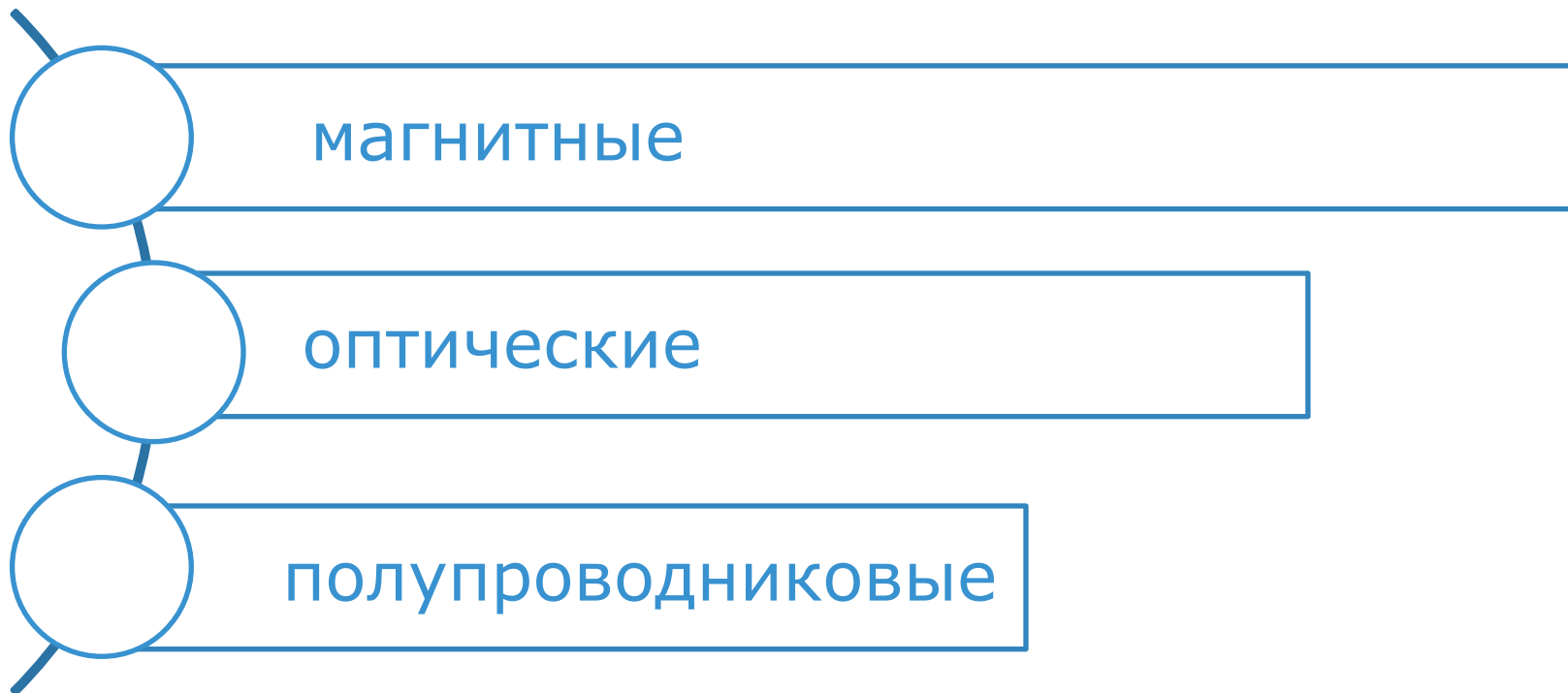
- Связь от пользователя к хосту (клавиатура, мышь, монитор и т.п.)
- Связь от хоста к хосту (NIC, modem)
- Связь от хоста к устройству хранения данных (HBA).

# Соединение устройств СХД

- Физические компоненты – платы, кабели
- Логические компоненты – программы, протоколы



# Устройства хранения



# Магнитные ленты



- Большой объем
- Дешевизна



- Последовательный доступ
- Один поток
- Изнашивание

# Оптические диски (CD, DVD, BD)

- Малый объем данных
- Дешевы
- Медленны
- Ненадежны



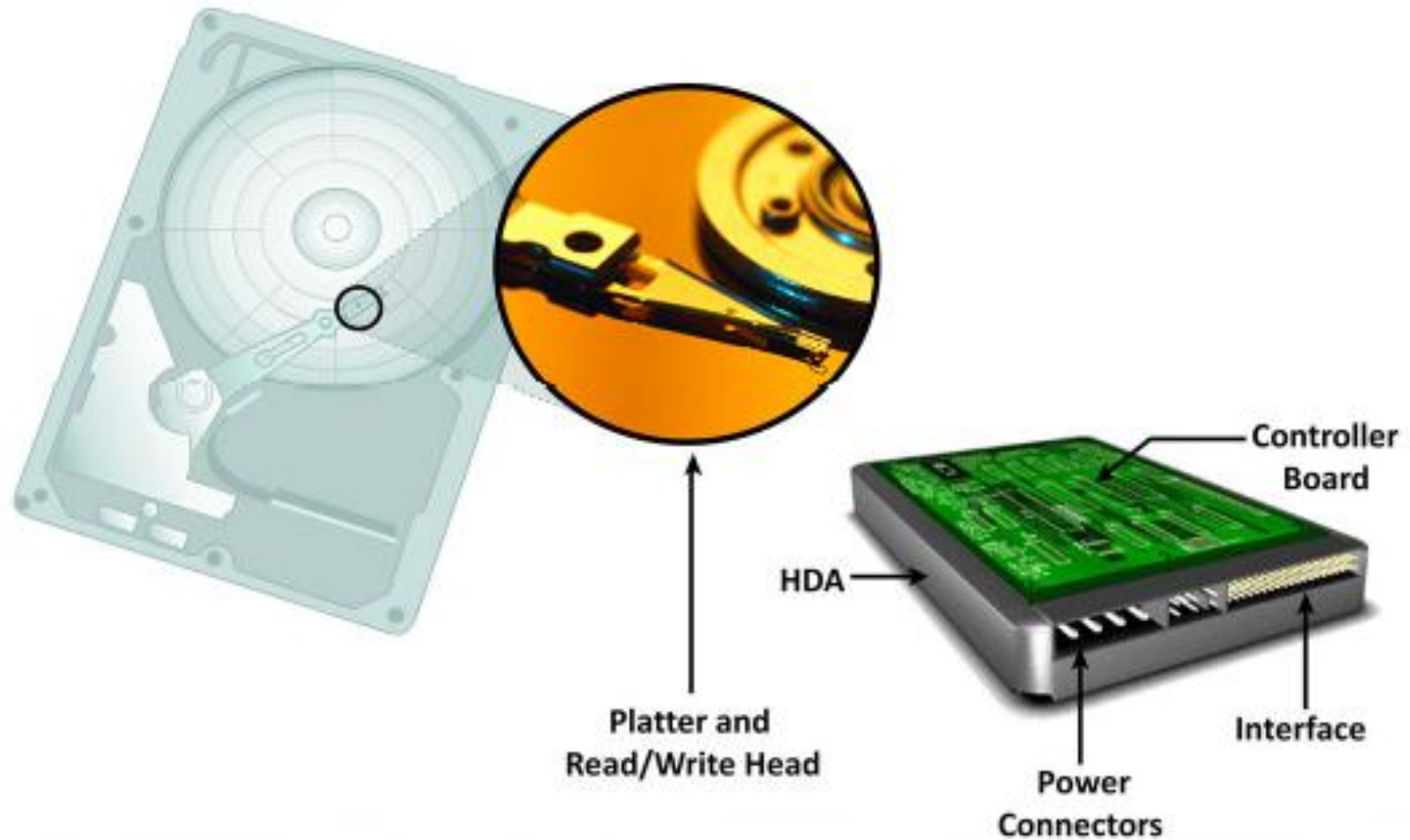
# Дисковые устройства (винчестеры, твердые диски, HDD)

- Большой объем
- Быстрота
- Высокая надежность
- Относительная дешевизна
- Нет лимитов на запись и чтение

# Жесткий диск



# Компоненты дискового устройства





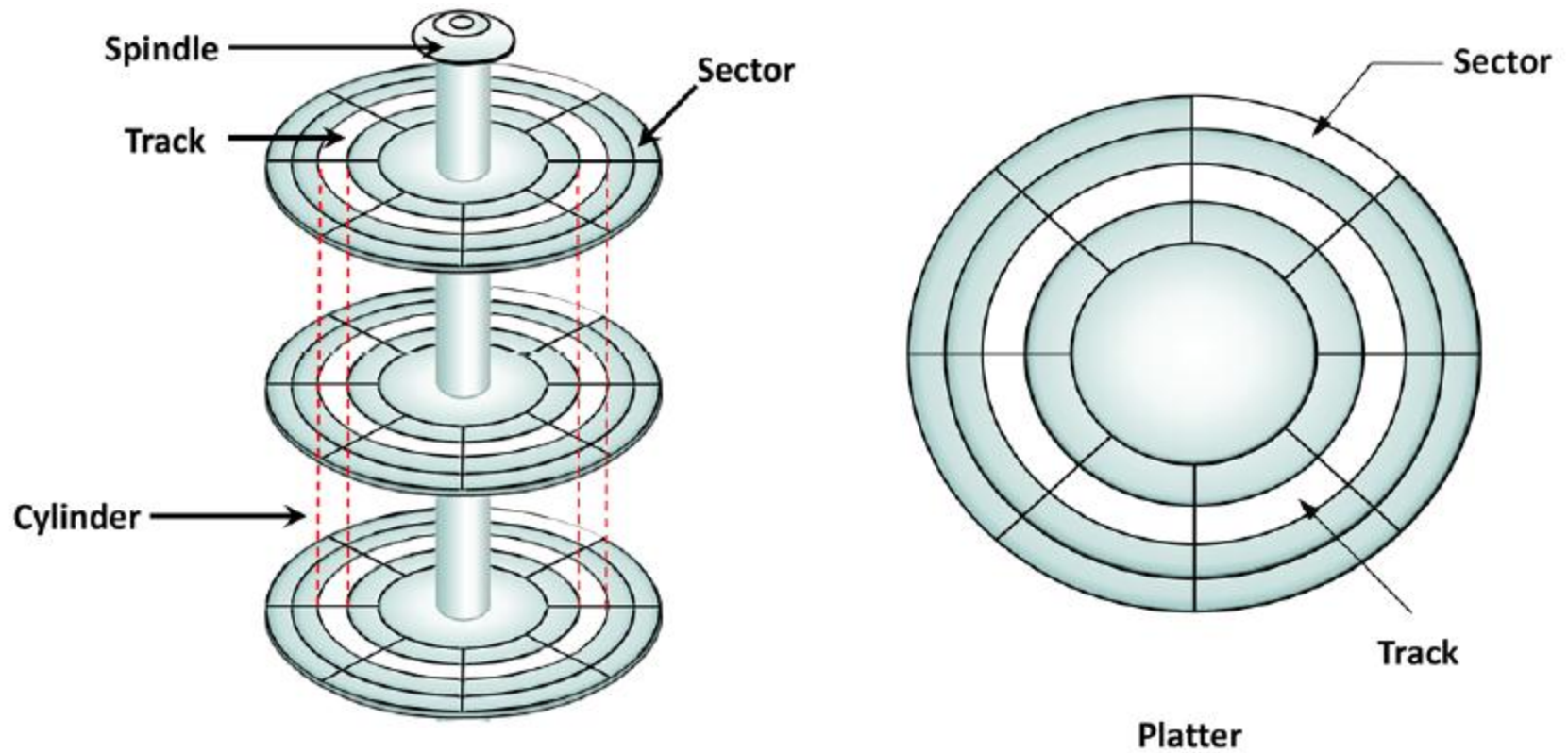
# Компоненты дискового устройства (2)

- Магнитные диски
  - двухсторонние
- Шпиндель
  - постоянная скорость вращения (7200, 10000, 15000)
  - до  $\frac{1}{4}$  скорости звука
- Головка чтения/записи
  - на каждую сторону диска
  - зазор над поверхностью диска во время вращения
  - зона парковки

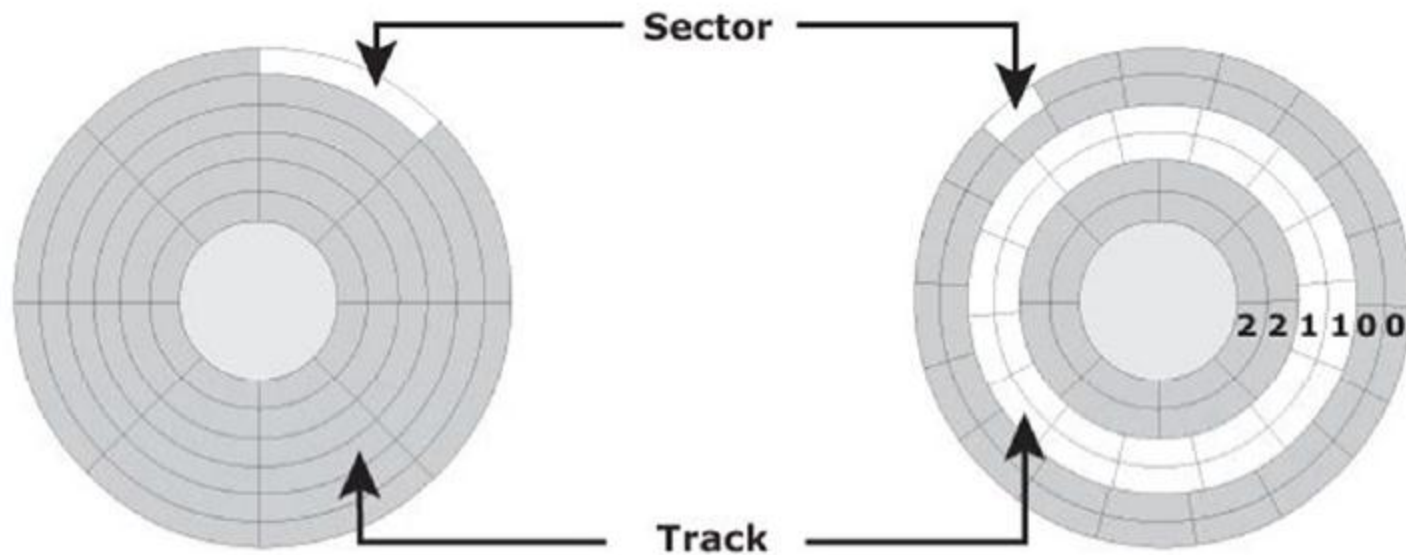
# Компоненты дискового устройства (3)

- Рычаг привода
  - Перемещает головки чтения/записи по диску
- Контроллер
  - микропроцессор, буфер, программы
  - поддерживает постоянную частоту вращения дисков
  - выполняет операции чтения/записи (интерфейс хоста к дискам)

# Структура жесткого диска

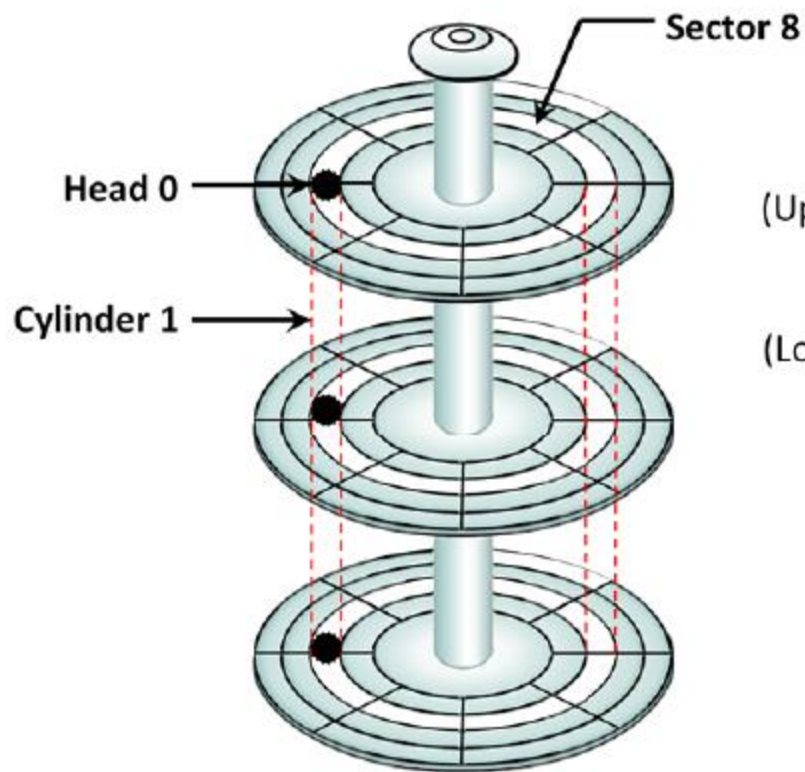


# Зонирование диска

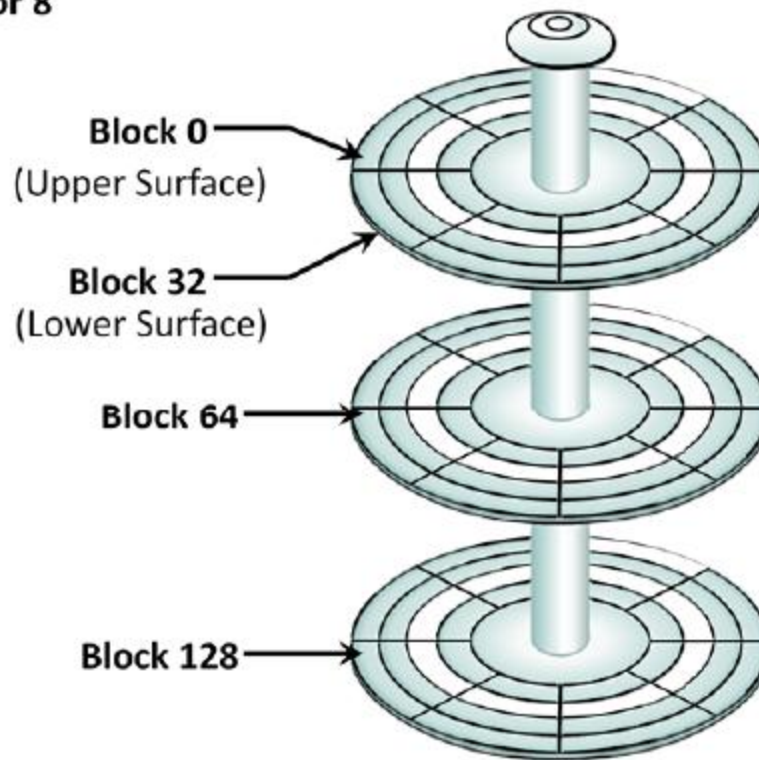


Q: где лучше разместить область подкачки страниц (swap)?

# Логическая адресация блоков диска



**Physical Address= CHS**



**Logical Block Address= Block#**

# Производительность жесткого диска

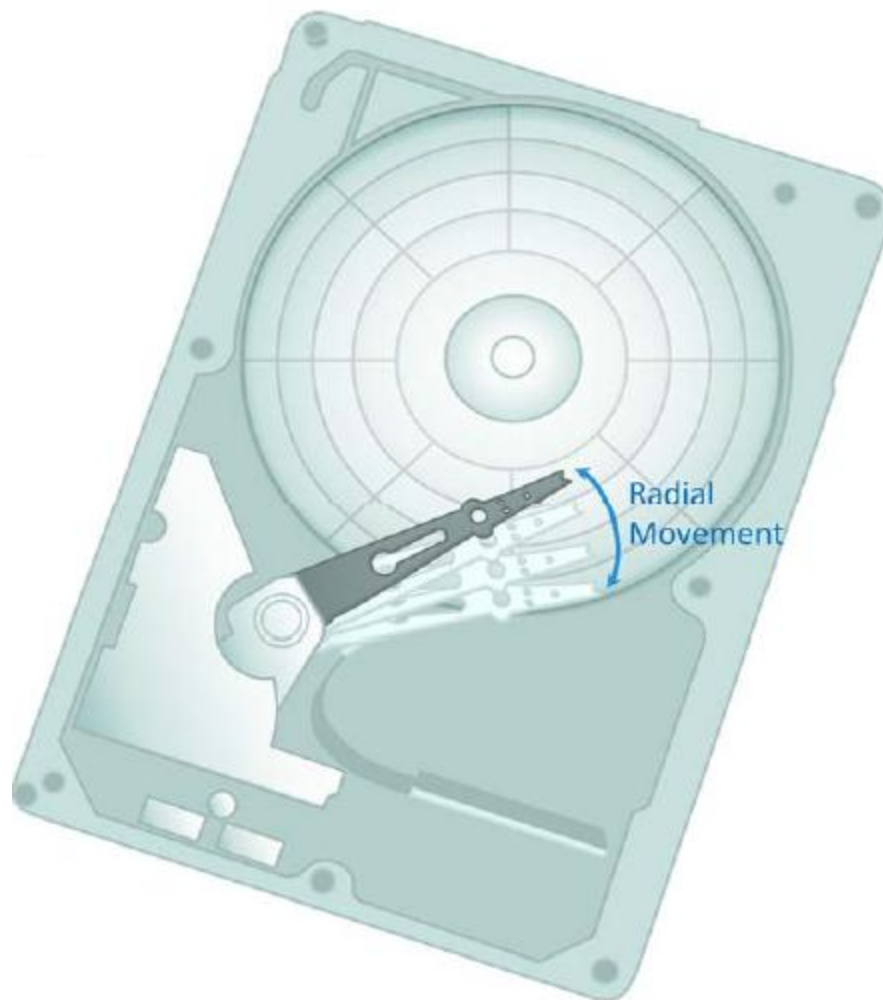
- Время обслуживания запроса диском  $R_s$  - время, необходимое диску на выполнение запроса чтения/записи

$$\begin{aligned} &= \\ &\text{время поиска} \\ &+ \\ &\text{время ожидания} \\ &+ \\ &\text{время передачи} \end{aligned}$$

# Время поиска (время доступа)

- Время размещения головок чтения/записи на нужном треке:
  - Максимальное время позиционирования = от самого края к самому центру
  - Среднее время =  $1/3$  от максимального
  - Время перехода между дорожками
- Среднее время у современных дисков от 3 до 15 ms
- Можно повысить производительность используя только внешние цилиндры

# Время поиска - иллюстрация

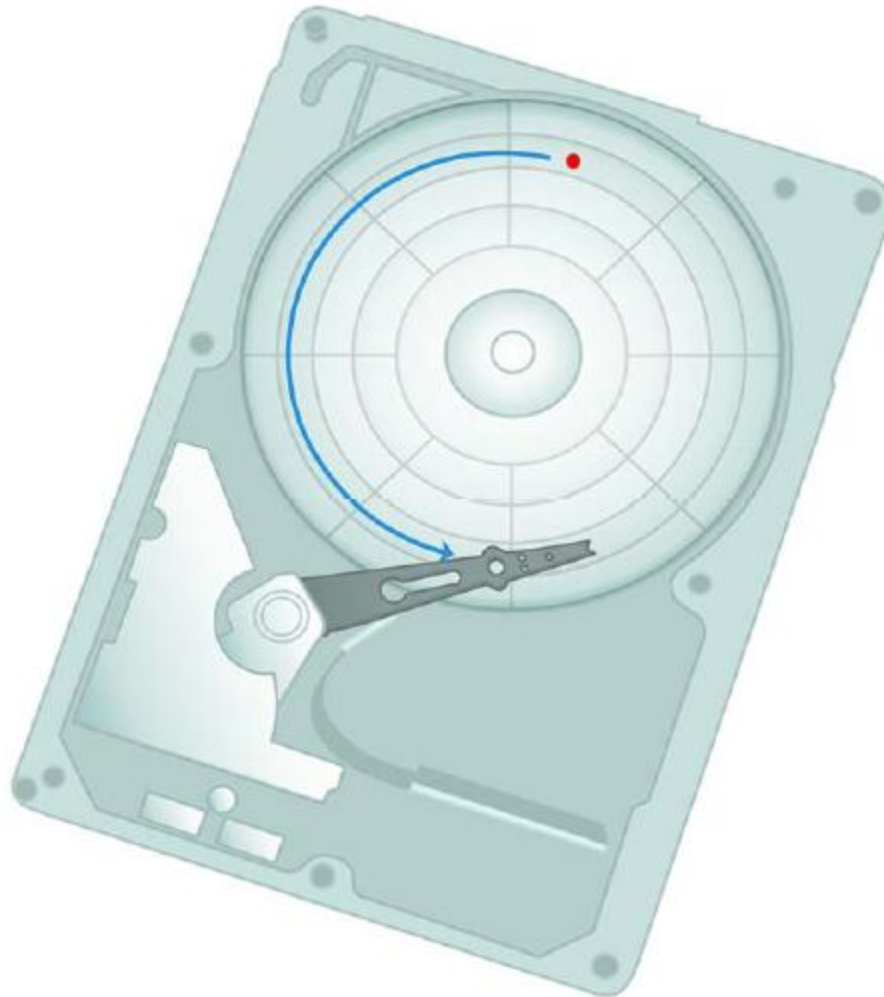




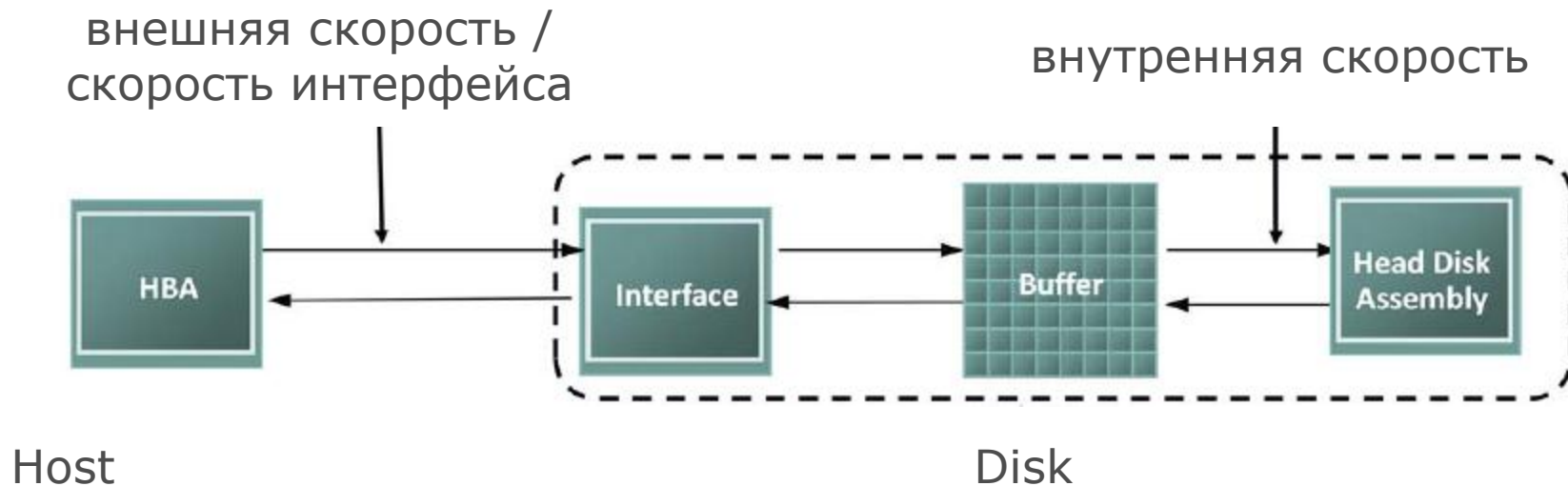
# Время ожидания

- Время, затрачиваемое на вращение
- Среднее время = время вращения диска на половину оборота, от 5 до 2 ms

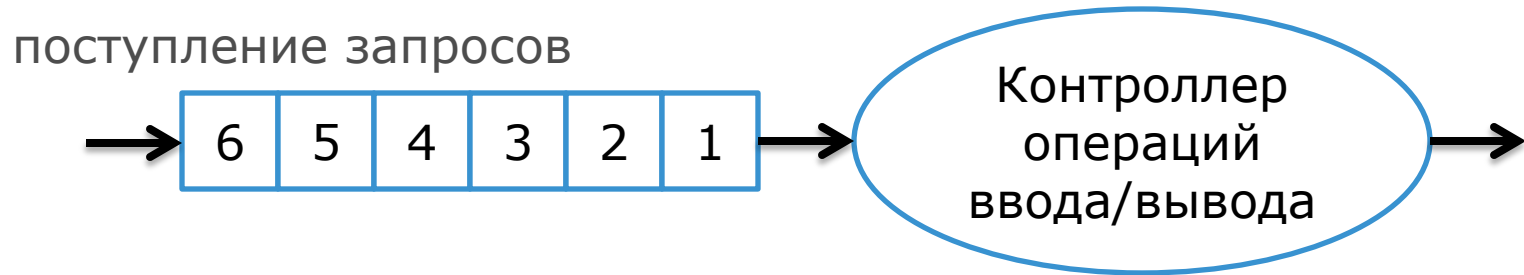
# Время ожидания - иллюстрация



# Скорость / время передачи данных



# Очереди запросов



Закон Литтла:

$$N = a \times R \quad (1)$$

где  $N$  = общее число запросов в системе  
 $a$  = частота поступления запросов  
 $R$  = среднее время отклика (от прибытия запроса до отправки из системы)

# Коэффициент загрузки жесткого диска

$$U = a \times R_s \quad (2)$$

где  $R_s$  = время обслуживания одного запроса контроллером,  $1/R_s$  = скорость обслуживания

Среднее время между запросами:

$$R_a = 1/a \quad (3)$$

# Коэффициент загрузки и время отклика

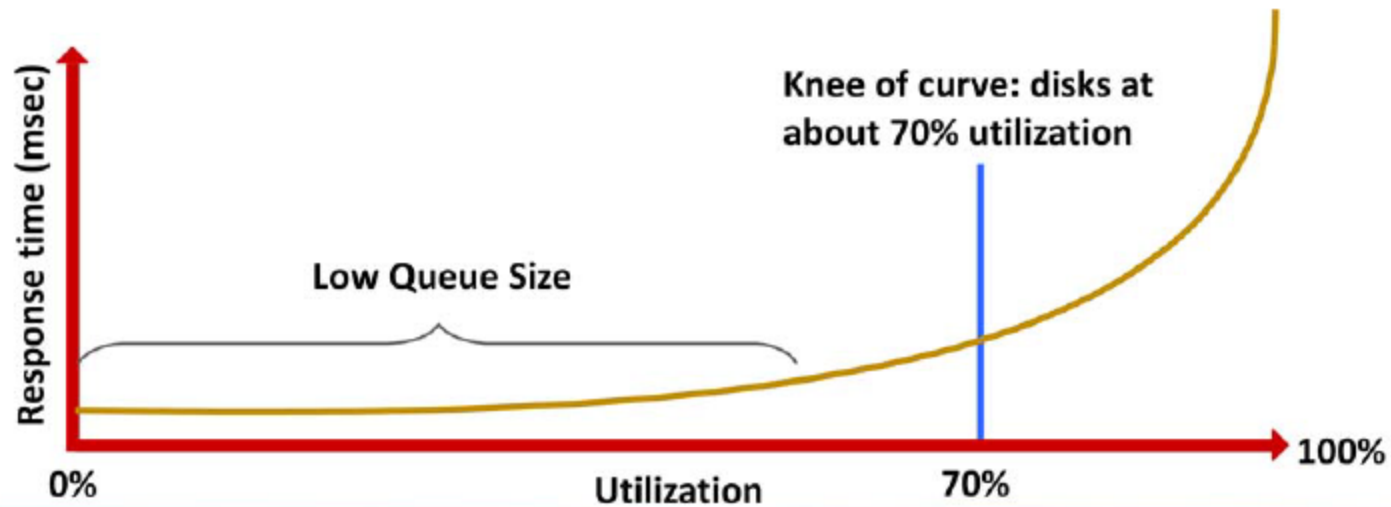
$$U = R_s / R_a \quad (4)$$

Скорость отклика  $S$  = скорость обслуживания – скорость поступления

$$R = 1 / S = 1 / (1 / R_s - 1 / R_a) =$$
$$R_s / (1 - a \times R_s)$$

$$R = R_s / (1 - U) \quad (5)$$

# Время отклика и предел реальной производительности диска

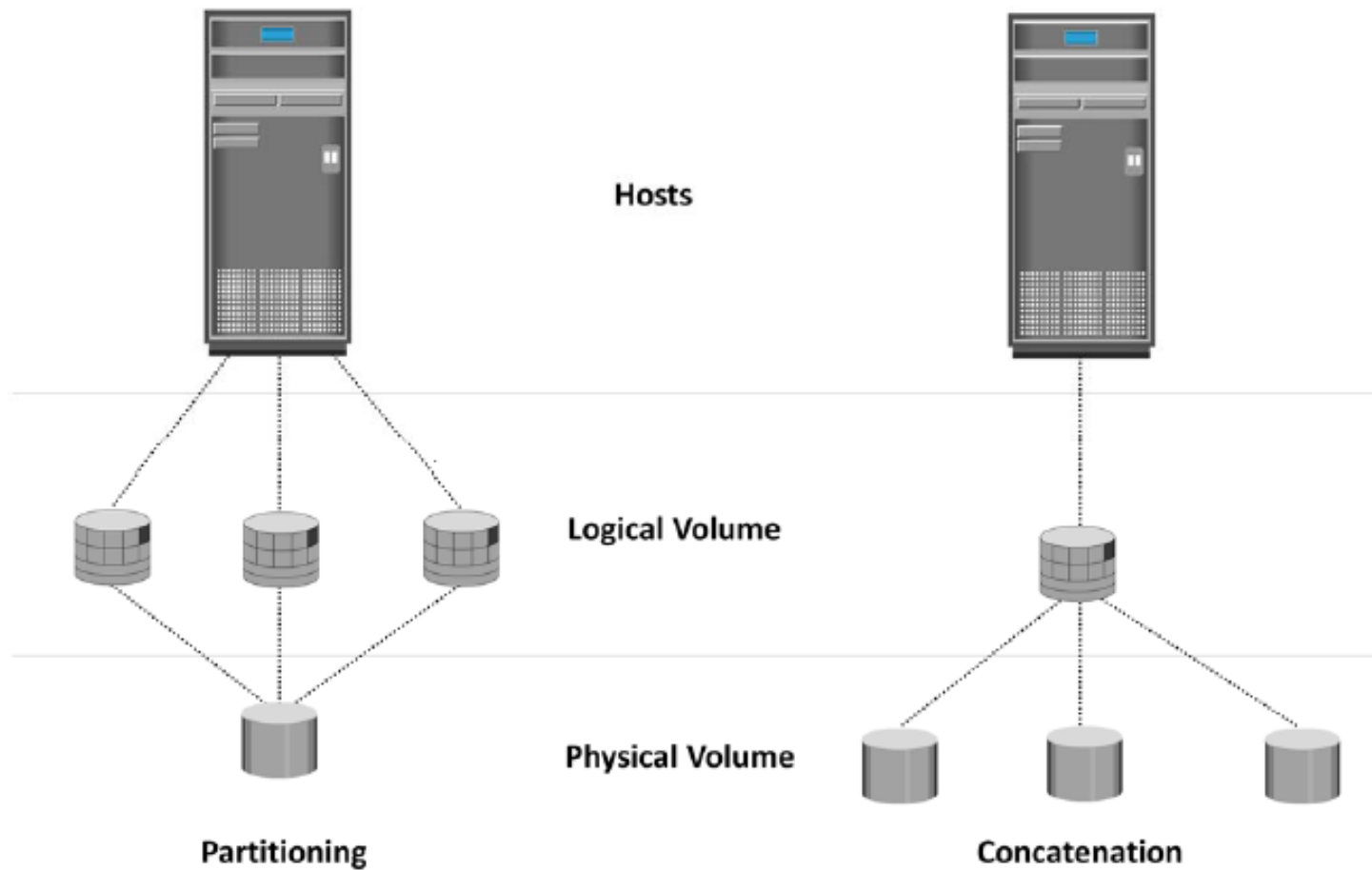


# Логические компоненты хоста

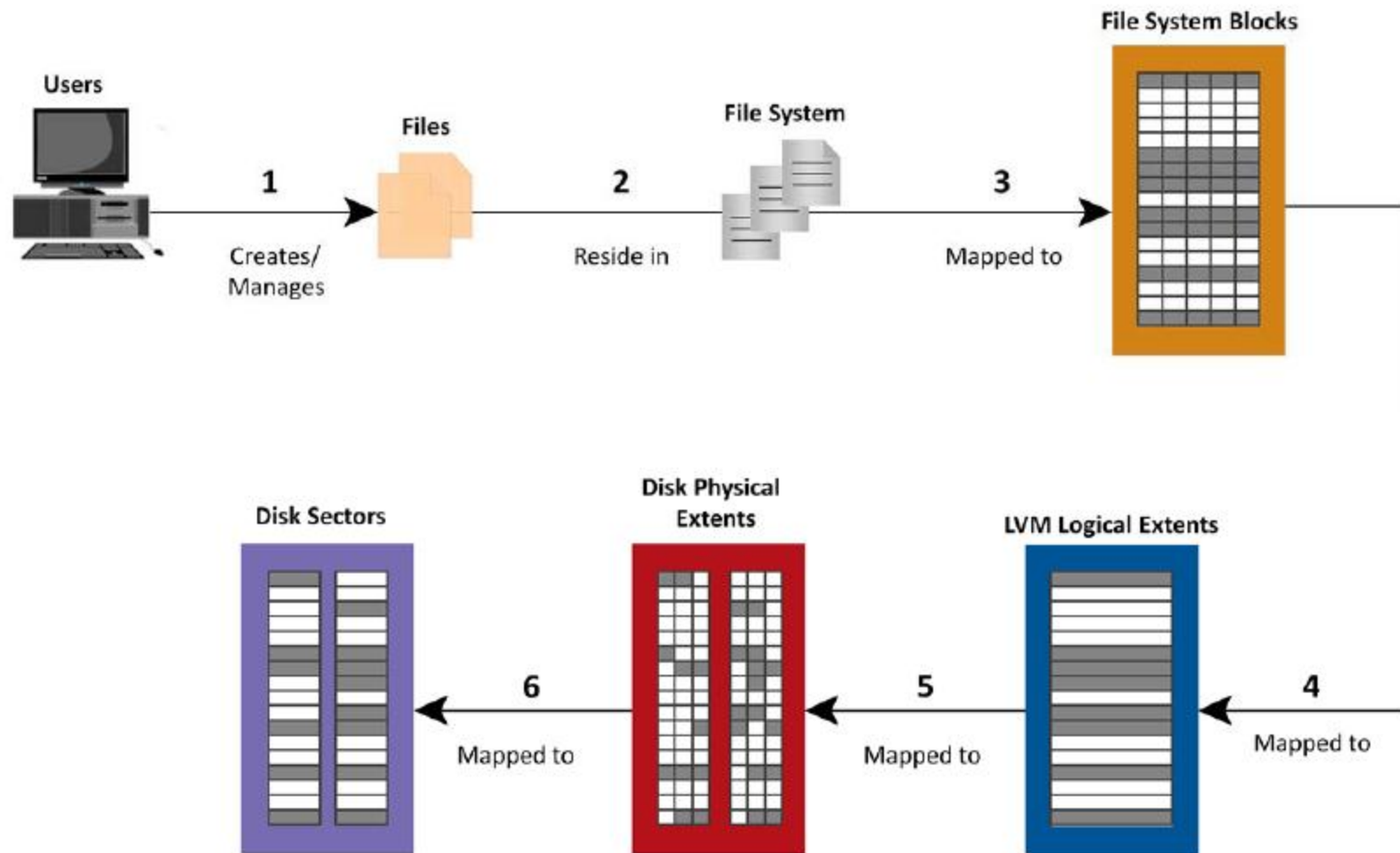
- Операционная система
- Драйверы устройств
- Менеджер логических томов (LVM)
- Файловая система
- Приложение



# Менеджер логических томов (LVM)



# Файловая система



# Формы доступа к данным

- Блочный – указание адреса логического блока при чтении/записи
- Файловый – указание имени и пути к файлу
- Объектный – по уникальному идентификатору

# Расчет числа дисков по требованиям к производительности и вместимости

- Расчет вместимости:

Кол-во дисков =  $\text{ceil} \left( \frac{\text{требуемый объем}}{\text{полезный объем одного диска}} \right)$

- Расчет IOPS:

Кол-во дисков =  $\text{ceil} \left( \frac{\text{IOPS на пике использования приложения}}{\text{IOPS диска}} \right)$

$$\text{IOPS} = 1 / R_s = 1 / (\text{время доступа} + \text{время ожидания} + \text{время передачи})$$

## Расчет числа дисков - итог

- Следует уменьшить IOPS диска до реального:  
$$\text{IOPS} = 0.7 \times \text{IOPS}$$
- Итоговое количество дисков =  
$$\text{MAX (расчет по объему, расчет по IOPS)}$$

# Твердотельные накопители

- NOR

- побайтовый произвольный параллельный доступ
- большие габариты
- дорог

- NAND

- доступ по блокам
- медленнее чем NOR
- нужно контролировать плохие сектора
- большой объем данных



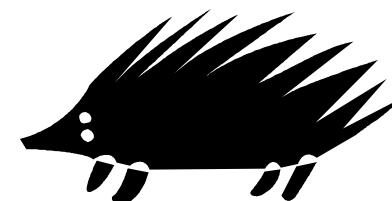
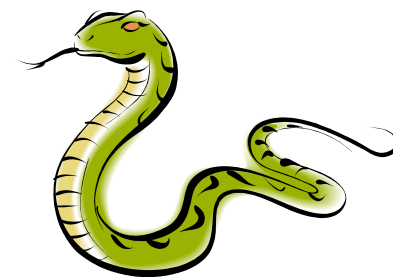
# По сравнению с жесткими дисками

- Промышленные SSD
  - громадные значения IOPS ( $>1\ 000\ 000$ )
  - малое время задержки ( $<1\ \text{ms}$ )
  - малое потребление энергии
  - высокая надежность
  - более выгодны



# Гибридные диски

- SSD + HDD
- Управление иерархией памяти
- Дешевы и удобны для некоммерческого использования
- Меньше энергозатраты
- Отличия от HDD:
  - Магнитные диски основное время «спят»
  - «Просыпание» происходит при чтении недостающих данных в SSD или переполнении SSD на запись (сброс данных из буфера)





# Недостатки гибридных дисков

- Очень медленное чтение, если данные не в SSD (жесткий диск «просыпается»)
- SSD плох для записи множества мелких файлов
- Менее долговечны
- Сложные драйверы, плохая поддержка среди операционных систем (работает в основном в MS Windows)

## Виртуализация

- Приложений
- Серверов
- Сетей
- Систем хранения данных



# Что такое виртуализация?

- Технология, абстрагирующая физические ресурсы вычислительных систем от программ, пользователей или других системам
- Метод получить больше из меньшего
- Позволяет видеть несколько небольших ресурсов как одно большое и наоборот



# Типы виртуализации - Серверная

- Серверная виртуализация
  - Часто называется просто виртуализацией
  - Скрывает физические ресурсы вычислительной системы от выполняемых ей программ
  - Может быть запущена в операционной системе или установлена на «голое» железо

# Типы виртуализации – ОС Контейнеры

- Виртуализация ОС – запускается поверх существующей операционной системы и представляет из себя независимые изолированные копии ОС (контейнеры), работающие одновременно

# Типы виртуализации – СХД

- Виртуализация хранилищ
  - Позволяет комбинировать и делить ресурсы для сетевого хранения данных в логические элементы, воспринимаемые хостами как настоящие физические устройства
  - В отличие от физических дисков структура и размер логических устройств можно менять «на лету»
  - Может работать с файлами и блоками (абстракция NAS или SAN)

# Типы виртуализации – Приложения

- Виртуализация приложений позволяет запускать программу в виртуальной машине на любой платформе и в любой ОС (например, Java VM).

# Типы виртуализации – Сети

- Виртуализация сети
  - Распределяет пропускную способность сети по независимым виртуальным каналам связи, которые могут быть выделены серверам в режиме реального времени
  - Абстрагирует адреса систем (например, виртуальных машин) от низлежащего оборудования
- VLAN, SDN, ..



# Спасибо!

EMC<sup>2</sup>®