Архитектура вычислительных систем

Лекция 4. Слой хранения
Часть 1

Artem Beresnev

t.me/ITSMDao

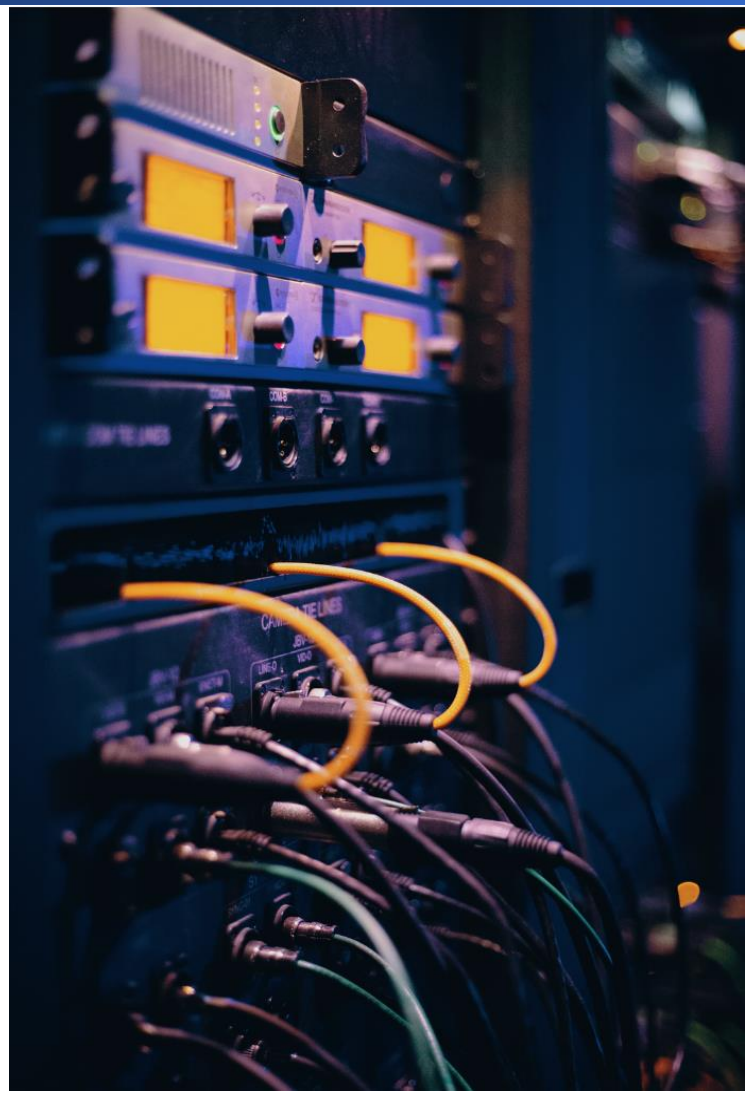
t.me/ITSMDaoChat

План

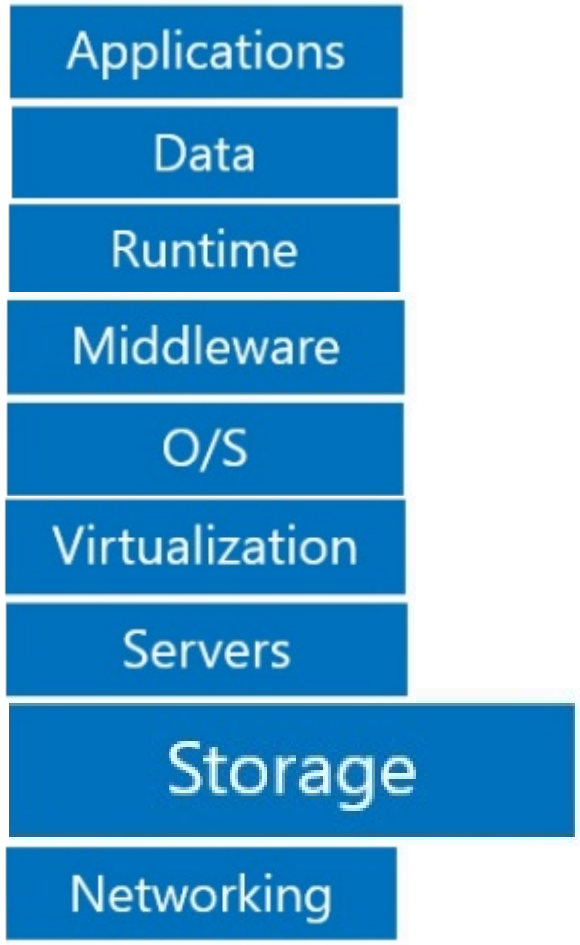
- Вспомним про слой Storage, обсудим его задачи
- Аппаратное обеспечение хранения
  - Диски
  - Интерфейсы
- Как мерить скорость подсистем

Слои ИТ-инфраструктуры

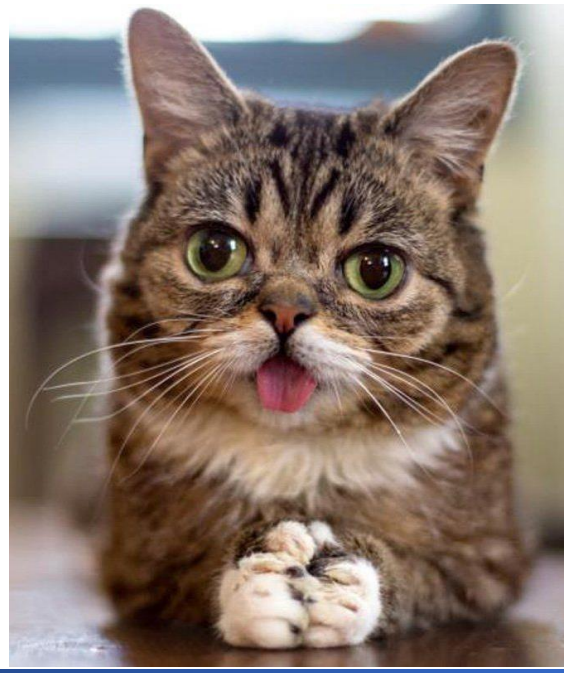
Что там было про хранение?



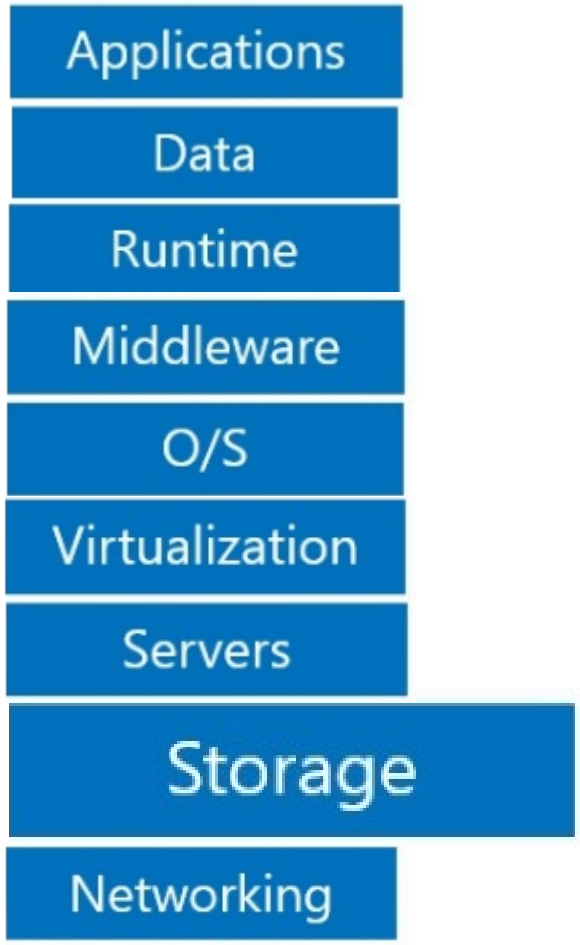
Слой Storage



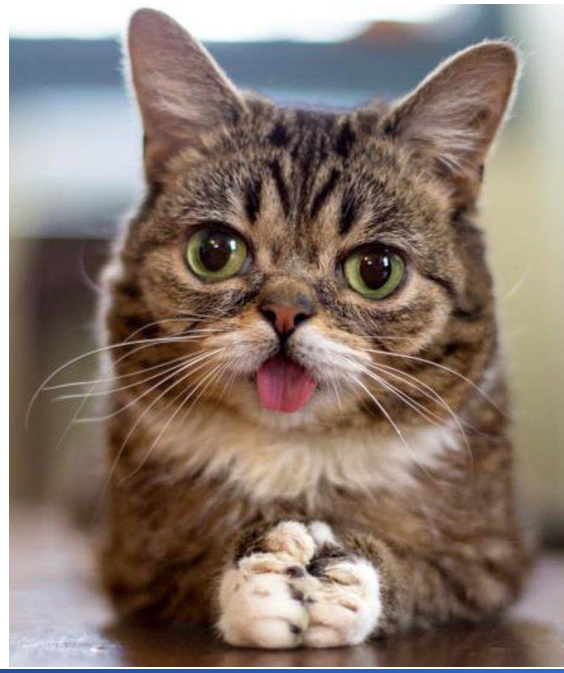
Storage – инфраструктура хранения данных.
Устройства хранения, DAS/SAS/NAS/SDN.
Абстрагированные, в том числе и облачные, сервисы хранения данных.
Этот слой управляет хранением данных в различных форматах и обеспечивает доступ к данным для различных приложений и служб.



Слой Storage. Задачи

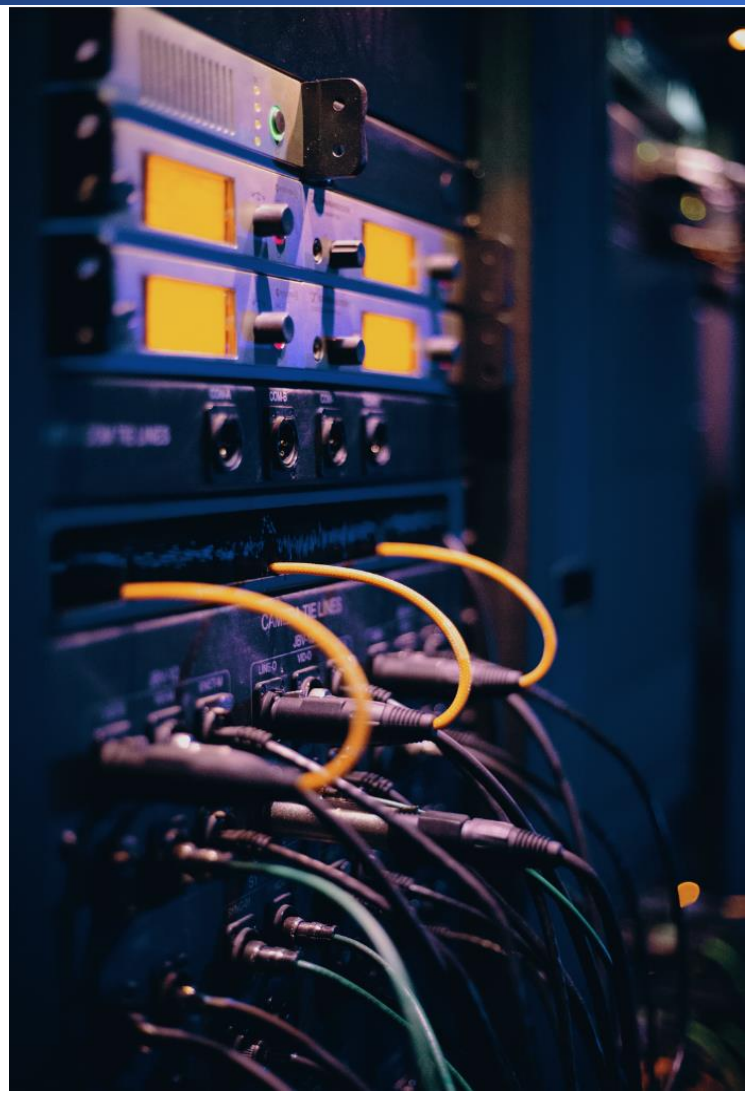


- Хранение данных
- Обеспечение:
  - надежности
  - производительности с учетом требований приложений
  - версионирования
  - масштабирование
  - безопасности
  - экономической эффективности

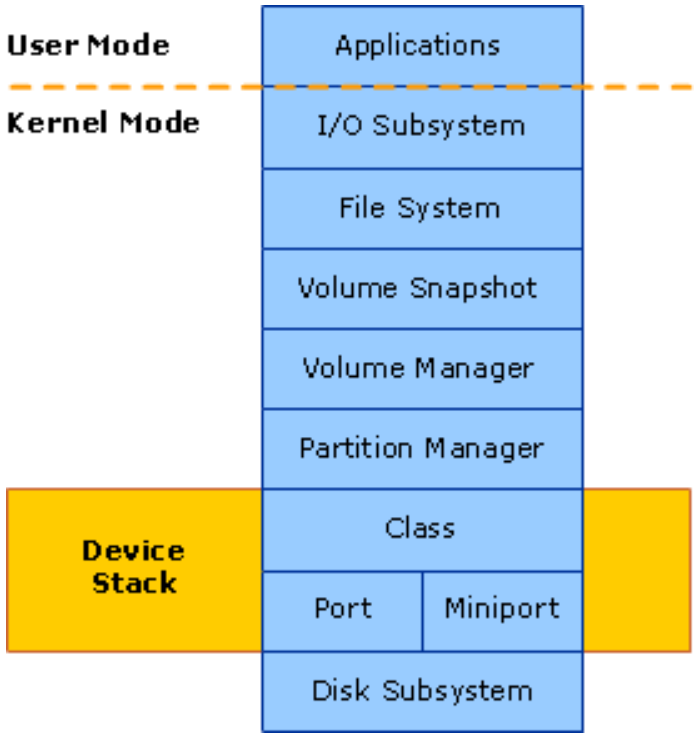


Архитектура слоя хранения

Все эти задачи не решить просто так… явно у этого слоя навороченная архитектура



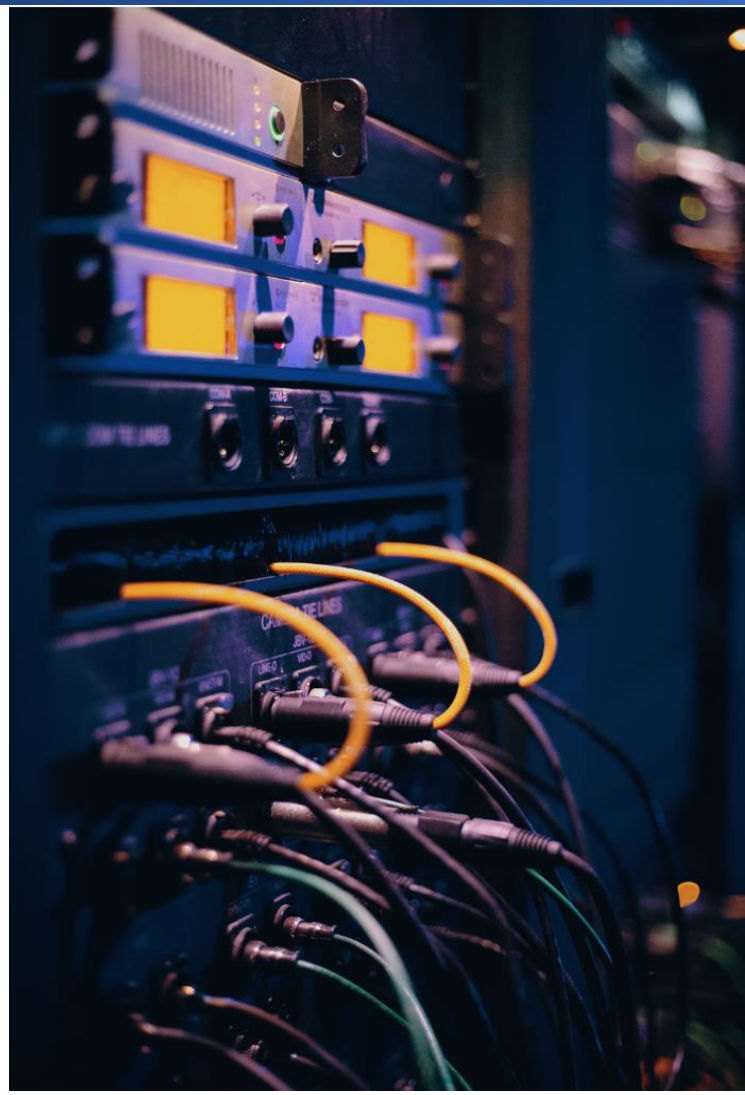
Классическая архитектура файлового хранения



- **I/O Subsystem** – интерфейс между приложением и хранением
- **File System** – механизм определяющий способ организации, хранения и именования данных. Содержит структуру и механизмы работы с именованными данными.
- **Volume Manager** – механизм, представляющий абстракцию тома.
- **Partition Manager** - управление разделами.
- **Class** – обеспечивает и унифицирует специфику работы устройств: диски, ленты, оптические носители.
- **Port** - управление «дисковыми протоколами» (SCSIport или SATAport)
- **Disk Subsystem** – аппаратное обеспечение чтения\записи

Аппаратное обеспечение хранения

Опишем

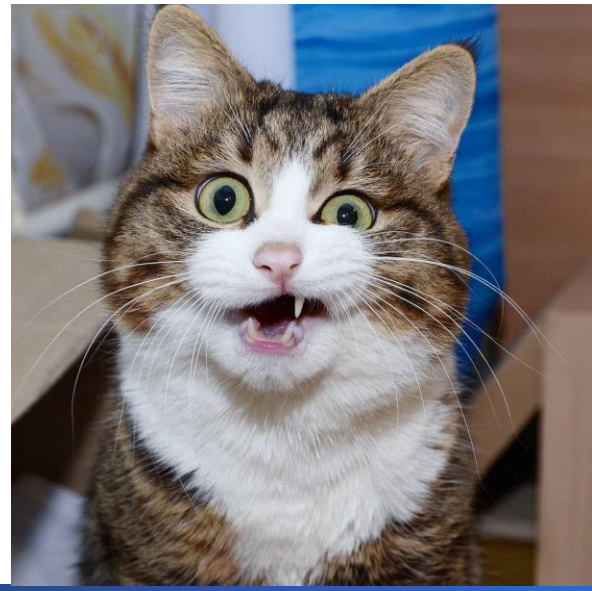


Что мы можем увидеть в аппаратных?

- Непосредственно подключенные устройства
- СХД DAS
- СХД NAS
- СХД SAN
- SDS (Software-Defined Storage)

- Облачное хранение?

"Нет облака, есть чужой компьютер"



Диски

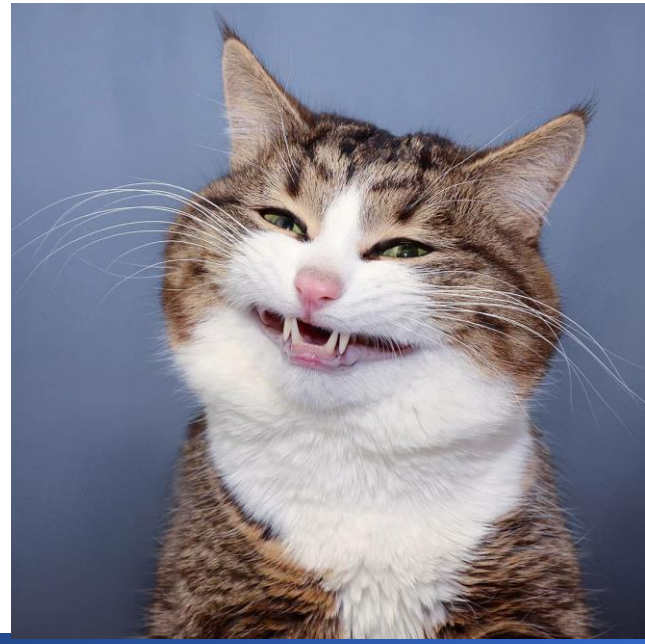
Сотрудники IBM грузят жесткий диск объемом 5 МБ, 1956 год.

Железо предназначалось для первого суперкомпьютера с жестким диском 305 RAMAC.

Весила система около тонны, — получается по 0,2 грамма за байт (или 5 килобайт в 1 кг) и состояла из 50-ти дисков диаметром в 24 дюйма (610 мм).

Устройства хранения

- Накопитель на жёстких магнитных дисках (HDD), жёсткий диск — запоминающее устройство произвольного доступа, основанное на принципе магнитной записи.
- Твердотельный накопитель (solid-state drive, SSD) — компьютерное энергонезависимое немеханическое запоминающее устройство на основе микросхем памяти.



HDD vs SSD

HDD vs SSD

Seagate ST14000DM001

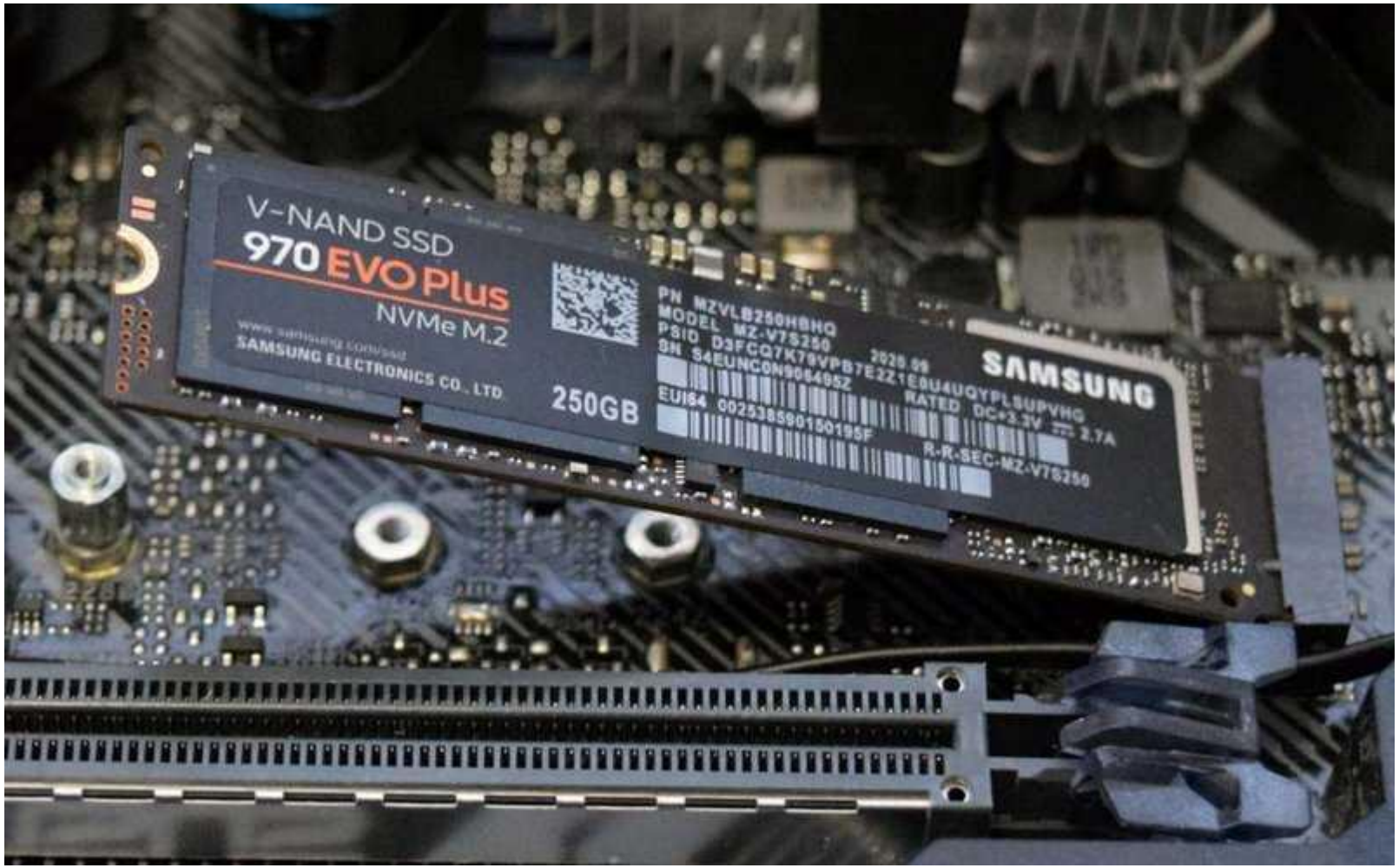
SSD 15,36 Тб компании Samsung

HDD vs SSD

INTEL Optane 900P
SSDPED1D480GASX 480Гб

NVMe (в форм-факторе M.2)

NVMe (в форм-факторе M.2)



HDD vs SSD

AWS Snowmobile

«жёсткий диск на колёсах» ёмкостью 100 петабайт
10 машин перевозять экзабайт примерно за полгода
(перекачка по 10 Gb\s каналу займет примерно 26 лет)

- Источник питания мощностью 350 киловатт (при чтении\записи),
- Защита от физического взлома,
- Шифрование,
- Система видеонаблюдения и GPS
- Вооруженная охрана

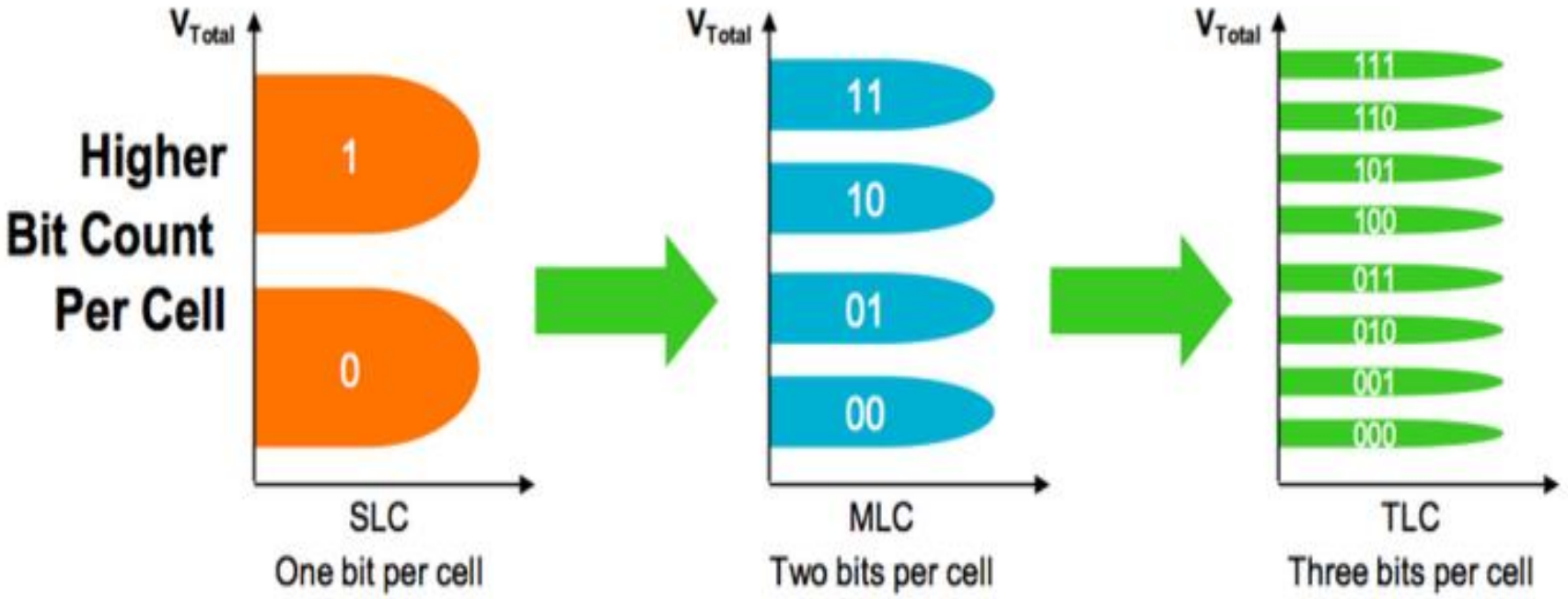
Диски HDD. Характеристики

- Форм-фактор
- Скорость вращения шпинделя
  - жесткие диски для ноутбуков имеют скорость вращения 4200, 5400 и 7200 оборотов в минуту,
  - для стационарных компьютеров 5400, 7200 и 10 000 об/мин.
  - для серверов 7200, 10 000 и 15 000 об/мин.
- Интерфейс подключения (Скоро…)
- Объем буфера (От 8 до 256 Мб)
- Наработка на отказ (MTBF)
- **IOPS** количество операций ввода-вывода в секунду.
- Уровень шума
- Ударостойкость
- Энергопотребление и тепловыделение

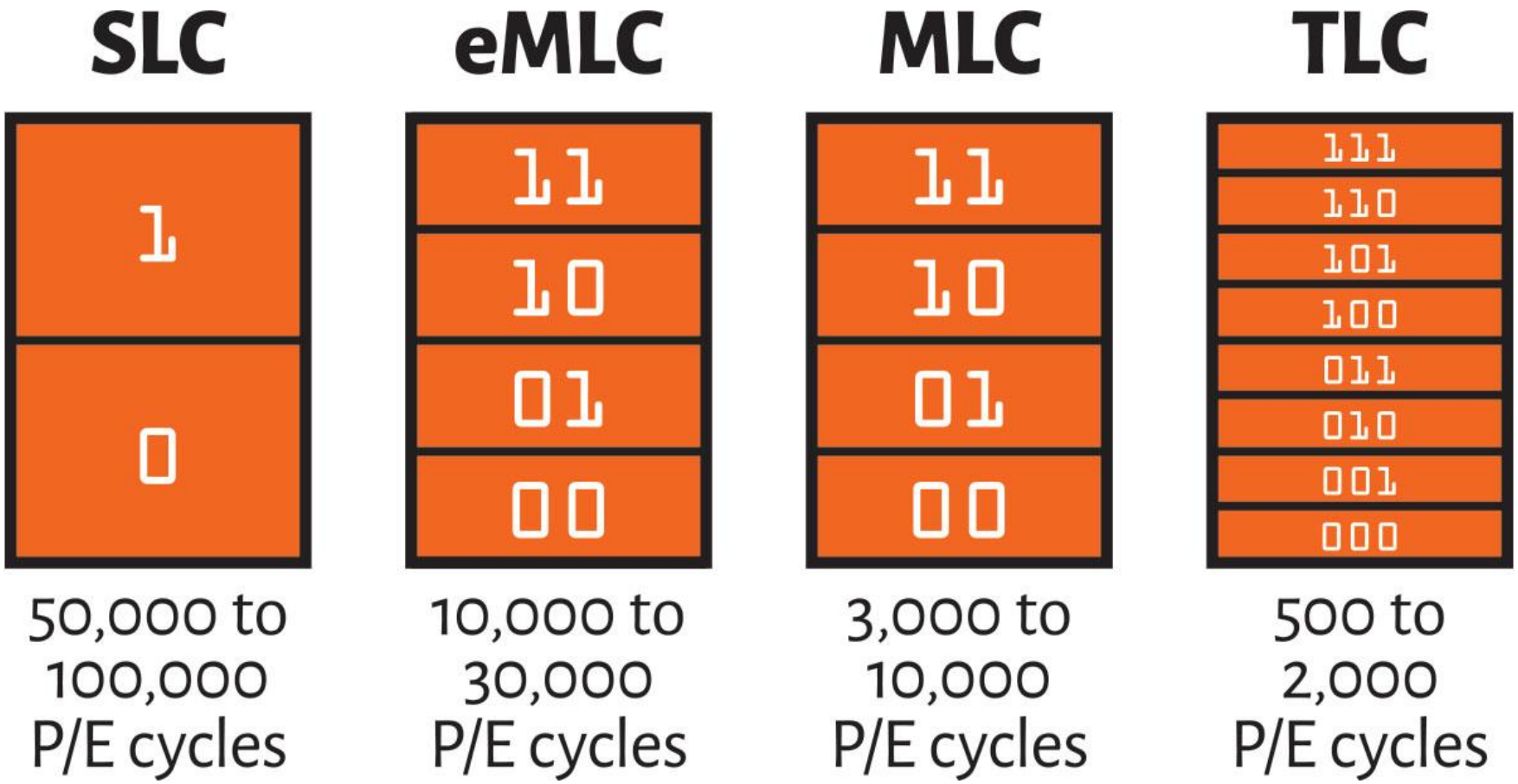
Диски SSD. Характеристики

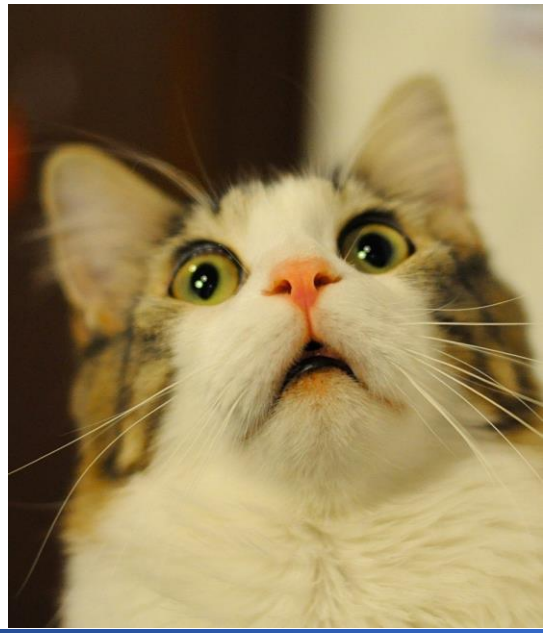
- Формфактор (2.5)
- Объем
- MTBF
- IOPS
- Энергопотребление \ тепловыделение
- Тип флеш-паямти
- **Количество циклов перезаписи**

Диски SSD. Тип флеш-паямти



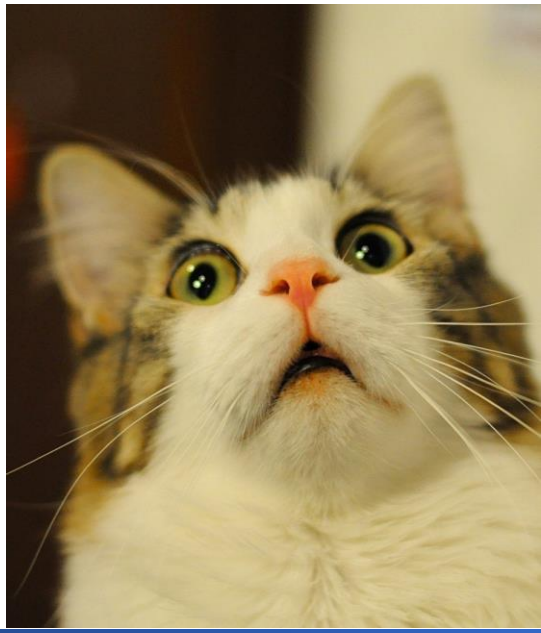
Диски SSD. Циклы перезаписи





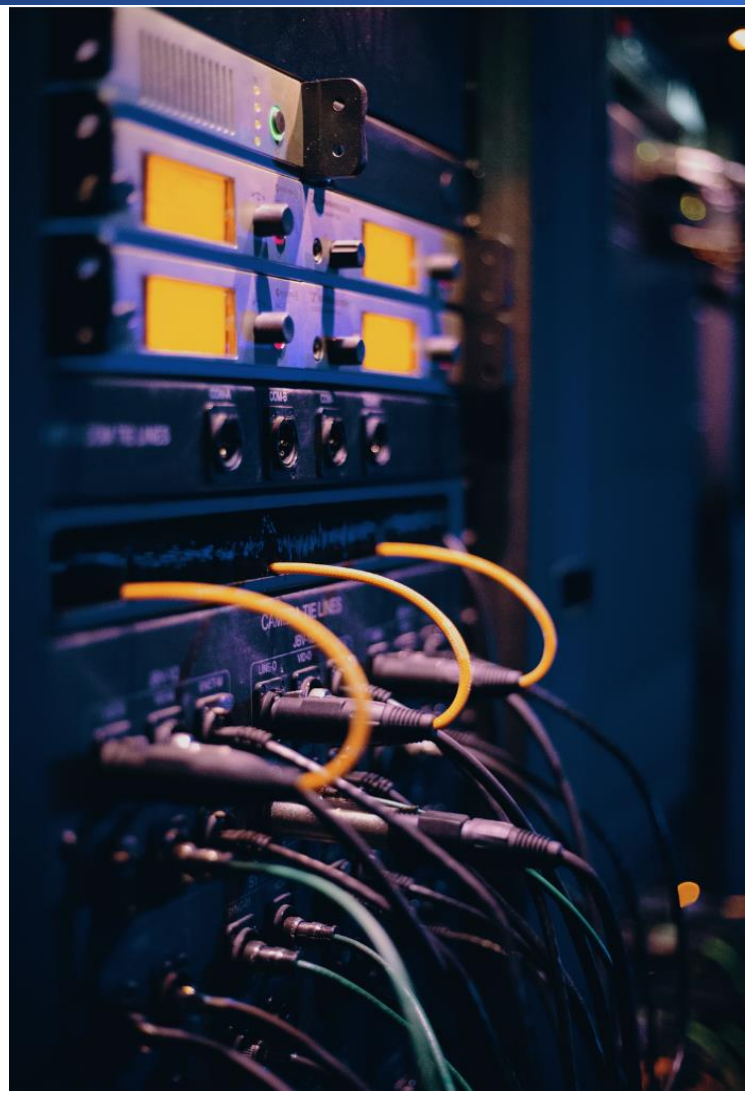
Диски SSD. Циклы перезаписи

Ресурс твердотельного накопителя		
Количество циклов перезаписи	**3 000**	**5 000**
Емкость памяти	120 Гб	120 Гб
Дневной объем записи	12 Гб	12 Гб
Увеличение объема записи	10x	10x
Циклов перезаписи в день	1 (10x12Гб)	1 (10x12Гб)
Оценка ресурса диска	**8 лет**	**13.5 лет**



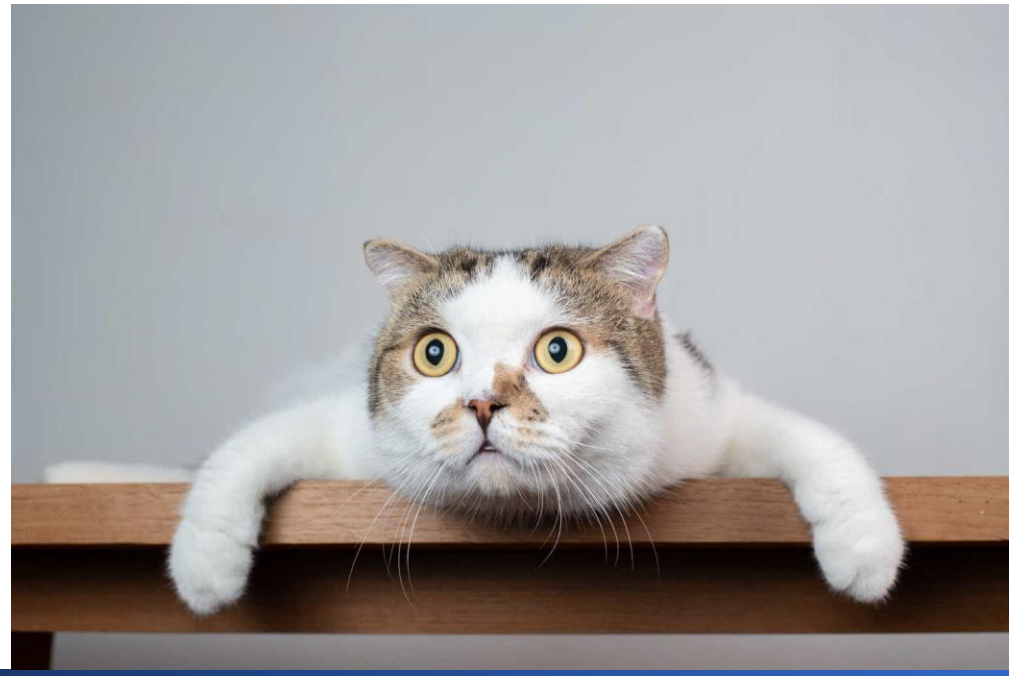
Интерфейсы подключения

Диски нужно подключать к компьютеру, опять много способов…



Интерфейсы подключения

- SATA
- SAS
- NVMe



SATA

- SATA (*Serial ATA*) — последовательный интерфейс обмена данными с накопителями информации

- **SATA Revision 1.0 - до 1,5 Гбит/с - 2003**
- **SATA Revision 2.0 - до 3 Гбит/с - 2005**
- **SATA Revision 3.0 - до 6 Гбит/с - 2008**

SATA

- последовательный интерфейс обмена данными с накопителями информации.
- SATA является развитием параллельного интерфейса ATA (IDE)
- SATA работает в полудуплексном режиме
- SATA поддерживает Hot Plug
- Работает по протоколу AHCI (Advanced Host Controller Interface)

NVMe

- NVMe 1.1b —2014
- NVMe 1.2 — 2014; NVMe 1.2a — 2015
- NVMe 1.3c — 2018
- Типичные скорости около 2000..2500 Мб/с

- для устройств потребительского класса - расширенное управление питанием и поддержка накопителей без динамической памяти;
- для устройств корпоративного класса - возможность обновления прошивки без остановки работы накопителя, снижение задержек в топологиях с большим количеством NVMe-накопителей и коммутаторами PCIe;

SATA vs NVMa

[AHCI]

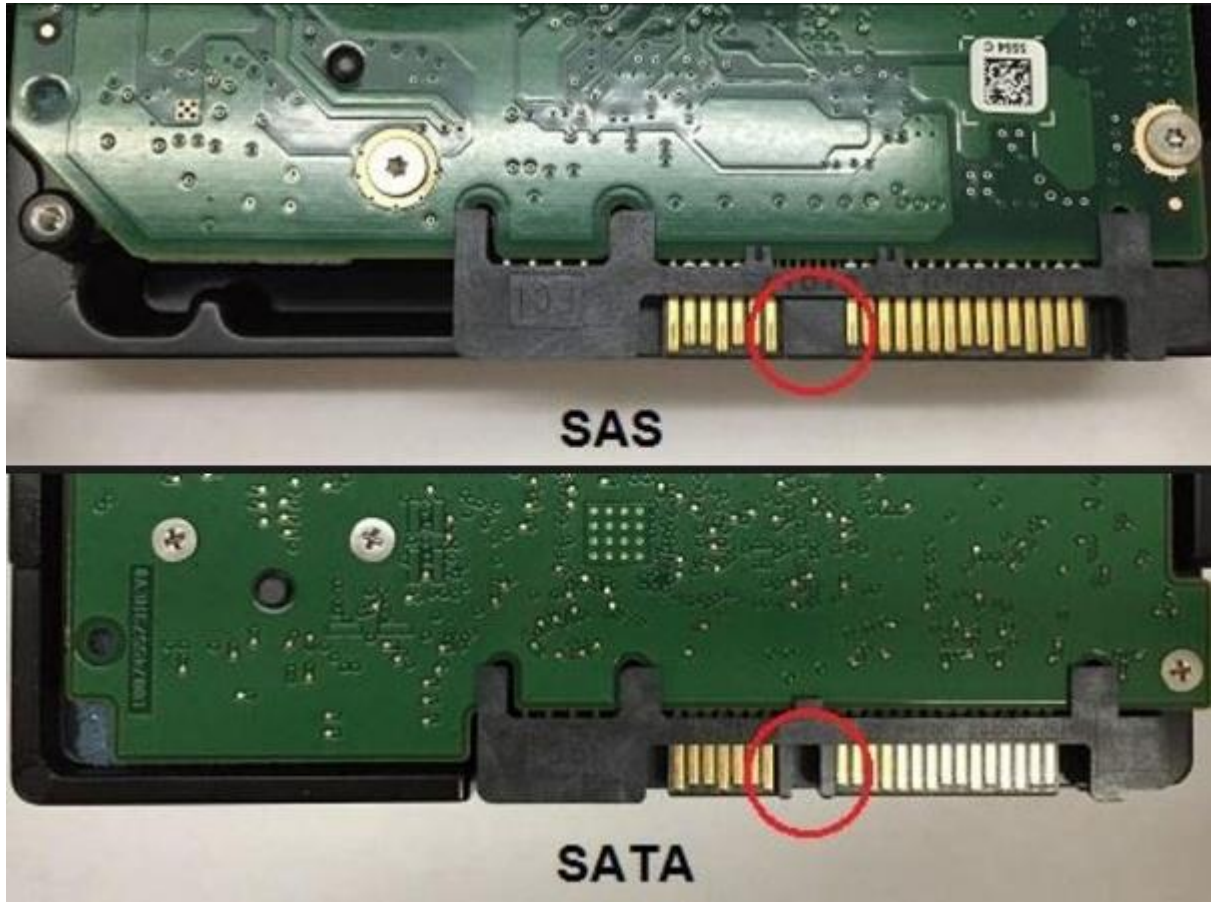
SAS (Serial Attached SCSI)

- последовательный компьютерный интерфейс, разработанный для подключения различных устройств хранения данных, например, жёстких дисков и ленточных накопителей.
- Протокол SAS обеспечивает полнодуплексную передачу данных.
- SAS разработан для замены параллельного интерфейса SCSI и основывается во многом на терминологии и наборах команд SCSI.
- Совместим с SATA
- SAS поддерживает большое количество устройств (> 16384), в то время как интерфейс SCSI поддерживает 8, 16, или 32 устройства на шине.
- SAS-1 - 3.0 Gbit/s – 2004 SAS-2 - 6.0 Gbit/s – 2009
  SAS-3 - 12.0 Gbit/s – 2013 SAS-4 - 22.5 Gbit/s (24G) - 2017

SAS

SAS контроллеры

SAS vs SATA

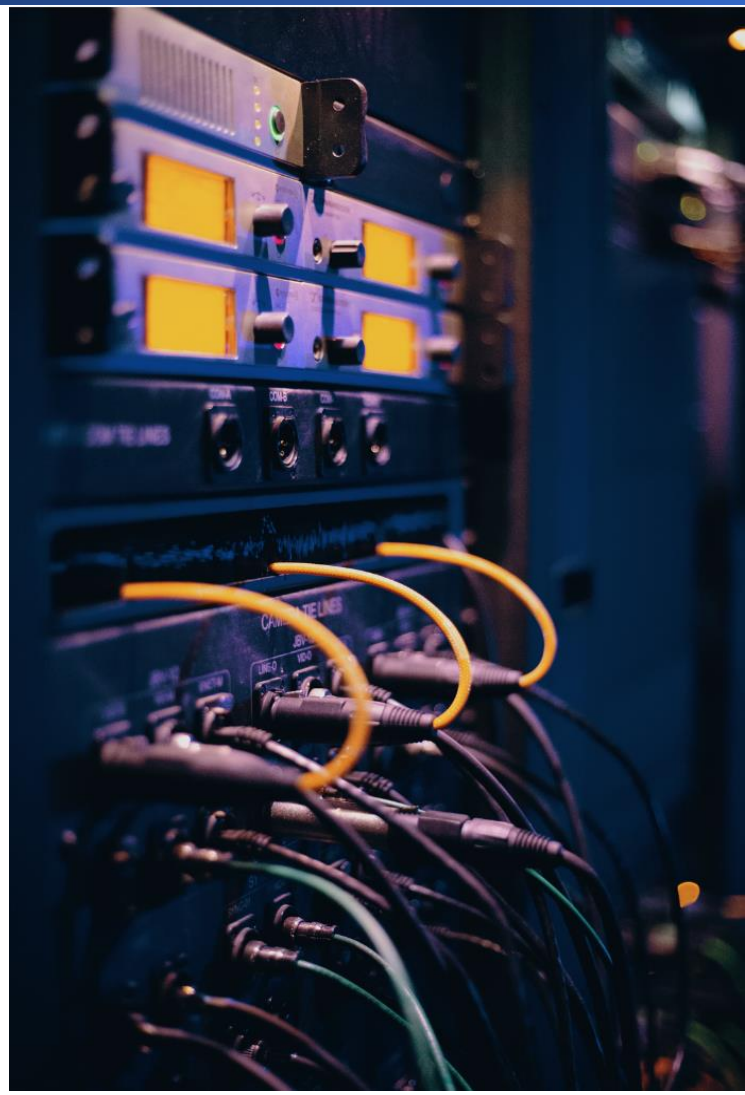


SAS (Serial Attached SCSI)

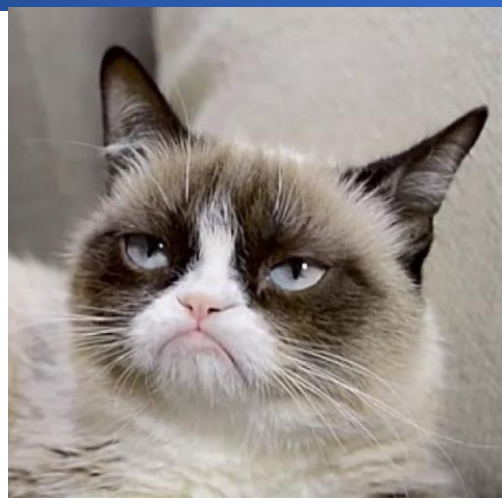
- Инициатор (**Initiator**) — устройство, которое порождает запросы на обслуживание для целевых устройств и получает подтверждения по мере исполнения запросов.
- Целевое устройство (**Targets**) содержит логические блоки и целевые порты, которые осуществляют приём запросов на обслуживание, исполняет их; после того, как закончена обработка запроса, инициатору запроса отсылается подтверждение выполнения запроса. Целевое устройство может быть как отдельным жёстким диском, так и целым дисковым массивом.
- Подсистема доставки данных (**Service Delivery Subsystem**) Является частью системы ввода-вывода, которая осуществляет передачу данных между инициаторами и целевыми устройствами. Обычно подсистема доставки данных состоит из кабелей, которые соединяют инициатор и целевое устройство. Дополнительно, кроме кабелей в состав подсистемы доставки данных могут входить расширители SAS.
- Расширители (экспандеры, **Expanders**) SAS — устройства, входящие в состав подсистемы доставки данных и позволяют облегчить передачи данных между устройствами SAS; например, расширитель позволяет подключить несколько целевых устройств SAS к одному порту инициатора. Подключение через расширитель является абсолютно прозрачным для целевых устройств.

Как измеряют производительность дисковой подсистемы?

Мерить нужно в единицах, учитывающий специфику устройств хранения



Метрики изменений



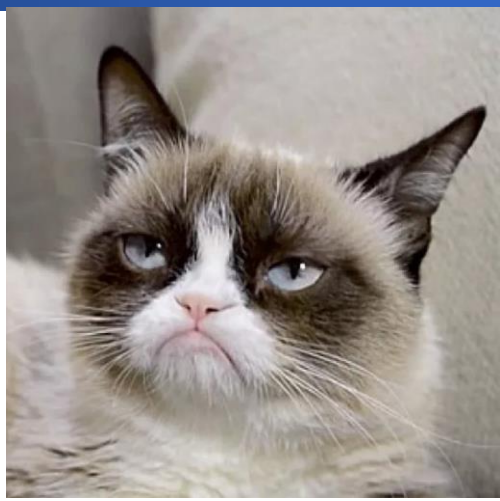
Скорости мерят в условиях:

- линейного чтения
- случайного чтения
- линейного записи
- случайной записи

Скорости мерять в:

- Мегабайты в секунду (MB/s) или Гигабайты в секунду (GB/s).
  Лучше характеризуют скорость **последовательного** чтения или записи данных. Важны для копирования больших файлов.
- IOPS (Operations Per Second, операций в секунду). Количество операций ввода-вывода, которые «диск» может выполнять за одну секунду. Особенно важно для оценки производительности в сценариях с произвольным доступом к данным, таких как работа с базами данных.
- Латентность (Latency). Это время ожидания между запросом данных и моментом их передачи. Латентность измеряется в миллисекундах (ms) или микросекундах (µs) и является критическим параметром для реакции системы на запросы ввода-вывода.

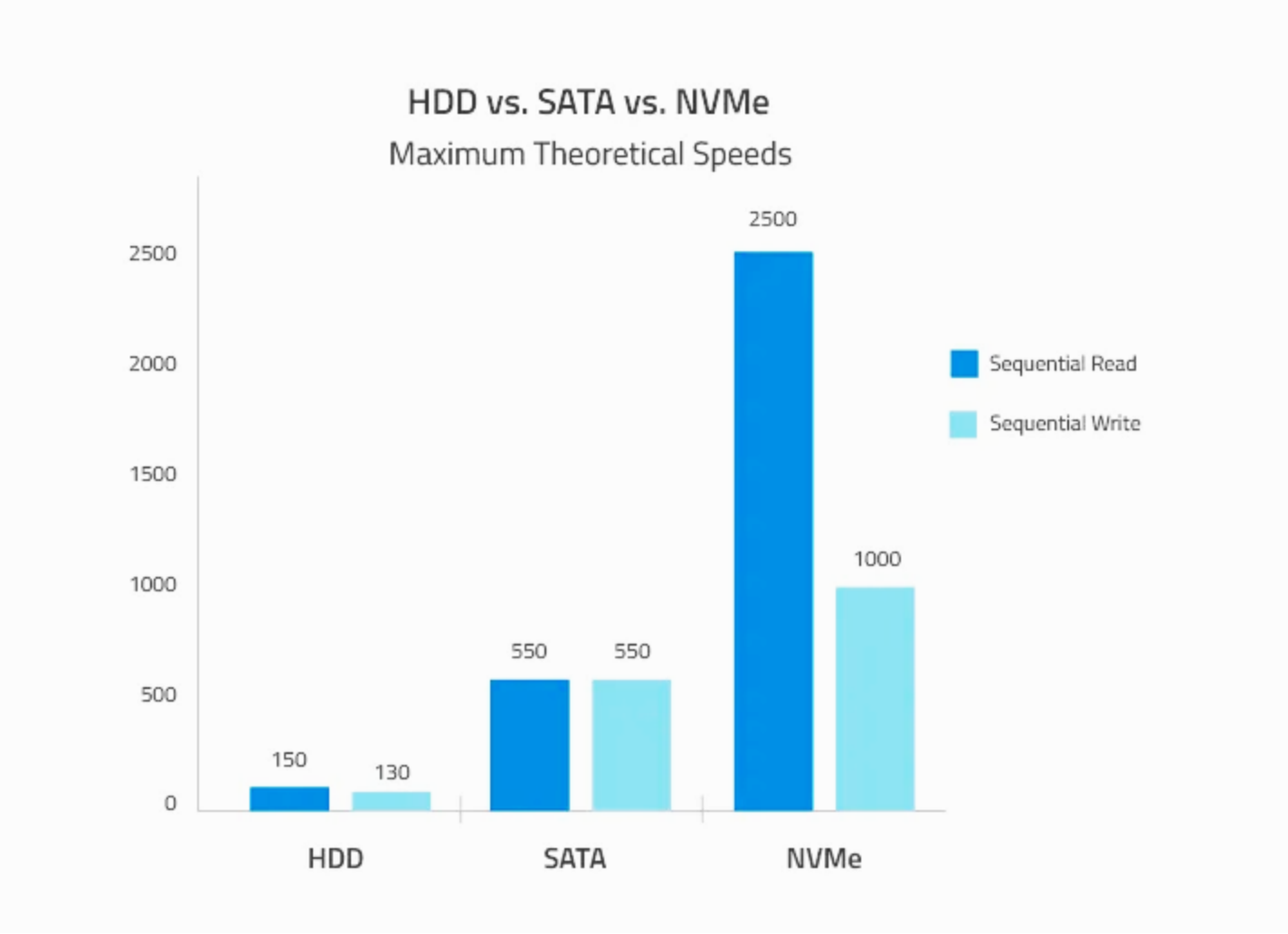
Скрытые факторы



Величина IOPS зависит от многих параметров:

- конструкция и настройки оборудования (дисков и RAID);
- устройство и настройки драйвера;
- устройство и настройки драйвера файловой системы;
- устройство и настройки операционной системы;
- условия запуска программы, выполняющей тестирование производительности (бенчмарка):
- отношение количества операций чтения к количеству операций записи;
- размеры блоков для чтения и записи при последовательном и случайном доступе;
- количество потоков, выполняющих чтение и запись;
- размеры очередей и буферов;
- наличие фрагментации в файловой системе;
- наличие приложений, работающих в фоновом режиме.

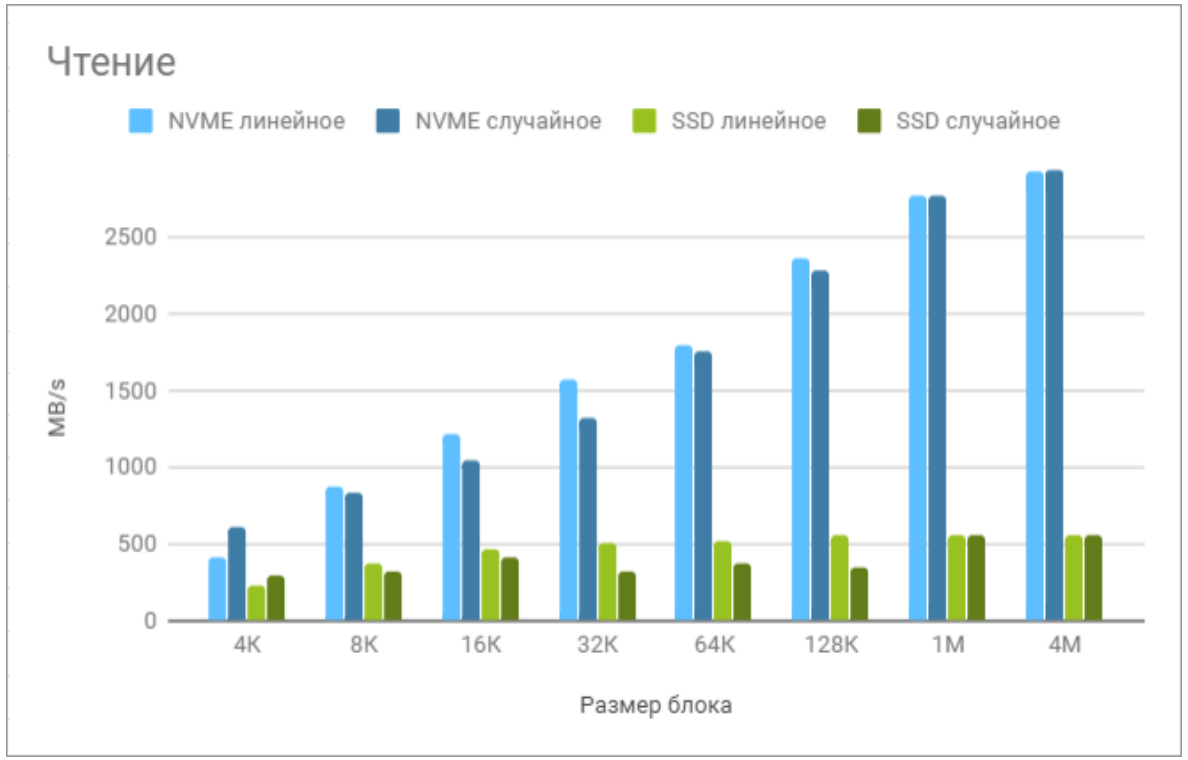
Скорости

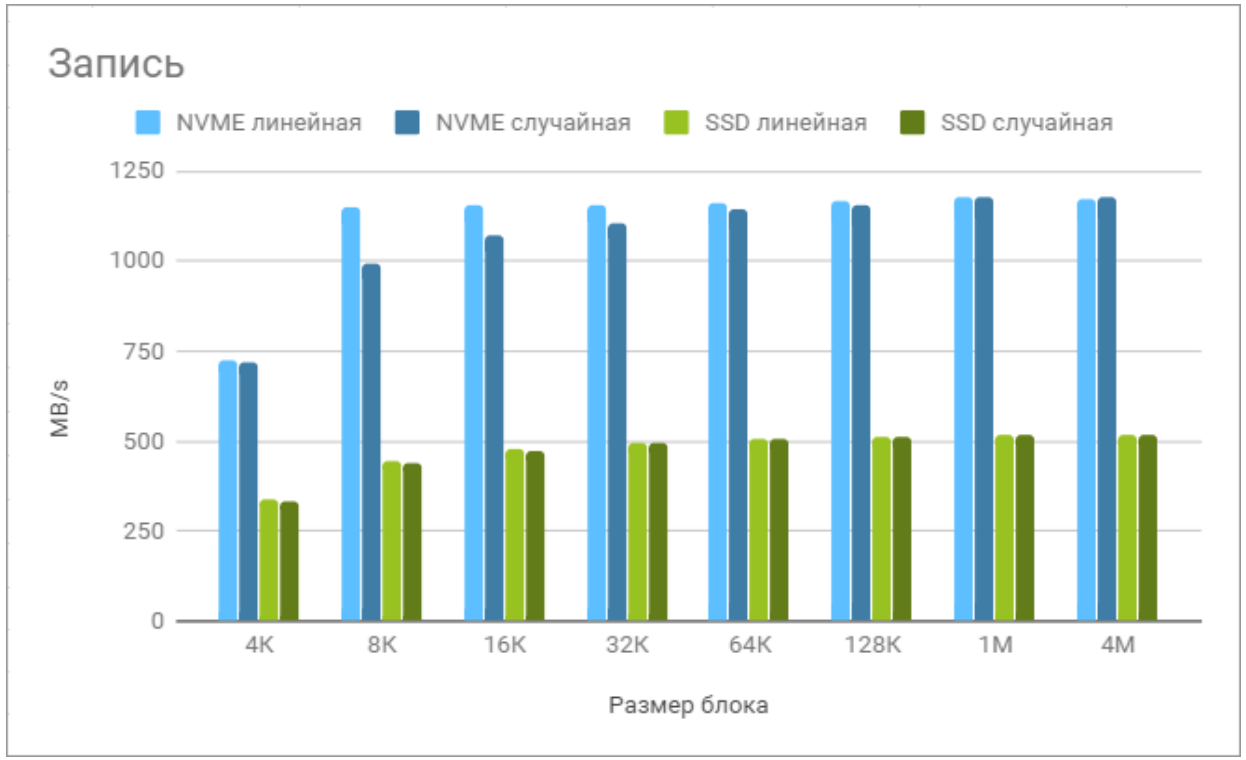

HDD vs. SATA vs. NVMe
Maximum Theoretical Speeds
2500
2000
1500
1000
500
0
150
130
550
550
2500
1000
Sequential Read
Sequential Write
HDD
SATA
NVMe

Зачем нужны HDD?

- Все еще дешевле за Мб хранения
- Многоярусное хранение
- Долговечное хранение
  (а еще есть стриммеры!!!)

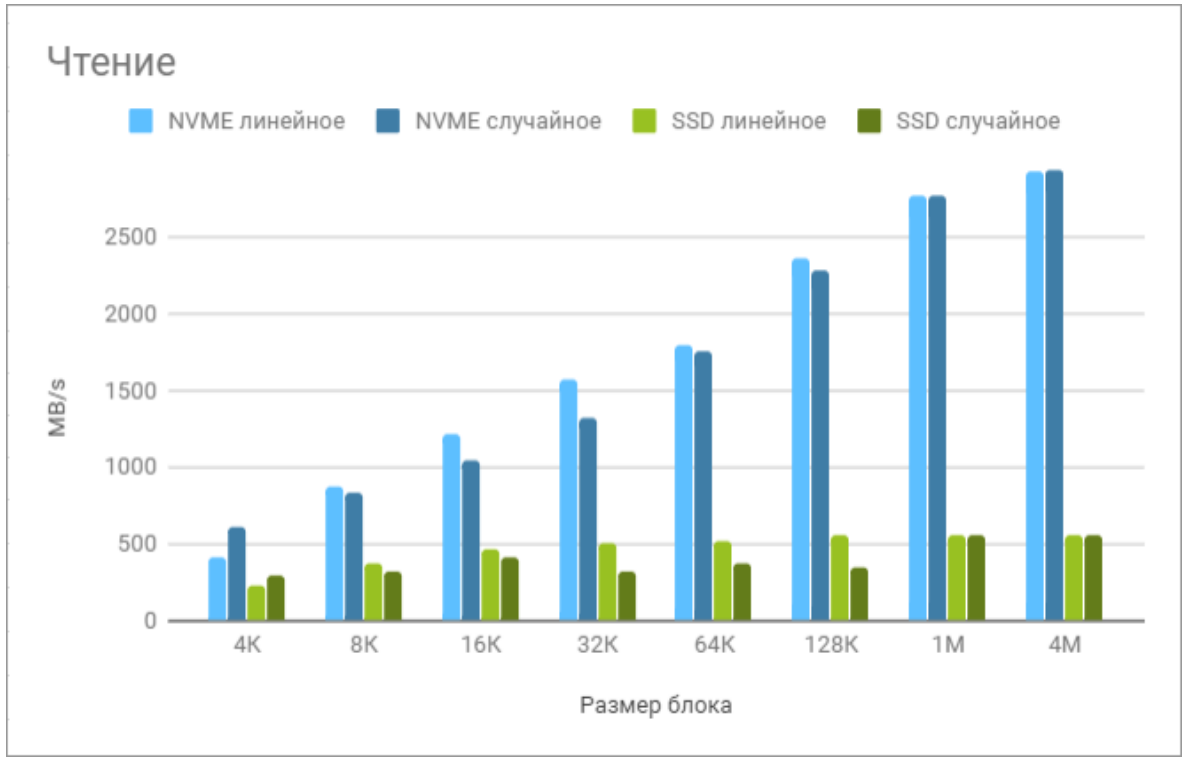
Скорости. SATA vs NVme

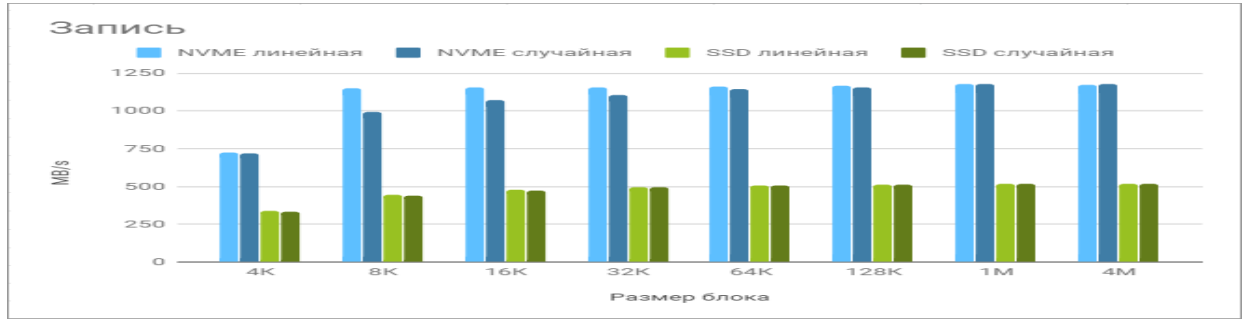




https://firstvds.ru/blog/nvme-vs-sata?ysclid=m1nygk2qnn505264642

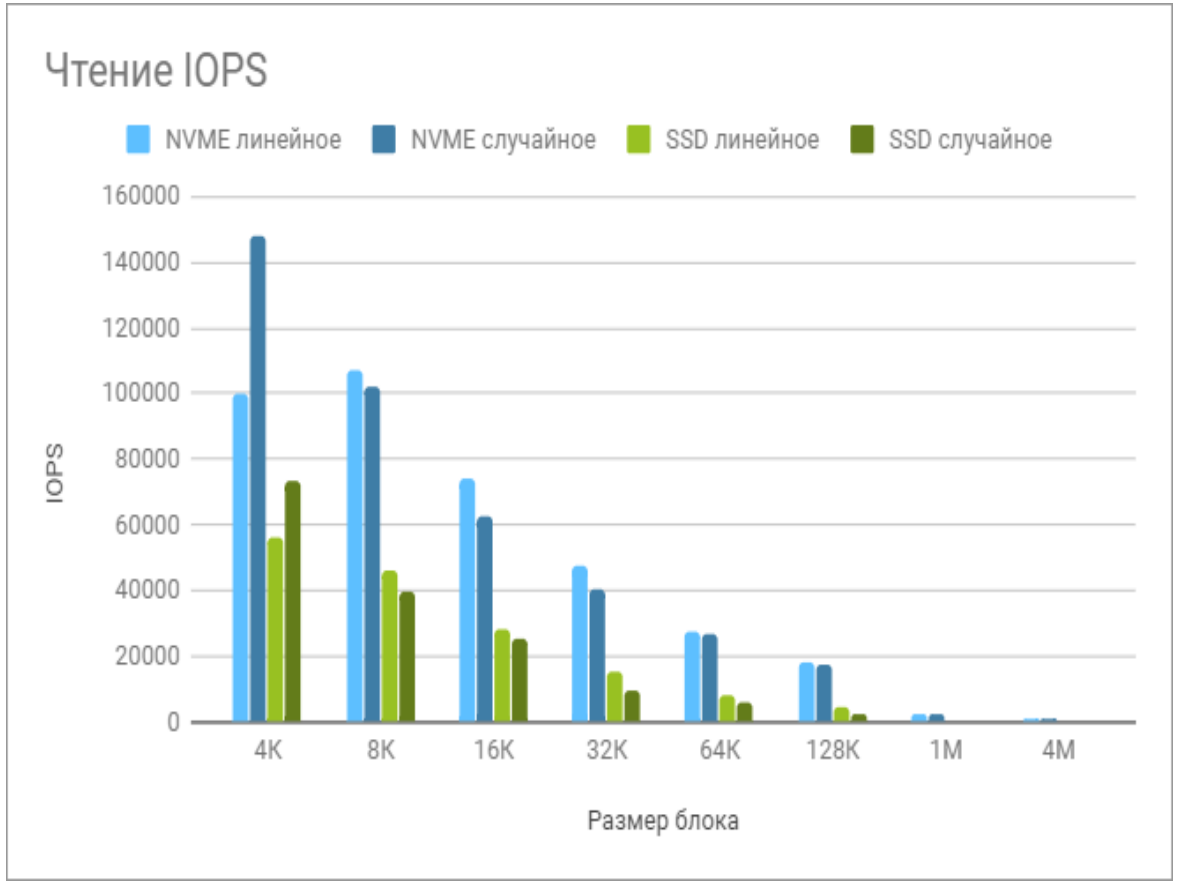
Скорости. SATA vs NVme

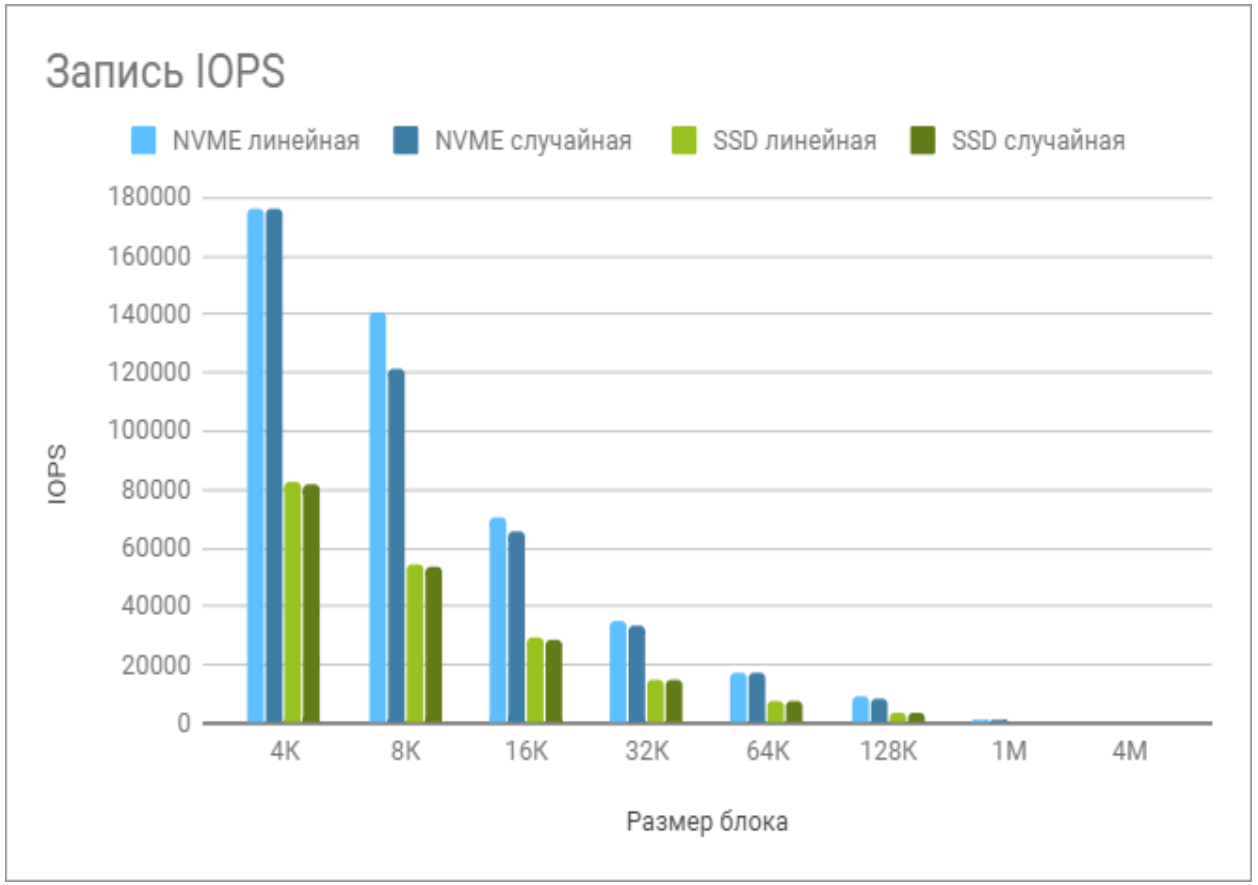




https://firstvds.ru/blog/nvme-vs-sata?ysclid=m1nygk2qnn505264642

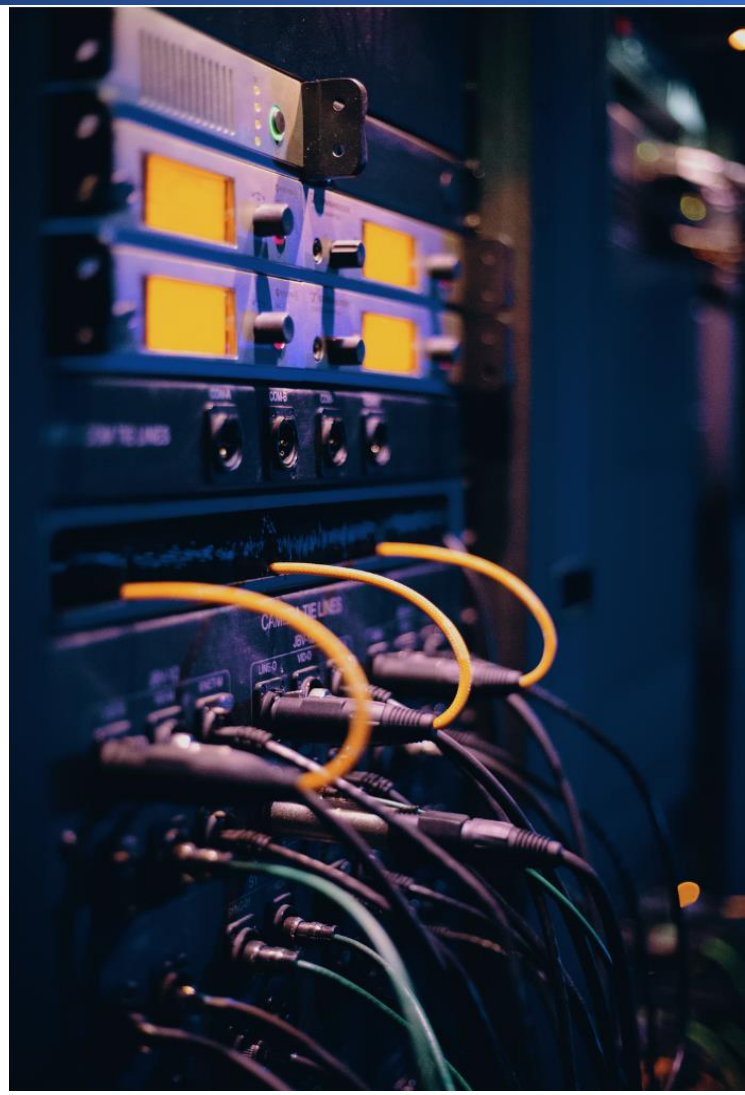
Скорости. SATA vs NVm





https://firstvds.ru/blog/nvme-vs-sata?ysclid=m1nygk2qnn505264642

Выводы



Выводы

- Опять слоеный пирог!
- Актуальные диски – флеш, но у HDD есть своя ниша «холодного хранения»
- Очень большой вклад вносят интерфейсы доступа
- Когда понимаешь, как устроено, понимаешь, как мерить ☺
- Живые котики лучше.