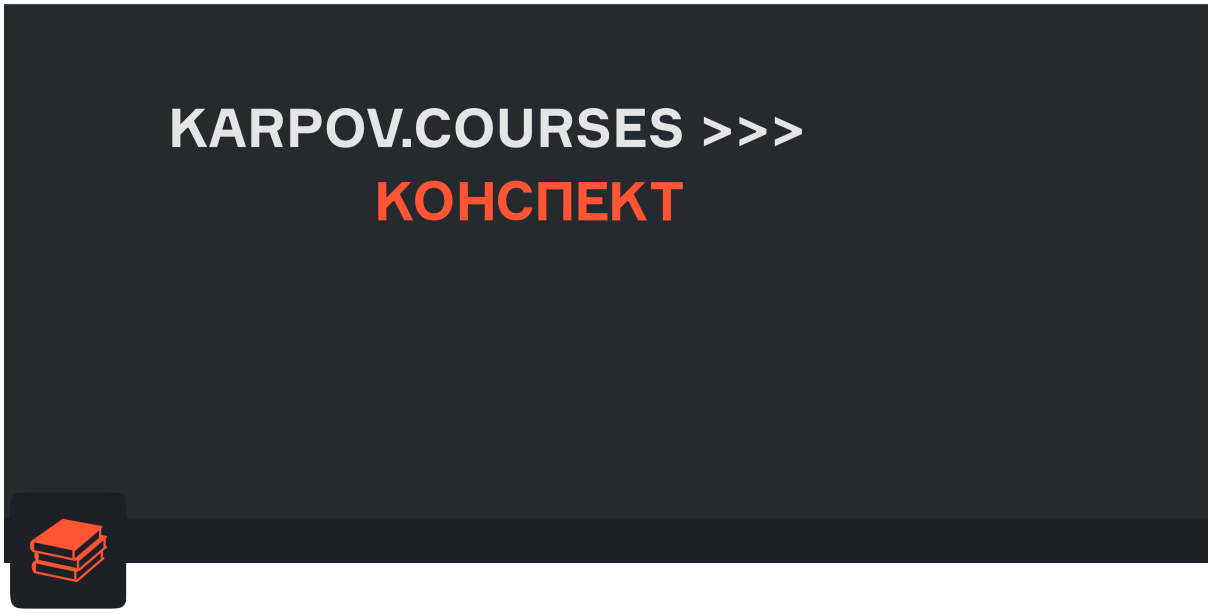


> Конспект > 3 урок > Особенности облаков для DE

> Оглавление

> Особенности дисков для DE

При разворачивании СУБД on-premise быстродействие напрямую зависит от типа диска, но не зависит от его объёма. В облаке производительность диска зависит не только от типа, но и от объёма.

Поясним на примере, почему это происходит: допустим, имеется облачный диск на 2 Тб, но мы создали базу на 50 Гб, соответственно от 2 Тб нарезали 50 Гб под нашу базу, а также по 50, 100 Гб и т.д. распределили между другими клиентами. Чтобы один клиент не занял весь диск по пропускной способности и

не влиял на работу других клиентов, то для каждого диска устанавливаются лимиты по производительности в зависимости от объёма. Соответственно, чтобы получить максимальную производительность от облачного диска, придётся брать у провайдера диск максимального объёма.

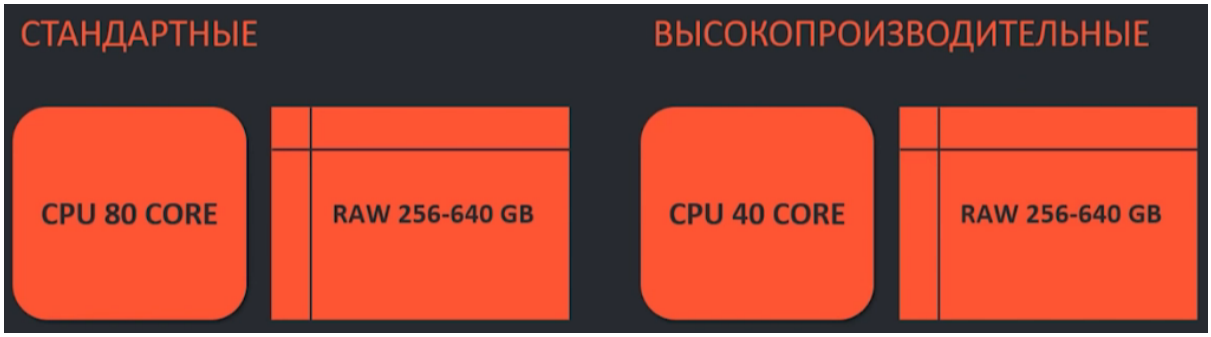
Типы дисков:

1. HDD
2. SSD
3. High IOPS SSD – высокопроизводительный SSD
4. Local NVMe – локальный высокопроизводительный SSD (отсутствует уменьшение производительности, которая присутствует при сетевом подключении диска).

Особенности:

- обычно на основе распределённых файловых систем производительность зависит от объёма
- объём диска можно только увеличивать
- для максимальной производительности используются локальные диски.

> Особенности VM для DE



Рассмотрим размеры виртуальных машин. Допустим, нужна нода с большим количеством ядер и памяти. Нужно уточнить у провайдера, какие максимальные по объёму машины/сервера можно создать, но будет ограничение по размеру стандартного сервера провайдера за минусом overhead-а. Чаще всего будет доступно около 80 ядер на машину и памяти до 500-600 Гб оперативной памяти.

Облачные провайдеры отдают виртуальные ядра, на которых может быть включён hyper-threading, что может сказаться на производительности. Но если мы хотим получить максимальную производительность, то можно запросить у провайдера высокопроизводительные ядра, которые представляют из себя реальные ядра, а не виртуальные, но такой вариант будет дороже и верхний лимит числа ядер на одну машину может быть ниже.

В облаке можно вертикально масштабироваться (увеличивать количество ядер) с downtime, т.е. для этого необходимо выключение машины на несколько минут с последующей перезагрузкой.

> Особенности сети

Помним о:

- сетевых задержках, когда строим распределённые по разным дата-центрам системы. Могут изменяться не только от провайдера к провайдеру, но и в зависимости от текущей нагрузки, от текущих обстоятельств внутри провайдера.
- пропускной способности. При миграции данных в облако и обратно будем упираться в ширину сетевого канала. Пропускная способность обычно варьируется от 1 Gbit/s до 10 Gbit/s и зависит от провайдера, типа VM, производительности ядер, способа размещения.

Пропускная способность

Используем:

- по умолчанию разворачиваем без доступа из внешней сети из-за слабой защищённости
- firewall, группы безопасности, чтобы предоставить доступ из внешней сети
- настраиваем доступ только с определённых ip
- VPNaaS (VirtualPrivateNetwork-as-a-Service) – внутри облака разворачиваем какую-либо систему (например, БД) и с помощью VPN соединяем private сеть облака с private сетью компании без публикации ip адресов во внешнюю сеть.

Не разворачивайте сервисы без тонкой настройки правил безопасности!

> GPU в облаке

Помним о различной функциональности:

- для графического режима

- для вычислений (например, для тренировки моделей ML)

При аренде GPU нужно уточнять, будет ли она поддерживать нужный режим работы.

Особенности и применение:

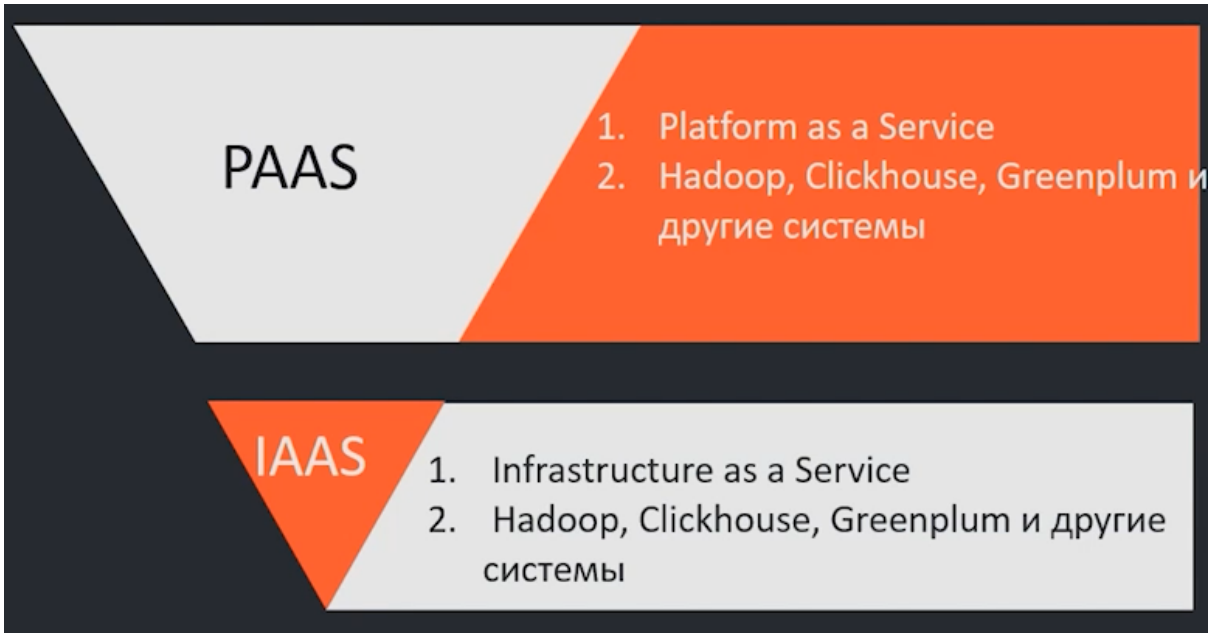
- фиксированные конфигурации (не всегда можно получить желаемое количество карточек GPU для своей конфигурации из-за особенностей облака у провайдера)
- подключение к кастомным конфигурациям VM
- использование GPU в рамках serverless сервисов.

> Сервисы для DE

Есть несколько моделей обслуживания в облаках:

IaaS – для максимального контроля над всеми настройками, нужно знать особенности настройки систем, администрирования.

PaaS – best practice: можно быстро стартовать, не нужно производить сложных настроек.



Преимущества облачных сервисов:

- нет проблем с установкой, мониторингом, бэкапами

- отказоустойчивые кластерные варианты разворачивания из коробки
- используйте S3 для хранения данных
- Kubernetes подойдёт для запуска data нагрузок, потоков

> S3 – simple storage service

- сервис представлен AWS 2006 году. Сейчас доступен у большинства провайдеров по совместительству с AWS S3 API
- по оценкам Databricks S3 дешевле HDFS в 5-10 раз
- практически безгранично эластичен и масштабируется по сравнению с HDFS
- SLA: 99.999999999% durabilityи 99.99% availability (единичные случаи потери данных)
- слои storage и compute разделены

> Kubernetes

- оркестризация контейнеризированных приложений (система для управления приложениями, которая упакована в контейнеры)
- управление жизненным циклом контейнеризированных приложений
- организация инфраструктуры для работы с приложениями
- берёт на себя compute нагрузку при разнесении на отдельные storage и compute слои

> Глоссарий

QoS (Quality of Service) – технология предоставления различным классам трафика различных приоритетов в обслуживании, также этим термином в области компьютерных сетей называют вероятность того, что сеть связи соответствует заданному соглашению о трафике, т.е. способность сети обеспечить необходимый сервис заданному трафику в определенных технологических рамках.

SLA (Service Level Agreement) – соглашение об уровне сервиса, это полноценный документ, в котором фиксируются параметры оказываемой

провайдером услуги.

Overhead – накладные расходы, т.е. это любая комбинация избыточного или косвенного времени вычислений, памяти, пропускной способности или других ресурсов, которые требуются для выполнения конкретной задачи.

Гипервизор – устройство или программа, которое создаёт и запускает виртуальные машины, управляют виртуализацией и ресурсами внутри облачного провайдера.

Hyper-threading – технология, разработанная компанией Intel, позволяющая обрабатывать на каждом ядре процессора несколько потоков. Чем больше потоков, тем больше задач может выполняться параллельно.

GPU (Graphics Processing Unit) – графический процессор, отдельное устройство персонального компьютера, выполняющее графическую визуализацию.

Serverless сервис – сервис, в котором мы не можем управлять сайзингом машины, а только отправляем на выполнение рабочую нагрузку, а он сам автоматически выделяет нужные ресурсы для выполнения задачи.

Kubernetes – открытое программное обеспечение
для оркестровки контейнеризированных приложений – автоматизации их развёртывания, масштабирования и координации в условиях кластера.