Архитектура вычислительных систем

Распределенные и сетевые файловые системы

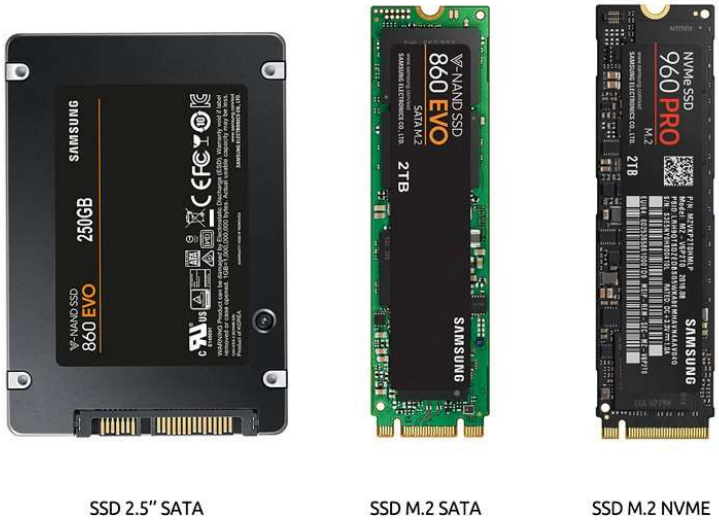
Романюта Алексей Андреевич

alexey-r.98@yandex.ru

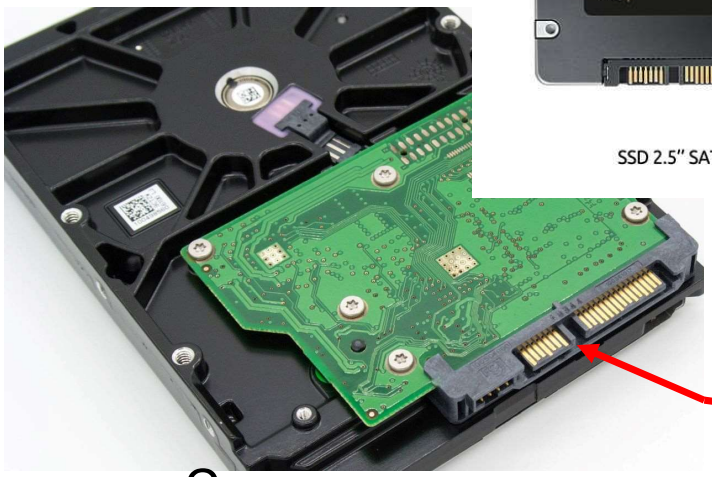
Кафедра вычислительных систем
Сибирский государственный университет телекоммуникаций и информатики

Диски

- HDD
  - SATA - последовательный интерфейс, на основе параллельного PATA (IDE)
  - SAS - последовательный интерфейс на основе параллельного
- SSD
  - SATA
  - SAS
  - M2



SSD 2.5" SATA | SSD M.2 SATA | SSD M.2 NVME



Основное внешнее отличие

В SAS можно воткнуть SATA диск, наоборот нет

Распределенные файловые системы и хранилища

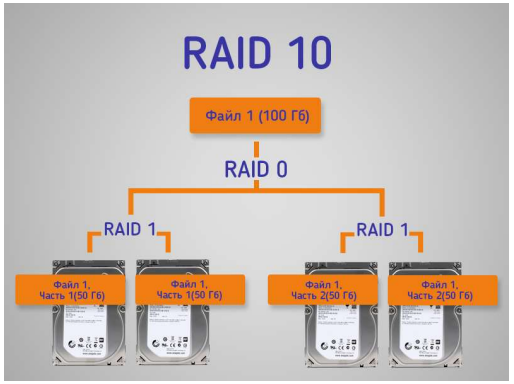
- Системы хранения данных
- ZFS
- GlusterFS
- LizardFS
- Cephfs
- HDFS
- GFS
- Lustre
- S3 – хранилище файлов

Распределенные файловые системы и хранилища

- Зачем это?
  - Когда одного диска мало
  - Когда нужна большая производительность
  - Когда требуется отказоустойчивость дисков
- Для каждого подхода разные технологии

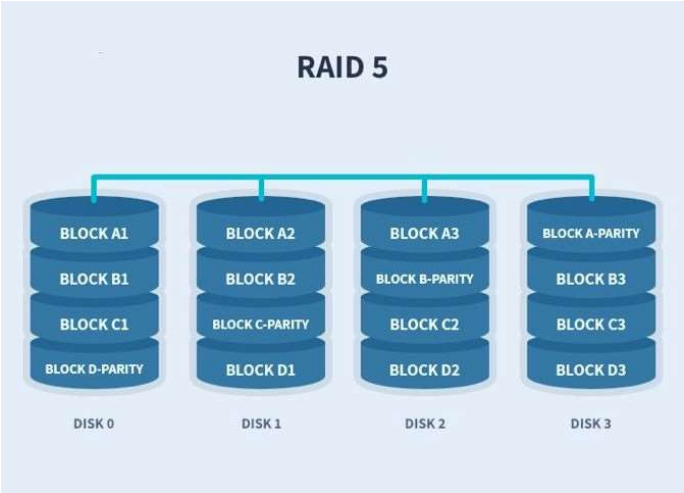
Raid

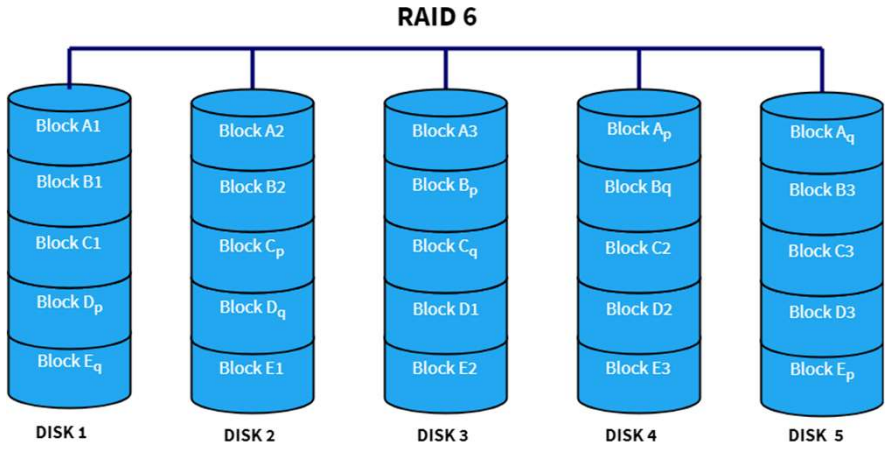
- Аппаратный – физический контроллер в системе
- Программный – mdadm, на уровне ОС
- Виды
  - Raid 0 – Объединение с чередованием записи по дискам. N*size вместимость, отказоустойчивости нет совсем. Хорошая производительность
  - Raid 1 – Зеркалирование. Строго говоря – только 2 диска, но современные системы поддерживают больше. Размер raid-а равен размеру минимального диска в нем
  - Raid 10 -> Зеркалирование + чередование. Добавляет отказоустойчивости



Raid

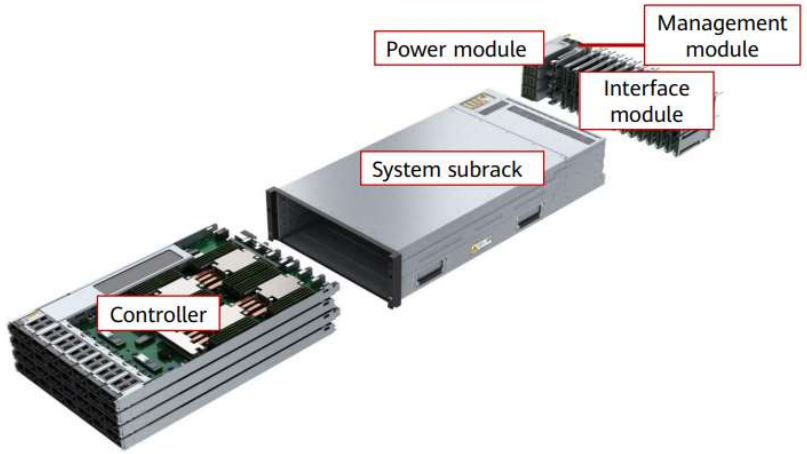
- Виды
  - Raid 5 – запись с чередованием, минимум 3 диска. Размер raid: (N – 1)size Может выдержать отказ одного диска
  - Raid 6 – идеология как у raid 5, но может выдержать отказ двух дисков. Минимум 4 диска. Медленнее raid 5
  - При отказе диска raid переходит в состояние degraded. На обеспечение целостности уходит процессорное время





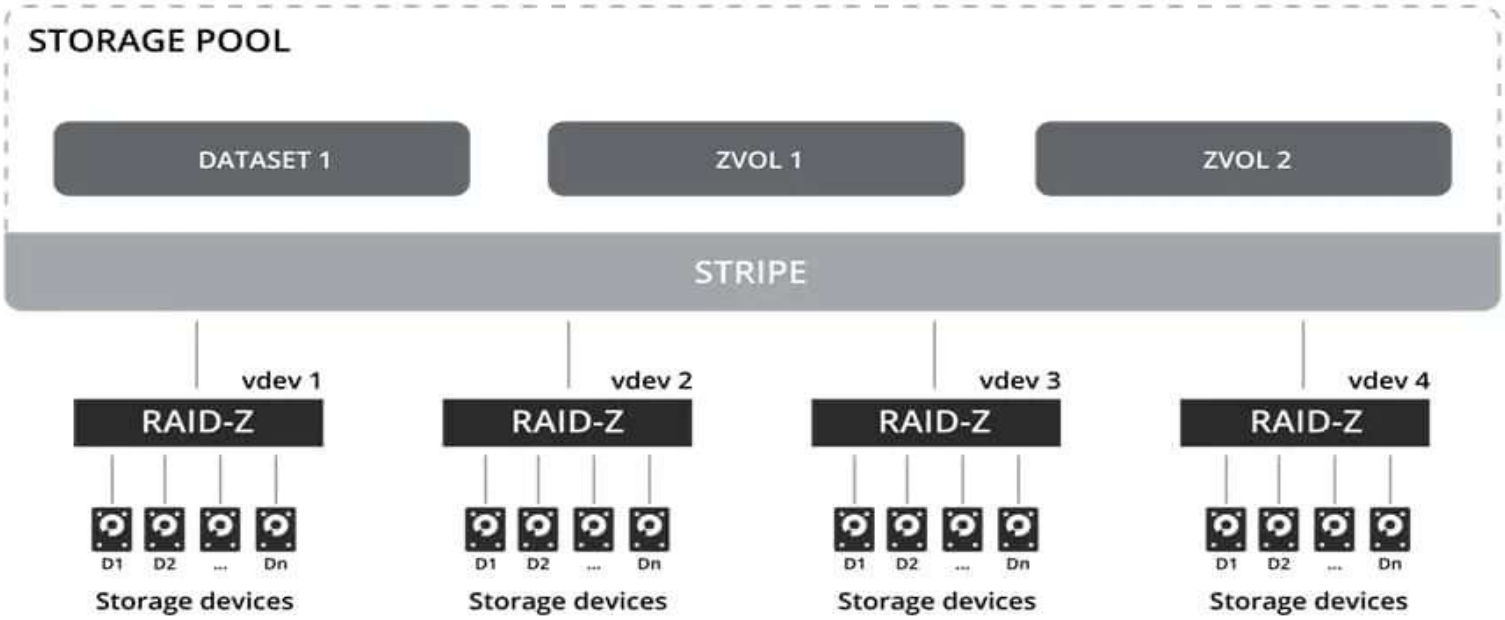
Системы хранения данных

- Отдельная система, физическая «железка»
- Встроено два контроллера и обеспечен failover
- В большинстве случаев, это аппаратный raid
- Подключаются к серверу через:
  - Direct – sas, m2: физические коннекторы
  - NAS – Ethernet, NFS, SMB/CIFS
  - SAN – FibreChannel/FCoE, iSCSI



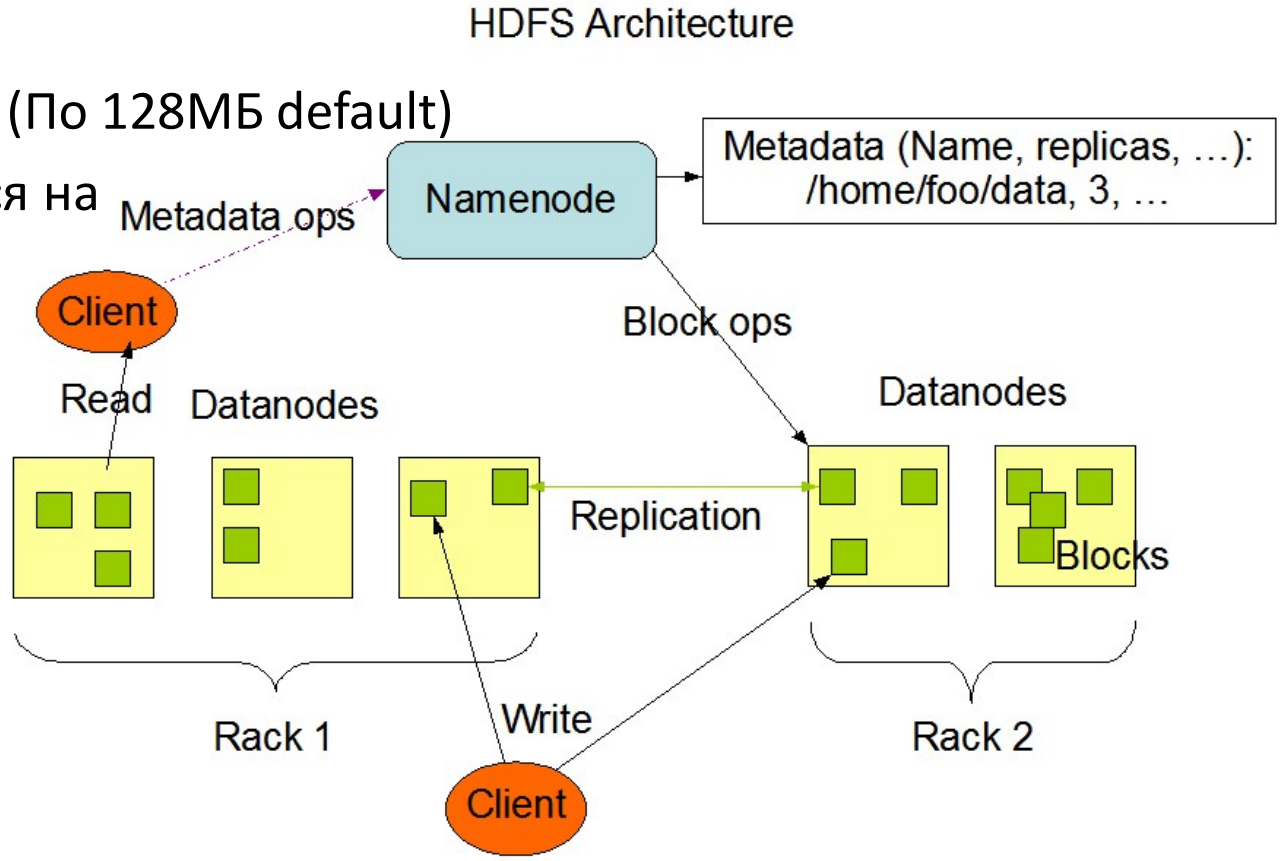
zfs

- Работает на одном сервере
- Использует несколько дисков/блочных устройств
- Отказоустойчив, может выдержать потерю диска (Или нескольких, при использовании Raid-Z*, до 3-ех дисков)



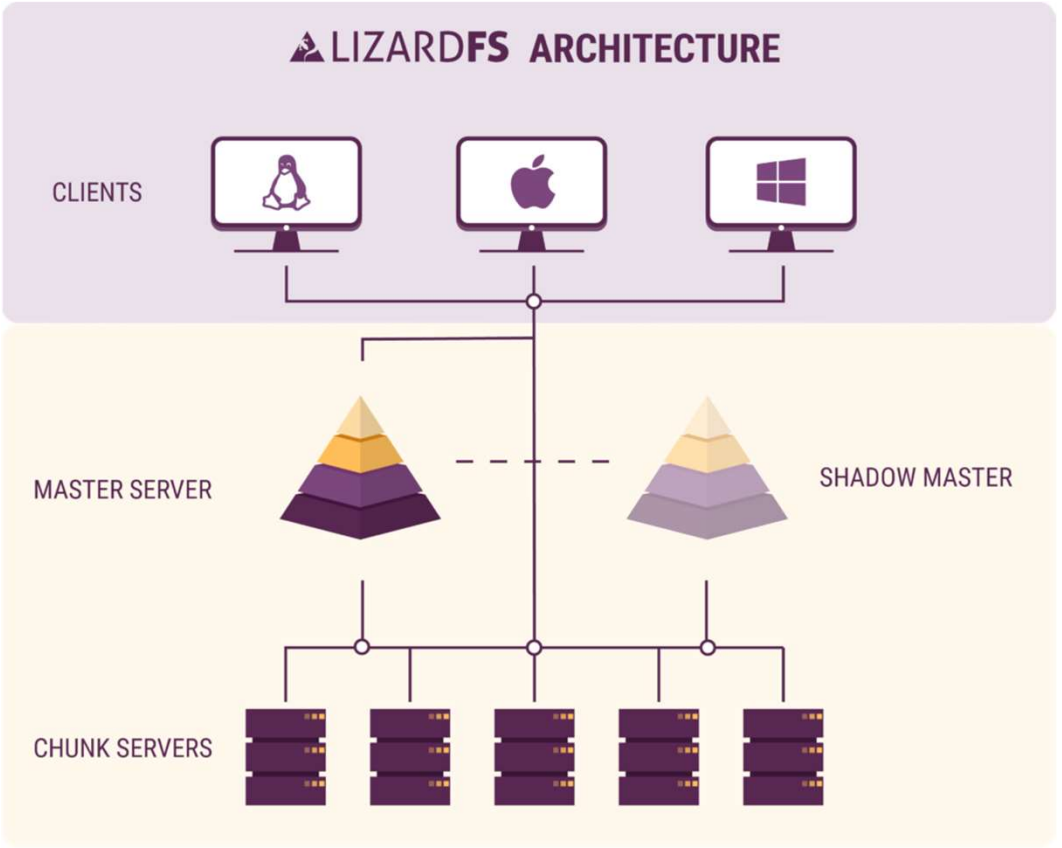
HDFS

- Namenode и Datanode
- Datanode масштабируются
- Файл разбивается на блоки (По 128МБ default)
- Каждый блок реплицируется на несколько datanode
- Можно настроить иерархию ДЦ – блоки будут распределяться с учетом ДЦ



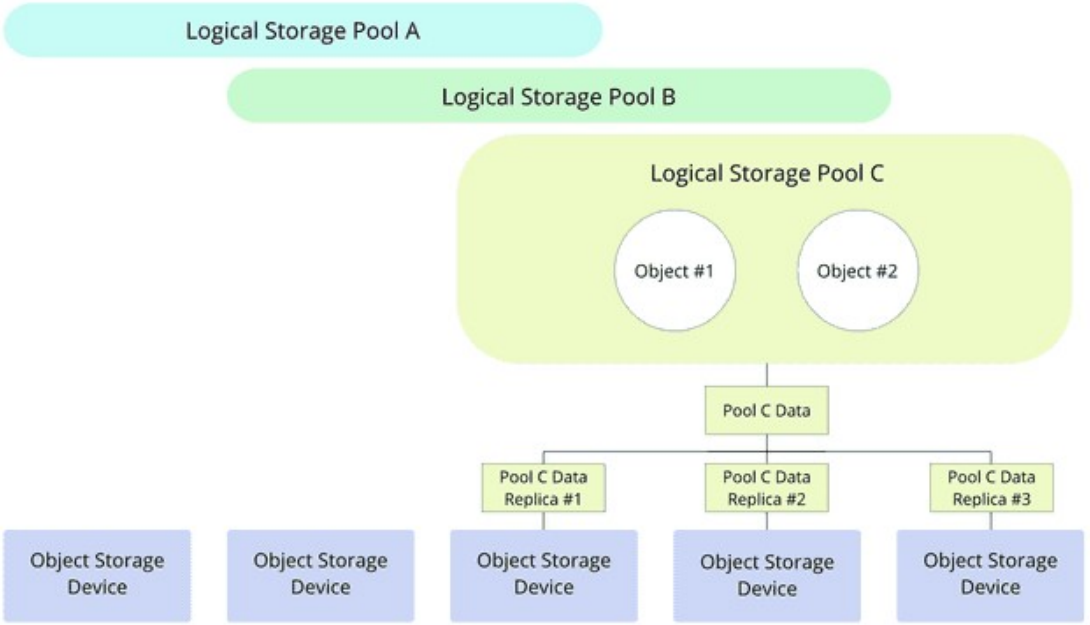
LizardFS

- Master server – метаданные
- Chunk server – данные
- 1 Chunk работает с 1 диском
- Также поддерживает иерархию /dc1/rack1/server1
  - dc1 — датацентр 1,
  - rack1 — стойка 1,
  - server1 — конкретный сервер.



S3

- S3 – Simple Storage Service, протокол работы с файлами
- Объектные хранилища + протокол S3, это одно из самых используемых решений для облачного хранения данных.
- MinIO – self-hosted s3
- Ceph – серьезное распределенное хранилище, s3 – часть функционала
- Amazon S3
- Yandex S3, Selectel, VK cloud



[s3] https://habr.com/ru/companies/selectel/articles/705908/

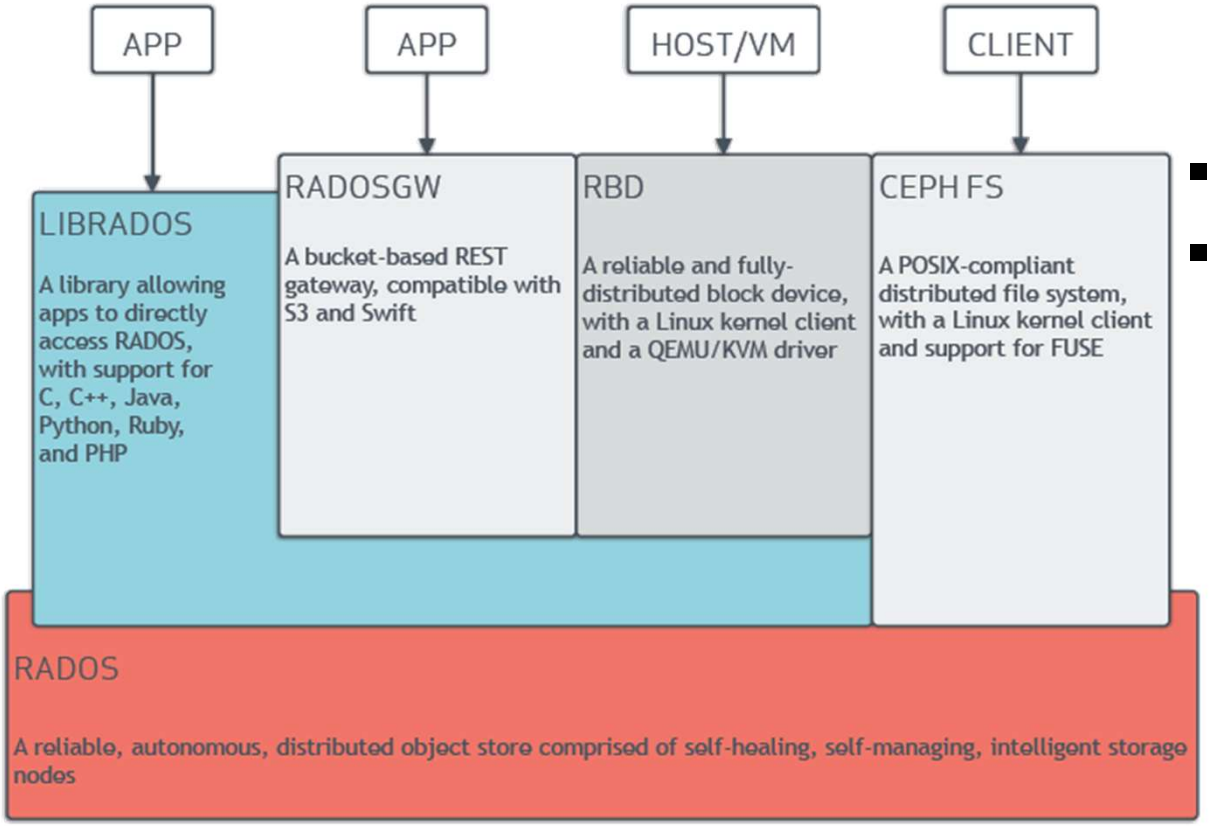
S3

- Каждый объект в S3 состоит из уникального идентификатора, метаданных и содержимого:
- Уникальный идентификатор — строка, с помощью которой можно однозначно идентифицировать хранимый объект и обращаться к нему в хранилище. URL-адрес с уникальным идентификатором — прямая ссылка на этот объект.
- Метаданные —значимые атрибуты объекта (размер, тип и другие пользовательские данные для отбора и сортировки). Необходимы, чтобы находить однотипные объекты и работать с ними.
- Содержимое — данные произвольного формата (цифровые документы, фото- и видеоматериалы, архивы, образы виртуальных систем). Пользователь определяет состав содержимого, а хранилище может накладывать на него технические ограничения, например на максимальный размер объекта. По сути это сам объект.

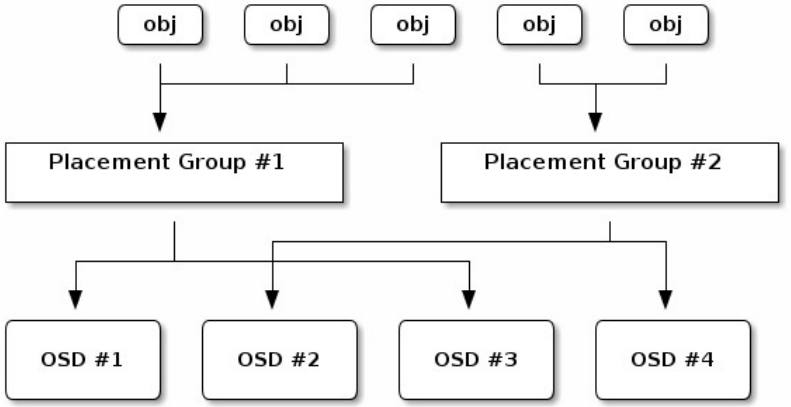
[s3] https://habr.com/ru/companies/selectel/articles/705908/

Ceph

- Mon – «монитор», знает про состояние кластера и знает про osd
- Osd – storage daemon, непосредственно работает с данными



- Умеет в иерархию по ДЦ
- Разворачивается в последних версиях через cephadm



Ceph

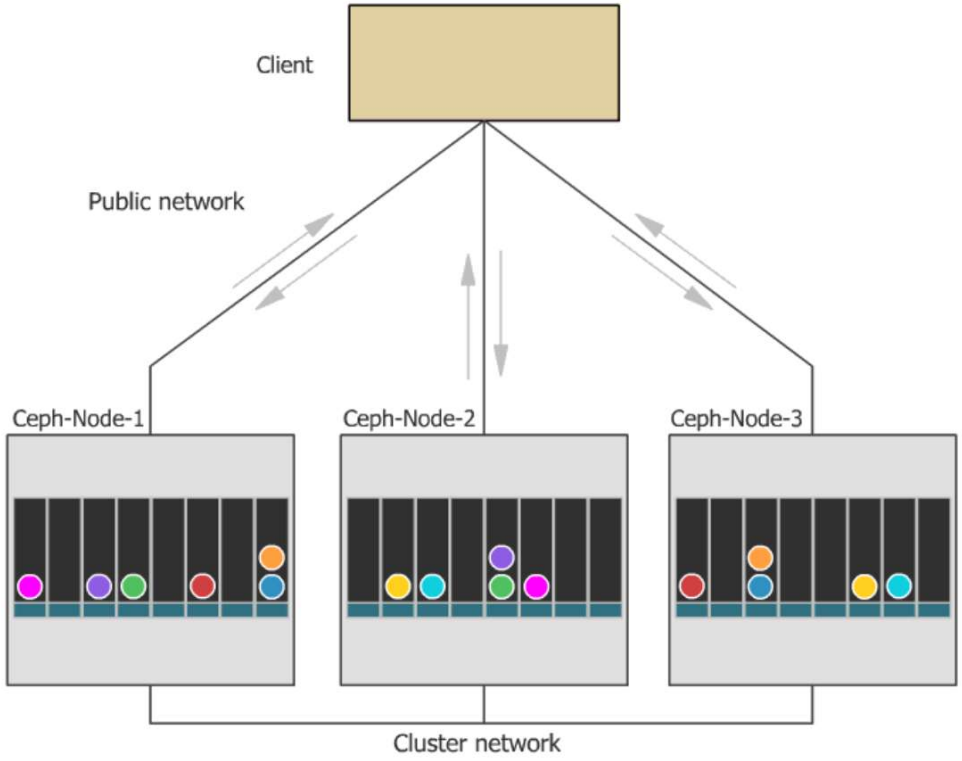
- Ceph MON (Ceph Monitor) — это компонент системы Ceph, который отвечает за мониторинг и управление состоянием кластера. MON обеспечивает согласованность и целостность данных.
- MON хранит и обновляет несколько "карт" (maps), таких как:
  - OSD Map — информация о всех OSD-демонах (состояние, конфигурация, расположение данных).
  - MON Map — информация о самих мониторах кластера (их количество, расположение и состояние).
  - PG Map — информация о PG (Placement Groups) — группах, в которые распределяются объекты.
  - CRUSH Map — карта, которая описывает, как данные распределяются по узлам с использованием CRUSH-алгоритма.

Ceph

- Ceph OSD (Object Storage Daemon) — отвечают за хранение данных, обработку запросов на чтение и запись, репликацию, восстановление, ребалансировку и проверку целостности данных в кластере Ceph.
- Хранение данных – OSD-демон управляет отдельным физическим или логическим диском.
- Балансировка: Ceph автоматически распределяет данные по OSD, чтобы равномерно распределить нагрузку и занимаемое дисковое пространство

Ceph

- Каждый объект может храниться в нескольких экземплярах. Как правило это минимум 2 экземпляра, по умолчанию 3



За кадром

- Физический тест СХД
- SAN/NAS, SAN Switch
- SeaweedFS
- Lustre
- GlusterFS

Live section

Романюта Алексей Андреевич

alexey-r.98@yandex.ru

Кафедра вычислительных систем
Сибирский государственный университет телекоммуникаций и информатики