Интенсив СЕРН

5ти-дневный интенсив
День №1

Авторы

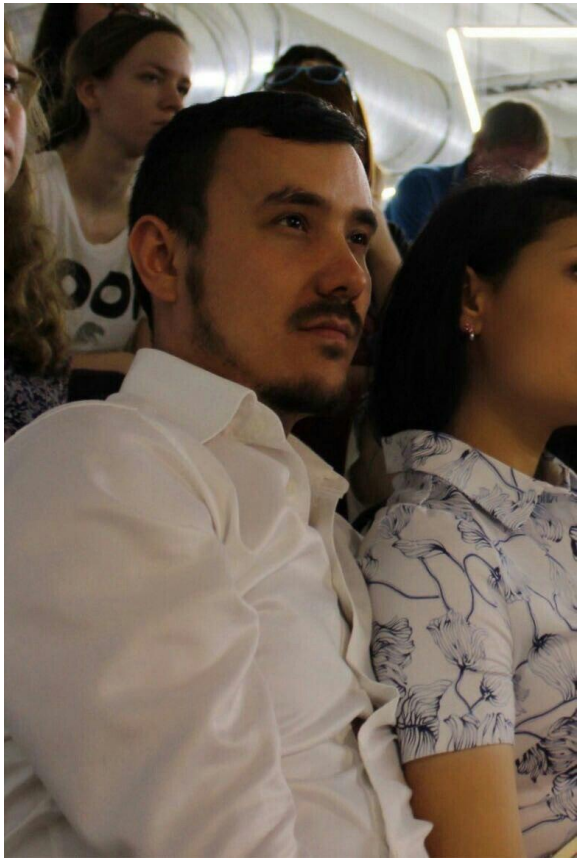
Михаил Жучков

- 10 лет опыта работы в системном администрировании.
- 1 год работы с CEPH с высоконагруженными кластерами
- размерами 0.6птб 1.2птб и 2.2Птб

Авторы

Радик Юсупов

- Со-разработчик курса Learning OpenStack Basic
- Со-разработчик курса OpenStack Programming
- Координатор OpenStack Russia Tatarstan
- Более 5 лет работы с OpenStack
- За время работы были созданы решения промышленного масштаба на основе OpenStack
- Работал заместителем директора по продукту в компании ООО "ТИОНИКС
- В прошлом активный участник ALTLinux Team. В рамках участия в сообществе разработчиков, были созданы дистрибутивы, которые до сих пор активно используются в школах России



Организационные моменты

- Мы делимся своим опытом
- После каждого блока вопросы и ответы на них
- Тайминг: 1,0-1,5 часа (до 21:30)
- За троллинг - бан в чате
- Оценка занятия

Программа занятия

- История CEPH, его задачи, границы применимости
- Архитектура CEPH (журналы, OSD, PG, мониторы)
- Внутреннее устройство (RBD RGW, MDS, пулы)
- О IOPS (Как измерять, о параметрах FIO)
- Кейсы, примеры того, где не нужно использовать Ceph
- Работа в гиперконвергентных конфигурациях - стоит ли игра свеч и если да - на что обращать внимание

Программно-определяемые СХД

- **SDS** – «интеллектуальная» часть сети хранения данных, не привязанная к оборудованию;
- **SDS** способна самостоятельно принимать решения относительно места хранения, методов защиты и перемещения данных
- Имеет линейно масштабируемую архитектуру CONTROL-PATH отделен от DATA-PATH

Архитектурные принципы

- Все компоненты должны быть масштабируемы
- Нет единой точки отказа
- Решение должно опираться на открытое программное обеспечение
- ПО должно работать на обычном железе (commodity hardware)
- Максимальная самоуправляемость, везде, где возможно

Преимущества архитектуры

Полная децентрализация:

- Узлы сами общаются, следят друг за другом и реплицируют данные
- Клиент сам вычисляет нужный узел
- Распределение данных по всем узлам кластера
- Восстановление и балансировка при изменении конфигурации: «многие ко многим»

Flash-кэш на чтение и на запись:

- SSD на каждом узле
- Отдельные full-flash узлы

Настраиваемая политика резервирования:

- Репликация объектов (быстрое восстановление, меньшая емкость)
- Erasure Coding (не быстрая запись, медленное восстановление, большая емкость)

Недостатки архитектуры

- Накладные расходы на сеть, емкость и производительность дисков
- На каждом узле все записи предварительно заносятся в журнал, только потом переносятся на диск
- Каждый диск обслуживает сразу несколько параллельных потоков (высокий seek time)
- Вся система работает поверх ЛВС:
  - Необходимость в отдельной backend-сети
  - Отсутствие стабильной поддержки Infiniband RDMA
- Собственные клиенты и протоколы:
  - Полное отсутствие поддержки Microsoft Windows, Vmware
  - Нет нативной поддержки Fibre Channel, сырая поддержка iSCSI

История и развитие Ceph

2003 Сейдж Вейл (Sage Weil) часть проекта докторской диссертации – ФС

2003—2007 Исследовательский проект, развивался сообществом

2007—2011 DreamHost, начало промышленного применения

2012 – Inktank, корпоративная подписка, саппорт

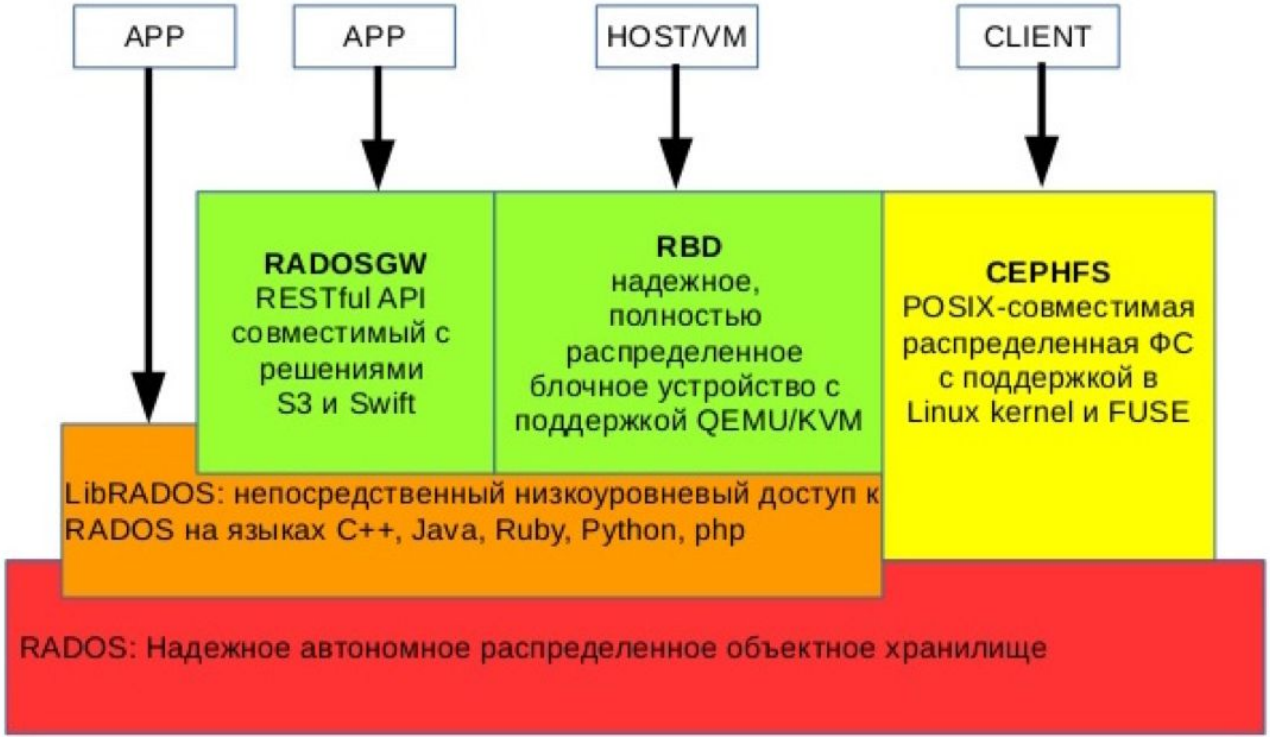
2014 – Red Hat Inc. (Cisco, CERN и Deutsche Telekom, Dell, Alcatel-Lucent, ...)

RH: A next-generation platform for petabyte-scale storage

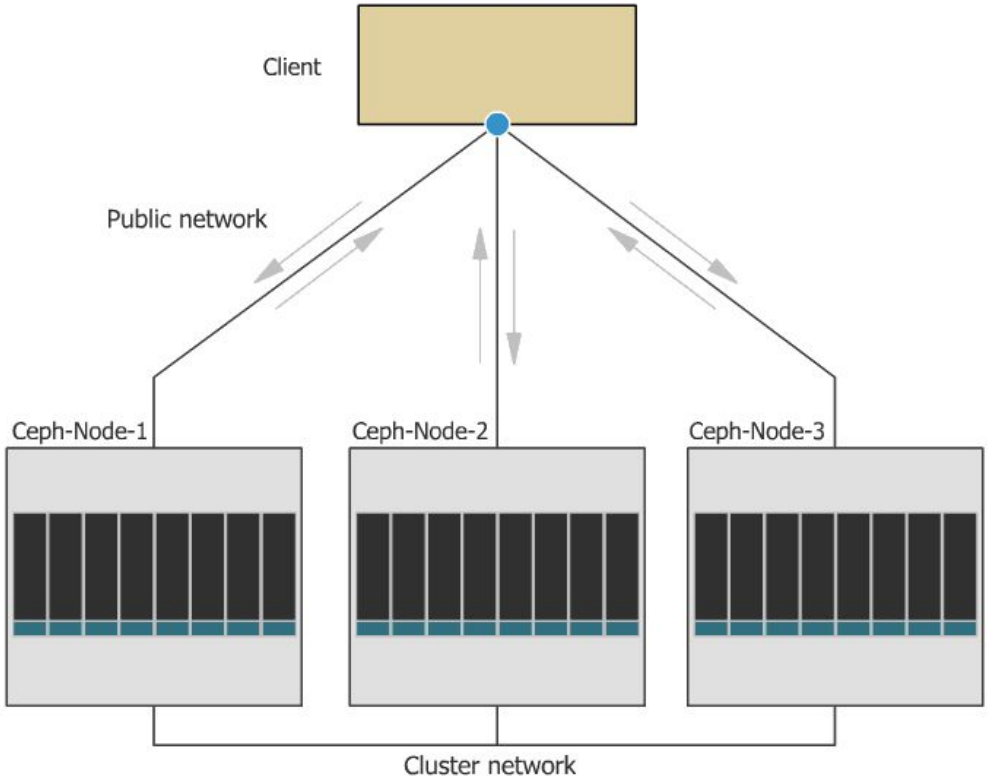
Статус проекта сейчас

- Более 15 разработчиков
- Инвестиции
- Ежедневные коммиты

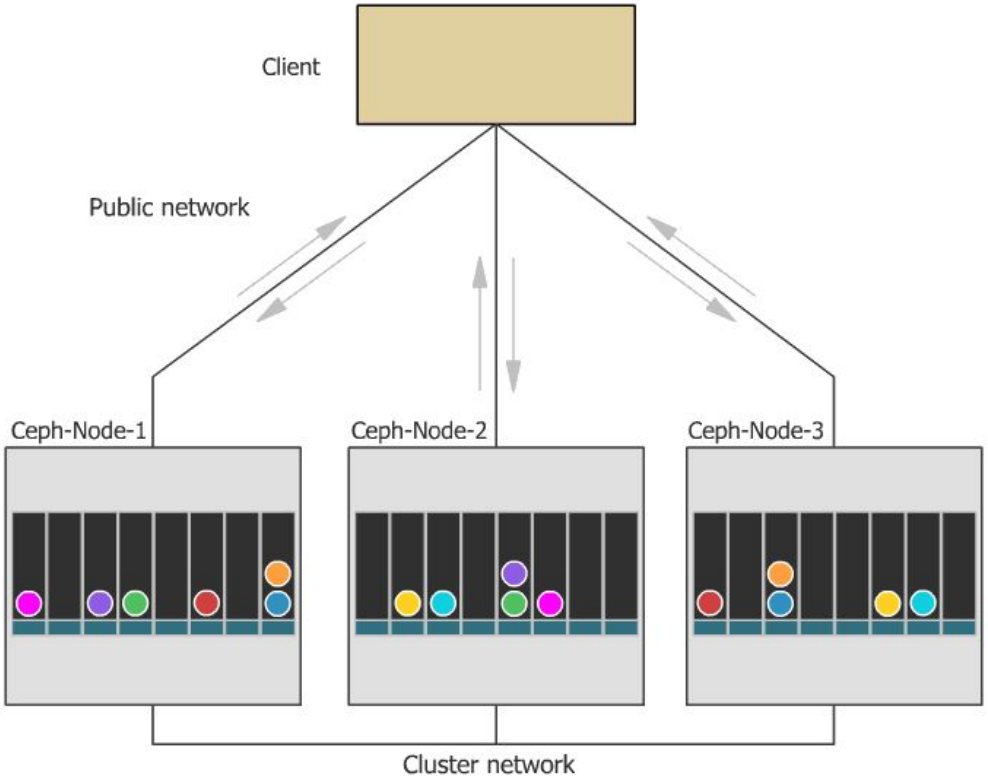
Архитектура



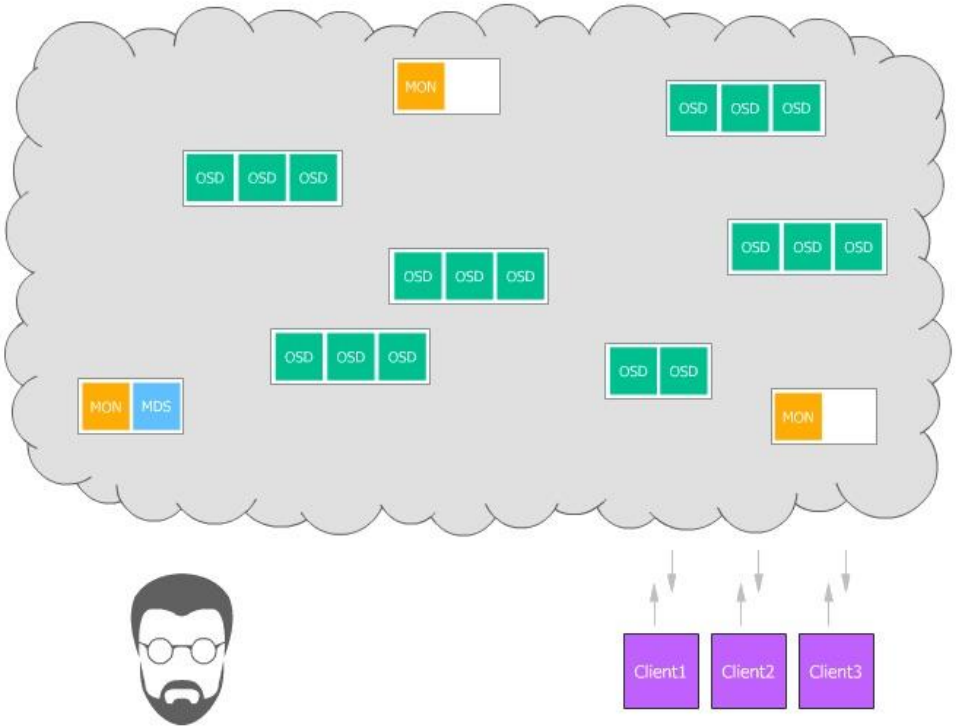
Общий вид работы Ceph



Если выходит из строя диск



3 вида демонов

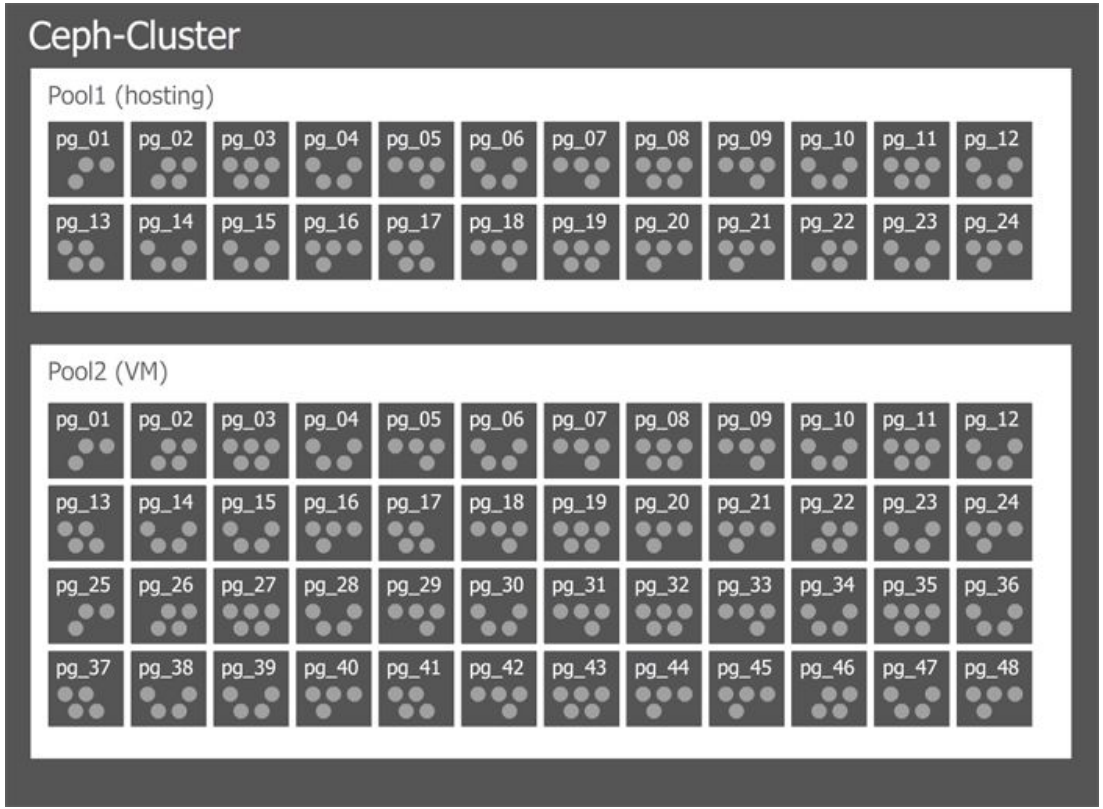


Сервер метаданных MDS Ceph

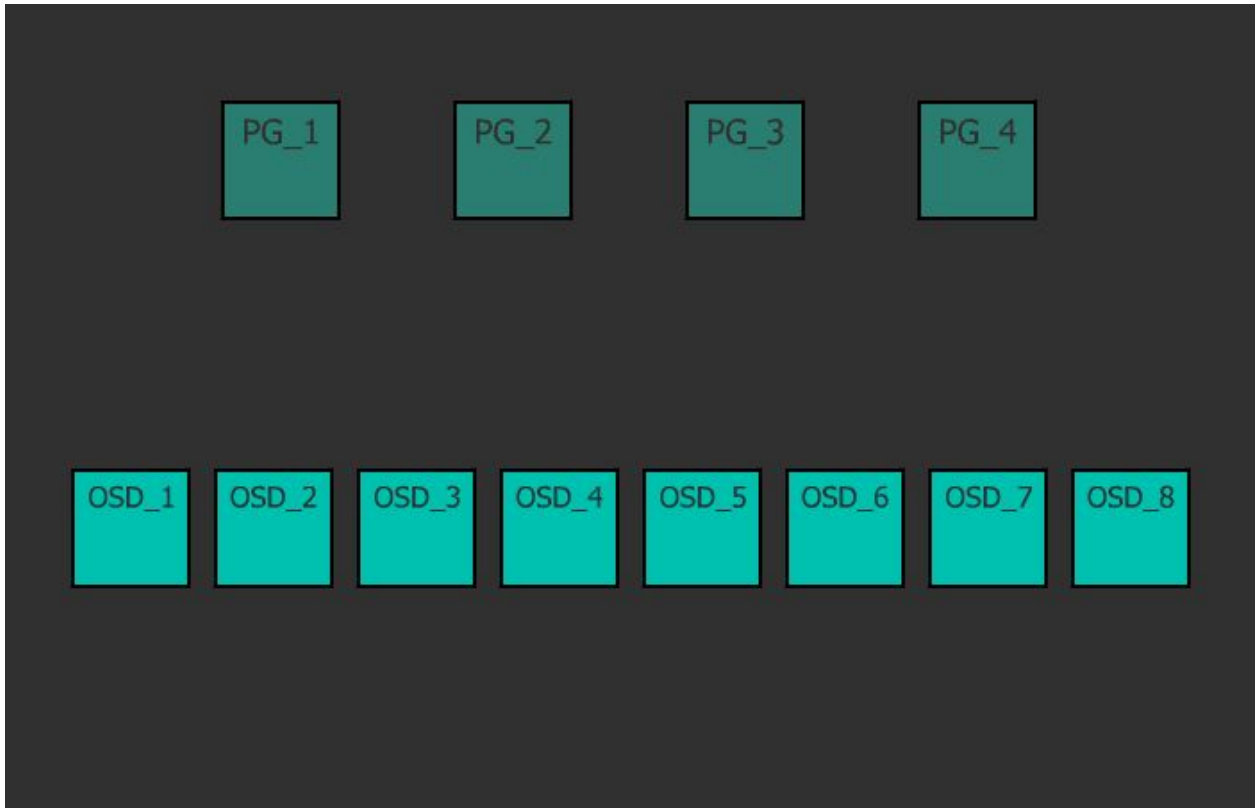
Ceph MDS – демон, обеспечивающий:

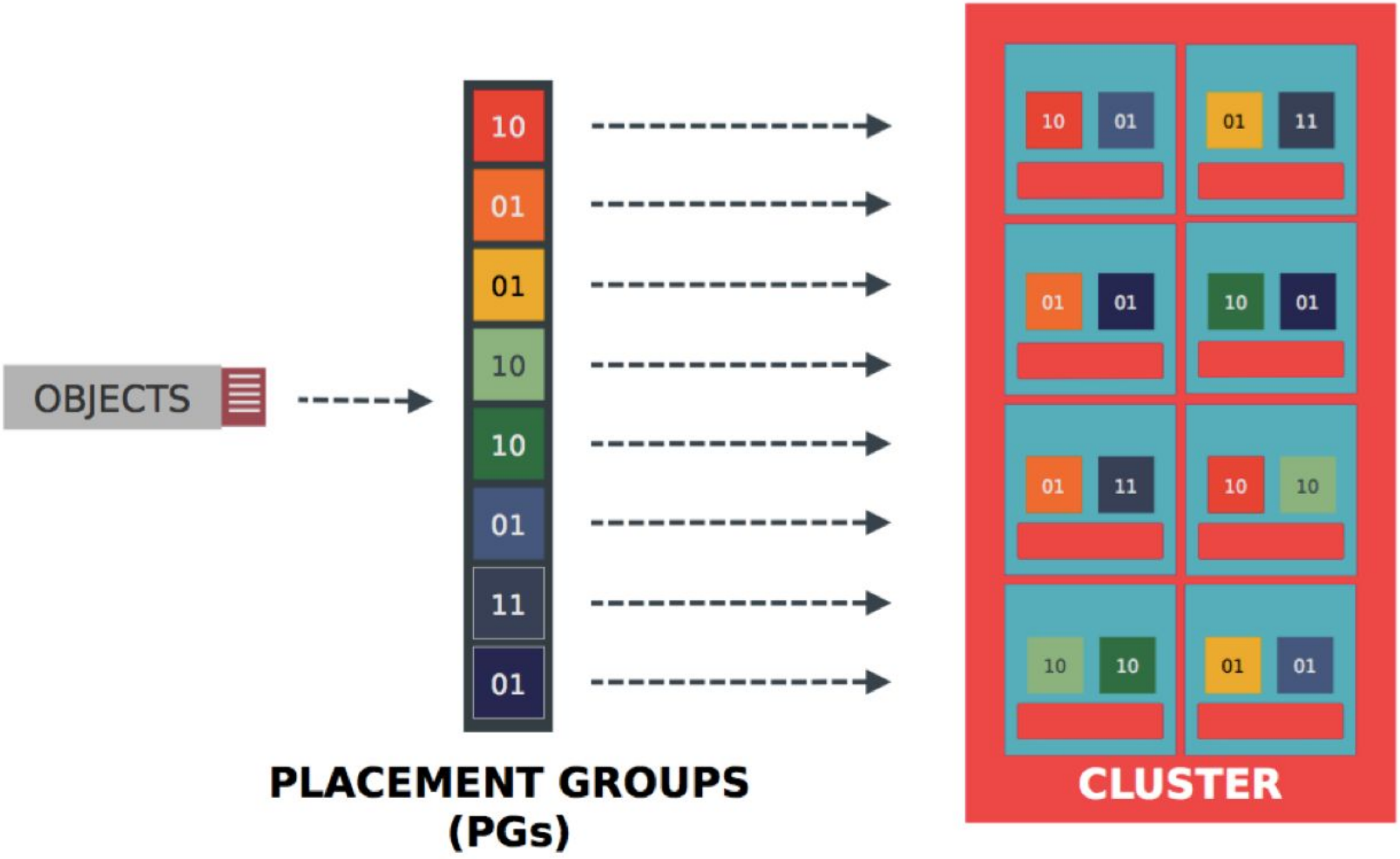
- возможность монтировать на клиентах POSIX ФС произвольного размера
- управление filesystem namespace
- координация доступа к OSD кластеру

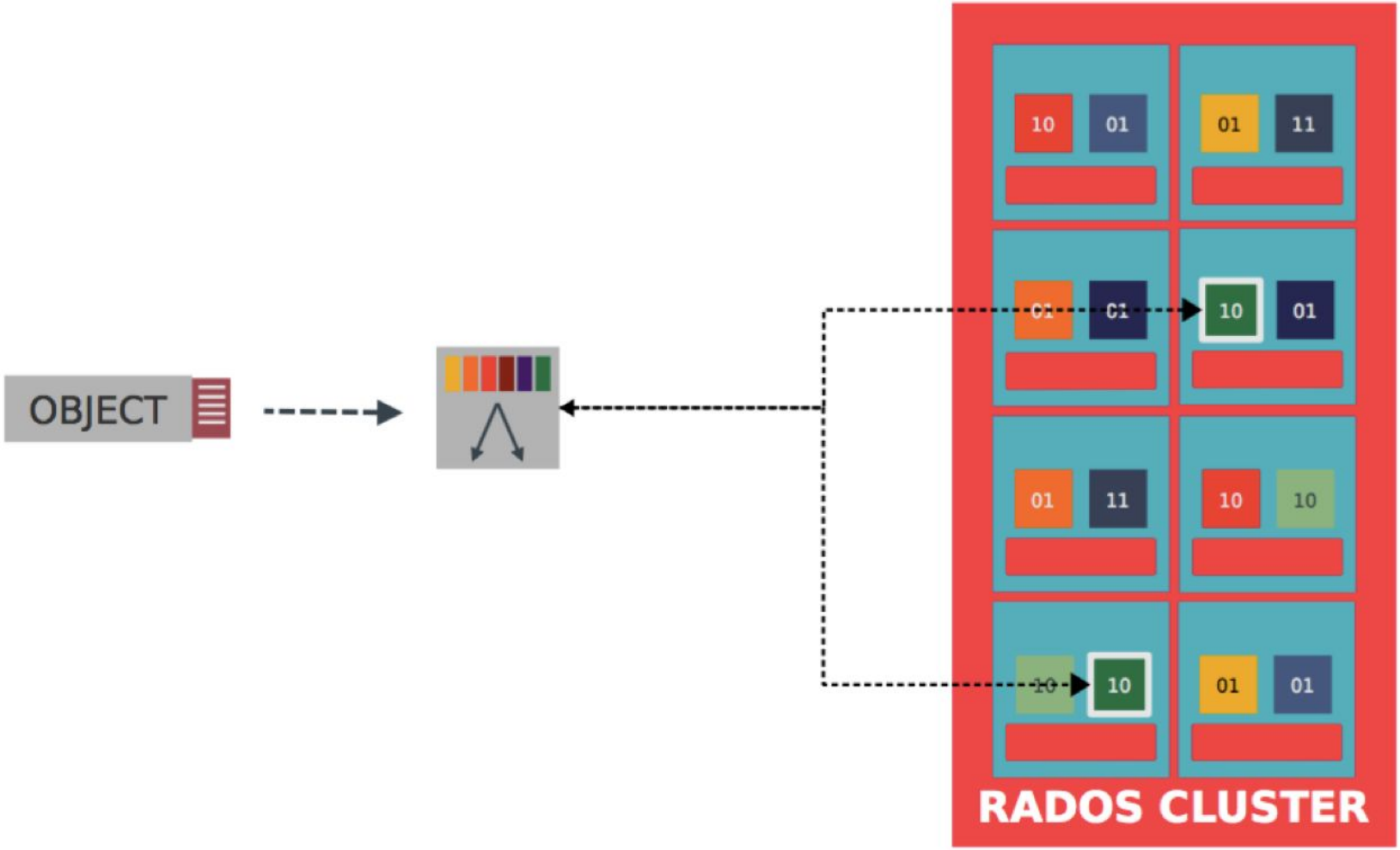
Структура хранения

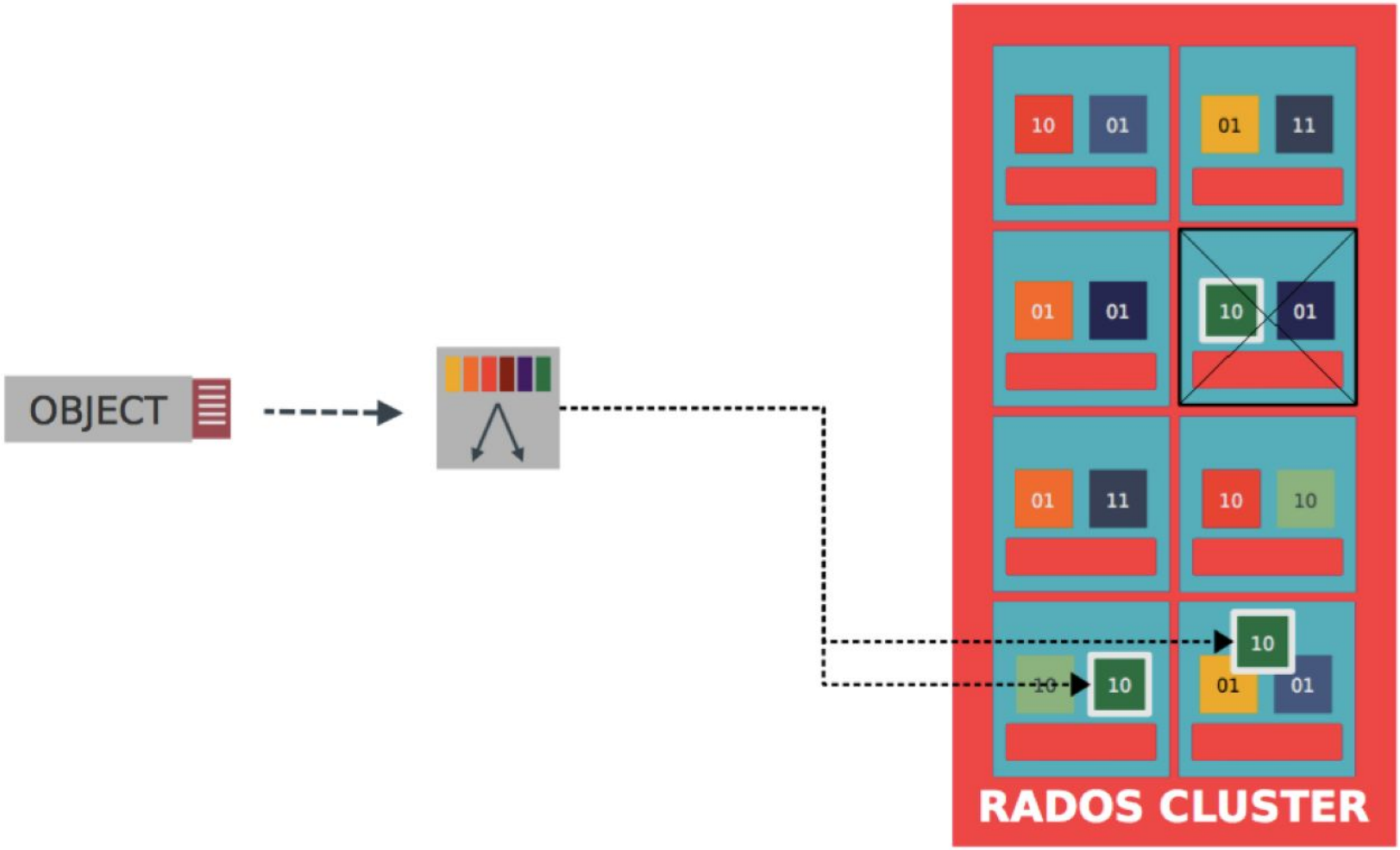


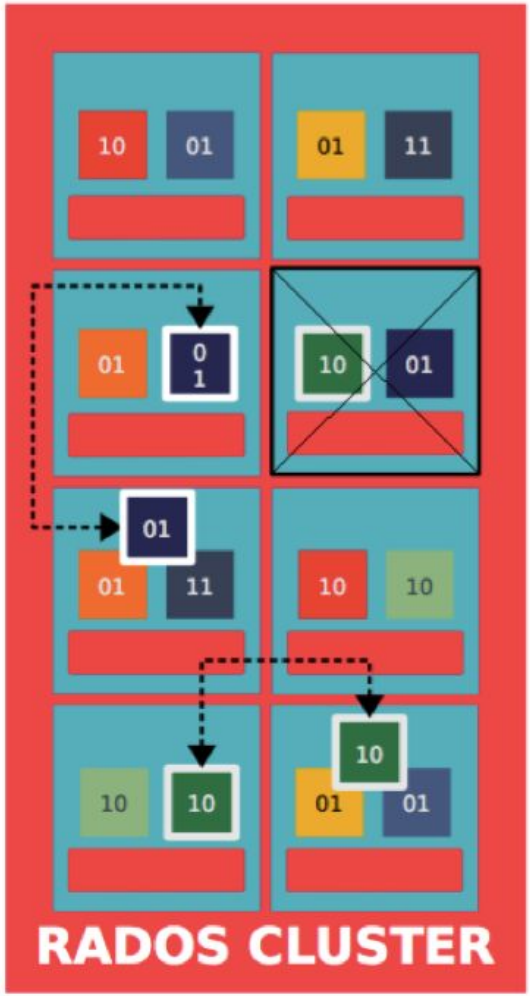
Плейсмент-группа (PG)

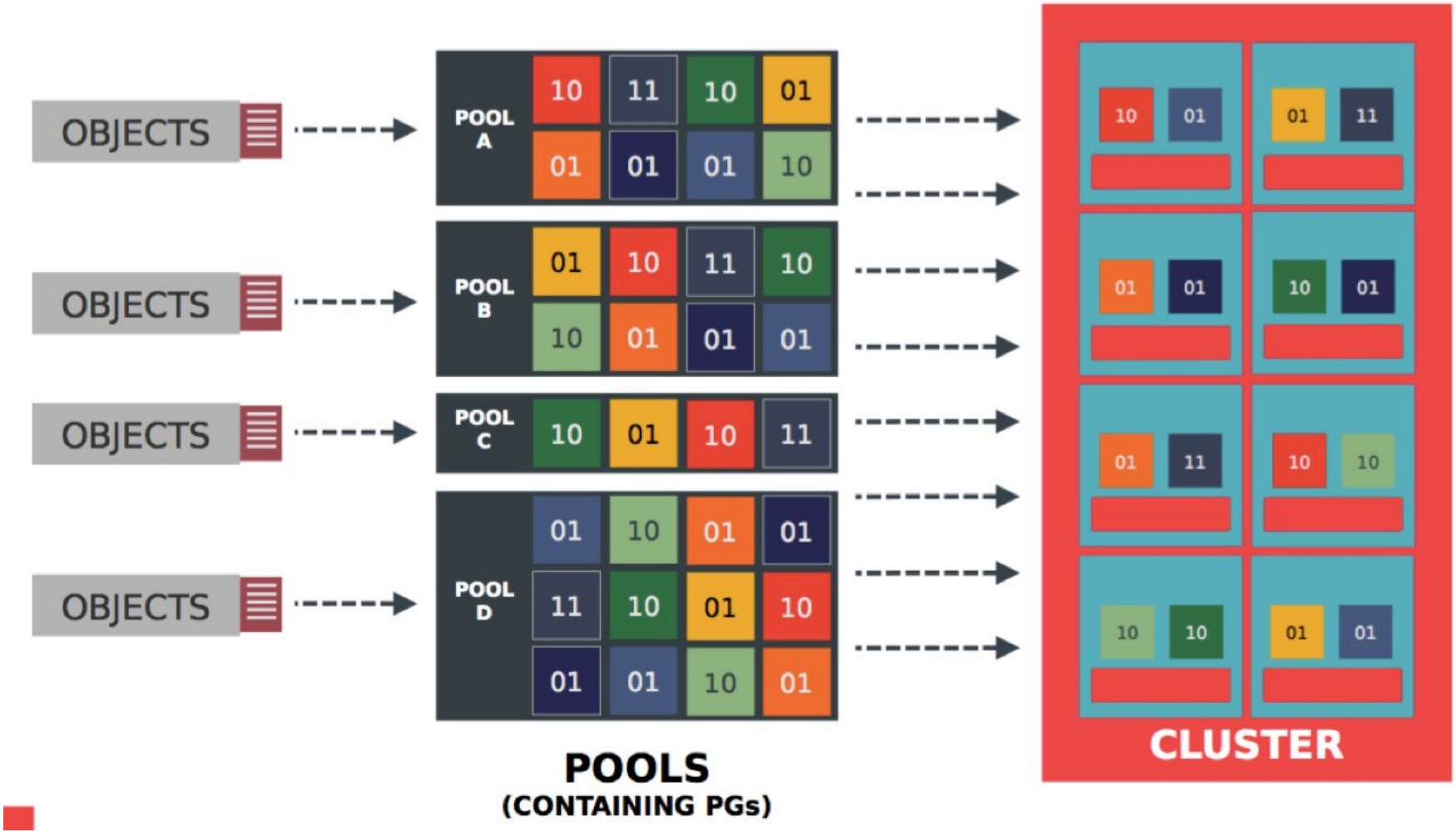



OBJECTS
10
01
01
10
10
01
11
01
PLACEMENT GROUPS
(PGs)
10 01
01 11
01 01
10 01
01 11
10 10
10 10
01 01
CLUSTER


OBJECT
10
01
01
11
01
01
10
01
01
11
10
10
10
10
01
01
RADOS CLUSTER


OBJECT
10
01
01
11
01
01
10
01
01
11
10
10
10
10
01
01
RADOS CLUSTER


10
01
01
11
01
0
1
10
01
01
01
11
10
10
10
10
10
01
01
RADOS CLUSTER

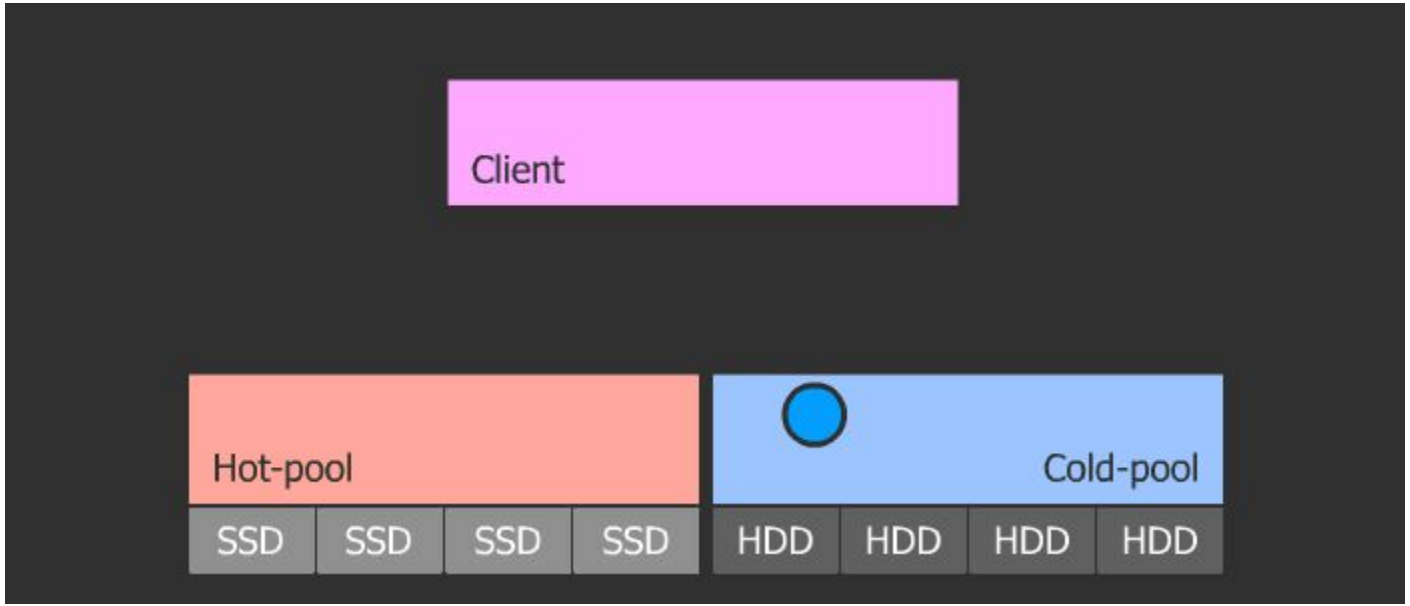

OBJECTS
OBJECTS
OBJECTS
OBJECTS
POOL A
10 11 10 01
01 01 01 10
POOL B
01 10 11 10
10 01 01 01
POOL C
10 01 10 11
POOL D
01 10 01 01
11 10 01 10
01 01 10 01
10 01
01 11
01 01
10 01
01 11
10 10
10 10
01 01
CLUSTER
POOLS
(CONTAINING PGs)

Пулы Ceph

- Устойчивость
  - установка количества копий объекта
  - количество кодированных блоков (chunks)
- Группы размещения
- CRUSH правила
  - для пула можно определить правила избыточности
- Снапшоты
- Установка владельца

Алгоритм CRUSH

Кеш-тиринг

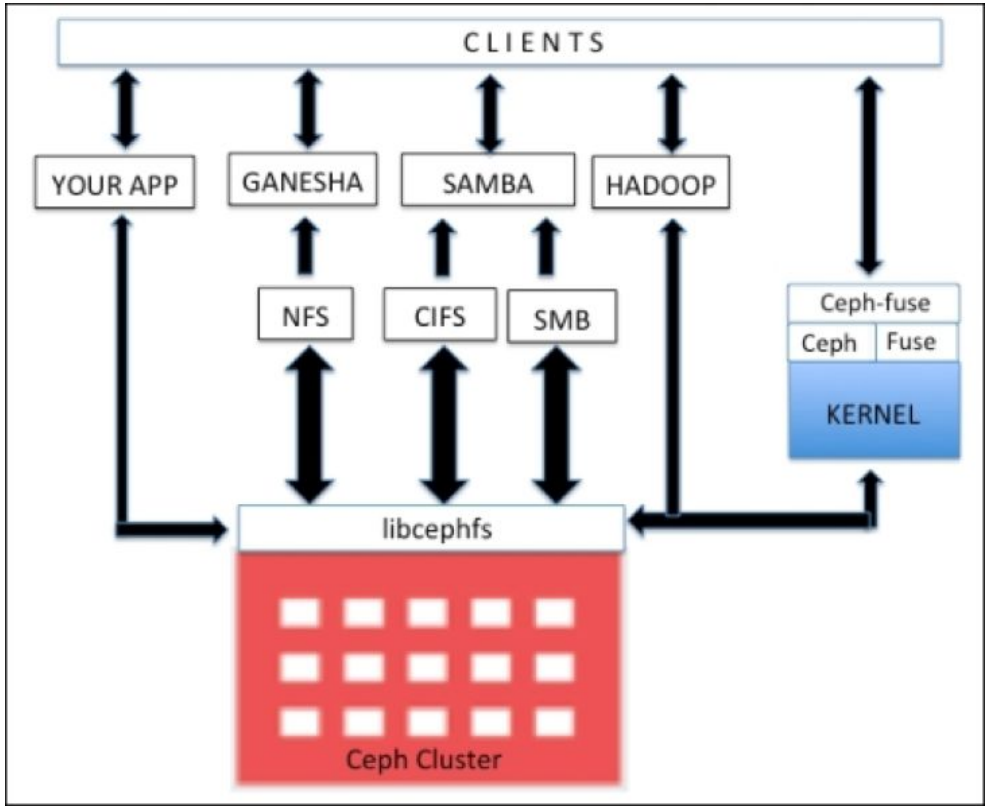


CephFS

Реализована в libcephfs

Поддержка:

- NFS
- CIFS
- SMB



Алгоритмы распределения

Uniform – все веса строго одинаковы. Подходит, когда кластер состоит из совершенно одинаковых машин и дисков

List – перемещаемые данные с некоторой вероятностью попадают в новое или старое хранилище. Expanding cluster

Tree – бинарные деревья, оптимизация скорости помещения объектов в хранилище

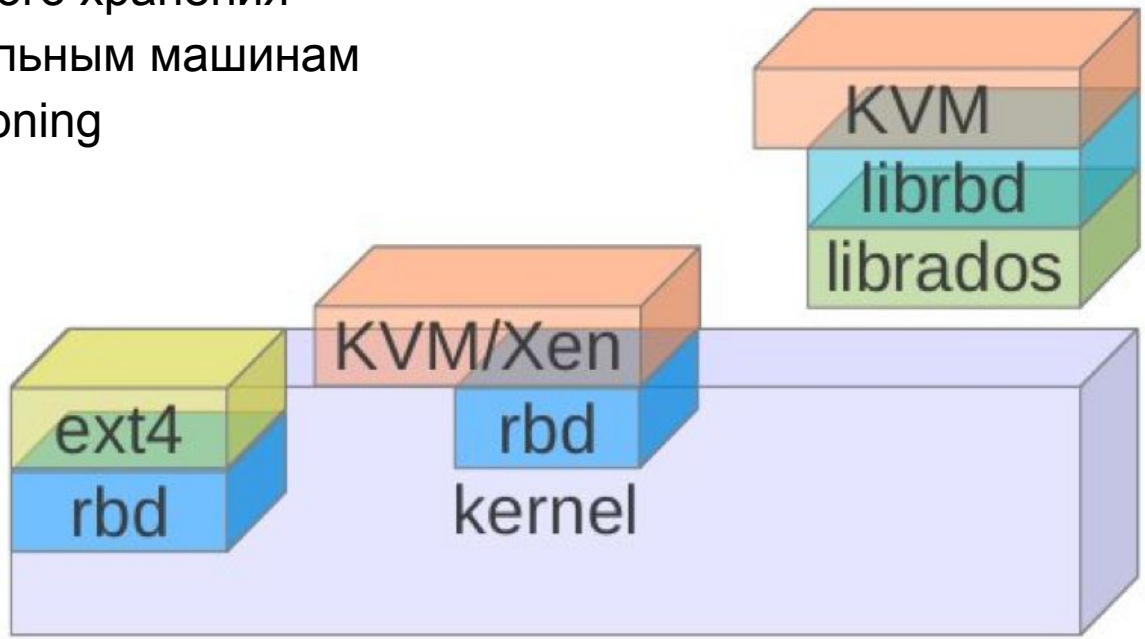
Straw – комбинация стратегий List и Tree для реализации принципа «разделяй и властвуй». Обеспечивает быстрое размещение, но иногда создает проблемы для реорганизации

Алгоритм PAXOS

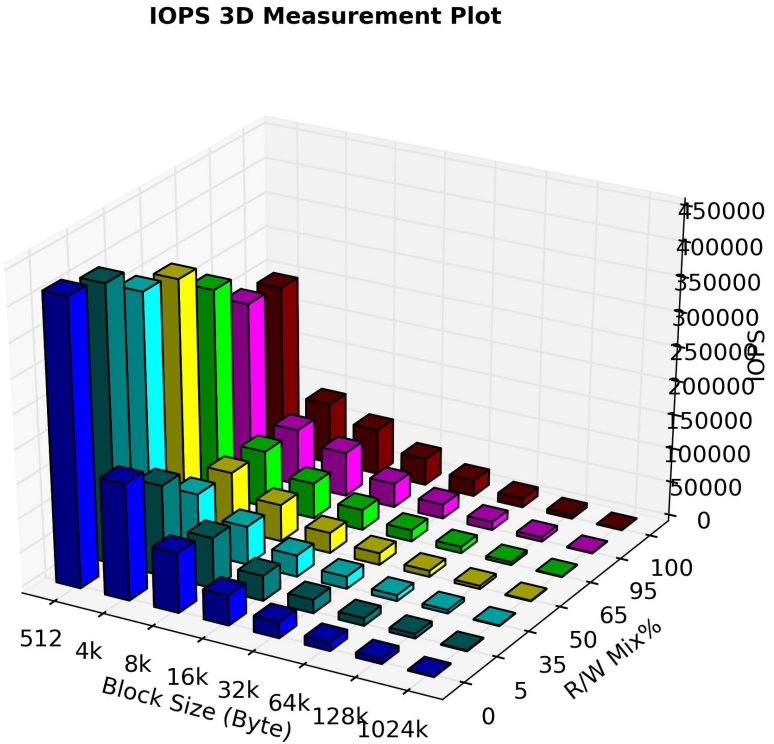
- Обеспечивает консенсус
- Соответствует показателям:
  - Согласованность →решение принимается только единогласно
  - Нетривиальность →количество вариантов решения известно заранее и больше 1
- Живучесть →если предлагается принять решение, то решение (не обязательно предложенное) рано или поздно будет принято.

RBD

- Предоставление блочного хранения гипервизорам и виртуальным машинам
- Реализация thin provisioning
- Поддержка:
  - XEN
  - KVM
  - QEMU



IOPS + latency



Когда вам **НЕ НУЖЕН** Ceph?

- Низкая латенси за недорого
- Когда у вас один сервер
- Когда вам не нужна отказоустойчивость

Когда вам **НУЖЕН** Ceph?

- Требуется большой S3 совместимый сторадж
- Большое хранилище под OpenStack
- Требуется большое **И** отказоустойчивое хранилище

Гиперконвергентные системы

- Рекомендуется отключать NUMA
- Другие настройки sysctl
- Другие настройки лимитов
- Количестве дисков > 6 на ноду

Формула расчета

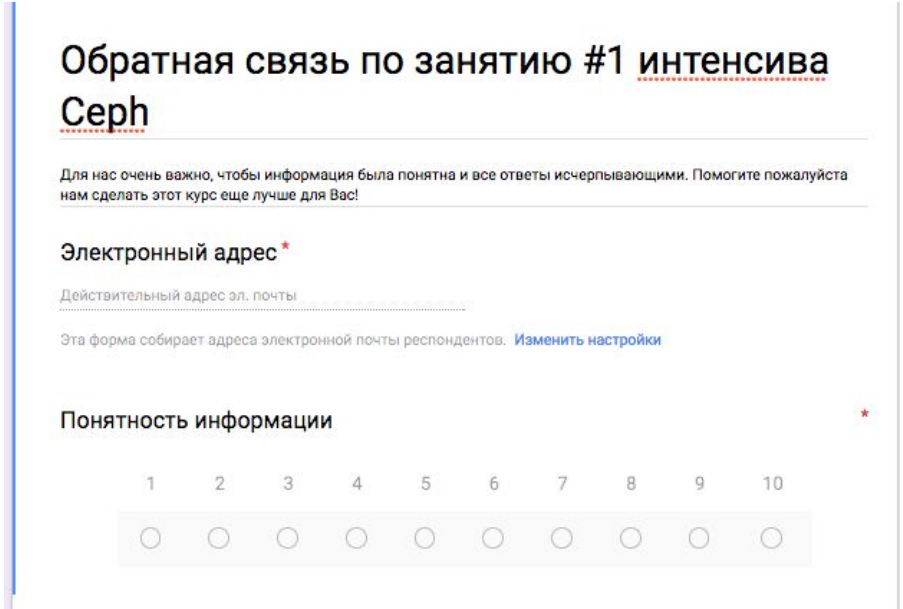
((разёмы ЦПУ * число ядер на разъём ЦПУ * тактовая частота ЦПУ в ГГц) / число OSD) >=1

Например, сервер с одним сокетом ЦПУ, 6 ядрами по 2.5ГГц должен быть достаточно хорош для 12 OSD Ceph, причём каждый OSD получит примерно 1.25ГГц вычислительной мощности: ((1*6*2.5)/12)= 1.25.

Вопросы?

Обратная связь

https://goo.gl/forms/DXEXZIUuxlYUKTBj2



Обратная связь по занятию #1 интенсива Ceph

Для нас очень важно, чтобы информация была понятна и все ответы исчерпывающими. Помогите пожалуйста нам сделать этот курс еще лучше для Вас!

Электронный адрес *

Действительный адрес эл. почты

Эта форма собирает адреса электронной почты респондентов. Изменить настройки

Понятность информации *

1 2 3 4 5 6 7 8 9 10