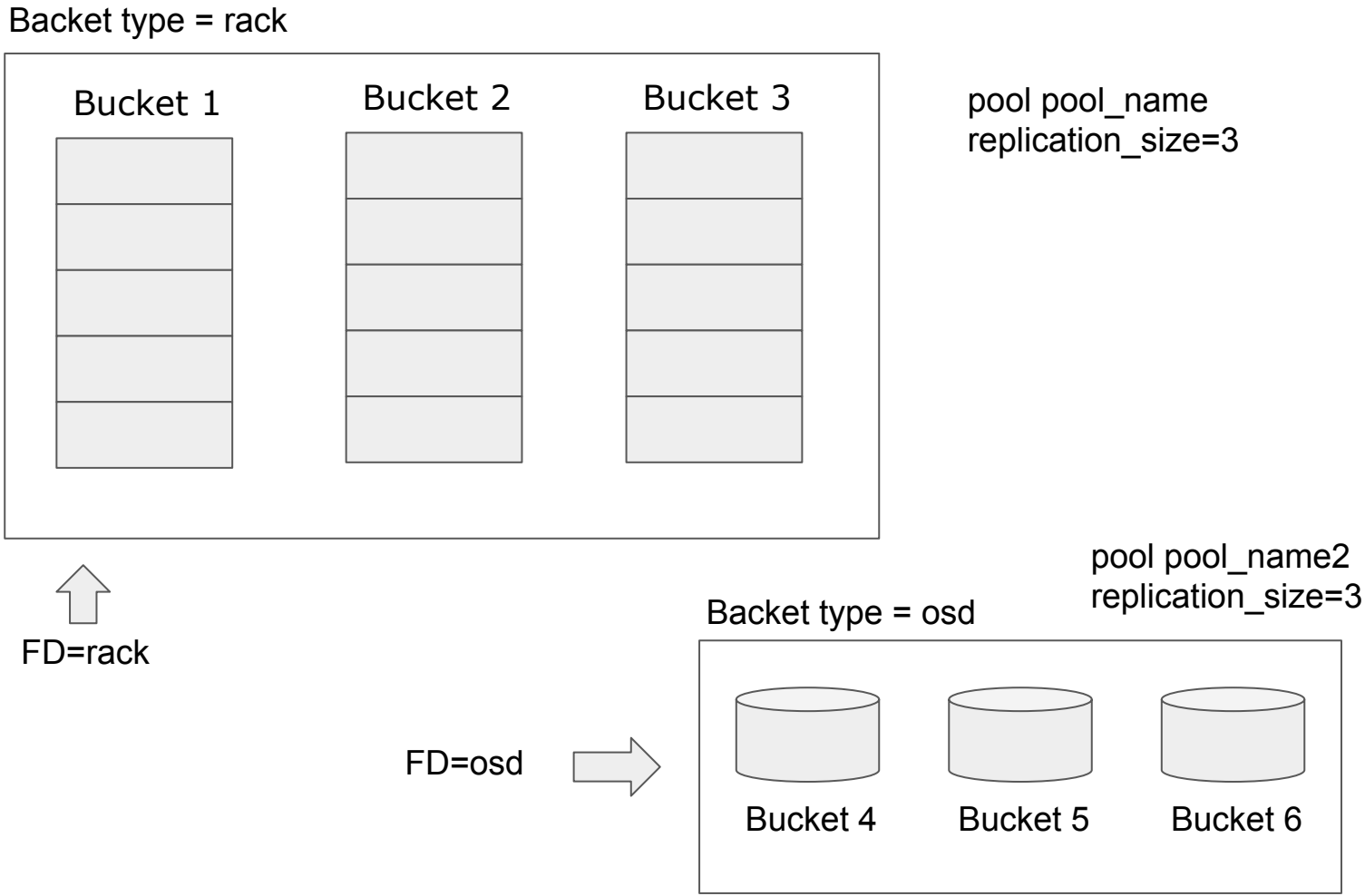
# Интенсив СЕРН

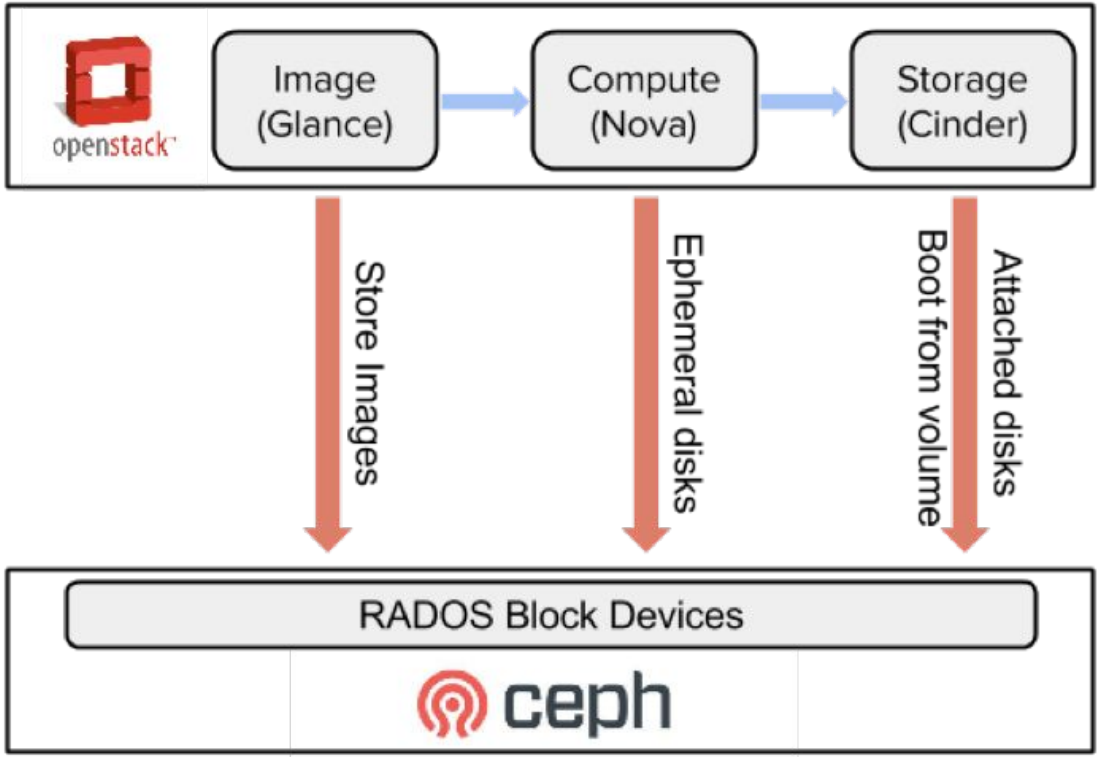
5ти-дневный интенсив
День №3

# Программа занятия

- Различные “крутилки” в ceph
- Связка с Openstack и варианты деплоя
- Ускорение производительности (разные параметры messenger, stripe, malloc)
- Мониторинг состояния
- Вывод ноды из кластера. Действия при падении ноды
- Ознакомление с авторизацией в CEPH
- Начальный дебаг проблем


Backet type = rack
Bucket 1
Bucket 2
Bucket 3
pool pool_name
replication_size=3
FD=rack
pool pool_name2
replication_size=3
Backet type = osd
FD=osd
Bucket 4
Bucket 5
Bucket 6

# OpenStack + Ceph



# OpenStack + Ceph

- 2 разных кластера не удобно
- 1 зона OpenStack - 1 Pool CEPH
- Ограничить IOPS. base image 200 IOPS block device 500 IOPS
- Снэпшоты != бекапы

# 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

# "Крутилки" scrub

osd_scrub_chunk_max = 1
osd_scrub_priority = 1
osd_scrub_begin_hour = 0
osd_scrub_end_hour = 4
osd_scrub_min_interval = 86400
osd_scrub_chunk_min = 1
osd_scrub_max_interval = 2419200.0
osd_scrub_sleep = 0.1
osd_deep_scrub_interval = 2419200.0

# Крутилки кешей

rbd_cache_max_dirty = 33554432
rbd_cache_size = 67108864
rbd_cache_max_dirty_age = 5
rbd_cache = True

# RBD object map feature

layering
striping
exclusive-lock
object-map

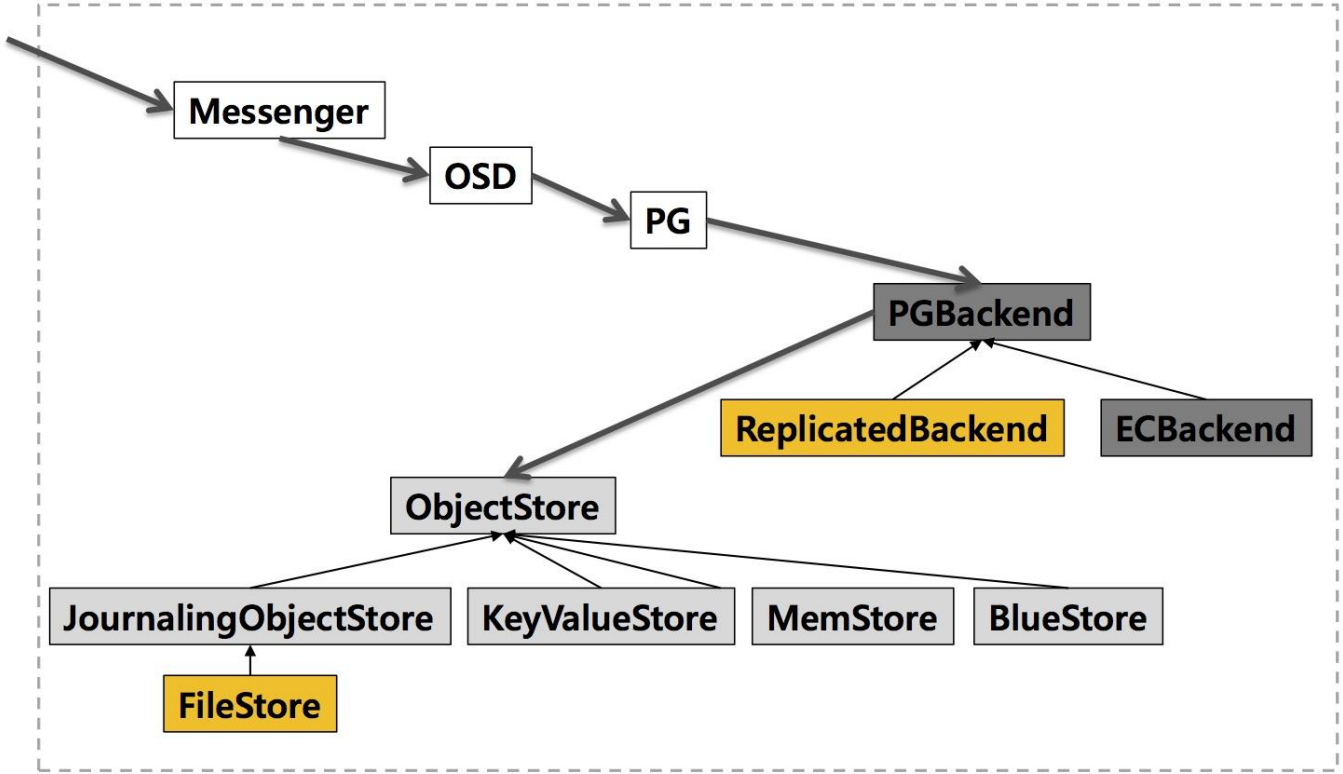
Включение через rbd_default_features = num

# Биты для установки rbd feature

В параметрах rbd_default_features указывается сумма бит.

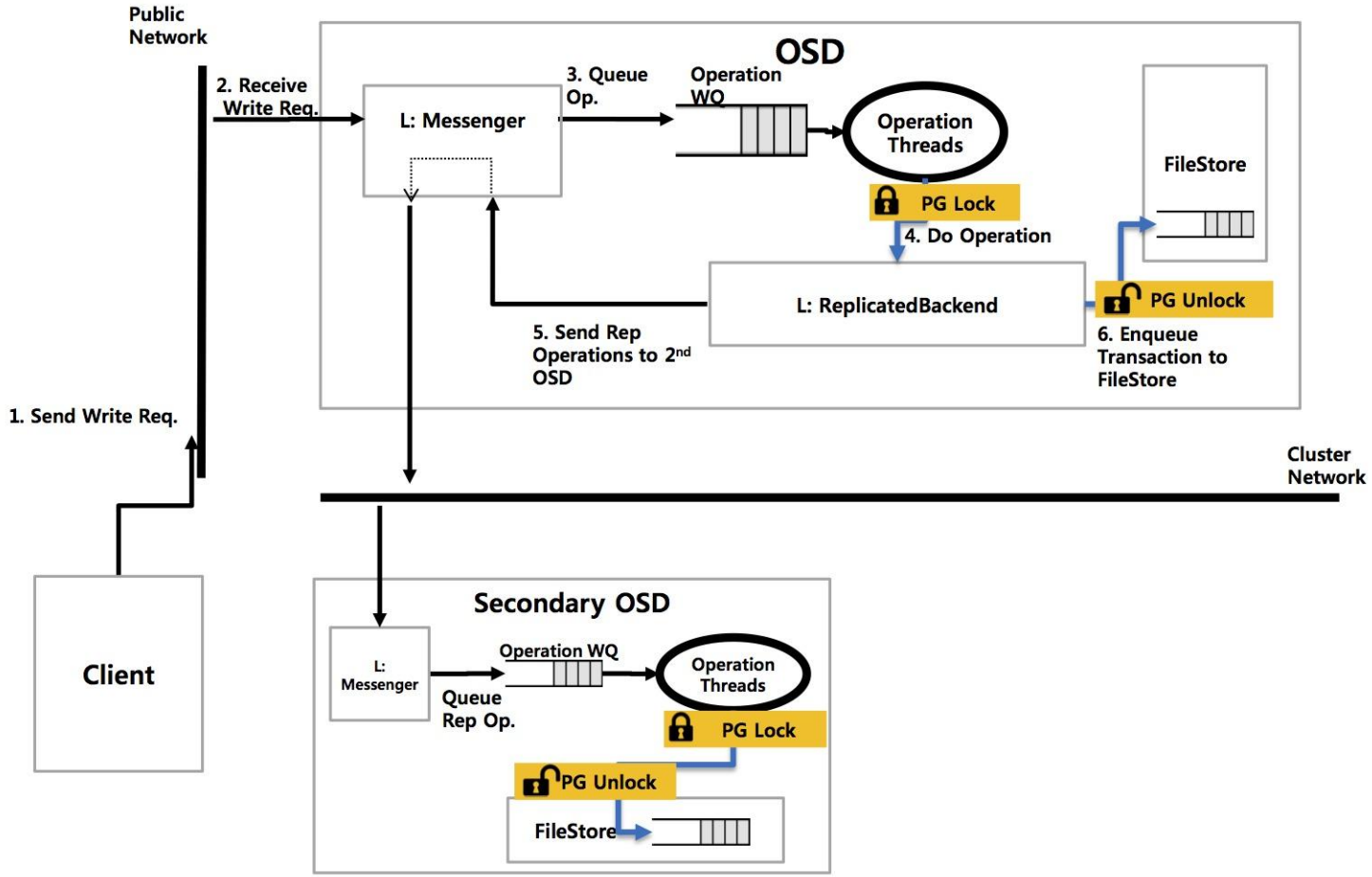
| 属性 | BIT码 |
| --- | --- |
| layering | 1 |
| striping | 2 |
| exclusive-lock | 4 |
| object-map | 8 |
| fast-diff | 16 |
| deep-flatten | 32 |

# Ceph IO Flow in OSD

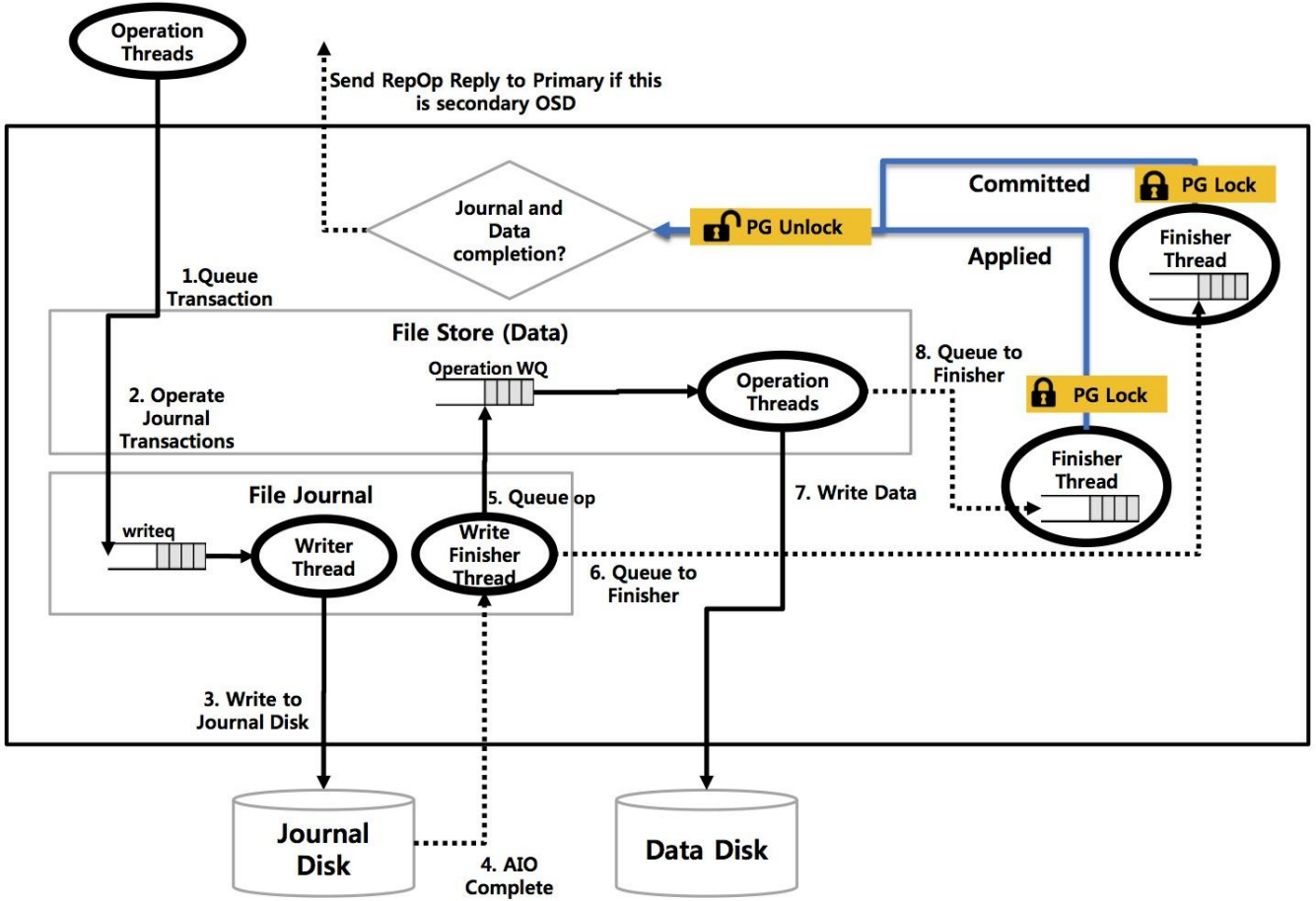



1. Journal: LIBAIO (O_DIRECT && O_DSYNC) ➔ Committed
2. Data: Buffered IO and syncfs() later ➔ Applied

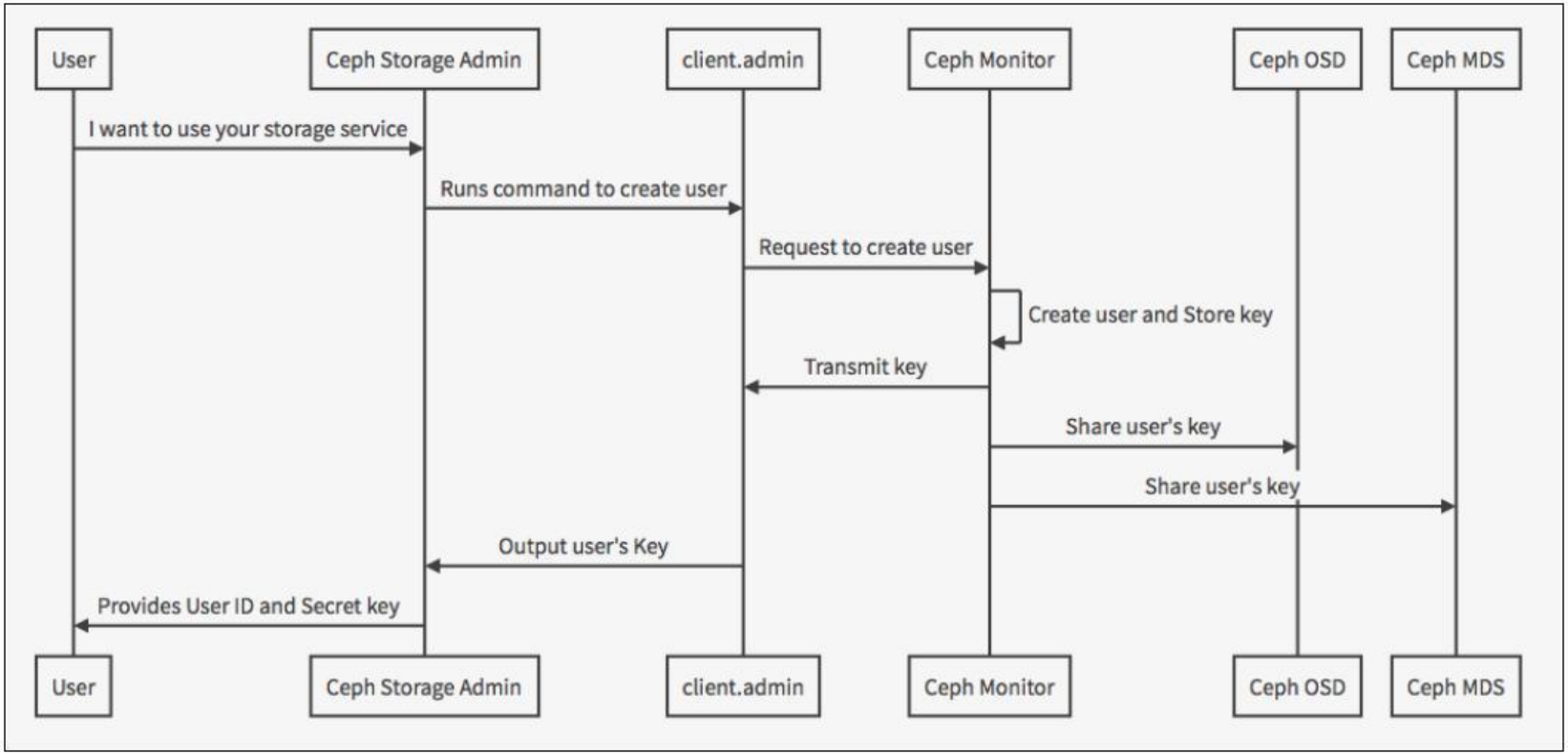
# Ceph Write IO Flow: Receiving Request



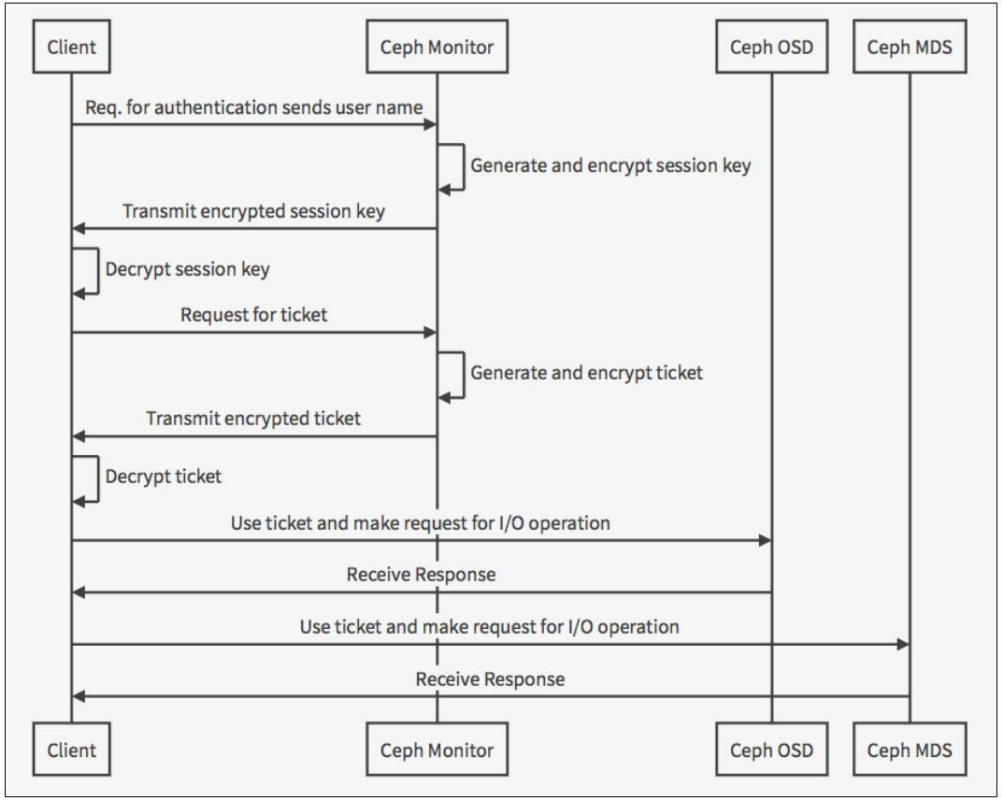
# Ceph Write IO Flow: in File Store



# Процесс создания пользователя и его ключа безопасности



# Процесс аутентификации



# Авторизация в CEPH

- Общий вид {daemon-type} 'allow {capability}' [{daemon-type} 'allow {capability}']
- Monitor caps: Включает параметры r, w, x, а также allow profiles {cap}
- OSD caps: Включает параметры r, w, x, class-read, class-write и profile osd.
- MDS caps: Требует только allow

# Что делать при падении

- Смотрим загрузку сети
- Смотрим потребление памяти
- Мониторим иопсы у клиентов
- Смотрим в логи ceph
- Регулируем
  - osd_recovery_op_priority
  - osd_recovery_threads
  - osd_client_op_priority

# Вывод ноды из кластера.

- Выставить weight на OSD ноды в 0
- Выполнить следующие команды:
  - ceph osd out osd.$i
  - systemctl stop ceph-osd.target
  - ceph osd crush remove osd.$i
  - umount /var/lib/ceph/osd/
  - ceph auth del osd.$i
  - ceph osd rm osd.$i

# Мониторинг CEPH

- Статус кластера OK Health Warn Health err
- warning на clock skew
- fs apply latency
- fs commit latency
- CEPH Cluster IOPS
- CEPH cluster IO (Mb\sec)
- Мониторить в логах наличие слова "slow request"

# Мониторинг сервера с CEPH

- Загрузка CPU
- Загрузка RAM
- Загрузка сетевых каналов.
- Заполненность дисков. warning на 85% Critical 92%


_write_bytes SGK4
Last 30 seconds
15:31:20
15:31:22
15:31:24
15:31:26
15:31:28
15:31:30
15:31:32
15:31:34
ceph_osd_perf_commit_latency_seconds
Last 30 seconds
1.3 ms
1.0 ms
750 µs
500 µs
250 µs
0 ns
15:31:05
15:31:10
15:31:15
15:31:20
15:31:25
15:31:30


CEPH-SGK4
Today R...
SGK4-write-iops
Last 30 seconds
40K iops
30K iops
20K iops
10K iops
0 iops
15:20:40
15:20:45
15:20:50
15:20:55
15:21:00
15:21:05
ceph_osd_perf_apply_latency_seconds
Last 30 seconds
125 ms
100 ms
75 ms
50 ms
25 ms
0 ns
15:20:40
15:20:45
15:20:50
15:20:55
15:21:00
15:21:05

# Начальный дебаг проблем. (FAQ)

**Логи:**

- /var/log/ceph/ceph.log
- /var/log/ceph/ceph-mon.NAME.log
- /var/log/ceph/ceph-osd.ID.log

# Начальный дебаг проблем. (FAQ)

**В) unable to mount OSD …**

О) Проверяем права на блочное устройство и существование точки монтирования

**В) unable to link journal…**

О) Проверяем права на симлинк journal в папке OSD. проверяем права на блочное устройство

**В) CEPH тормозит.**

О) Смотрим мониторинг, slow request, правильную работу журналов, iostat на нодах.

**В) У меня тут демон OSD падает с жутким трейсом!**

О) Вспоминаем, что меняли перед этими событиями, смотрим не было ли OOM, гуглим, идем в чат ceph_ru

Вопросы?

# Обратная связь

https://goo.gl/forms/rkW0knrPAWr4fRMB3