Интенсив СЕРН

5ти-дневный интенсив
День №4

ТЗ

- Описать задачу, которую хотите достичь
- Описать проблему, с которой столкнулись
- Прислать схему решения с описаниями количества СХД, дисков и тд
- Прислать конфиг: ceph.conf
- Все это прислать на radyus@yandex.ru

Программа занятия

- Ответы на вопросы
- RBD+Erasure coding
- Оптимизация ввода-вывода для хостов-инициаторов, multipathing и иже с ним
- Перекрестное журналирование. Как жить вообще без журналов?
- Bluestore. Плюсы и минусы (заглянем в будущее)
- CEPH in Docker

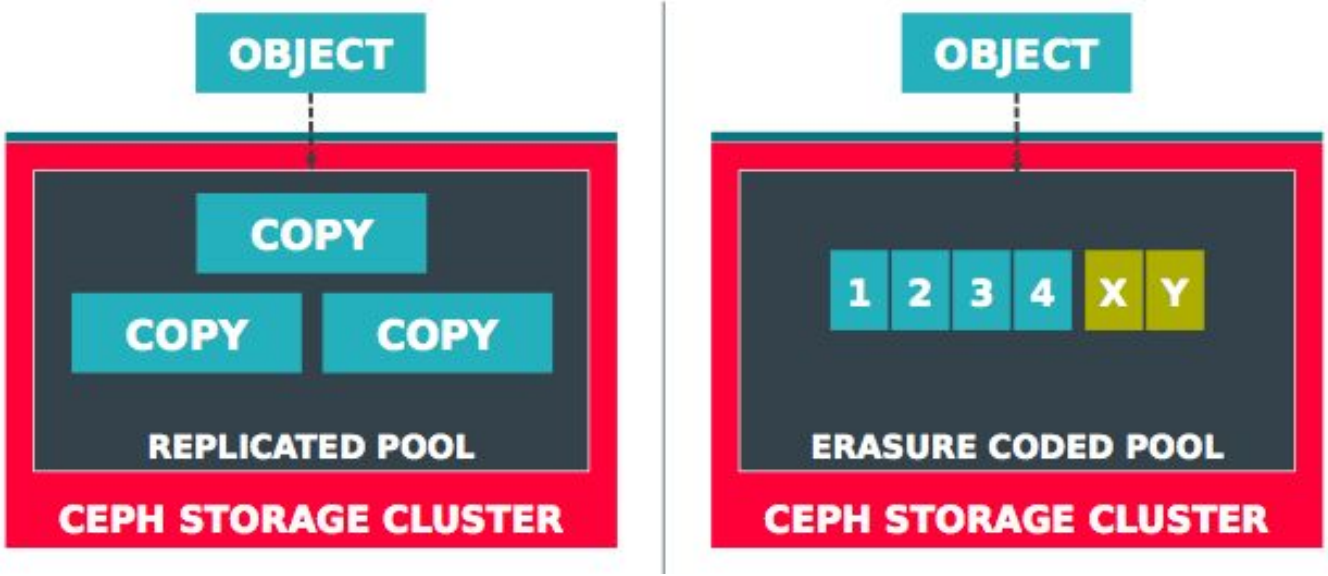
Вопрос для ответа на следующем занятии

Каким образом вы снимаете данные о загрузке сети на нодах(сервера, порты на свиче). Насколько потом эти данные реалистичны?

1. Верно ли что Ceph определяет которую из CRUSH map использовать по ее версии(epoch) среди мониторов? если да, то как достать CRUSH map из монитора, который вылетел из кластера (например, если текущая активная map у нас неверная а резервной копии нет)? 2. Были ли у Вас инциденты связанные с мониторами, не описанные в http://docs.ceph.com/docs/master/rados/troubleshooting/troubleshooting-mon/? 3. Есть ли какие-либо сторонние продукты по backup/restore для Ceph?

Звучала фраза что при апгрейде кластера, надежнее переезжать на соседний кластер, хотелось бы более подробно про переезд кластер-кластер, неужели надо иметь еще почти один набор такого же железа для подобных апгрейдов ?
Как удобнее переносить данные между кластерами ?
Дело в том что я собираюсь использовать изначально kraken с bluestore и подобные апгрейды мне вероятно предстоят относительно часто :)
Мой кейс - около 30T в geo-distributed object storage.
Еще вопрос, как ceph реагирует на например "зависший контроллер" ?
Есть например нода с подобной проблемой и иногда она может минут 20-30(пока дежурные реагируют) лежать, сразу ли начнется rebalance ?
Какие есть рекоммендации к настройке, что б возможно rebalance начинался не сразу ?

Erasure coding



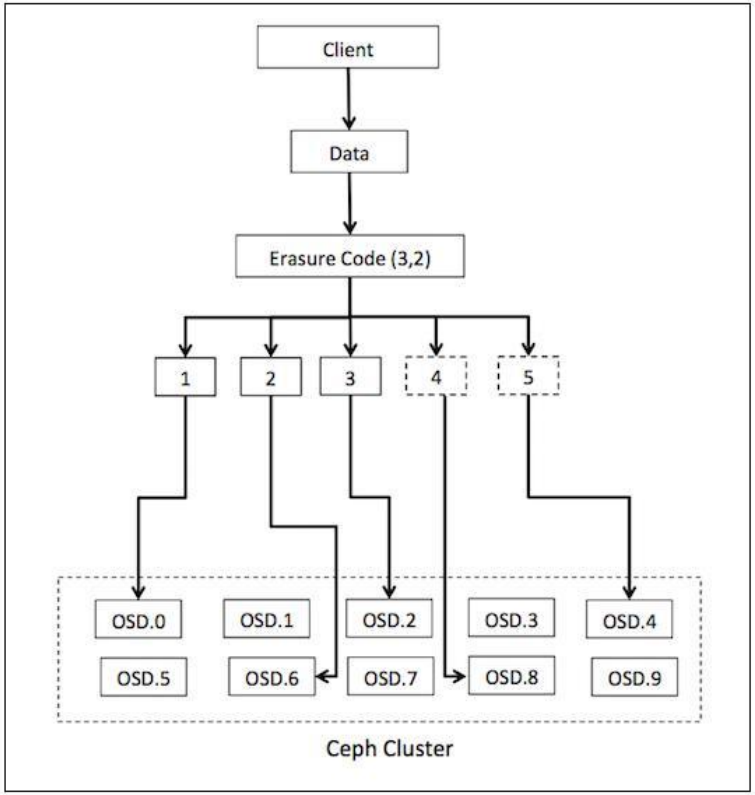
Полные копии хранимых объектов

- Очень высокая прочность
- 3x (200% накладных расходов)
- Более быстрое восстановление

Одна копия + паритет

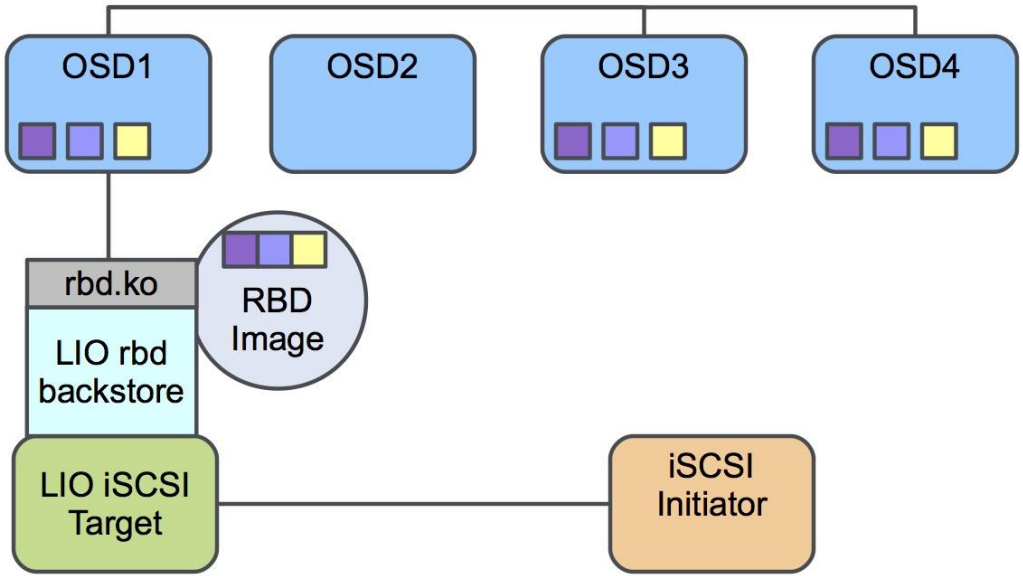
- Экономичная долговечность
- 1,5x (50% накладных расходов)
- Дорогое восстановление

RBD+Erasure coding



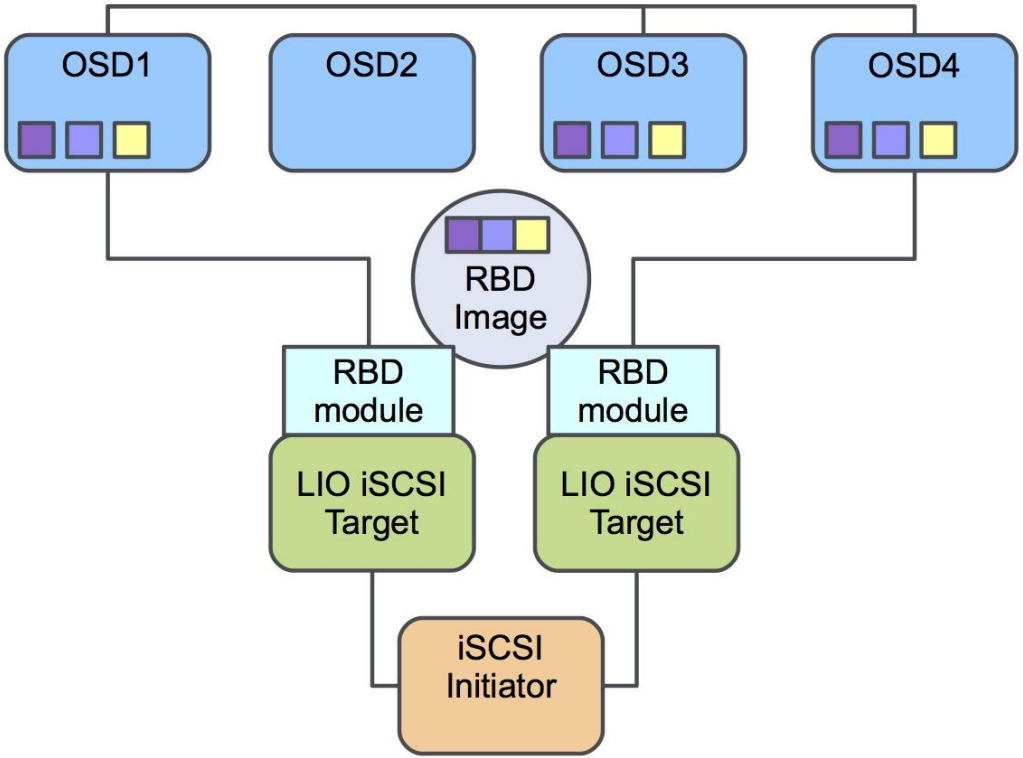
RBD iSCSI gateway

The Ceph View



LIO using RBD iSCSI gateway

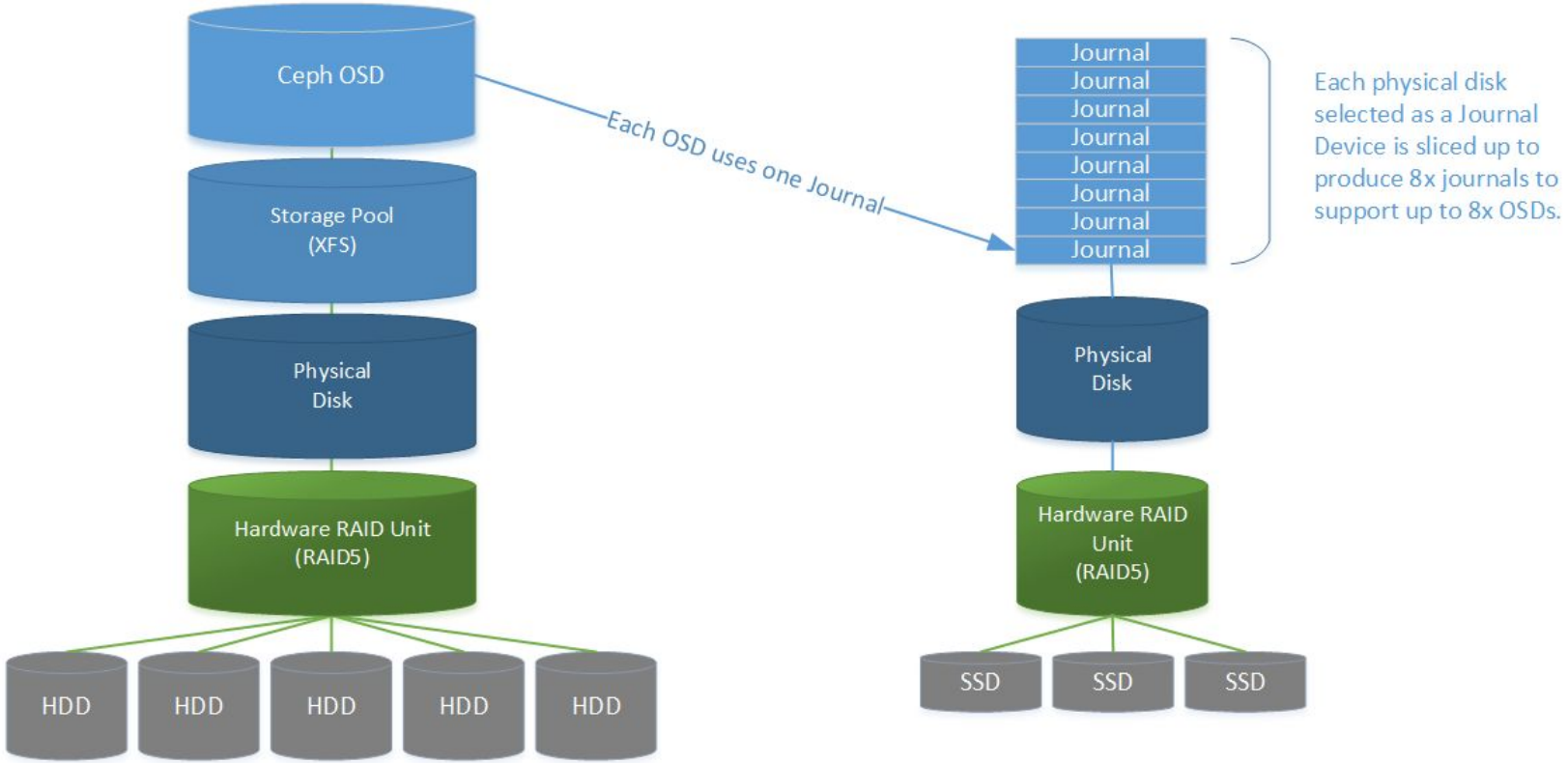
Multipath Support



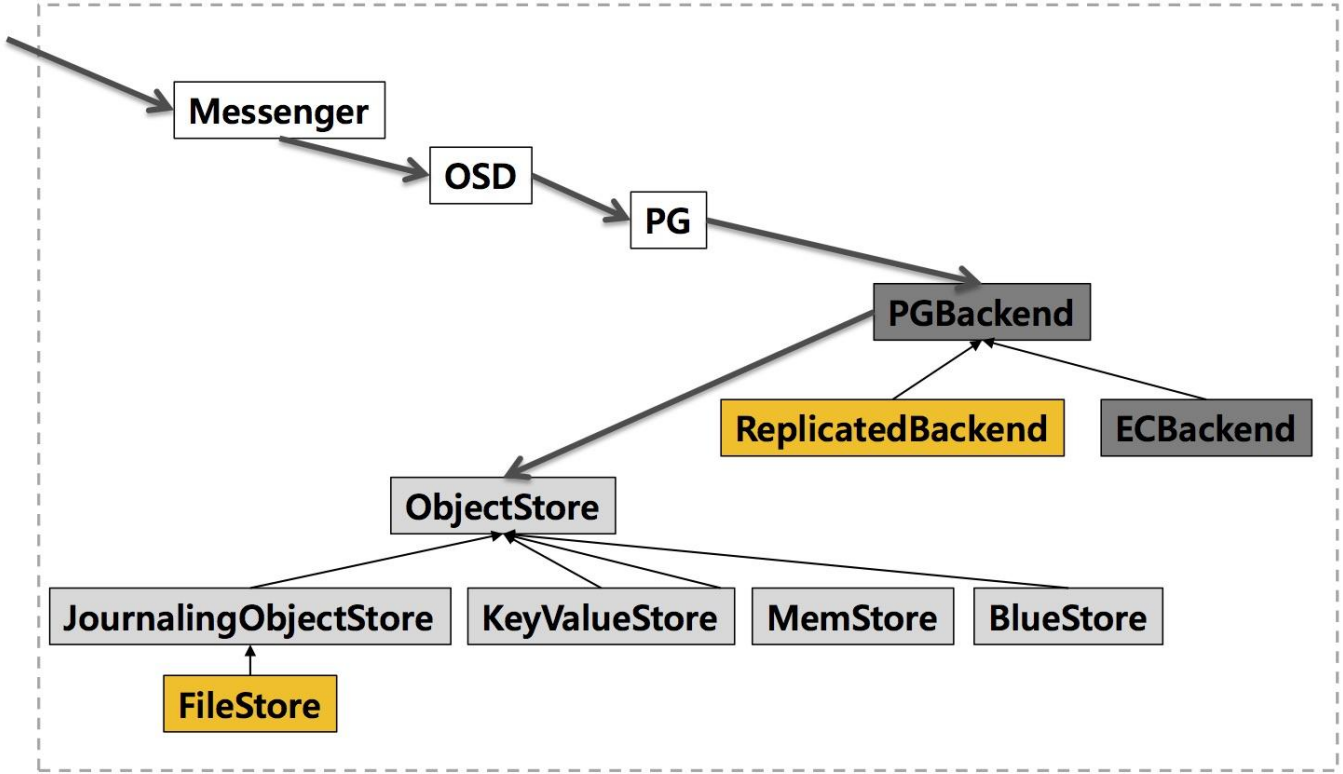
Варианты ISCSI

- tgtd + RBD - медленно.
- LIO + RBD - нет в centos.
- SPDK + RBD - вообще не понятно как ставить.

Журналы Ceph

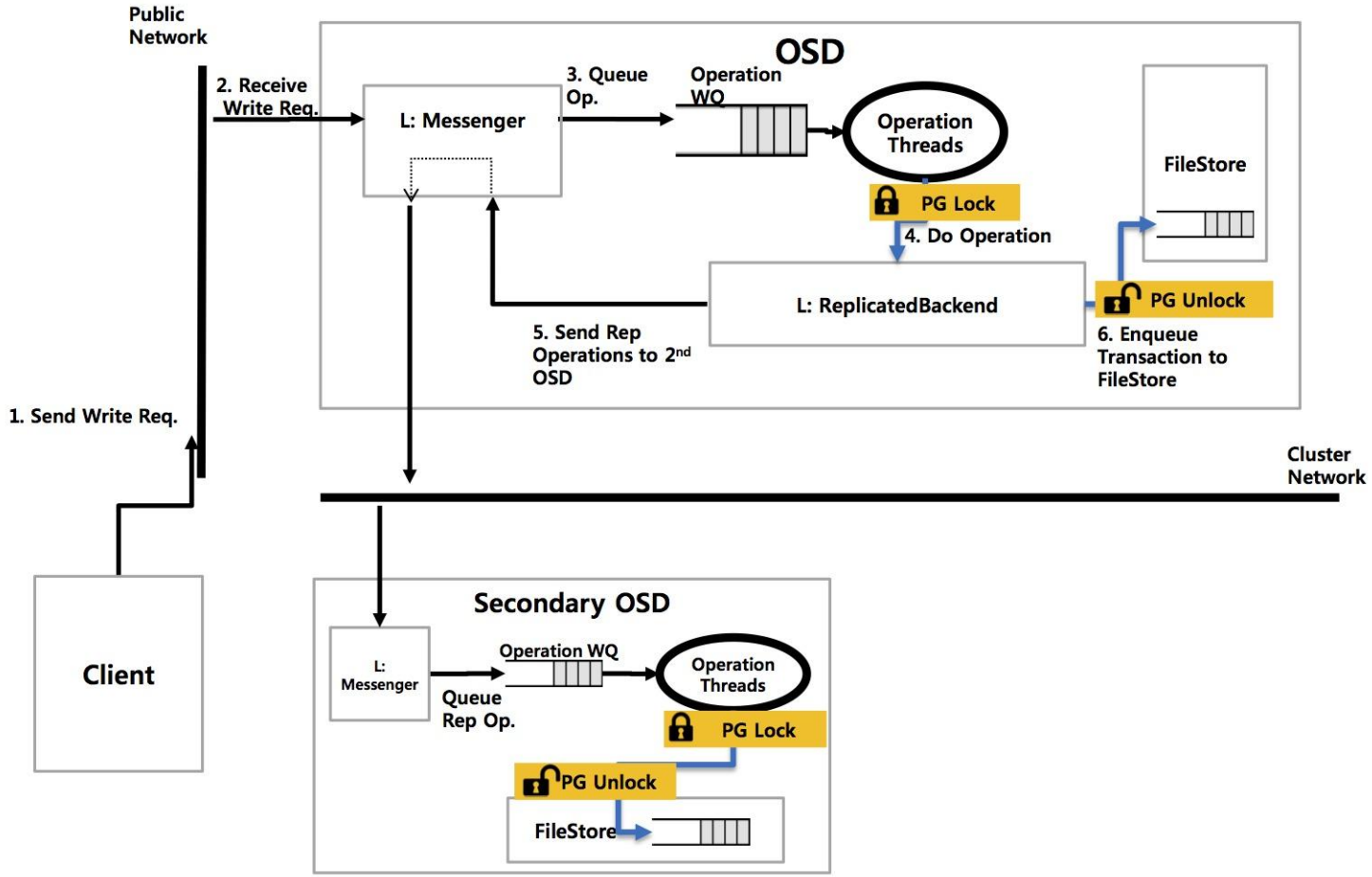


Ceph IO Flow in OSD

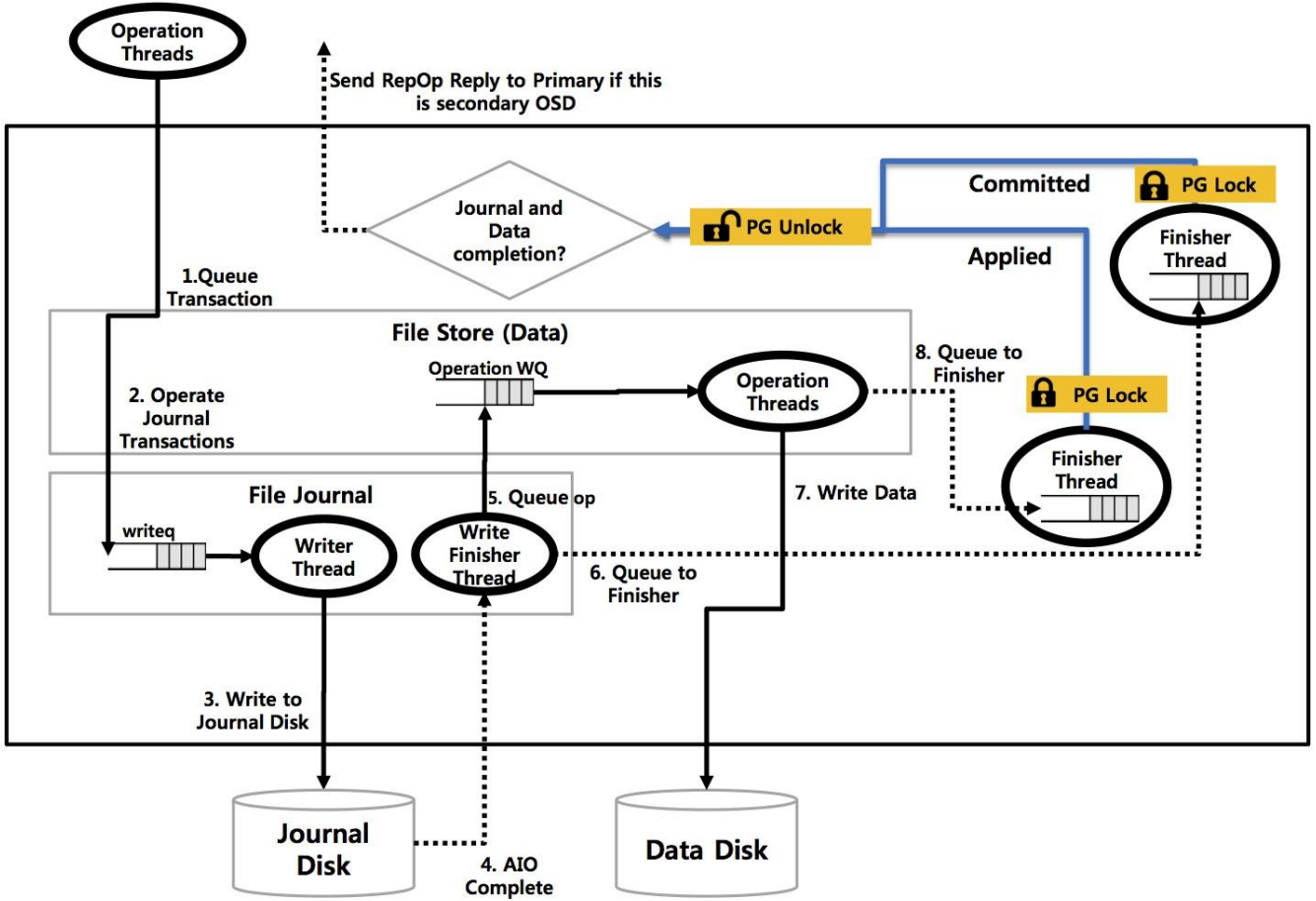


1. Journal: LIBAIO (O_DIRECT && O_DSYNC) ➔ Committed
2. Data: Buffered IO and syncfs() later ➔ Applied

Ceph Write IO Flow: Receiving Request



Ceph Write IO Flow: in File Store



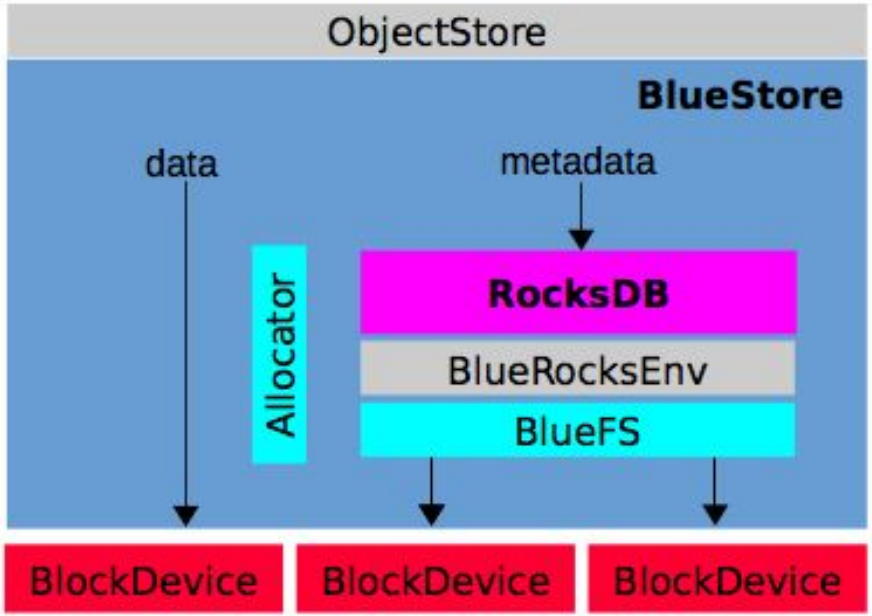
Bluestore

BlueStore = Block + NewStore

- consume raw block device(s)
- key/value database (RocksDB) for metadata
- data written directly to block device
- pluggable block Allocator

We must share the block device with RocksDB

- implement our own rocksdb::Env
- implement tiny “ le system” BlueFS
- make BlueStore and BlueFS share



CEPH in Docker

ceph-docker

Containerizing Ceph daemons

The project:

- Launched on Jan 18, 2015
- Upstream project: https://github.com/ceph/ceph-docker
- Support from Hammer to the latest version of Ceph (currently Jewel)
- Wide range of distros: Ubuntu (14.04 and 16.04), Fedora (24), CentOS (7)
- Automated builds on the Docker Hub
- More than 500K+ pulls!

Вопросы?

Обратная связь

https://goo.gl/forms/viX0fw45H4I2z9W33