



Онлайн-образование

Не забыть включить запись!





Меня хорошо видно && слышно?

Ставьте ☐+, если все хорошо
Напишите в чат, если есть проблемы

Правила вебинара



Активно участвуем



Задаем вопрос в чат или голосом



Off-topic обсуждаем в Slack #канал группы или #general



Вопросы вижу в чате, могу ответить не сразу



СЕРН

Викирюк Павел

Системный инженер

Маршрут вебинара

СЕРН: архитектура



СЕРН: развертывание кластера



CephFS

Цели занятия | После занятия вы сможете

- 1 Познакомиться с распределенной файловой системой CERN
- 2 Разобраться с основными моментами развертывания кластера CERN
- 3 Понять аспекты использования файловой системы CephFS

Смысл | Зачем вам это уметь

1 Чтобы получить минимальный набор знаний для работы с файловой системой CERN

2 Чтобы уметь самостоятельно развернуть кластер CERN

3 Чтобы научиться использовать распределенные файловые системы для любых нужд



Основные свойства СХД

Основные свойства СХД

Требования, которые предъявляются к современным СХД:

- высокая надежность хранения
- высокая доступность данных, то есть минимальное время простоя при авариях
- высокая скорость доступа и минимальные задержки
- низкая удельная стоимость хранения
- масштабирование
- различные прикладные возможности: клонирование, снимки состояния и т.д.



Вопрос к аудитории:
Какие системы хранения данных
отвечают условиям выше?



СЕРН: архитектура

СЕРН: архитектура

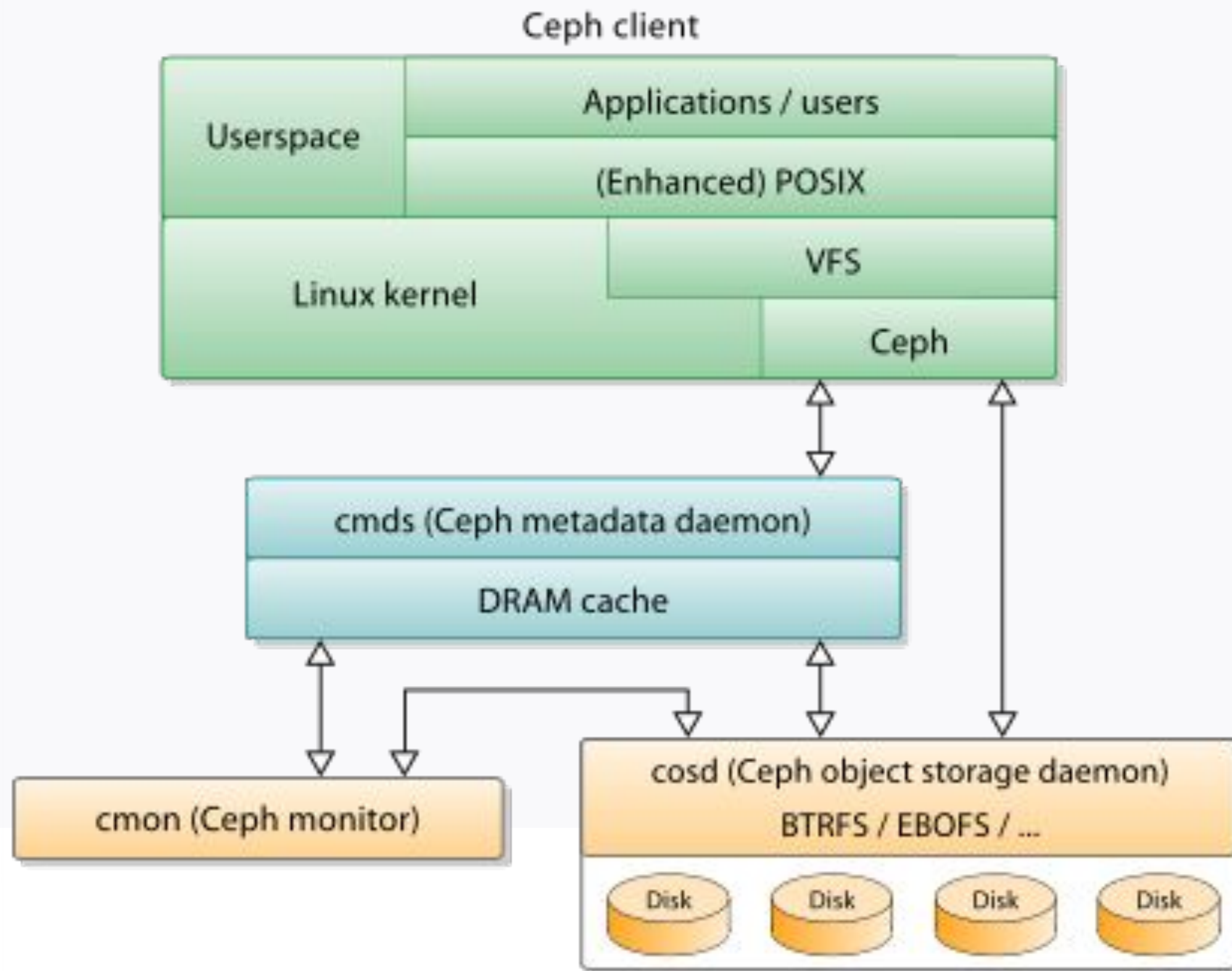
Ceph — свободная программная объектная сеть хранения (англ. object storage), обеспечивающая как файловый, так и блочный интерфейсы доступа

<https://ru.wikipedia.org/wiki/Ceph>

Особенности:

- может использоваться на системах, состоящих как из нескольких Linux-машин, так и из тысяч узлов
- известно об эксплуатации систем на Ceph как размером в сотни петабайт, так и скромнее (Yahoo!, Mail.ru, КРОК)
- встроенные механизмы продублированной репликации данных обеспечивают высокую живучесть системы
- при добавлении или удалении новых узлов массив данных автоматически перебалансируется с учётом изменений

СЕРН: архитектура



Ceph: архитектура

Абстракции для работы с хранилищем Ceph:

Абстракция объектного хранилища (RADOS Gateway, или RGW) при вместе с FastCGI-сервером позволяет использовать Ceph для хранения пользовательских объектов и предоставляет API, совместимый с S3/Swift

Абстракция блочного устройства (RADOS Block Device, или RBD) предоставляет пользователю возможность создавать и использовать виртуальные блочные устройства произвольного размера. Программный интерфейс RBD позволяет работать с этими устройствами в режиме чтения/записи и выполнять служебные операции — изменение размера, клонирование, создание и возврат к снимку состояния итд.

Гипервизор QEMU содержит драйвер для работы с Ceph и обеспечивает виртуальным машинам доступ к блочным устройствам RBD. Поэтому Ceph сейчас поддерживается во всех популярных решениях для оркестрации облаков — OpenStack, CloudStack, ProxMox

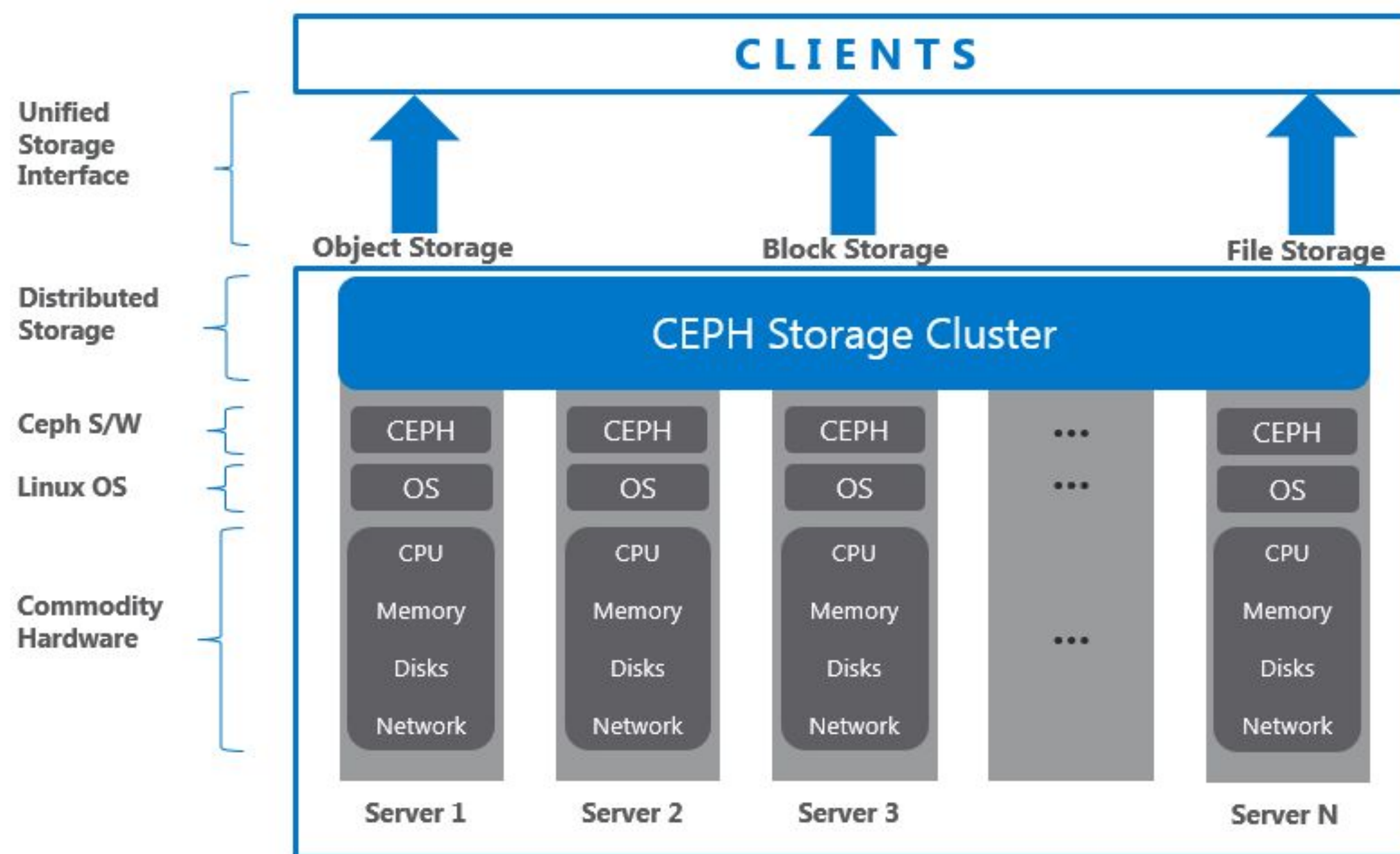
CEPH: архитектура

Абстракции для работы с хранилищем CEPH:

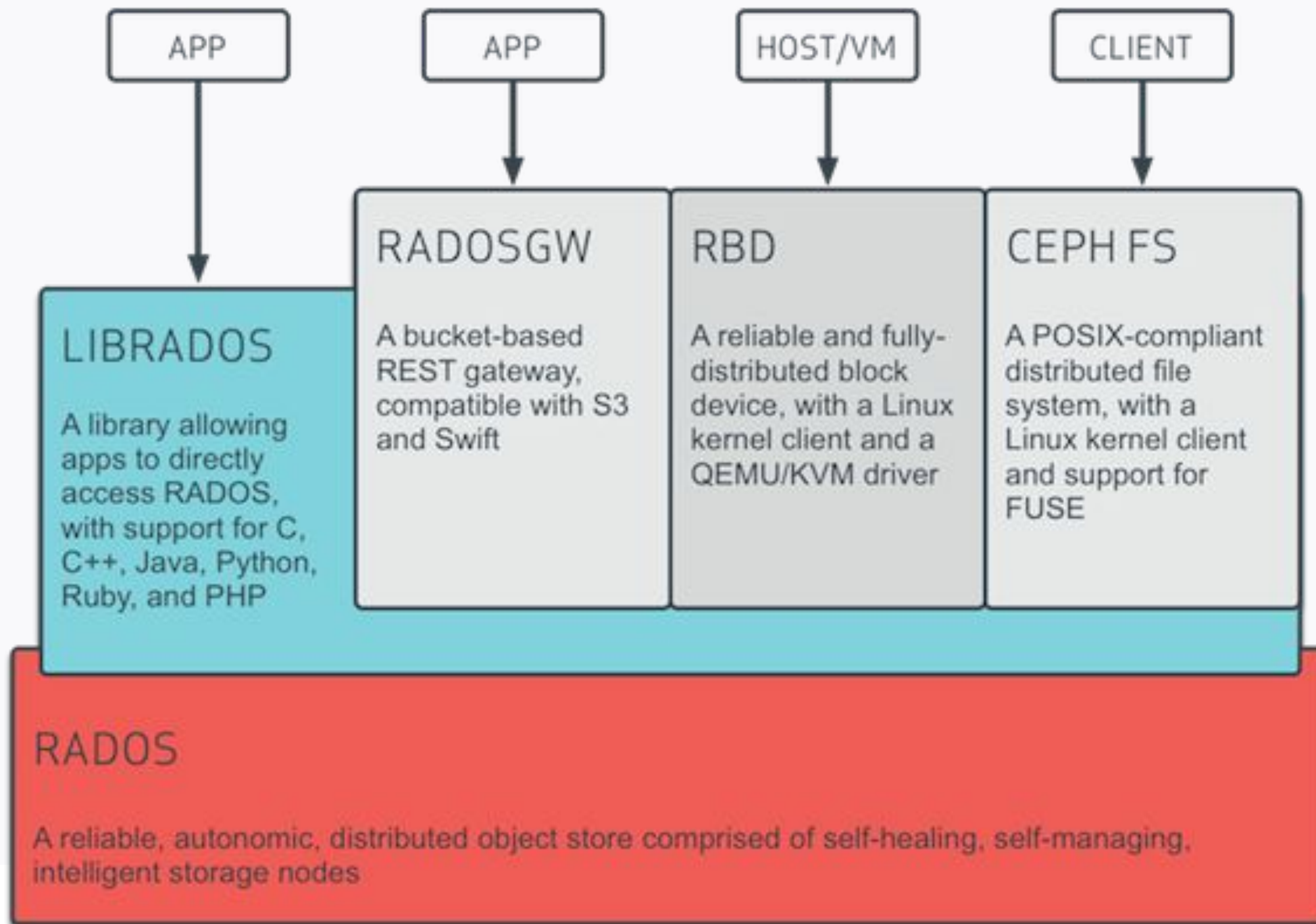
Абстракция POSIX-совместимой файловой системы

CephFS — POSIX-совместимая файловая система, использующая Ceph в качестве хранилища

СЕРН: архитектура



СЕРН: архитектура



СЕРН: архитектура

Основные термины и элементы:

Metadata server (MDS) - вспомогательный демон для обеспечения синхронного состояния файлов в точках монтирования CephFS

- работает по схеме активная копия + резервы
- активная копия в пределах кластера только одна

Mon (Monitor) - демон, выполняющий роль координатора и формирующий кластер

- хранит информацию о состоянии кластера
- обменивается с другими мониторами информацией о том, на какие OSD писать и с каких читать данные
- может быть один, но лучше поднимать несколько мониторов
- количество мониторов должно быть нечетным (split-brain)
- если потеряна связь с половиной мониторов - кластер блокируется для предотвращения рассогласованности данных

СЕРН: архитектура

Основные термины и элементы:

OSD (Object Storage Device) - основной элемент хранения данных в СЕРН

- хранит данные и обменивается ими с другими OSD
- обычно OSD - это диск
- за диск отвечает отдельный OSD демон, работающий на сервере с ЭТИМ ДИСКОМ

Объект (Object) - блок данных фиксированного размера

- по-умолчанию равен 4 Мб
- не имеет ничего общего с пользовательскими объектами из Object Storage

СЕРН: архитектура

Основные термины и элементы:

Карта OSD (OSD Map) - карта, ассоциирующая каждой **placement group** набор из одной Primary OSD и одной или нескольких Replica OSD

- распределение placement groups (PG) по нодам хранилища OSD описывается срезом карты osdmap, в которой указаны положения всех PG и их реплик
- каждое изменение расположения PG в кластере сопровождается выпуском новой карты OSD, которая распространяется среди всех участников кластера

СЕРН: архитектура

Основные термины и элементы:

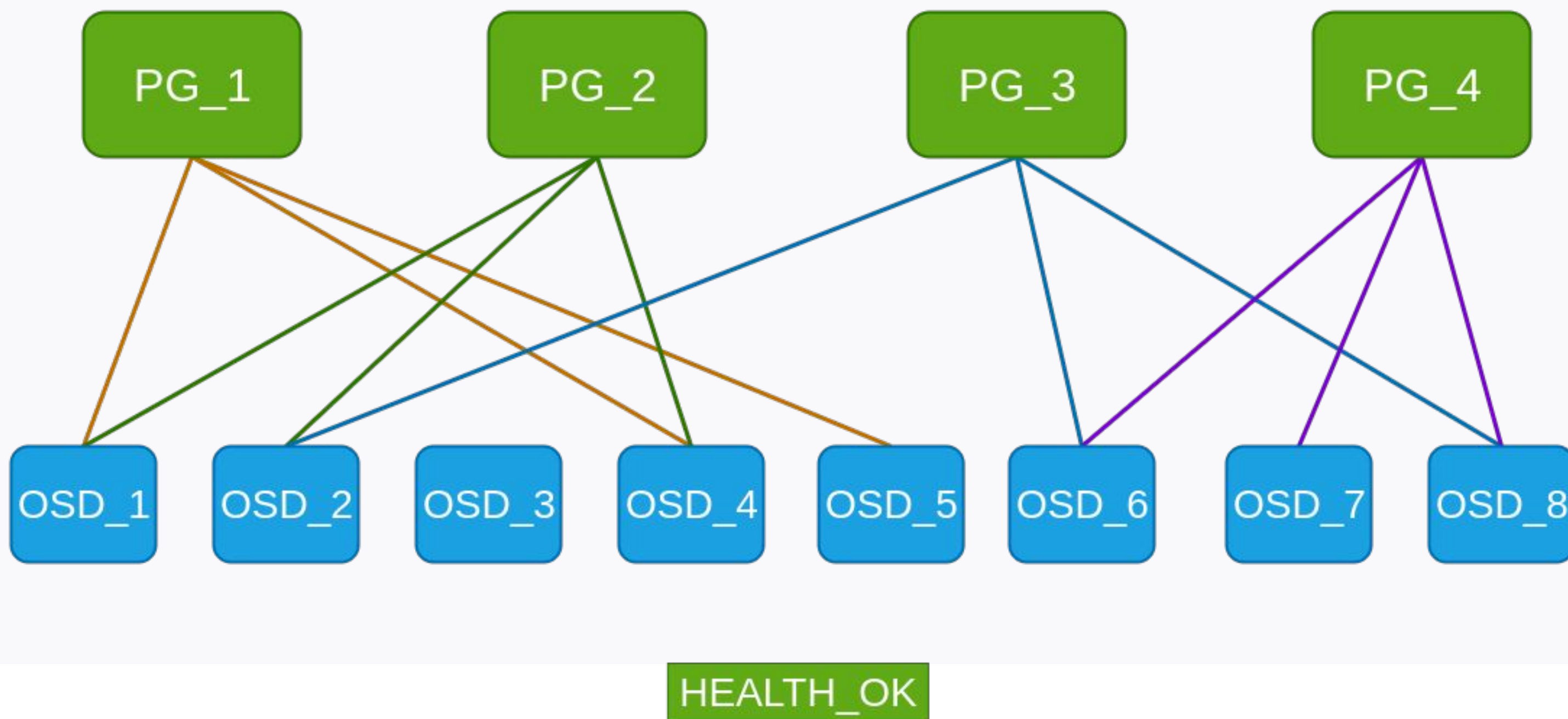
Placement Group (PG) - логическая группа, объединяющая множество объектов, предназначенная для упрощения адресации и синхронизации объектов

- каждый объект состоит лишь в одной плейсмент группы
- количество объектов, участвующих в плейсмент-группе, не регламентировано и может меняться.

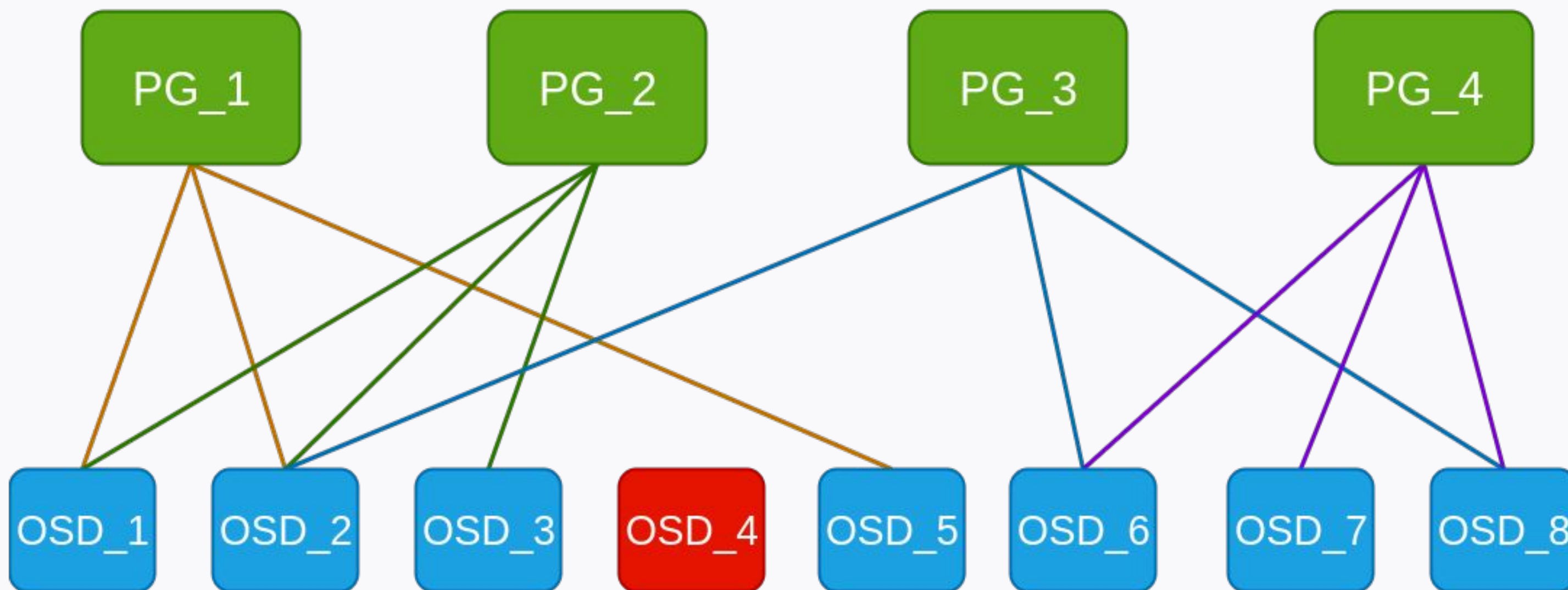
Primary OSD - OSD, выбранный в качестве **Primary** для данной **PG**

- клиент всегда обслуживается тем OSD, который является Primary для **PG**, в которой находится интересующий клиента блок данных (объект)
- Primary OSD в асинхронном режиме реплицирует все данные на **Replica OSD**

СЕРН: архитектура



СЕРН: архитектура



HEALTH_WARN

СЕРН: архитектура

Основные термины и элементы:

Replica OSD (Secondary) - OSD, которая не является Primary для данной PG и используется для репликации. Клиент никогда не общается Replica OSD напрямую

Фактор репликации (RF) - избыточность хранения данных. Фактор репликации является целым числом и показывает, сколько копий одного и того же объекта хранит кластер



CEPH: CRUSH

Ceph: CRUSH

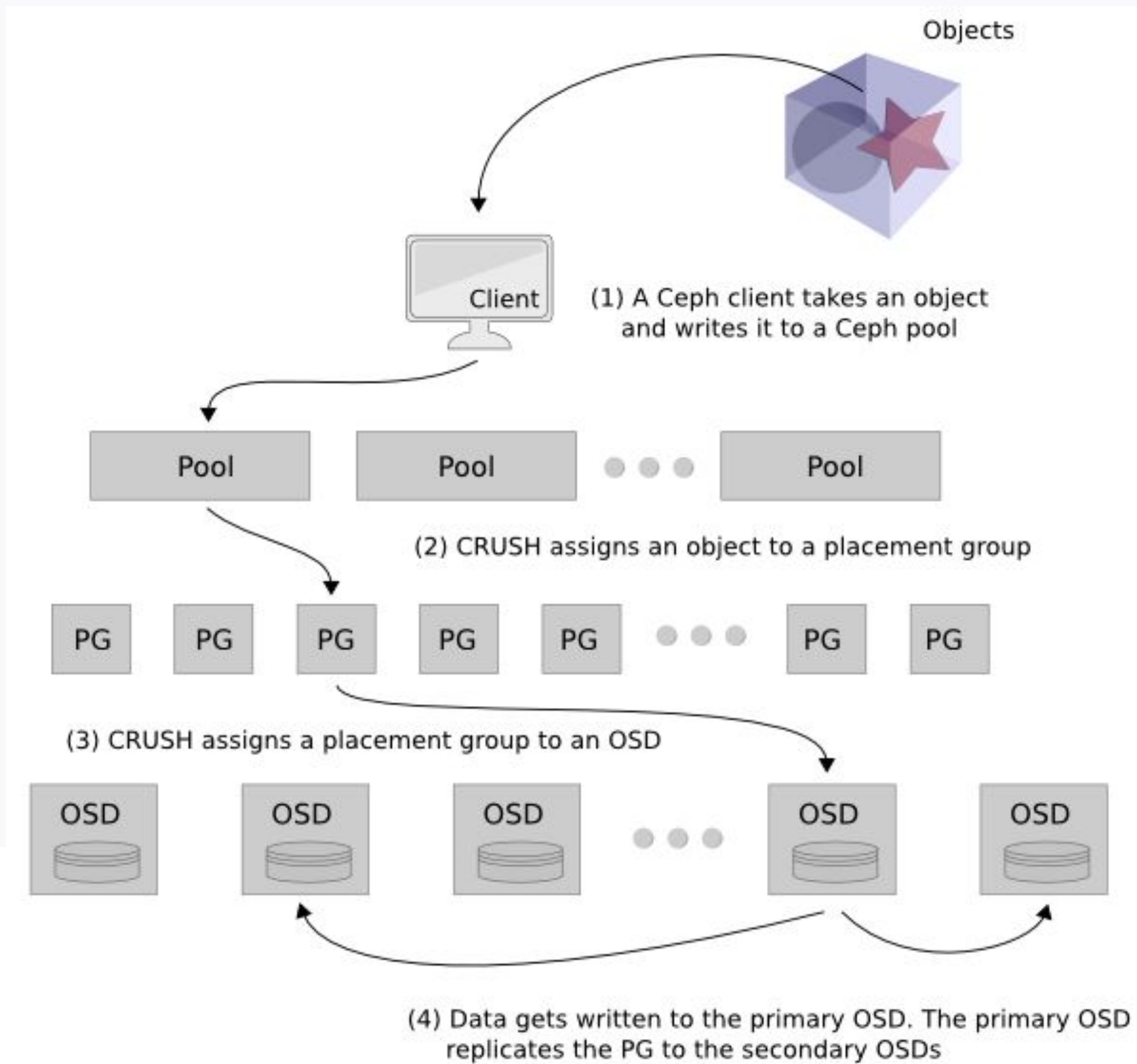
Основные термины и элементы:

CRUSH (Controlled Replicated Under Scalable Hashing) - алгоритм, позволяющий однозначно определить местоположение объекта на основе хэша имени объекта и определенной карты, которая формируется исходя из физической и логической структур кластера

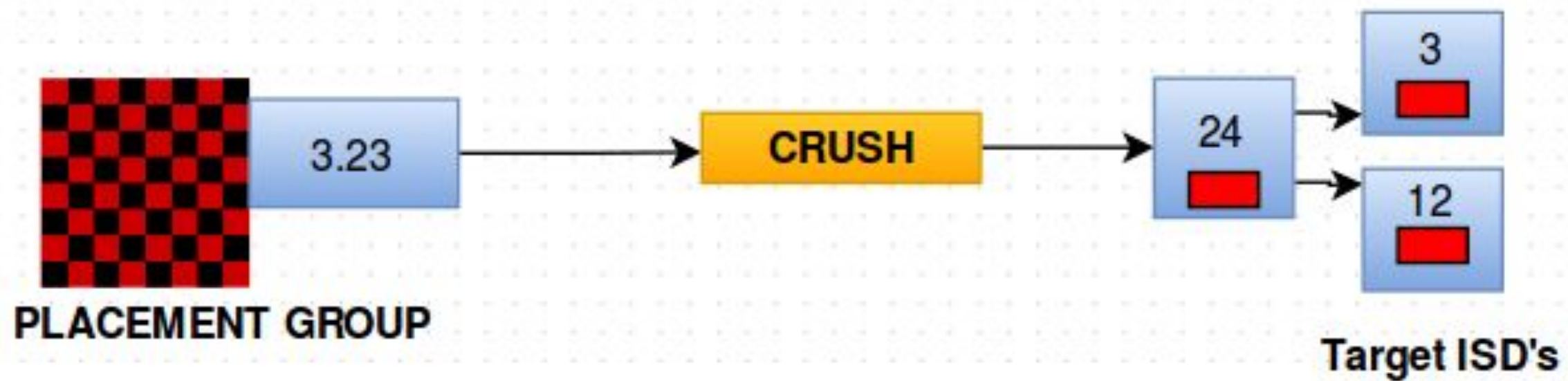
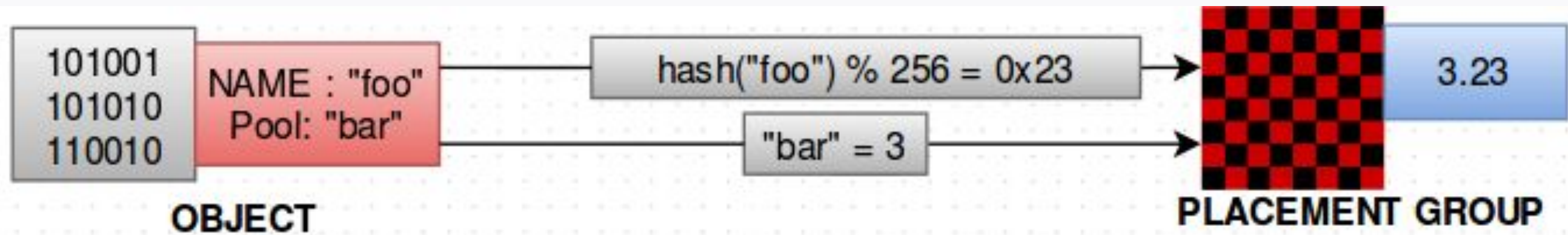
Особенности:

- карта не включает в себя информацию о местонахождении данных
- путь к данным каждый клиент определяет сам с помощью CRUSH и актуальной карты, которую отдает демон монитора
- при добавлении или падении элементов кластера карта обновляется автоматически
- по-умолчанию карта плоская, однако ее можно редактировать и менять некоторые параметры в конфигурации

CEPH: CRUSH



CEPH: CRUSH

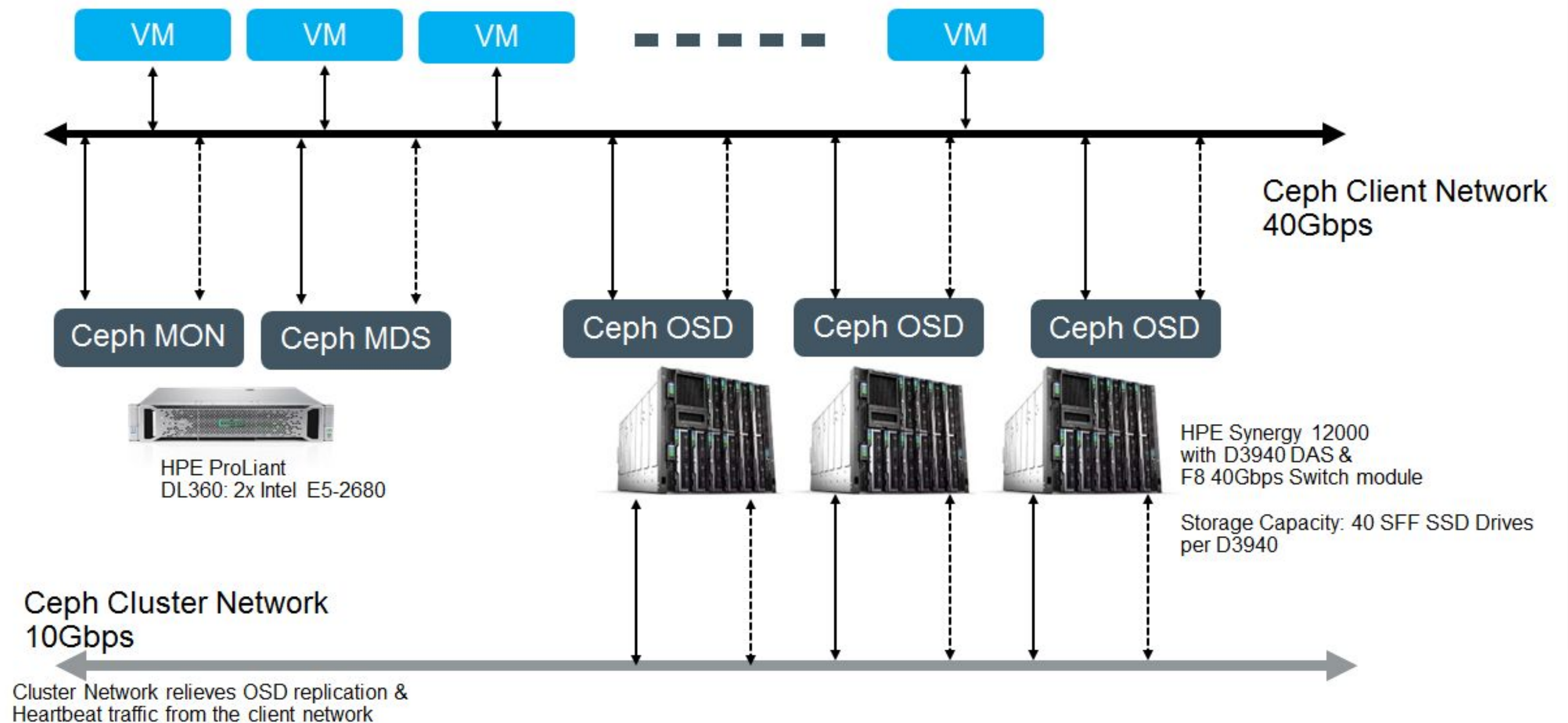


The background of the slide is a blue-tinted aerial photograph of a dense city skyline, likely New York City. Overlaid on this image is a network diagram consisting of numerous small blue dots connected by thin, light blue lines, creating a web-like pattern across the center of the slide. The text "СЕРН: сеть" is centered within this network area.

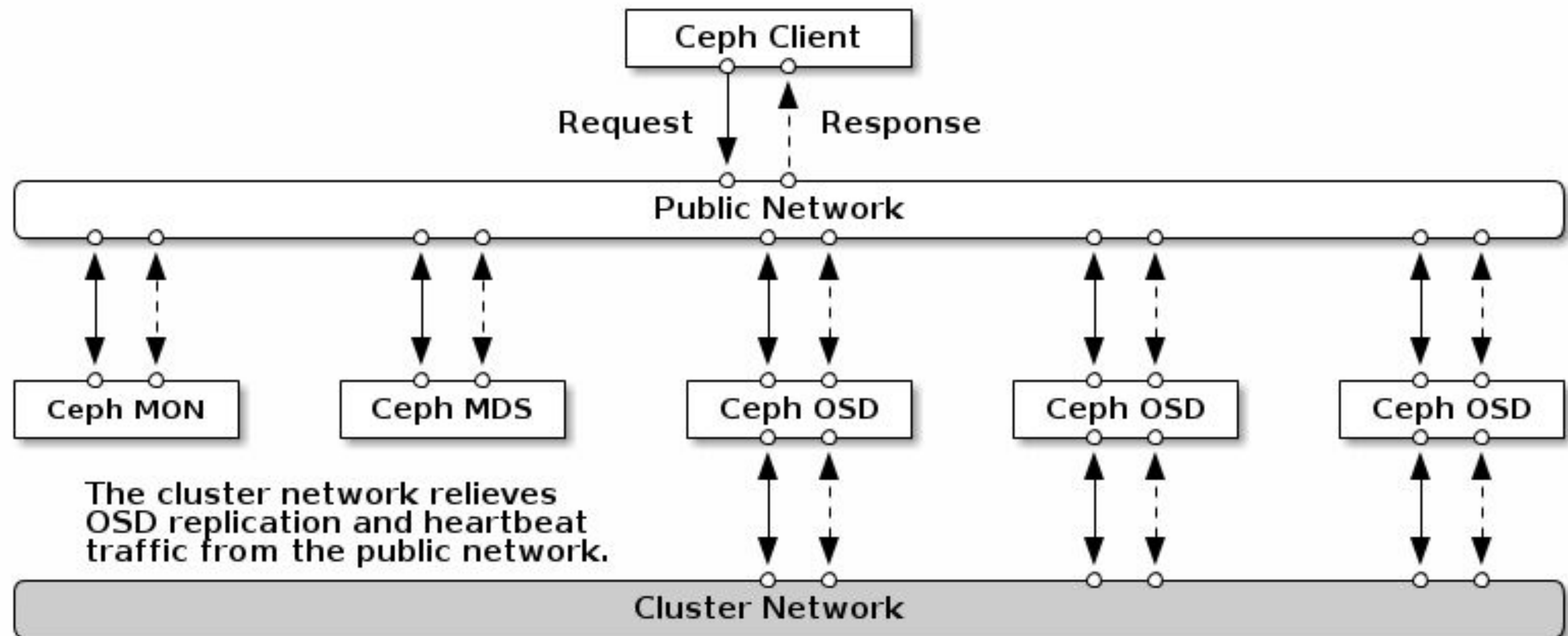
СЕРН: сеть

CEPH: сеть

Network Architecture for Ceph Storage



СЕРН: сеть



СЕРН: сеть

Особенности работы СЕРН с сетью:

- принято разделять сети на **public network** (сеть через которую с СЕРН взаимодействуют пользователи) и **cluster network** (сеть через которую взаимодействуют между собой элементы кластера)
- рекомендуемая производительность **cluster network** - не менее 10 Гбит/сек, а лучше порядка 40 Гбит/сек



Ваши вопросы?

Маршрут вебинара

СЕРН: архитектура



СЕРН: развертывание кластера



CephFS

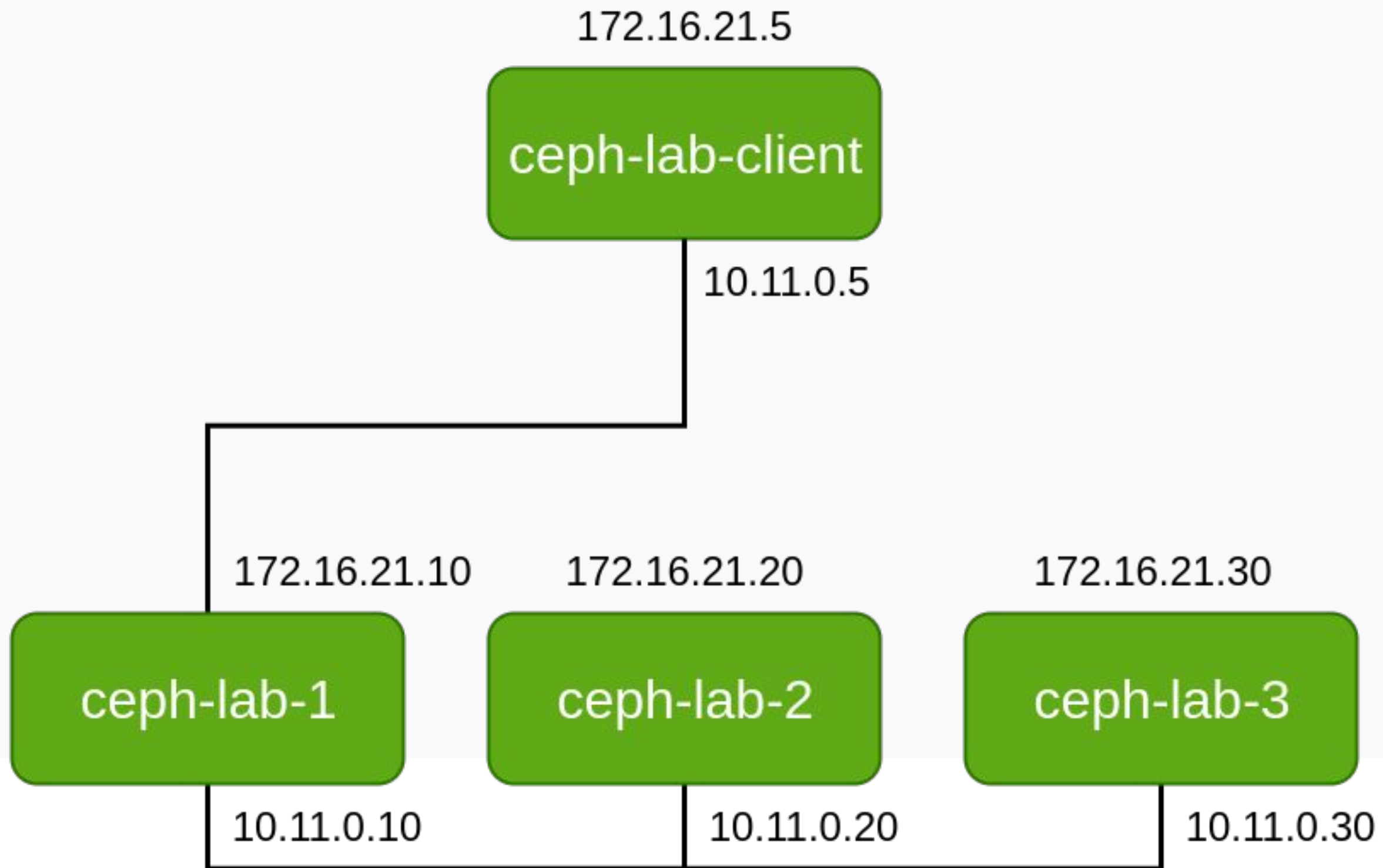


СЕРН: развертывание кластера



Схема тестового стенда

Схема тестового стенда



СЕРН: развертывание кластера

Подготовка:

- установка репозитория EPEL и СЕРН
- создание пользователя serh (или любого другого)
- деплой ssh ключей и редактирование sudoers
- DNS или hosts (хосты должны резолвить друг друга)
- NTP с единым сервером - для продакшна обязательно

СЕРН: развертывание кластера

Разворачиваем кластер:

Логинимся под пользователем ceph:

```
sudo -u ceph -i  
mkdir test-cluster && cd test-cluster
```

Формируем конфиг для развертывания кластера:

```
ceph-deploy new ceph-lab-1 ceph-lab-2 ceph-lab-3
```

Добавляем параметры задающие размер пула osd и позволяющие игнорировать отклонения нод кластера по времени:

```
echo 'osd_pool_default_size = 2' >> ceph.conf  
echo 'mon_clock_drift_allowed = 1' >> ceph.conf
```


СЕРН: развертывание кластера

Разворачиваем кластер:

Устанавливаем необходимую версию СЕРН на ноды кластера:

```
ceph-deploy install --release=nautilus ceph-lab-client ceph-lab-1 ceph-lab-2  
ceph-lab-3
```

Инициализация кластера и создание мониторов:

```
ceph-deploy mon create-initial
```

Деплоим конфигурацию кластера на все сервера кластера:

```
ceph-deploy admin ceph-lab-1 ceph-lab-2 ceph-lab-3 ceph-lab-client
```


СЕРН: развертывание кластера

Настраиваем управление кластером:

Меняем права на файл keyring на всех серверах кластера:

```
ssh ceph-lab-1 sudo chmod +r /etc/ceph/ceph.client.admin.keyring  
ssh ceph-lab-2 sudo chmod +r /etc/ceph/ceph.client.admin.keyring  
ssh ceph-lab-3 sudo chmod +r /etc/ceph/ceph.client.admin.keyring  
ssh ceph-lab-client sudo chmod +r /etc/ceph/ceph.client.admin.keyring
```

Создаем ceph manager демонов:

```
ceph-deploy mgr create ceph-lab-1 ceph-lab-2 ceph-lab-3
```


СЕРН: развертывание кластера

Создаем хранилище в кластере:

Создаем OSD:

```
ceph-deploy osd create --data /dev/sdb ceph-lab-1  
ceph-deploy osd create --data /dev/sdb ceph-lab-2  
ceph-deploy osd create --data /dev/sdb ceph-lab-3
```

Просмотр информации по OSD:

```
ceph osd status  
ceph osd df  
ceph osd tree
```

Создаем пул и указываем количество PG:

```
ceph osd pool create storhd 100 100
```

Включаем режим работы с кластером (в данном случае - rbd):

```
ceph osd pool application enable storhd rbd
```


СЕРН: развертывание кластера

Создаем блочное устройство на клиенте:

Создаем блочное устройство:

```
rbd create --pool storhd --size 4096 -m ceph-lab-1 storage
```

Необходимые настройки для устройства:

```
rbd feature disable storage -p storhd deep-flatten,fast-diff,object-map
```

Маппинг блочного устройства storage к созданному пулу storhd:

```
sudo rbd map storage --pool storhd --name client.admin -m ceph-lab-1
```

Создаем файловую систему и монтируем устройство в системе:

```
sudo mkfs.ext4 -m0 /dev/rbd/rbd0  
sudo mkdir /mnt/ceph-block-device  
sudo mount /dev/rbd/rbd0 /mnt/ceph-block-device
```


The background of the slide features an aerial view of a dense city skyline, likely New York City, with numerous skyscrapers. The image is overlaid with a semi-transparent blue and teal gradient. A network of white lines and dots, resembling a digital or social network, is superimposed over the cityscape. The text "Ваши вопросы?" is centered in the middle of the slide in a large, white, sans-serif font.

Ваши вопросы?

Маршрут вебинара

СЕРН: архитектура



СЕРН: развертывание кластера

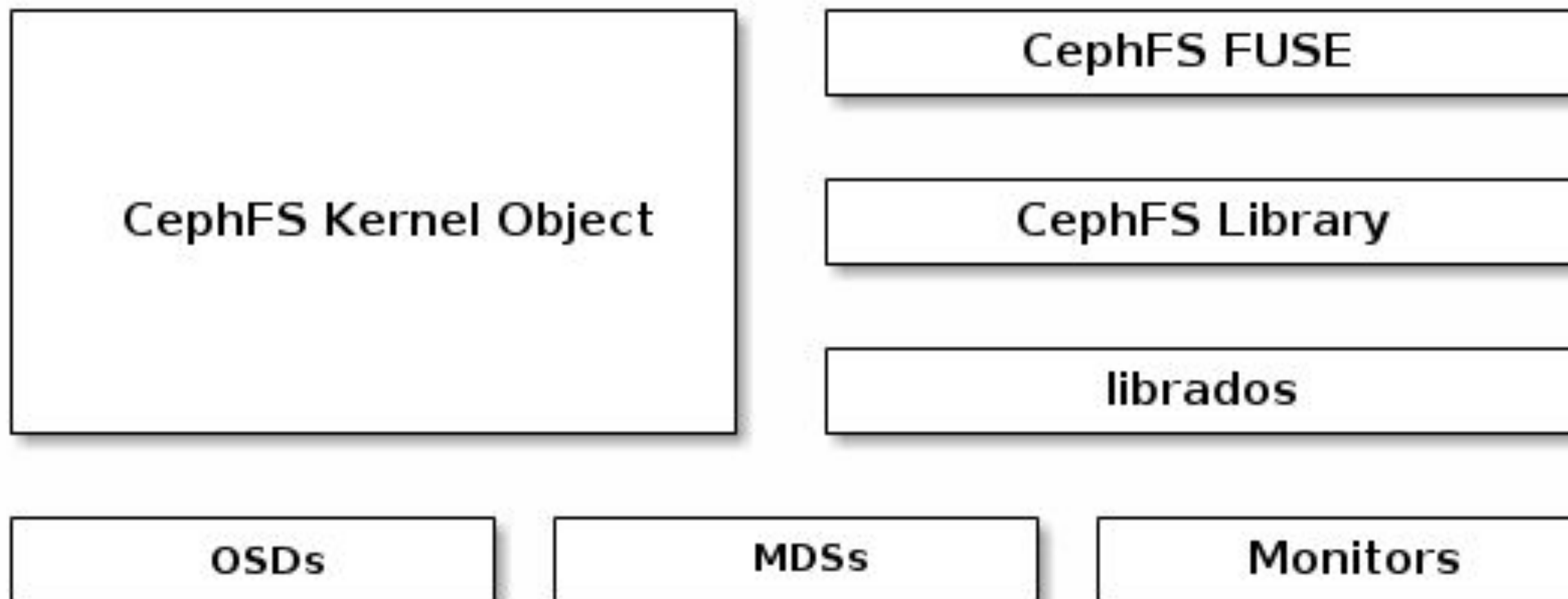


CephFS

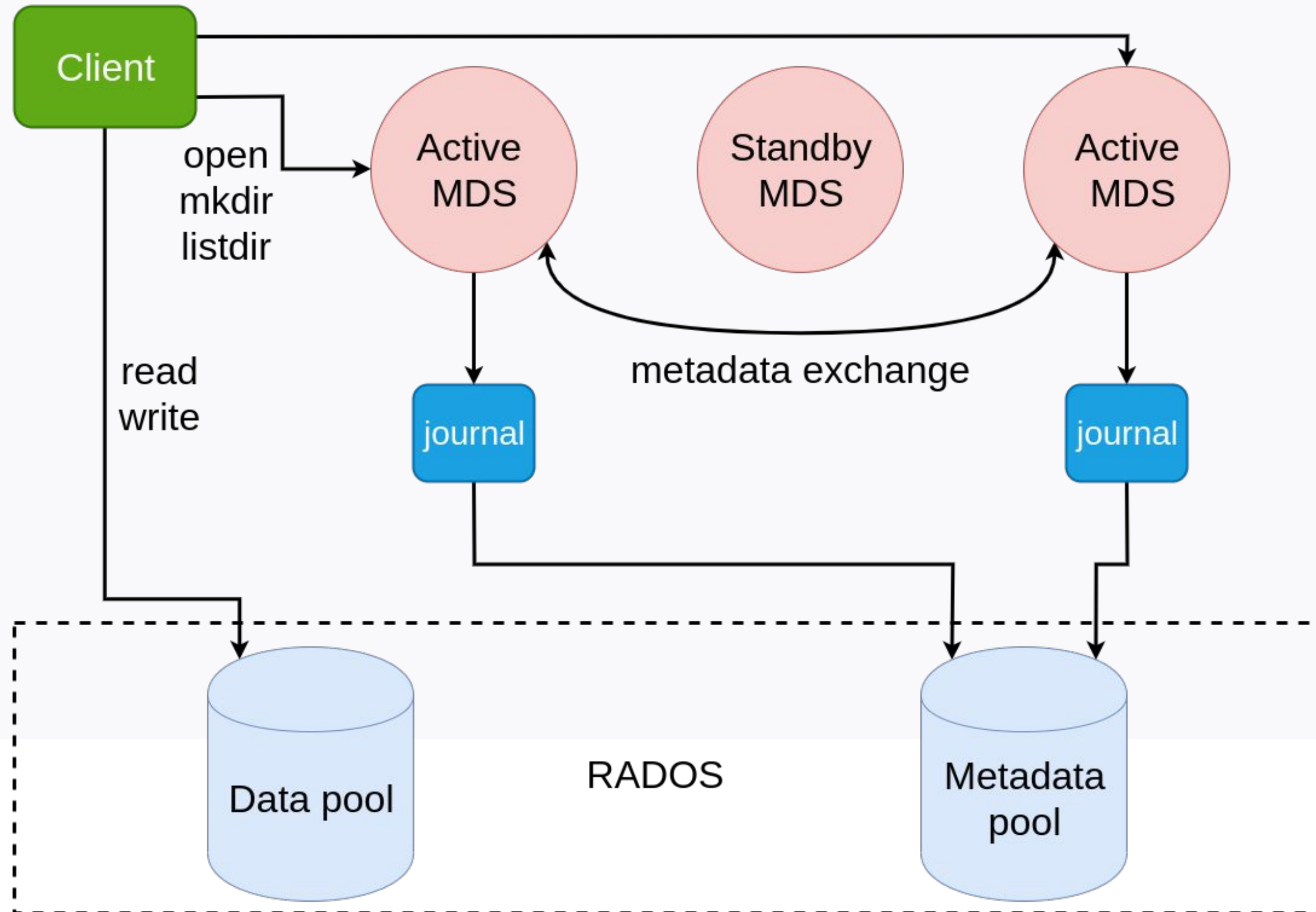


CephFS

CephFS



CephFS



CephFS

CephFS

Особенности:

- масштабируемая
- совместно используемая (shared)
- высокодоступная (HA)
- умеет ACL
- умеет квоты

CephFS

Конфигурирование CephFS

Создаем сервер MDS:

```
ceph-deploy mds create ceph-lab-1:mds-storage
```

Создаем пулы под данные и под метаданные:

```
ceph osd pool create cephfs_data 100  
ceph osd pool create cephfs_metadata 100
```

Создаем файловую систему:

```
ceph fs new otusfs cephfs_metadata cephfs_data  
ceph fs ls
```


Конфигурирование CephFS

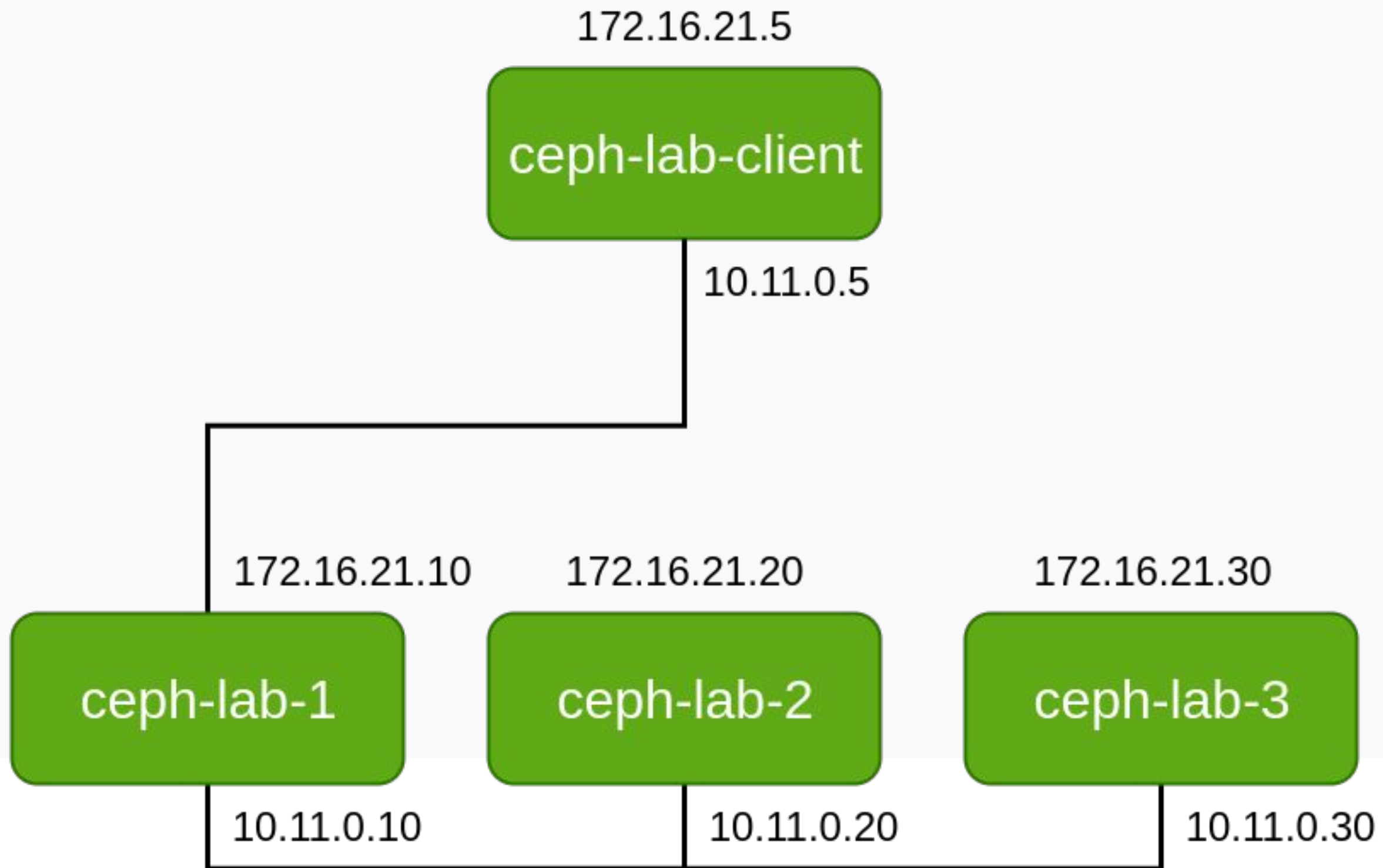
Монтируем файловую систему на клиенте:

```
mkdir /mnt/mycephfs  
mount -t ceph ceph-lab-3:6789:/ /mnt/mycephfs -o \  
name=admin,secret=`ceph-authtool -p /etc/ceph/ceph.client.admin.keyring`
```


The background of the slide features an aerial photograph of a city skyline, likely New York City, with numerous skyscrapers. The image is overlaid with a semi-transparent blue layer that contains a white network pattern of interconnected dots and lines. The text "Схема тестового стенда" is centered in this blue area in a large, white, sans-serif font.

Схема тестового стенда

Схема тестового стенда





Ваши вопросы?

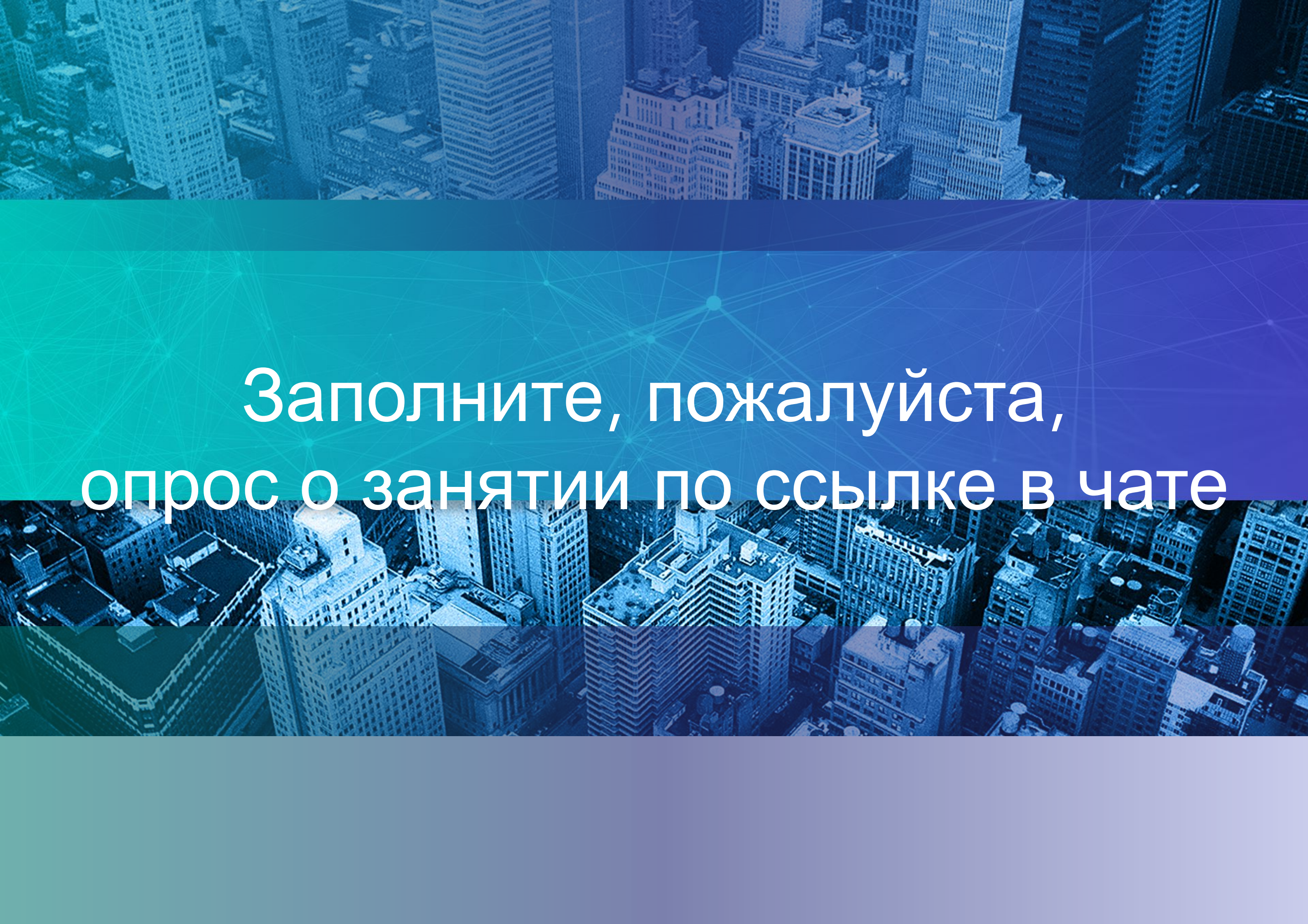
Рефлексия



Назовите 3 момента, которые вам запомнились в процессе занятия



Что вы будете применять в работе из сегодняшнего вебинара?



Заполните, пожалуйста,
опрос о занятии по ссылке в чате



Спасибо за внимание!
Приходите на следующие вебинары

Викирюк Павел

Системный инженер