# Интенсив СЕРН

5ти-дневный интенсив
День №2

# Программа занятия

- О выборе железа под проект
- Подготовка нод. Правка лимитов
- Составление конфига для деплоя
- Использование ceph deploy и параметров prepare & activate
- Как добавить ноду без потери производительности в кластере
- Горизонтальное и вертикальное масштабирование кластеров - на что обращать внимание
- Настройка отказо- и катастрофоустойчивости - как настроить Failure Domain, чтобы, например, кластер выдерживал отказ целой стойки или целого ЦОД


ВИДИМОСТЬ
Пик завышенных ожиданий
Плато производительности
Склон просвещения
Нижняя точка разочарования
Запуск технологии
ВРЕМЯ

# Формула расчета железа

**((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.

# Подготовка нод

- Сеть 10 гигабит, желательно контроллеры с поддержкой DPDK - на будущее
- Диски для журналов: небольшой объем, максимум количество циклов перезаписи, средние IOPS.
- RAID контроллеры - любые, без кешей,которые умеют работать в HBA режиме.
- JBOD- любой, но см по процессорной мощности

# Настройки сети, конфиг

net.ipv4.tcp_max_tw_buckets = 1000000
net.ipv4.udp_rmem_min = 16384
net.ipv4.tcp_tw_reuse = 1
net.core.wmem_max = 8388608
net.ipv4.tcp_max_syn_backlog = 2048
net.core.netdev_max_backlog = 8192
net.ipv4.tcp_rmem = 8192 87380 8388608
net.ipv4.udp_mem = 8388608 12582912 16777216
net.ipv4.tcp_tw_recycle = 0
net.nf_conntrack_max = 1073741824
net.core.rmem_max = 8388608
net.ipv4.tcp_mem = 8388608 12582912 16777216
net.ipv4.udp_wmem_min = 16384
net.ipv4.tcp_wmem = 8192 87380 8388608
net.core.somaxconn = 65535

# Настройки sysctl

```
kernel.msgmax = 65536
kernel.msgmnb = 65536
kernel.shmmax = 68719476736
kernel.shmall = 4294967296
kernel.pid_max = 4194303

fs.aio-max-nr = 4294967296
fs.file-max = 4294967296

vm.swappiness = 0
```

# Прочие настройки хоста

- Выключить NUMA или запинить процесс на ядро
- Правка лимитов nproc nofile
- Меняем планировщики SSD = noop OSD=deadline
- Если вы решились на btrfs, то используйте последнее ядро

# ceph.conf перед деплоем

- Определитесь с размером журнала
- С сетями
- С авторизацией

# Чем деплоить?

- Чем хочется!
- ceph-deploy
- ceph-disk
- ansible +ceph
- salt-stack + ceph

# Деплой новой ноды в текущий кластер

- Задать по вкусу
  - osd_recovery_op_priority
  - osd_recovery_threads
  - osd_client_op_priority
  - osd_max_backfills

- Задеплоить ноду, выставить weight в 0 и потихоньку поднимать

# Деплой новой ноды в текущий кластер

- Смотрим загрузку сети
- Смотрим потребление памяти
- Мониторим иопсы у клиентов
- Смотрим в логи ceph

# Типы сегментов ceph

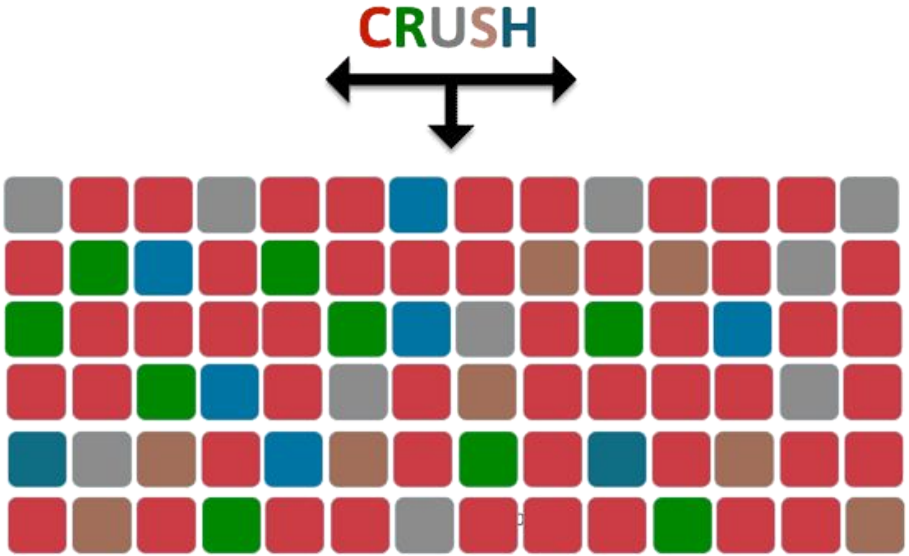
| | | |
|---|---|---|
| 0 | **OSD** | Демон OSD (например, **osd.1**, **osd.2** и так далее). |
| 1 | **Host** | Имя хоста, содержащего одно или более OSD. |
| 2 | **Rack** | Вычислительная стойка, содержащая один или более хостов. |
| 3 | **Row** | Ряд в последовательности стоек. |
| 4 | **Room** | Помещение, содержащее стойки и их ряды с хостами. |
| 5 | **Data Center** | Физический центр обработки данных, состоящий из помещений. |
| 6 | **Root** | Это начало иерархии сегментов. |

# Отказоустойчивость для всего кластера

osd crush chooseleaf type = {n}
#0 for a 1-node cluster.
#1 for a multi node cluster in a single rack
#2 for a multi node, multi chassis cluster with multiple hosts in a chassis
#3 for a multi node cluster with hosts across racks, etc.

# 2 варианта редактирования CRUSHMAP

1. ceph osd getcrushmap -o
2. CLI



# Редактирование CRUSHMAP с помощью CLI

```
# ceph osd crush add-bucket zone1 root
# ceph osd crush add-bucket rack1 rack
# ceph osd crush move rack1 root=zone1
# ceph osd crush move server-ceph-osd1 rack=rack1
# ceph osd crush rule create-simple zone1 zone1 rack
# ceph osd pool set volumes.zone1 crush_ruleset number
```

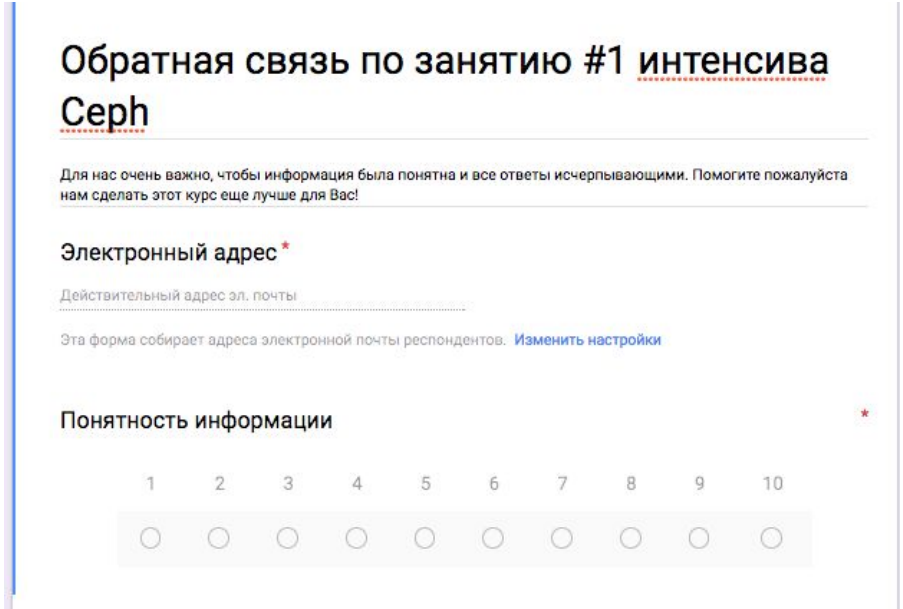
# Масштабирование

Ceph - горизонтально масштабируемое хранилище.
Каноничное вертикальное масштабирование в Ceph не поддерживается

Вопросы?

# Обратная связь

https://goo.gl/forms/6uHjEzvRGcAfUQ4L2



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