Интенсив СЕРН

5ти-дневный интенсив
День №5

Программа занятия

- Обсуждаем ваши инсталляции

Вопрос №1

Ceph для меня решение новое, поэтому кейсов у меня по нему нет, реализации его под рукой так же нет.

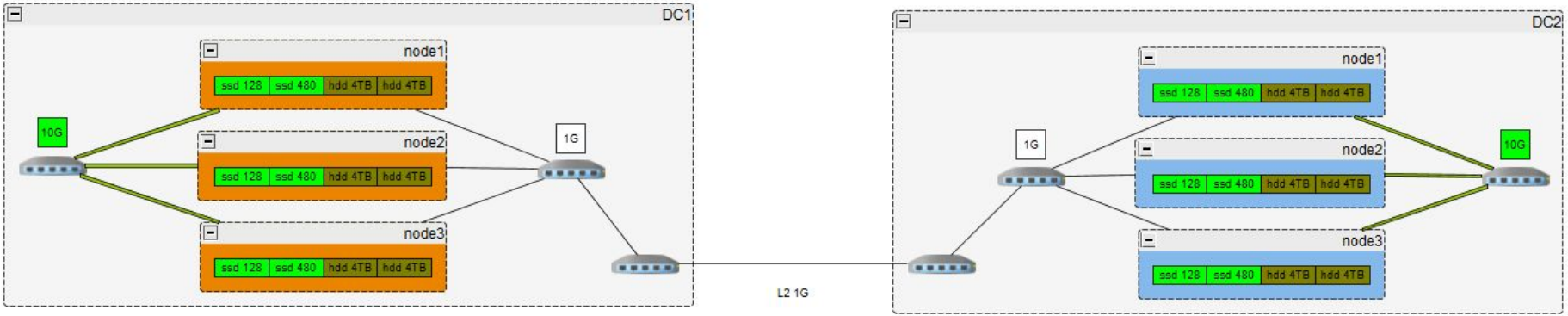
Поэтому, если Вы позволите, моя задача будет стоять в том, чтобы разработать эскизное решение под следующие требования:

- планируется 500 виртуальных машин, работающих под Гипервизором KVM
- требуется блочное хранение для 250ТБ
- 100 IOPS на ВМ, 50 000 IOPS агрегированных
- требуется 3 копии

Ответ

- Предлагаю такой конфиг 14 серверов с JBOD по 30 дисков 1.8 SAS 10k или 15к
- Если планируется использование нод и под CEPH и под виртуализацию, то 28 серверов по 15 дисков 1.8 Тб SAS
- Прибивать гвоздями OSD демон к ядрам

Вопрос №2



Имеем 2 дата центра. DC1 и DC2

В каждом ЦОДе по 3 узла со следующими характеристиками:

- CPU(s) Intel(R) Xeon(R) CPU E5-2665 0 @ 2.40GHz (2 Sockets)

- RAM 128Gb
- Сеть:
  - 2x10G интерфейса
  - 2x1G интерфейса

4 диска:

- sata ssd 128GB,
- sata ssd 480GB,
- sata hdd 4TB,
- sata hdd 4TB

Каждая нода будет и монитором и osd и kmv

Хочется собрать среду виртуализиции с отказоустойчивостью на 2 цода.

Вопрос №2

1) Как бы Вы распределили osd с таким набором дисков

2) При создании пула указал 512 PG. И все хорошо. Ceph helf ok. Но когда создал еще один пул количество pg на новый пул указал 256. Ceph helf ушел в warning - слишком много PG. Он что складывает PG по разным пулам?

Можно еще раз пояснить о количестве PG в контексте нескольких пулов.

3) Как все-таки реализовать отказоустойчивость на 2 цода? Где поискать инфу.

4) Возможно что-то еще посоветуете.

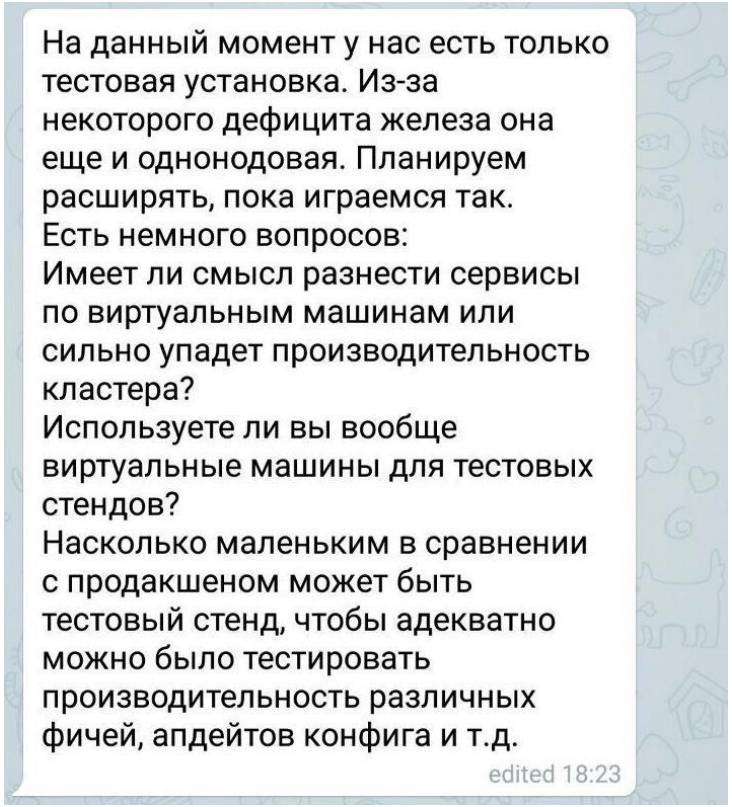
Ответ

1) По доному журналу на каждом SSD
2) Да, подсчет идёт по пулам и количество PG складывается. Для такого маленького количества OSD 512 PG слишком много
3) Для RBD если только RBD mirror и очень долго и вдумчиво тестировать.

Вопрос №3

Не пробовали LizardFS?

Вопрос №4


На данный момент у нас есть только тестовая установка. Из-за некоторого дефицита железа она еще и однонодовая. Планируем расширять, пока играемся так.
Есть немного вопросов:
Имеет ли смысл разнести сервисы по виртуальным машинам или сильно упадет производительность кластера?
Используете ли вы вообще виртуальные машины для тестовых стендов?
Насколько маленьким в сравнении с продакшеном может быть тестовый стенд, чтобы адекватно можно было тестировать производительность различных фичей, апдейтов конфига и т.д.
edited 18:23

Ответ

- Виртуальные машины для тестовых стендов не используем
- Тестовый стенд должен быть сопоставим с продом. У нас 48 серверов в проде и 3 на стенде.

Наш конфиг:

- 24 сервера в каждом по JBOD 700+ OSD
- разделено по стойка по 4 стойке в пуле
- 5 журналов на ссд с системой и по 10 журналов на доп OSD
- один 10гб поделенный VLAN

Вопрос №5

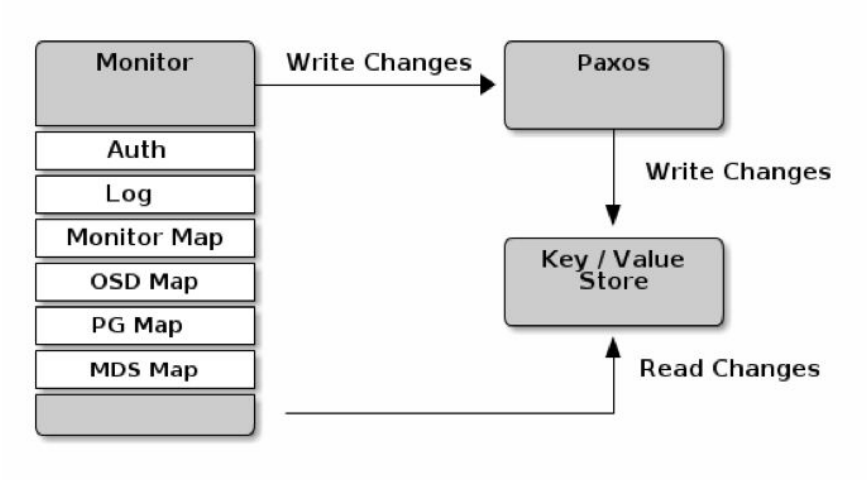
Не совсем понял как должны быть настроены мониторы.

Ответ

Монитор – демон обеспечивающий поддержание режима членства в кластере, хранение настроек и состояния.
Карты:

- Монитора
- OSD
- PG
- CRUSH
- MDS

Согласованность принятия решений обеспечивает paxos → число мониторов нечетно, >=3



Вопросы?

Обратная связь

https://goo.gl/forms/Xm5Tj8oPTB1ACJqE3