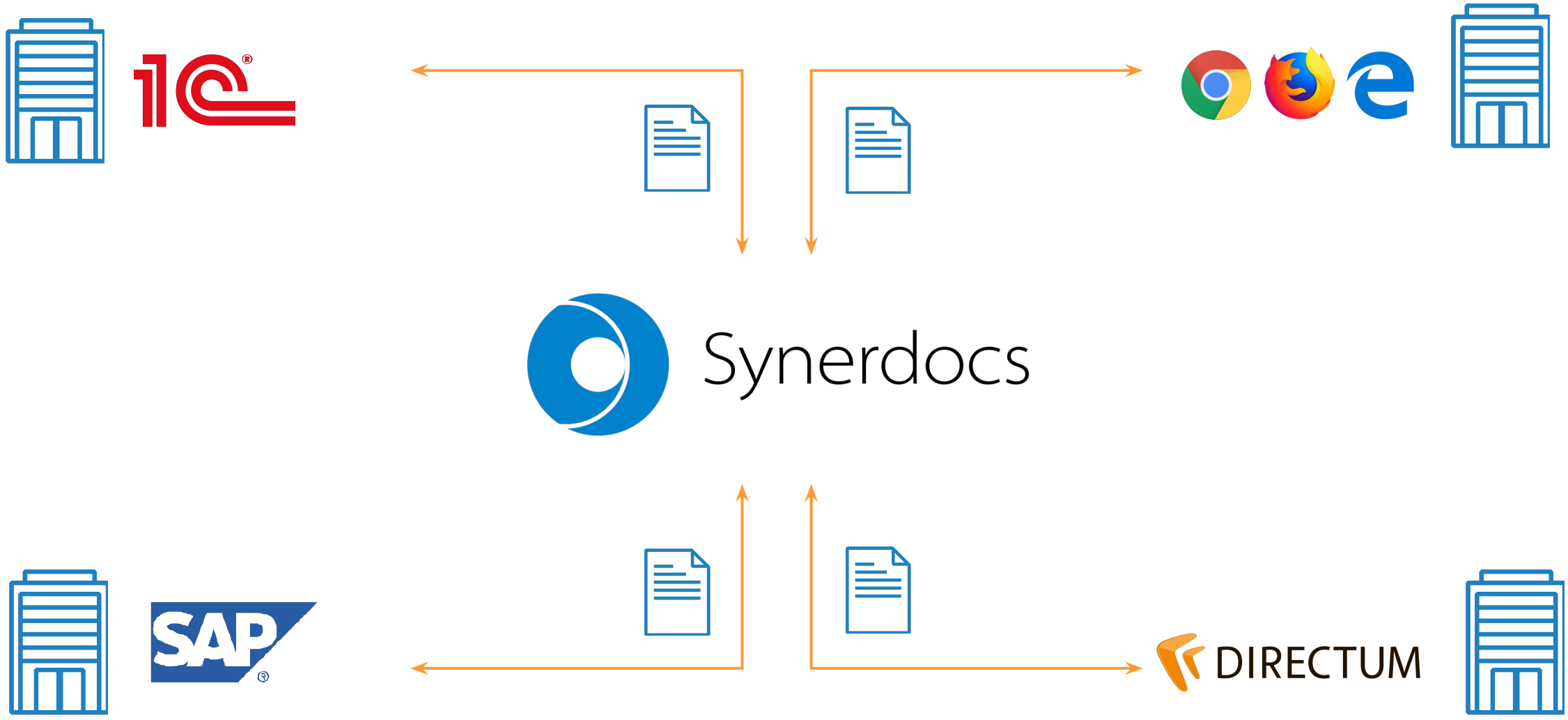
Synerdocs
электронный обмен документами

Использование Ceph для организации файлового хранилища

www.synerdocs.ru

О продукте

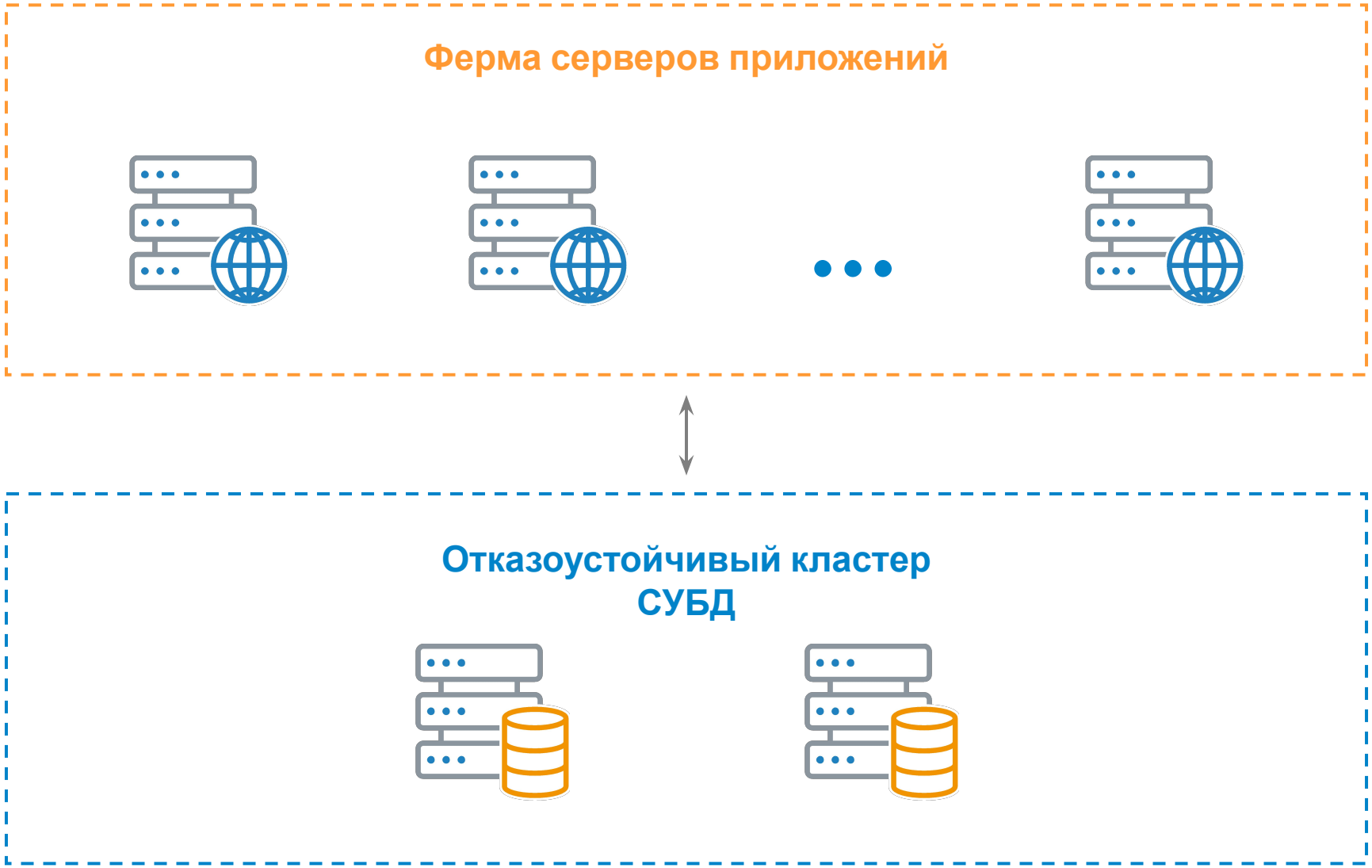

1С
Synerdocs
SAP
DIRECTUM

Содержание доклада

- Предпосылки внедрения ФХ
- Варианты решений для ФХ
- Ceph:
  - кратко о системе;
  - варианты установки;
  - принцип работы;
  - администрирование кластера.
- Резюме

Предпосылки внедрения ФХ в архитектуру

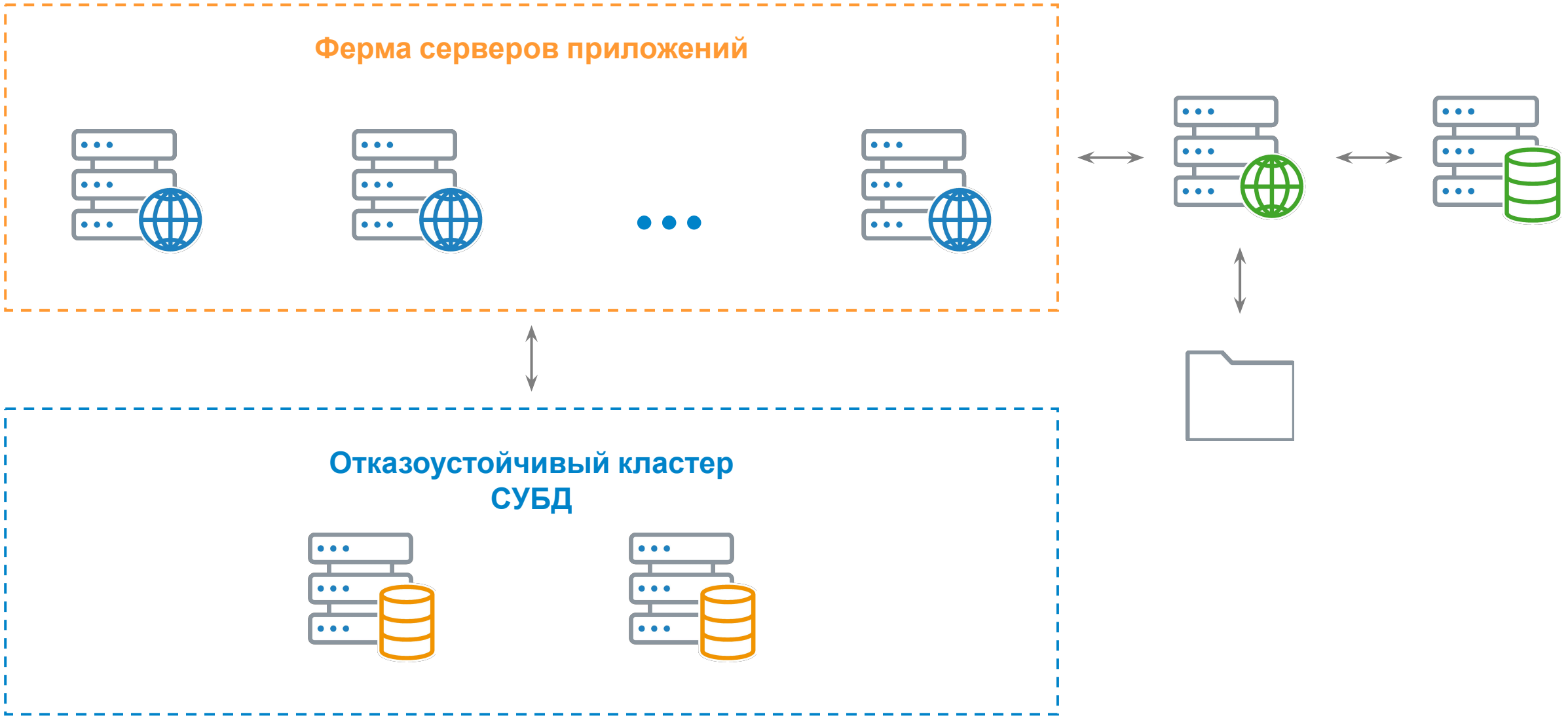
Инфраструктура до



БД сервиса

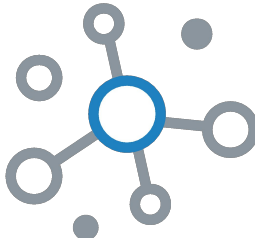
60% объема – тела документов

Инфраструктура после



Варианты решений для ФХ

Требования



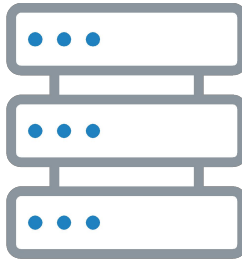
Распределенность

Отказоустойчивость

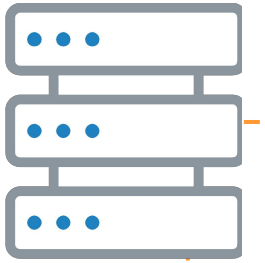
Репликация данных

Скорость

Сравнение скоростей



Физический сервер – источник файлов

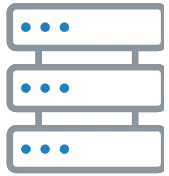


Физический сервер

SSD диск

Анализируемый кластер

vm1 vm2 vmN



vm – SMB-шара

GlusterFS

Последовательное копирование нескольких файлов по ~10 Гб:

- копирование в GlusterFS в 5-10 раз медленнее копирования в SMB-шару.

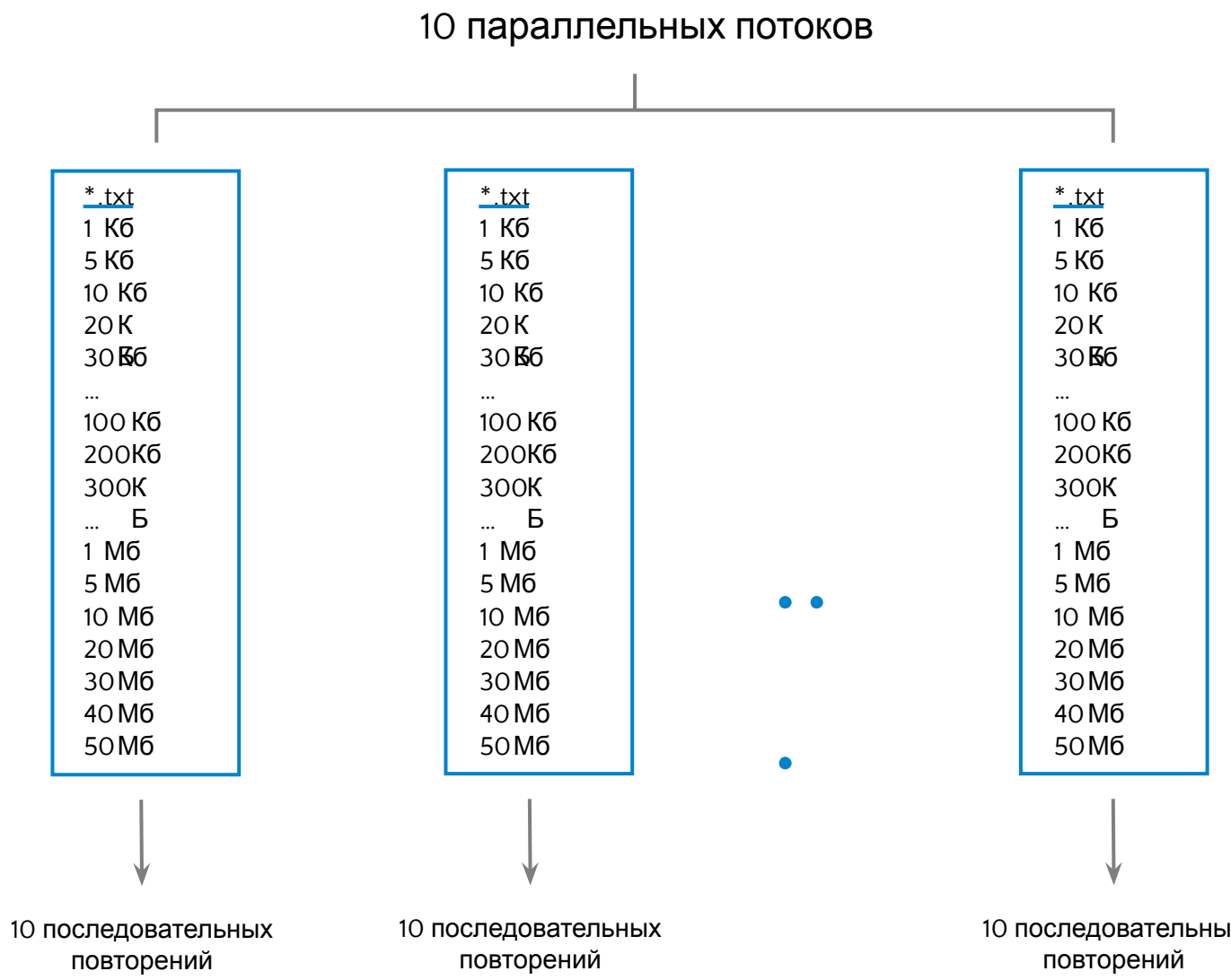
Копирование большого количества мелких файлов не проверяли.

rsync

Последовательное копирование нескольких небольших (мегабайты) и нескольких мелких (килобайты) файлов:

rsync успевал реплицировать данные с задержкой менее минуты.

rsync



До окончания цикла
из 10 повторений rsync
ни разу не доживал

Ceph

Варианты доступа к данным в Ceph:

- Ceph Object Gateway (S3/Swift-совместимое API);
- CephFS – POSIX-совместимая файловая система;
- Ceph Block Device (RBD – RADOS Block Device).

Ceph + Object Gateway

Не было времени для реализации поддержки S3.

Ceph + CephFS

Последовательное копирование нескольких файлов по ~10 Гб:

- копирование в CephFS в 2-3 раза медленнее копирования в SMB-шару.

Копирование большого количества мелких файлов не проверяли.

CephFS не была заявлена как решение для продуктивных сред из-за сырой реализации.

Ceph + RBD

Скорость записи в кластер соизмерима со скоростью записи в SMB-шару.

Ceph + RBD

Что такое Ceph

- **Ceph** – сеть хранения данных. На каждом узле сети используются свои вычислительные ресурсы для управлением данными.

- **RADOS** (Reliable Autonomic Distributed Object Store) – утилита для взаимодействия компонентов кластера Ceph, его неотъемлемая подкапотная часть.

Основные сущности Ceph

MON – демон монитора, серверы с MON – мозги кластера. MON должно быть минимум 3 штуки. Кластер работает, пока доступных MON > 50%.

OSD – демон хранения данных, управляет пространством на диске.

Pool – виртуальная группа для «определения» данных «одного хранилища».

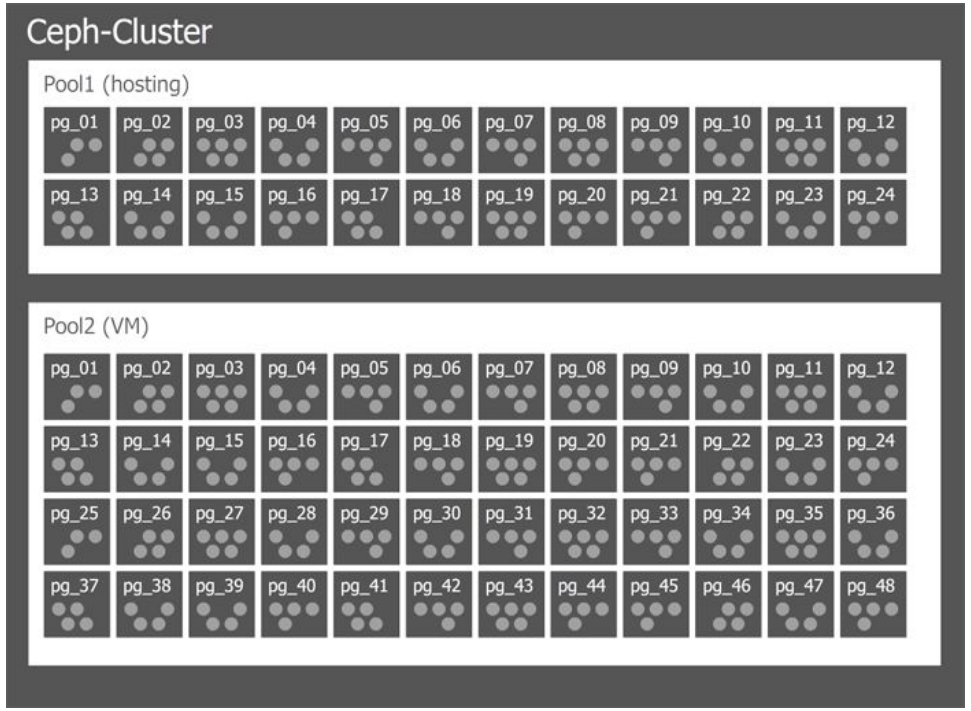
Object – набор данных фиксированного размера. Все попадающие в Ceph данные распределяются по Objects.

PG (Placement Group) – ячейка кластера Ceph, Objects хранятся в PG, а уже PG распределяются по OSD.

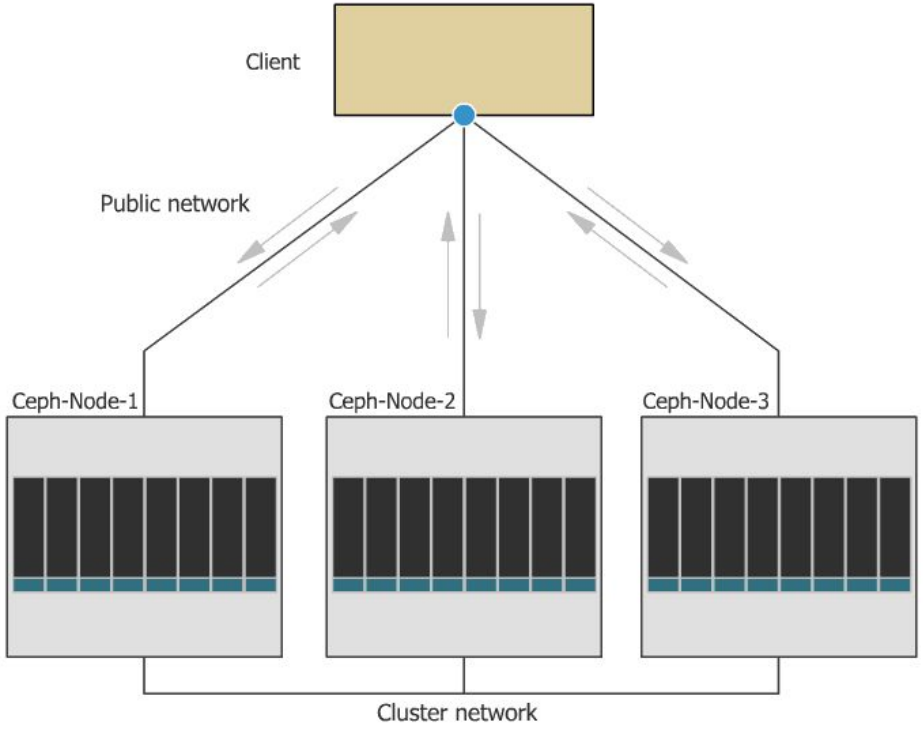
Synerdocs
электронный обмен документами

Картинки из Интернета

Все записываемые данные «складируются» в PG.
Пулы данных состоят из PG.



PG распределяются по OSD.



Описание с картинками: https://habr.com/post/313644/

Варианты установки

Независимая установка компонентов на серверы, ручная настройка кластера.

Сценарии Ansible.

Утилита ceph-deploy.

ceph-deploy

На отдельном сервере **admin-node** папка **/opt/my_cluster**

Создание кластера (объявление мониторов):

ceph-deploy new admin-node node01 node02

ceph.mon.keyring
ceph.conf:

```
[global]
mon_initial_members = admin-node, node01, node02
mon_host = 192.168.1.10:6789, 192.168.1.11:6789, 192.168.1.12:6789
auth_cluster_required = cephx
auth_service_required = cephx
auth_client_required = cephx
```

ceph-deploy

Установка ПО Ceph на все будущие ноды и мониторы кластера:

```
ceph-deploy install admin-node node01 node02 node03 node04
```

Инициализация кластера:

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

ceph-deploy

Добавление нод в кластер:

```
ceph-deploy admin admin-node node01 node02 node03 node04
```

Добавить OSD:

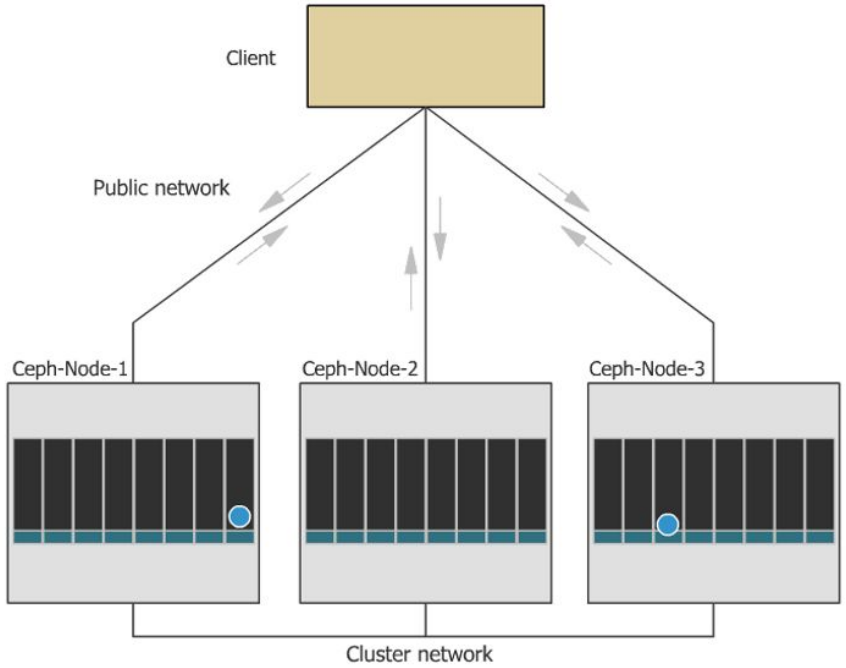
```
ceph-deploy osd create --data /dev/vda node01
ceph-deploy osd create --data /dev/vdb node01

ceph-deploy osd create --data /dev/vda node02
...
```

Реплики PG

CRUSH (controlled replication under scalable hashing) – алгоритм распределения реплик PG по OSD.

Репликация:

- другое здание/помещение;
- другая стойка;
- другая нода;
- OSD с наибольшим весом*;
- OSD с наибольшим свободным местом;
- случайный выбор.



* Данный параметр рассчитывается автоматически исходя из распределения размеров OSD в рамках одной ноды. Его можно корректировать, но не стоит.

Советы по работе с OSD

- Иметь равное количество OSD на всех нодах.
- «Набор» размеров OSD на каждой ноде должен быть одинаковым.
- Для увеличения размера кластера лучше добавлять новые OSD, чем увеличивать размер текущих.

Пул и образ данных

RBD (RADOS Block Device) – абстракция блочного устройства, через которое реализован доступ к данным внутри кластера Ceph.

Для настройки RBD в пуле данных используется сущность Image.
Маппинг RBD происходит именно в Image.
В примере: имя пула = documents, имя образа = data

```
ceph osd pool create documents 128 128
```

```
rbd create --size 1000G documents/data
```

Подключение RBD

Для подключения RBD лучше использовать отдельную машину (client).
Установить на нее Ceph и ввести в кластер:

```
ceph-deploy install client
ceph-deploy admin client
```

Создать RBD:

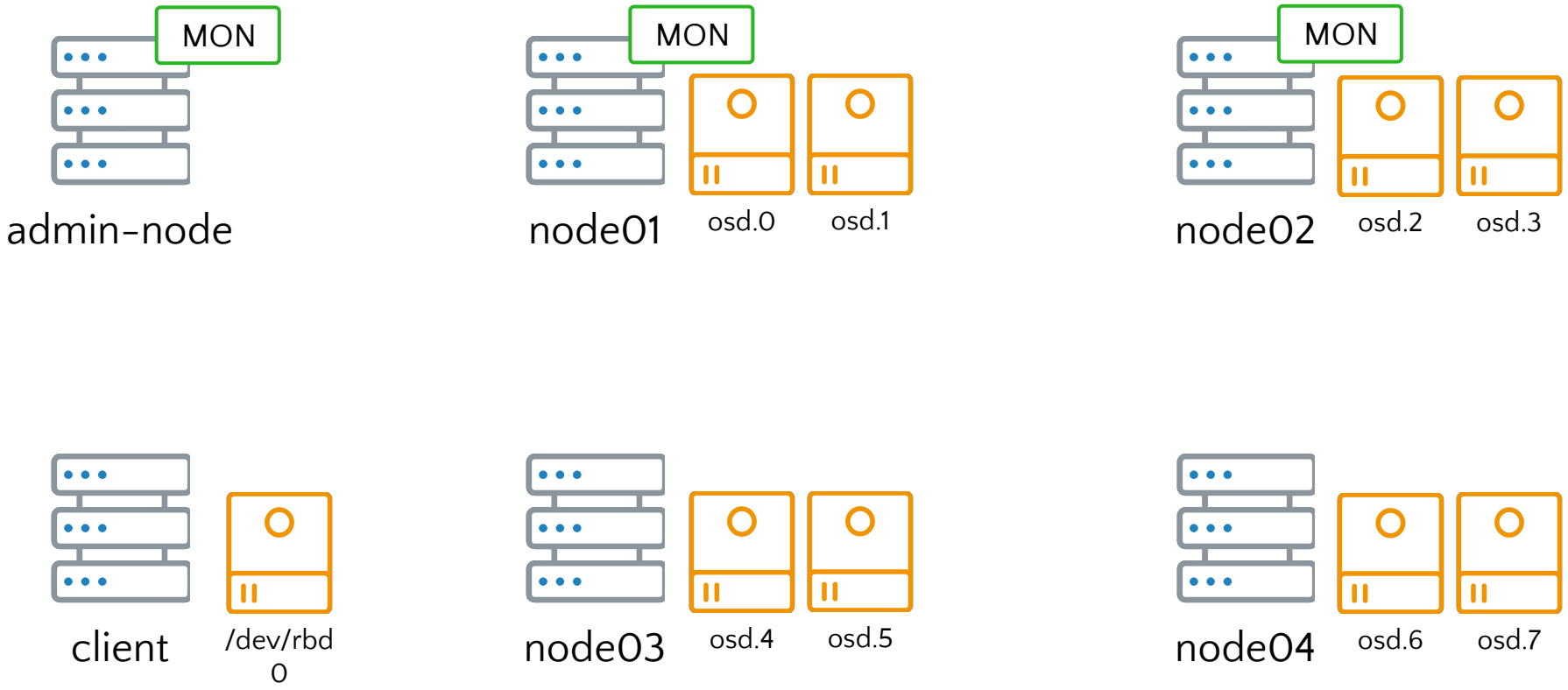
```
rbd map --pool documents --image data
```

```
/dev/rbd0
```

```
mkfs.xfs  /dev/rbd0
```

```
mount /dev/rbd0 /mnt/documents
```

Схема кластера



Размеры пулов (образов)

Размер RBD-образа не зависит от фактического суммарного объема OSD.

Заданный размер образа достигается постепенно по мере наполнения хранилища данными.

Если при итерации увеличения размера образа фактическое суммарное свободное место на всех OSD закончится -> данные утеряны навсегда.

Фактический суммарный объем всех OSD должен быть **ВСЕГДА** больше суммарного заданного объема всех пулов (образов).

Увеличение максимального размера RBD-образа необходимо делать заранее: при 80% заполнения или раньше.

Режим обслуживания

При выходе из строя или недоступности OSD кластер автоматически начинает перебалансировку потерянных PG.

Перед работами на нодах необходимо включать режим обслуживания кластера:

```
ceph osd set noout
```

Выход из режима обслуживания:

```
ceph osd unset noout
```

Увеличение RBD-образа

Подключить новые диски к нодам, добавить их в Ceph (например 4 диска по 300 Гб):

```
ceph-deploy osd create --data /dev/vdc node01
ceph-deploy osd create --data /dev/vdc node02
...
```

Указать новый максимальный размер образа (только вверх и надо учесть количество реплик):

```
rbd resize --size 1390G --pool documents --image data
```

Проверка:

```
rbd info --pool documents --image data
fdisk –l #на client
```

Актуализировать размер на точке монтирования:

```
xfs_growfs /mnt/documents/
```

Мониторинг места на OSD

Цель – всегда иметь запас места на перебалансировку в случае смерти одной ноды.

Легенда: 4 ноды по N OSD на каждой.

Шаг 1. Определить ноду, на которой самый большой объем занятого места на всех ее OSD (V_1).

Шаг 2. Среди оставшихся 3 нод определить ту, на которой меньше всего свободного места (V_2).

Шаг 3. $D = V_2 - V_1/3$

Если $D < 50$ Гб, то пора добавлять OSD.

Резюме по Ceph

Плюсы:

- легко устанавливается и настраивается;
- быстрый;
- легко масштабируется;
- достаточное количество команд CLI для диагностики состояния кластера и управления им.

Минусы:

- огромное количество недокументированных параметров для тонкой настройки (~1500 штук);
- много противоречивой информации о настройках в разных источниках;
- необходимо тщательно следить за статистикой наполнения пулов и свободным местом на OSD.

Резюме по ФХ

Хранить бинарный контент в БД считается моветоном.

Подавляющую часть нашего бинарного контента (тела документов) мы вынесли в ФХ.

Основная цель переноса тел документов в ФХ – разгрузка БД:

- увеличилась производительность СУБД на прежних мощностях;
- уменьшилось время конвертации таблиц.

Тонкая настройка производительности.

Не надо платить за лицензии на ядрышки ЦП для СУБД.

Где искать помощь

https://docs.ceph.com/docs/master/

http://onreader.mdl.ru/MasteringCeph/content/Ch01.html

http://onreader.mdl.ru/MasteringCeph.2ed/content/Ch01.html

https://t.me/ceph_ru

Google + Яндекс

Synerdocs
электронный обмен документами

Ceph есть в Synerdocs и Yahoo!, а у Вас?

Евгений Корляков
korlyakov_es@synerdocs.ru

www.synerdocs.r
u