Ceph multisite 구성과 bucket sync 기본값을 disabled로 변경한 multisite 운영과 문제 해결

박범석

2020.11

LINE

Contents

LINE에서 Ceph (1)

Ceph의 활용

Object Storage	RGW, Nginx(auth_request, geoip) + Openstack Keystone clone Dashboard 제공 S3 호환 API 제공
Block Storage	RBD, Openstack Cinder Dashbaord 제공 VM, Kubernetes에 볼륨 제공
File System	CephFS, Openstack Manila Dashbaord 제공 Fuse, Kernel mount 제공

LINE에서 Ceph (2)

Ceph 버전

Luminous (v12.2.x)

- RHCS 공개 소스 기반
- Upstream backport
- Upstream PR[1]
- 사내 전용 패치
- Jewel (v10.2.x) 부터 업그레이드

Nautilus (v14.2.x)

- Ceph community 소스 기반
- Upstream backport
- Upstream PR[1]
- 사내 전용 패치

[1] https://github.com/ceph/ceph/pulls?q=is%3Apr+author%3AIlsooByun

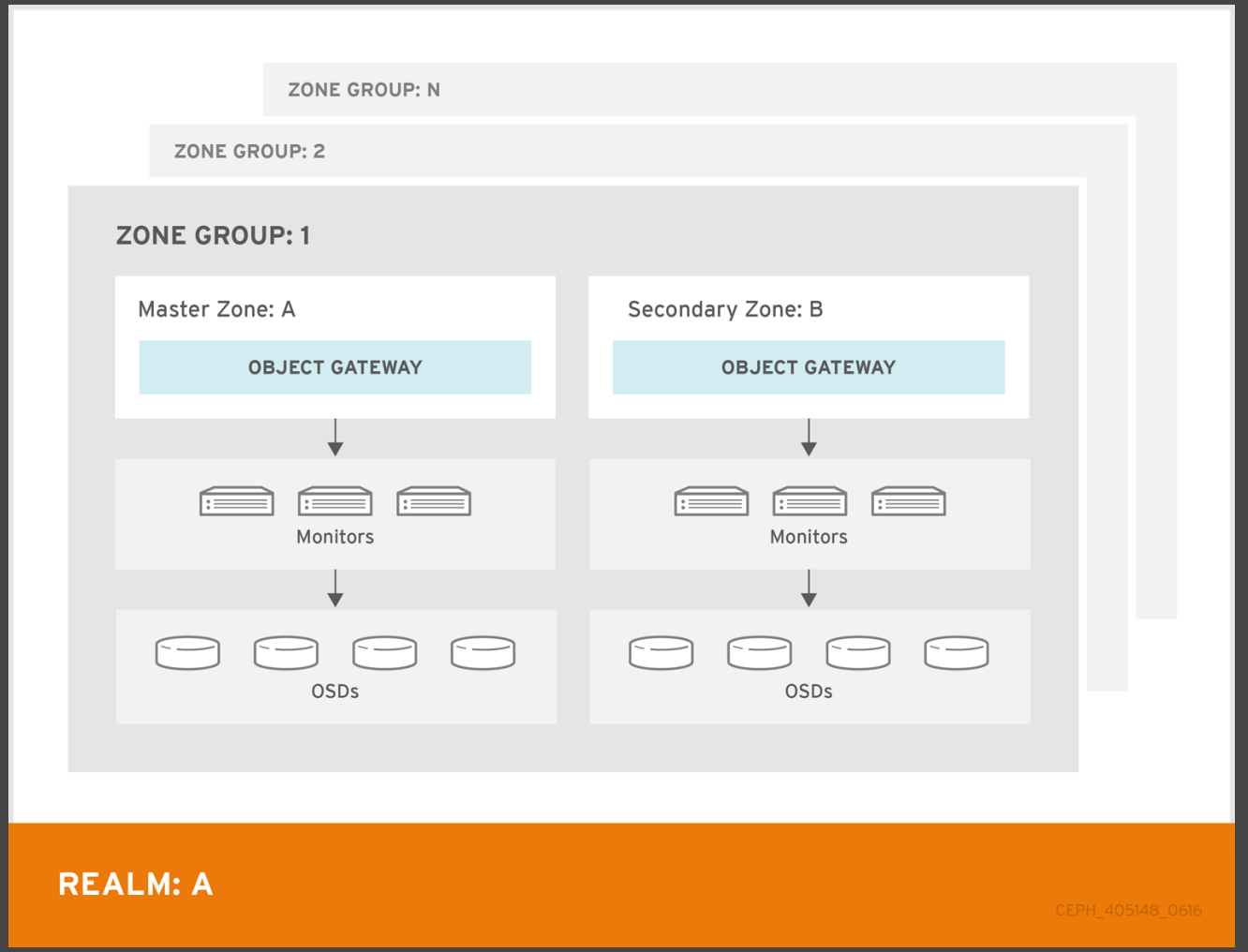
Multisite란?

Zone

Zone Group

Realm

Period



※출처 : https://access.redhat.com/documentation/en-us/red_hat_ceph_storage/3/html-single/object_gateway_guide_for_red_hat_enterprise_linux

Multisite 구성 방법 (1)

사전에 결정

Realm	realm1 (s3)
Zone Group	zonegroup1 (us)
Master Zone	zone1 (us-east-1)
Secondary Zone	zone2 (us-west-1)
Master Zone RGW1	rgw1 (http://192.168.100.1:80)
Master Zone RGW2	rgw2 (http://192.168.100.2:80)
Secondary Zone RGW1	rgw3 (http://192.168.200.1:80)
Secondary Zone RGW2	rgw4 (http://192.168.200.2:80)

Multisite 구성 방법 (2)

Master Zone

Realm 생성

```
# radosgw-admin realm create --rgw-realm=realm1 --default
{
    "id": "3c4004f9-194c-4284-adf8-2172cd62c55f",
    "name": "realm1",
    "current_period": "042a14cb-7409-4451-8714-03bc65af4463",
    "epoch": 1
}
```

Multisite 구성 방법 (3)

Master Zone

Master Zone Group 생성

신규 클러스터

```
# radosgw-admin zonegroup create --rgw-zonegroup=zonegroup1 \
                                 --endpoints=http://192.168.100.1:80 \
                                 --rgw-realm=realm1 --master --default
```

기존 클러스터

```
# radosgw-admin zonegroup rename --rgw-zonegroup=default --zonegroup-new-name=zonegroup1
# radosgw-admin zonegroup modify --rgw-zonegroup=zonegroup1 \
                                 --endpoints=http://192.168.100.1:80 \
                                 --rgw-realm=realm1 --master --default
```

Multisite 구성 방법 (4)

Master Zone

Master Zone 생성

신규 클러스터

```
# radosgw-admin zone create --rgw-zonegroup=zonegroup1 --rgw-zone=zone1 \
                            --master --default \
                            --endpoints=http://192.168.100.1:80,http://192.168.100.2:80
```

기존 클러스터

```
# radosgw-admin zone rename --rgw-zone=default --zone-new-name=zone1 --rgw-zonegroup=zonegroup1
# radosgw-admin zone modify --rgw-zonegroup=zonegroup1 --rgw-zone=zone1 \
                            --master --default \
                            --endpoints=http://192.168.100.1:80,http://192.168.100.2:80
```

endpoints에 로드 밸런서 사용시 참고 : https://www.spinics.net/lists/ceph-users/msg54005.html

Multisite 구성 방법 (5)

Master Zone

Default Zone Group과 Default Zone 삭제

신규 클러스터

```
# radosgw-admin zonegroup remove --rgw-zonegroup=default --rgw-zone=default
# radosgw-admin period update --commit
# radosgw-admin zone delete --rgw-zone=default
# radosgw-admin period update --commit
# radosgw-admin zonegroup delete --rgw-zonegroup=default
# radosgw-admin period update --commit
```

기존 클러스터는 이부분을 skip

Multisite 구성 방법 (6)

Master Zone

Default pool 삭제

신규 클러스터

```
# ceph tell mon.\* injectargs '--mon-allow-pool-delete=true'
# rados rmpool default.rgw.control default.rgw.control --yes-i-really-really-mean-it
# rados rmpool default.rgw.data.root default.rgw.data.root --yes-i-really-really-mean-it
# rados rmpool default.rgw.gc default.rgw.gc --yes-i-really-really-mean-it
# rados rmpool default.rgw.log default.rgw.log --yes-i-really-really-mean-it
# rados rmpool default.rgw.meta default.rgw.meta --yes-i-really-really-mean-it
# rados rmpool default.rgw.users.uid default.rgw.users.uid --yes-i-really-really-mean-it
# ceph tell mon.\* injectargs '--mon-allow-pool-delete=false'
```

기존 클러스터는 이부분을 skip

Multisite 구성 방법 (7)

Master Zone

System User 생성

```
# radosgw-admin user create --uid="synchronization-user" --display-name="Synchronization User" --system
# radosgw-admin zone modify --rgw-zone=zone1 --access-key={access_key} --secret={secret_key}
```

Multisite 구성 방법 (8)

Master Zone

Period 업데이트

```
# radosgw-admin period update --commit
```

Multisite 구성 방법 (9)

Master Zone

ceph.conf 변경

신규 클러스터

[client.rgw.rgw1]

host = rgw1

rgw zone = zone1

[client.rgw.rgw2]

host = rgw2

rgw zone = zone1

기존 클러스터는 ceph.conf 변경 skip

systemctl restart ceph-radosgw@rgw.`hostname -s`

Multisite 구성 방법 (10)

Secondary Zone

Realm 가져오기

System User 생성시의 access_key와 secret_key를 이용해서 realm을 가져온다.

```
# radosgw-admin realm pull --url=http://192.168.100.1:80 --access-key={access_key} --secret={secret_key}
# radosgw-admin realm default --rgw-realm=realm1
```

Multisite 구성 방법 (11)

Secondary Zone

Period 가져오기

System User 생성시의 access_key와 secret_key를 이용해서 period를 가져온다.

radosgw-admin period pull --url=http://192.168.100.1 --access-key={access_key} --secret={secret_key}

Multisite 구성 방법 (12)

Secondary Zone

Secondary Zone 생성

System User 생성시의 access_key와 secret_key를 이용해서 Secondary Zone을 생성한다.

```
# radosgw-admin zone create \
                --rgw-zonegroup=zonegroup1 \
                --rgw-zone=zone2 \
                --access-key={access_key} --secret={secret_key} \
                --endpoints=http://192.168.200.1:80,http://192.168.200.2:80 \
                [--read-only]
```

Multisite 구성 방법 (13)

Secondary Zone

Default Zone 삭제

```
# radosgw-admin zone delete --rgw-zone=default
```

Multisite 구성 방법 (14)

Secondary Zone

Default pool 삭제

```
# ceph tell mon.\* injectargs '--mon-allow-pool-delete=true'
# rados rmpool default.rgw.control default.rgw.control --yes-i-really-really-mean-it
# rados rmpool default.rgw.data.root default.rgw.data.root --yes-i-really-really-mean-it
# rados rmpool default.rgw.gc default.rgw.gc --yes-i-really-really-mean-it
# rados rmpool default.rgw.log default.rgw.log --yes-i-really-really-mean-it
# rados rmpool default.rgw.meta default.rgw.meta --yes-i-really-really-mean-it
# rados rmpool default.rgw.users.uid default.rgw.users.uid --yes-i-really-really-mean-it
# ceph tell mon.\* injectargs '--mon-allow-pool-delete=false'
```

Multisite 구성 방법 (15)

Secondary Zone

Period 업데이트

```
# radosgw-admin period update --commit
```

Multisite 구성 방법 (16)

Secondary Zone

ceph.conf 변경

```
[client.rgw.rgw3]
host = rgw3
rgw zone = zone2
[client.rgw.rgw4]
host = rgw4
rgw zone = zone2
```

```
# systemctl restart ceph-radosgw@rgw.`hostname -s`
```

Multisite 구성 시 주의점 (1)

bucket reshard

dynamic resharding 불가

bucket index의 dynamic resharding이 불가능[1] 하다.

manual resharding이 가능하긴 하지만 모든 RGW를 중지해야 하기 때문에 어려움이 있다.

[1] master branch에서 dynamic resharding이 가능하도록 개발중이다.

Multisite 구성 시 주의점 (2)

luminous 이전에 생성된 bucket (1)

explicit_placement 수정

수정이 필요한 bucket 찾기

```
# radosgw-admin bucket stats | jq -r '.[] | select( .explicit_placement.data_pool | contains("default")) | .bucket'
```

rgw_bucket.cc 2개 라인[1]이 실행되지 않도록 수정 후 radosgw-admin 빌드

```
/* existing bucket, keep its placement */
bci.info.bucket.explicit_placement = old_bci->info.bucket.explicit_placement;
bci.info.placement_rule = old_bci->info.placement_rule;
```

```
# ./radosgw-admin metadata get bucket.instance:<bucket>:<bucket id> > <bucket>.json
# ./radosgw-admin metadata put bucket.instance:<bucket>:<bucket id> < <bucket>.json
```

[1] https://github.com/ceph/ceph/blob/e5775fee5c3940ac7578e0bba898296685375212/src/rgw/rgw_bucket.cc#L3007-L3008

Multisite 구성 시 주의점

luminous 이전에 생성된 bucket (2)

explicit_placement 수정

AS IS

```
"explicit_placement": {
    "data_pool": "default.rgw.buckets.data",
    "data_extra_pool": "default.rgw.buckets.non-ec",
    "index_pool": "default.rgw.buckets.index"
},
```

TO BE

```
"explicit_placement": {
    "data_pool": "",
    "data_extra_pool": "",
    "index_pool": ""
},
```

Bucket sync 기본값을 disabled로 변경한 이유

- 모든 bucket을 sync할 필요가 없다. DR이 필요한 bucket만 나중에 enable 하자.
- Secondary Zone을 백업 서버로도 활용하고 싶다.

기본값 변경으로 나타난 문제점 (1)

bilog

- large omap objects 경고 발생
- Master Zone과 Secondary Zone의 bucket meta의 syncstopped flag는 정상적으로 동기화 되었지만 bucket shard의 syncstopped flag는 동기화 되지 않아서 bilog가 계속 쌓임.
- bilog가 쌓여도 RGW에서 bilog trim이 실행되기 때문에 주기적으로 trim이 되어야 했지만 bucket meta의 syncstopped flag는 true인 상태여서 trim이 되지 않음
- bucket shard당 수백만개의 bilog가 쌓이는 경우도 있었음.
- bucket sync가 disabled인 bucket에 대해서 bilog trim을 시도함

기본값 변경으로 나타난 문제점 (2)

syncstopped flag 불일치

rados getomapheader –p {INDEX_POOL} .dir.{BUCKET_ID}.{SHARD_ID}

Master Zone

00000040 00 00 00 00 00 00 ff ff ff ff 01

Secondary Zone

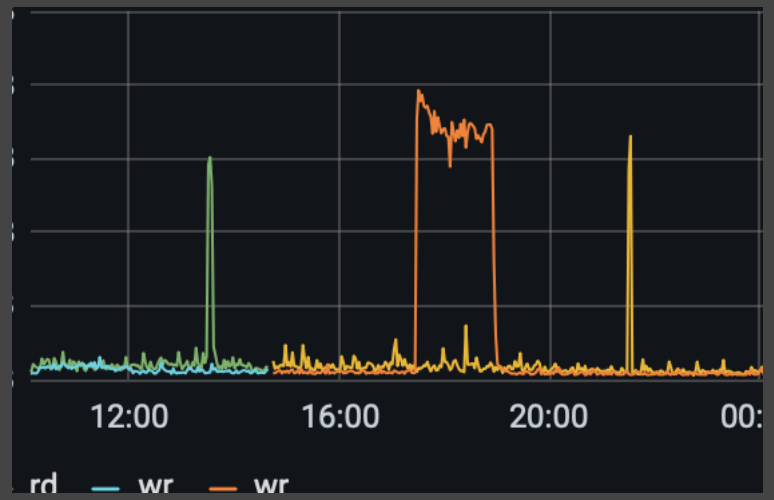
00000090 00 00 00 ff ff ff ff 00

'ff ff ff ff'가 syncstopped field

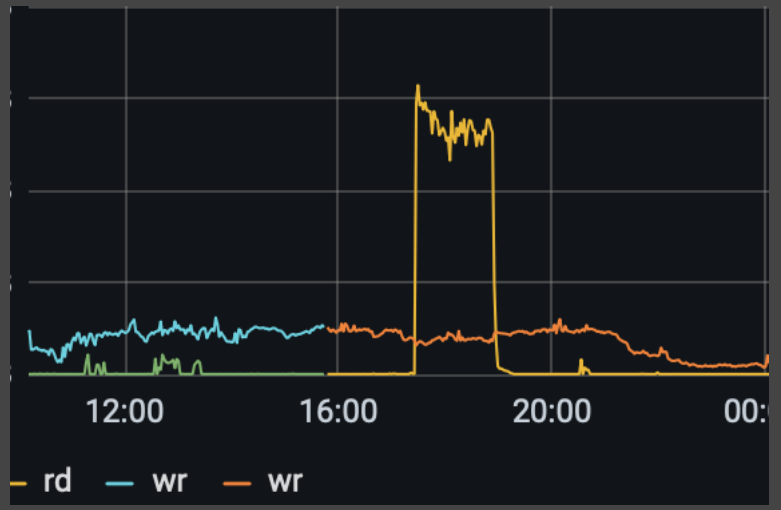
문제 해결 (1)

syncstopped flag 일치 시키기

- Secondary Zone에 bucket을 생성
- bucket 생성 중 Secondary Zone에서 Master Zone으로 대규모 동기화 트래픽 발생



Master Zone



Secondary Zone

문제 해결 (2)

syncstopped flag 일치 시키기

- Secondary Zone으로의 bucket 생성을 중지
- radosgw-admin bucket check --fix시 syncstopped flag가 유지되지 않는 문제[1] 발견 후 패치
- cli를 개발해 문제가 있는 bucket의 syncstopped flag를 일치 시킴
- bucket의 bilog trim
- datasync가 disabled인 버킷은 bilog가 쌓이지 않도록 패치
- 기존에 존재하는 bucket은 stop_bi_log_entries를 호출하지 않도록 변경
- 근본적인 해결을 위해서는 osd쪽 코드도 수정해야 함

[1] https://github.com/ceph/ceph/pull/37892

유용한 패치

rgw: Added caching for S3 credentials retrieved from keystone

https://github.com/ceph/ceph/pull/26095

https://github.com/ceph/ceph/pull/27100

THANK YOU