

Cinder: configurazioni avanzate

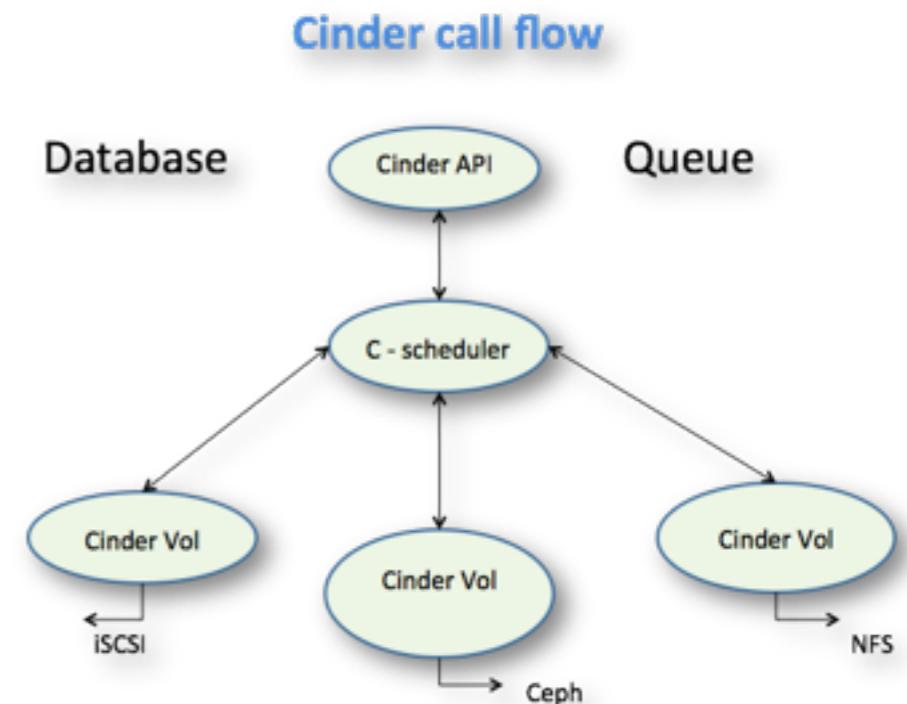
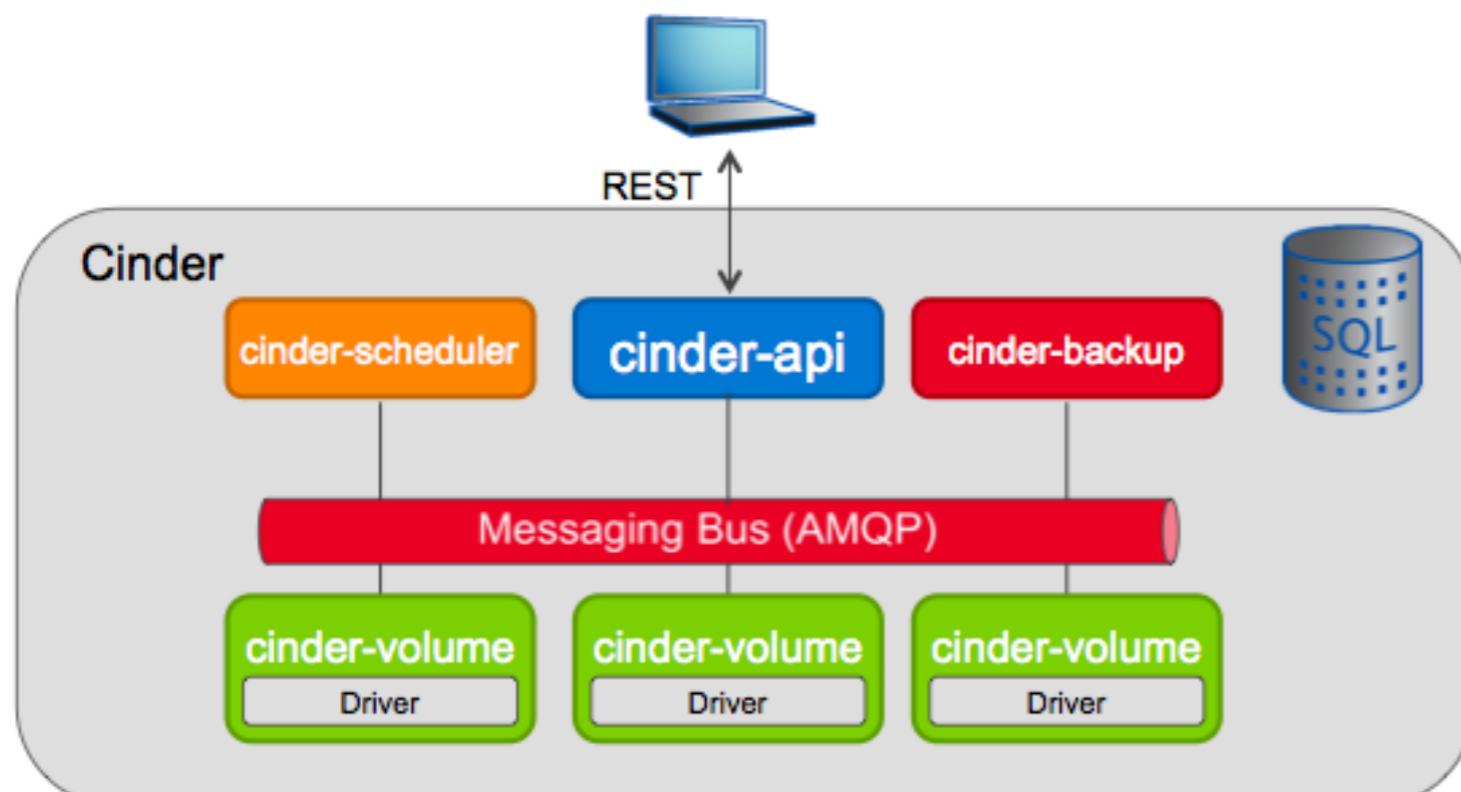
Marica Antonacci - INFN Bari

*Scuola di Cloud Computing
Bari, 27-30 Aprile 2015*

Outline

- Multi-backend
- QoS & Rate-limiting
- Encryption
- Backup & Disaster-Recovery

Cinder: backend multiplexing



Un nodo cinder-volume può gestire uno o più backend

La configurazione

- backend diversi (driver diversi)

cinder.conf

```
enabled_backends=lvm1,nfs1
[lvm1]
volume_driver=cinder.volume.drivers.lvm.LVMISCSIDriver
volume_backend_name=LVM_iSCSI
volume_group=cinder-volumes
[nfs1]
nfs_shares_config=${PATH_TO_YOUR_SHARES_FILE}
volume_driver=cinder.volume.drivers.nfs.NfsDriver
volume_backend_name=NFS
```

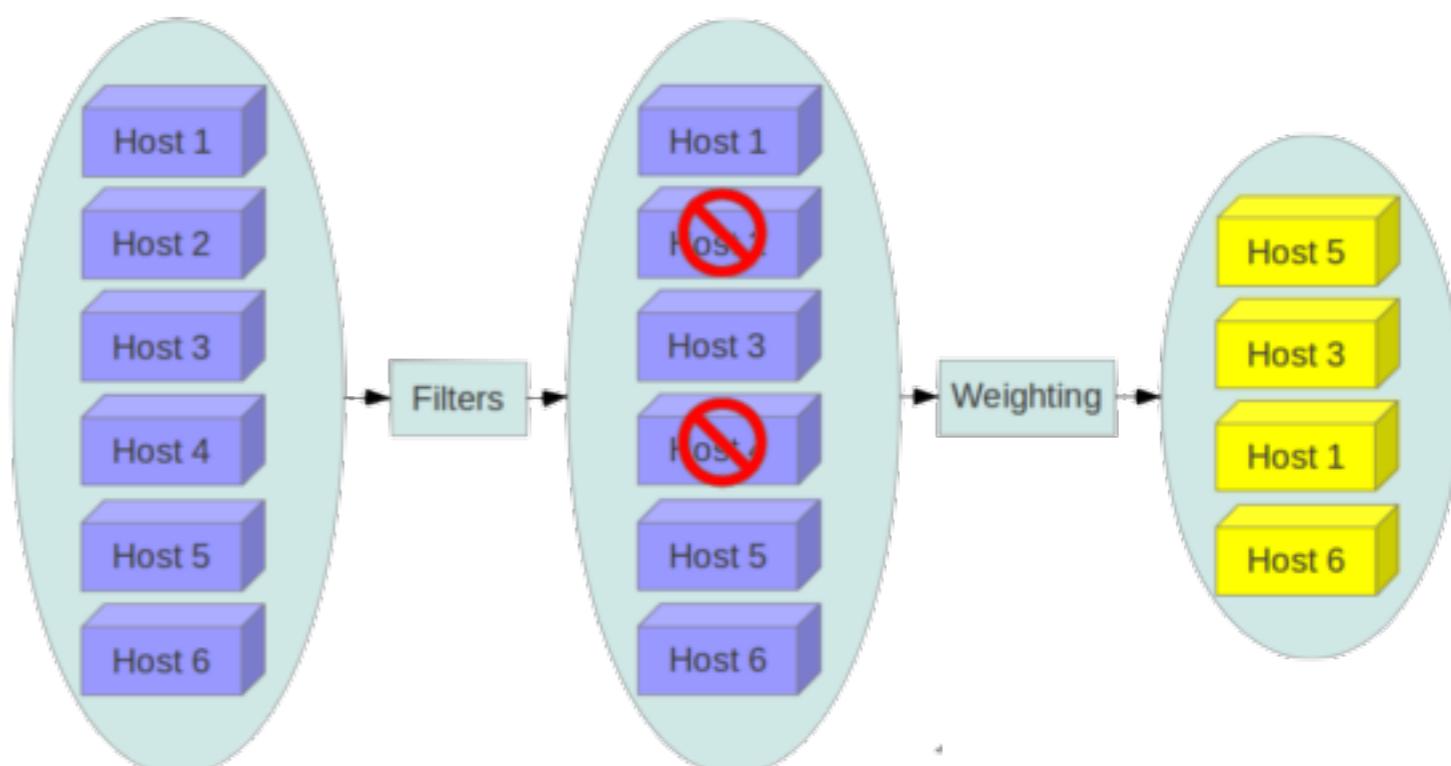
- backend dello stesso tipo (driver)

cinder.conf

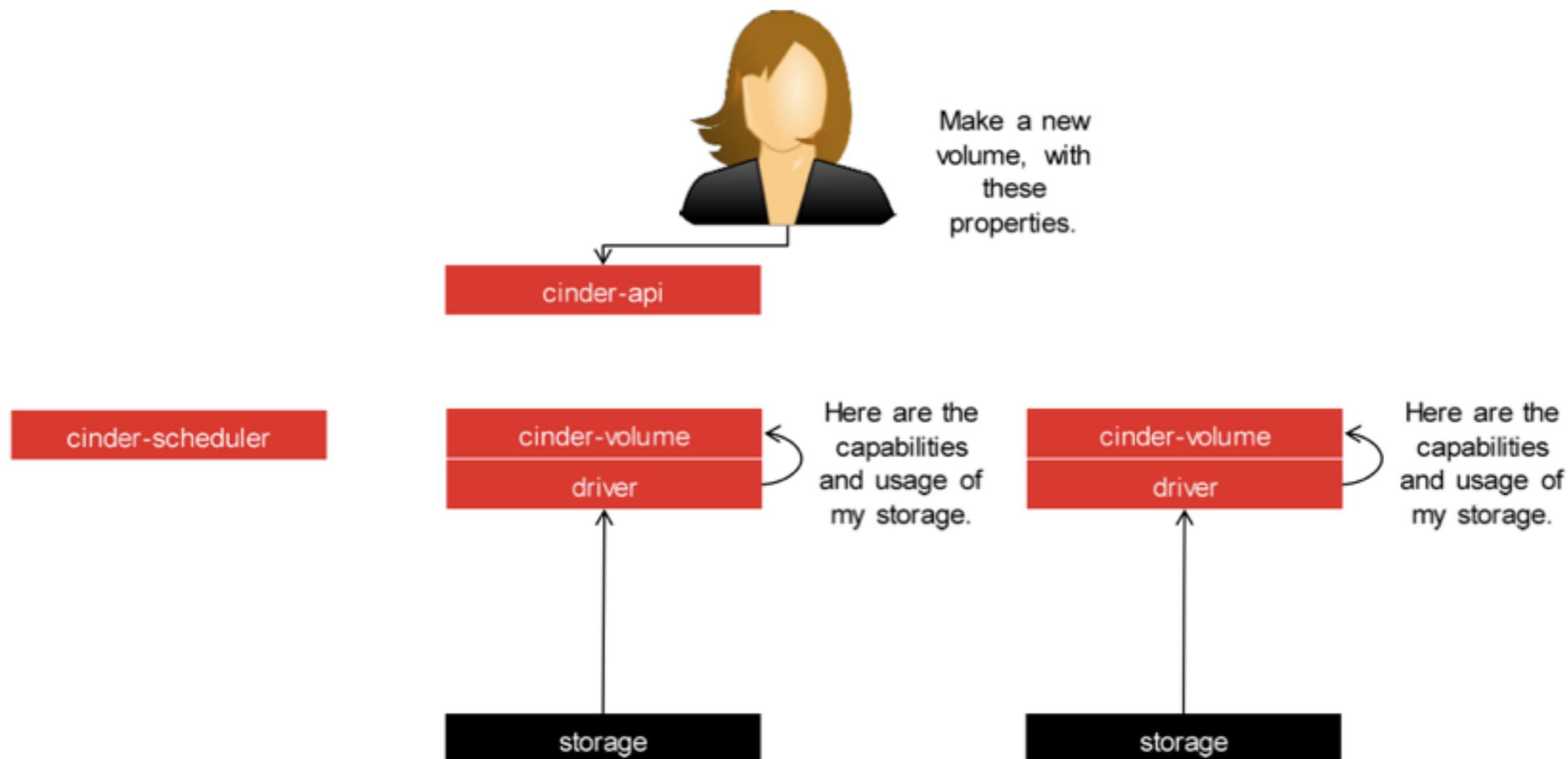
```
enabled_backends=lvmdriver-1,lvmdriver-2,lvmdriver-3
[lvmdriver-1]
volume_group=cinder-volumes-1
volume_driver=cinder.volume.drivers.lvm.LVMISCSIDriver
volume_backend_name=LVM_iSCSI
[lvmdriver-2]
volume_group=cinder-volumes-2
volume_driver=cinder.volume.drivers.lvm.LVMISCSIDriver
volume_backend_name=LVM_iSCSI
[lvmdriver-3]
volume_group=cinder-volumes-3
volume_driver=cinder.volume.drivers.lvm.LVMISCSIDriver
volume_backend_name=LVM_iSCSI_b
```

Cinder scheduler

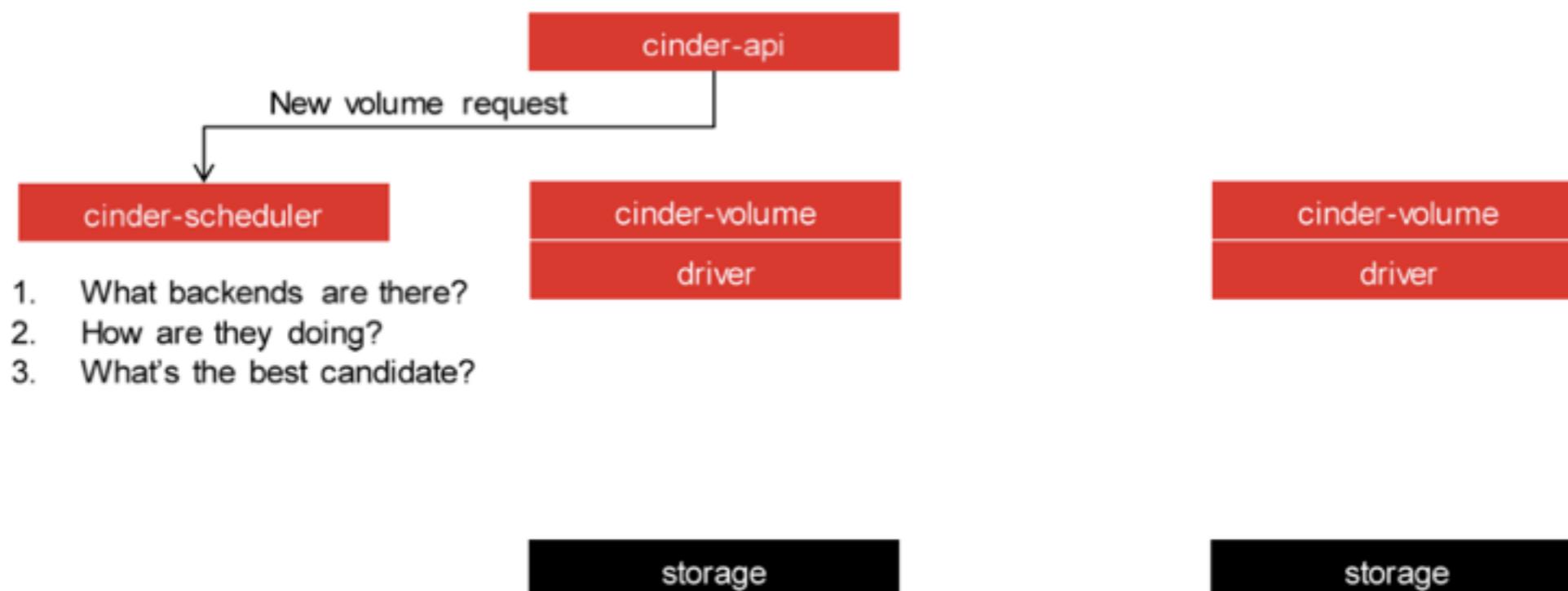
- Per utilizzare i backend multipli, va abilitato il Filter scheduler:
`scheduler_driver=cinder.scheduler.filter_scheduler.FilterScheduler`
 - ▶ Filtra i backend disponibili. Default:
AvailabilityZoneFilter, CapacityFilter e CapabilitiesFilter.
 - ▶ Associa un peso ad ogni backend filtrato. Default:
CapacityWeigher option.



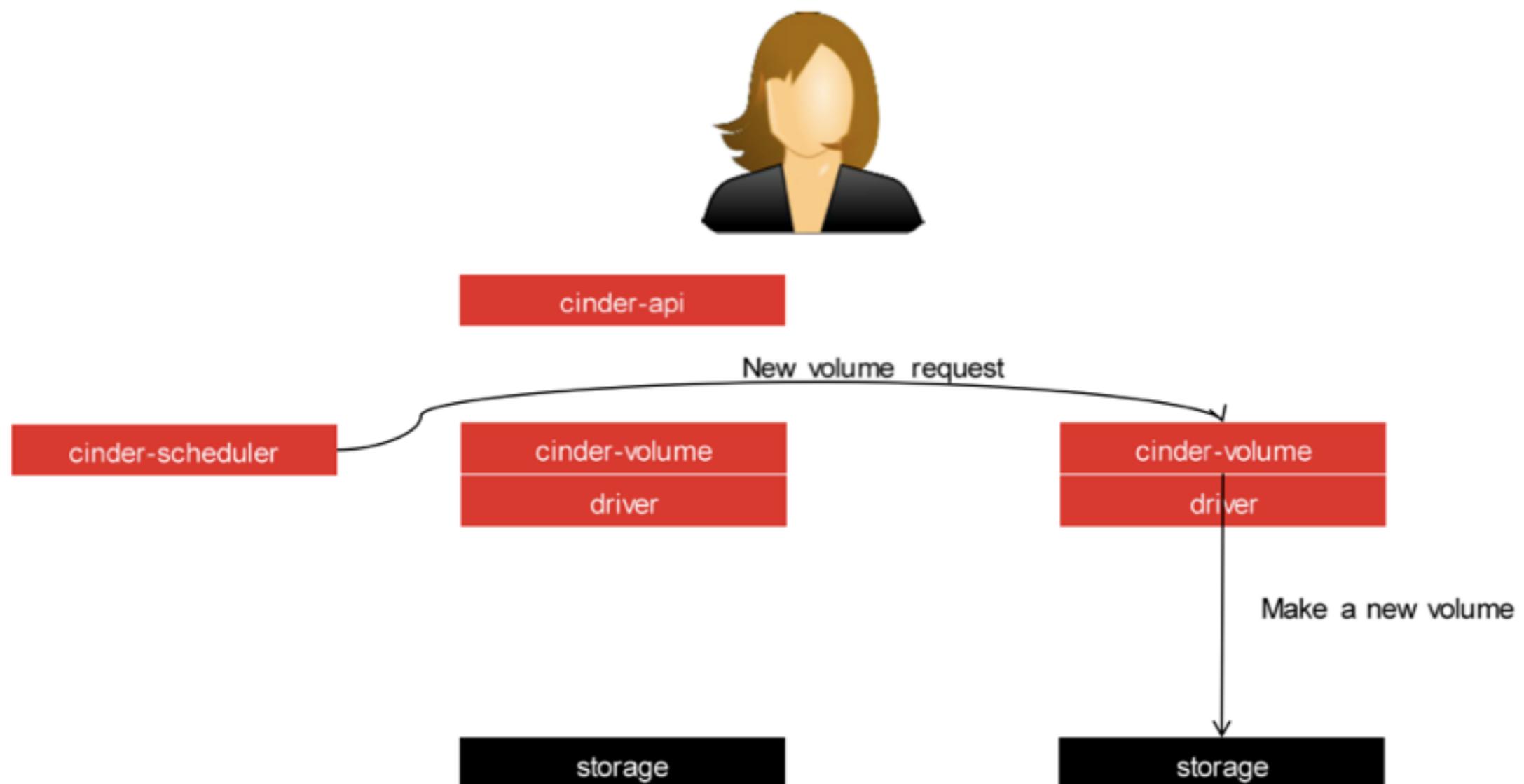
Cinder Scheduler



Cinder Scheduler



Cinder Scheduler



Uso dei volume_type

- I volume-type possono essere utilizzati per controllare dove i volumi verranno allocati:

```
# cinder create --volume_type lvm --display_name my-test 1
```

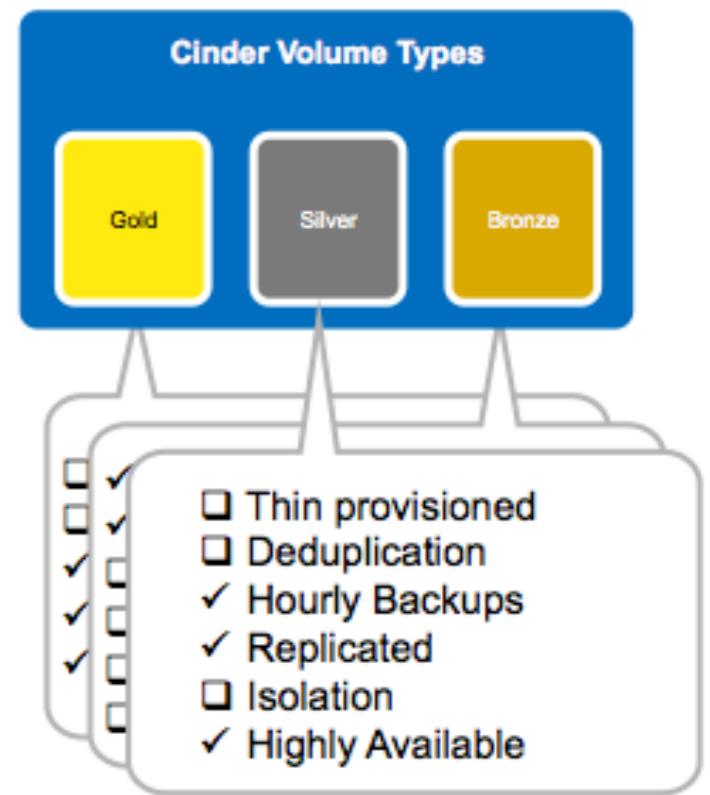
- Ogni volume-type contiene un set di coppie chiave-valore chiamati **extra-specs**. Queste informazioni sono usate dal Cinder-Scheduler per prendere decisioni sul placement dei volumi in base alle capabilities dei backend disponibili

```
# cinder type-create lvm
# cinder type-key lvm set volume_backend_name=LVM_iSCSI
# cinder extra-specs-list
```

ID	Name	extra_specs
0c91c16f-f3ab-493c-960b-75eb3f10d90e	services	{u'volume_backend_name': u'LVM_SERVICES'}
143b23ca-e47c-401f-aa0b-8b9a408c72b7	data	{u'volume_backend_name': u'CEPH_DATA'}
5c1d93c7-64a9-496a-b0dc-2f675bddb057	encrypted-data	{u'volume_backend_name': u'CEPH_DATA_ENCR'}

Quality of Service

- I Volume type possono essere usati per fornire agli utenti differenti livelli (tier) di storage:
 - ✓ performance diverse (p.e. HDD tier, mixed HDD-SDD tier, o SSD tier),
 - ✓ resilienza (selezionando differenti livelli di RAID, o replica)
 - ✓ specifiche features (p.e. compressione, data-deduplication, etc.).



Esempio di configurazione multi-tier

- Per esempio, assumiamo di avere 2 pool su Ceph che utilizzano storage device differenti:
- il pool “cinder-sata” usa un rack SATA
- il pool “cinder-ssd” usa un rack SSD

```
# Multi backend options

# Define the names of the groups for multiple volume backends
enabled_backends=rbd-sata,rbd-ssd

# Define the groups as above
[rbd-sata]
volume_driver=cinder.volume.driver.RBDDriver
rbd_pool=cinder-sata
volume_backend_name=RBD_SATA
# if cephX is enable
#rbd_user=cinder
#rbd_secret_uuid=<None>
[rbd-ssd]
volume_driver=cinder.volume.driver.RBDDriver
rbd_pool=cinder-ssd
volume_backend_name=RBD_SSD
# if cephX is enable
#rbd_user=cinder
#rbd_secret_uuid=<None>
```

Rate limiting

- Feature introdotta in **Havana**
- Implementa il supporto QoS in Nova e Cinder (sfruttando il rate limiting già supportato in KVM e QEMU attraverso libvirt) - utile nel caso in cui lo storage non espone questa funzionalità
- Il limiting può quindi essere realizzato dal “frontend” (hypervisor) o dal “backend” (storage subsystem) o entrambi
- **Backend:** campi specifici definiti dal vendor:
 - ❖ HP 3PAR (IOPS, tput: min, max; latency, priority)
 - ❖ Solidfire (IOPS: min, max, burst)
 - ❖ NetApp* (QoS Policy Group)
 - ❖ Huawei* (priority)

*defined through extra specs

Rate limiting

- **Frontend** QoS options:
 - throughput
 - total_bytes_sec: the total allowed bandwidth for the guest per second
 - read_bytes_sec: sequential read limitation
 - write_bytes_sec: sequential write limitation
 - IOPS
 - total_iops_sec: the total allowed IOPS for the guest per second
 - read_iops_sec: random read limitation
 - write_iops_sec: random write limitation
- Il file di definizione della VM a cui viene agganciato il volume con qos-specs conterrà un campo xml extra “**<iotune>**”. Es.

```
<iotune>
    <read_iops_sec>2000</read_iops_sec>
    <write_iops_sec>1000</write_iops_sec>
</iotune>
```

Rate limiting: cinder CLI

create qos specs

```
$ cinder qos-create <name> <key=value>  
[<key=value> ...]
```

```
$ cinder qos-create high-iops consumer="front-end" read_iops_sec=2000  
write_iops_sec=1000  
+-----+  
| Property | Value  
+-----+  
| consumer | front-end  
| id | c38d72f8-f4a4-4999-8acd-a17f34b040cb  
| name | high-iops  
| specs | {u'write_iops_sec': u'1000', u'read_iops_sec': u'2000'}  
+-----+
```

Associate qos specs with specific volume type

```
$ cinder qos-associate <qos_specs> <volume_type_id>
```

Esempi: extra-specs + qos-specs

Mettiamo insieme un po' tutto:

- volume-types,
- extra-specs,
- qos-specs

Volume Type	Extra Specs	QoS Specs
Gold	{netapp:disk_type=SSD, netapp_thick_provisioned=True}	{}
Silver	{}	{total_iops_sec=500}
Bronze	{volume_backend_name=lvm}	{total_iops_sec=100}

Creazione di volumi da dashboard

Create Volume

Volume Name: * vol-01

Description:

Type: (circled in purple)

Size (GB): * 100

Volume Source: No source, empty volume

Availability Zone: Any Availability Zone

Description:
Volumes are block devices that can be attached to instances.

Volume Limits

Total Gigabytes (20 GB)
1,000 <django.utils.functional.__proxy__ object at 0x7f040417ce90> Available

Number of Volumes (2) 10 Available

Cancel **Create Volume**

La dashboard consente la creazione dei volume-type, ma non permette al momento l'associazione con i backend.

Non sono neanche implementate le funzioni relative alla gestione degli encryption-type.

QoS “dinamico”

- **Volume-Retype**: consente di cambiare il tipo di volume dopo la sua creazione.
 - Questa funzionalità è utile per esempio per modificare il livello di QoS dinamicamente (nel caso in cui un volume sia sottoposto ad utilizzo pesante nel tempo e si renda necessario il passaggio ad un tier che offra un servizio migliore).
- Icehouse bug: <https://bugs.launchpad.net/python-cinderclient/+bug/1316939>
 - patch: <https://review.openstack.org/#/c/92768/>

Volume encryption

- Questa funzionalità è stata introdotta nella release Havana
- Utilizzabile nel caso di backend LVM-iSCSI
- Semplice da configurare, trasparente per l'utente finale (creare un volume cifrato richiede le stesse operazioni di un volume non cifrato)
- Il transito dei dati è sicuro
 - p.e. non è necessario usare IPsec per proteggere il traffico iSCSI
- Supporta:
 - le funzionalità esistenti in cinder (p.e. snapshot)
 - boot da volumi criptati
 - possibilità di scegliere il key-manager da usare per gestire le chiavi

Key Manager

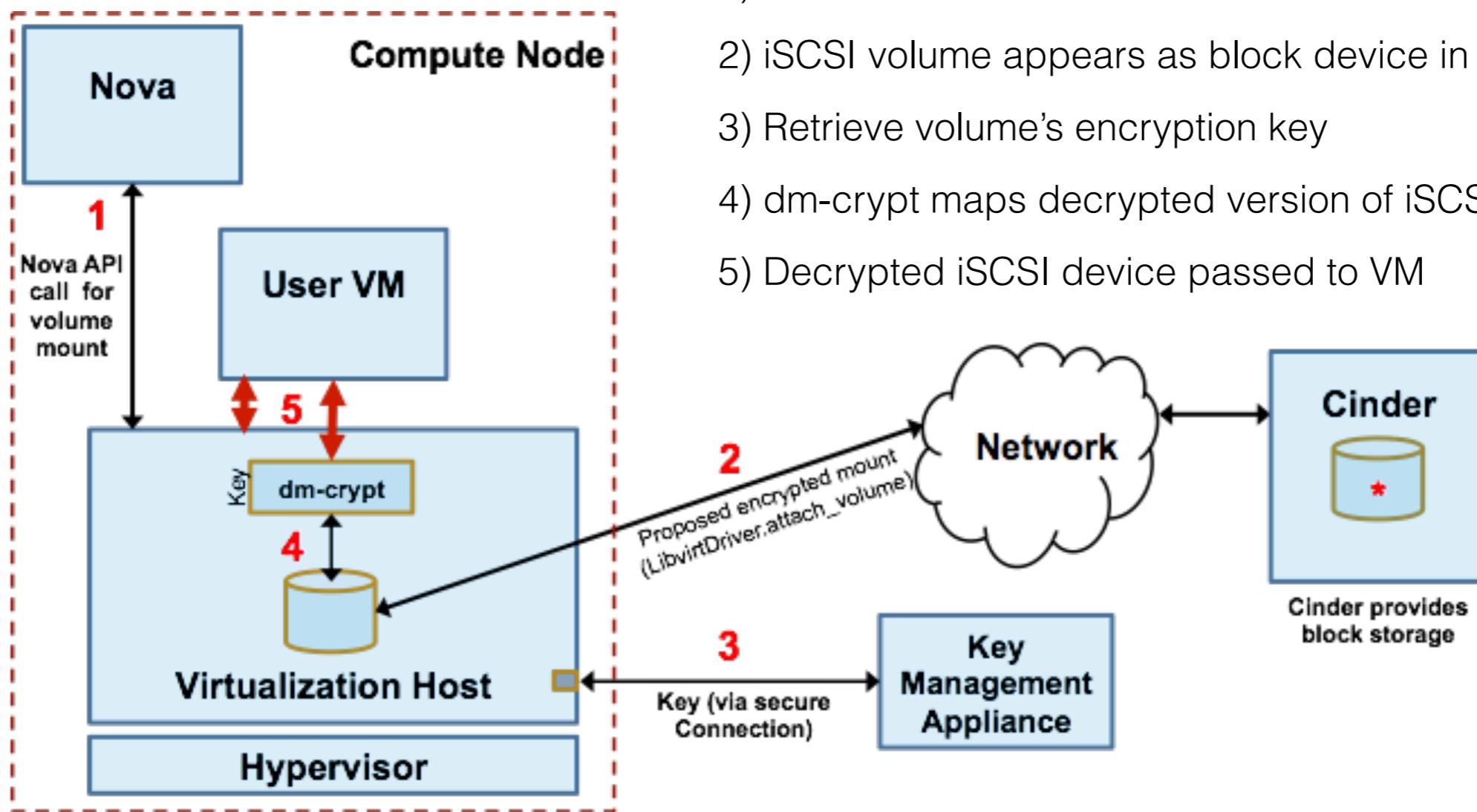
- Il key manager di default è “configuration-based”
 - supporta singola chiave statica usata per tutti i volumi
 - da NON usare in produzione. La sicurezza dei dati dipende dalla segretezza della chiave
 - consigliato utilizzare un key-manager esterno (e.g. Barbican)
- Il key-manager espone un’interfaccia astratta che consente di integrare qualunque key-manager (incluso sistemi commerciali, p.e. Safenet, IBM, HP, etc.)

Block encryption

Steps for block encryption:

Preparatory Step: Cinder creates encrypted data volume

- 1) Nova API call
- 2) iSCSI volume appears as block device in compute host
- 3) Retrieve volume's encryption key
- 4) dm-crypt maps decrypted version of iSCSI volume
- 5) Decrypted iSCSI device passed to VM



Encrypted volume-types

- estensione dell'astrazione del volume-type
- utilizzo di metadata predefiniti
 - cipher: modalità di cifratura
 - e.g. aes-cbc-essiv:sha256 o aes-xts-plain64
 - key-length: dimensione della chiave in bits
 - e.g. 128 o 256
- Provider: classe responsabile dell'attachment/detachment del volume criptato
 - nova.volume.encryptors.cryptsetup.CryptsetupEncryptor: uses “raw” cryptsetup
 - nova.volume.encryptors.luks.LuksEncryptor: uses LUKS extensions to cryptsetup
- Control location: servizio che esegue l'encryption
 - ‘front-end’ → Nova; ‘back-end’ → Cinder
 - ‘back-end’ (i.e., encryption by Cinder) not yet implemented

Ceph disk-encryption

- Ceph supporta dm-crypt
 - # ceph-deploy osd --dmcrypt [--dmcrypt-key-dir KEYDIR] create|prepare HOST:DISK
- creare pool su OSD criptati
- configurare in cinder.conf un nuovo backend associandolo al pool encrypted
- creare un volume-type specifico

Cinder backup

- Un backup è una copia del volume archiviata nell'Object Store
- Gestito da un servizio a parte: **cinder-backup** (non attivo di default)
- Driver configurabili:
 - ➔ Ceph
 - ➔ Swift
 - ➔ IBM Tivoli Storage Manager

Backup driver Swift

Modificare il file cinder.conf - sezione DEFAULT

```
backup_driver=cinder.backup.drivers.swift

# The URL of the Swift endpoint (string value)
backup_swift_url=http://localhost:8080/v1/AUTH_

# Swift authentication mechanism (string value)
backup_swift_auth=per_user

# Swift user name (string value)
#backup_swift_user=<None>

# Swift key for authentication (string value)
#backup_swift_key=<None>

# The default Swift container to use (string value)
backup_swift_container=volumebackups

# The size in bytes of Swift backup objects (integer value)
backup_swift_object_size=52428800

# The number of retries to make for Swift operations (integer
# value)
#backup_swift_retry_attempts=3

# The backoff time in seconds between Swift retries (integer
# value)
#backup_swift_retry_backoff=2

# Compression algorithm (None to disable) (string value)
#backup_compression_algorithm=zlib
```

Backup driver Swift

```
root@wn-recas-uniba-30:~# cinder backup-create --display-name test-bck 4b849af0-f989-4e95-9d79-60aede80a4ca
+-----+
| Property |          Value          |
+-----+
|   id     | 0542b982-45c5-4b39-8caf-930c05c12654 |
|   name   |           test-bck           |
| volume_id | 4b849af0-f989-4e95-9d79-60aede80a4ca |
+-----+
```

ID	Volume ID	Status	Name	Size	Object Count	Container
0542b982-45c5-4b39-8caf-930c05c12654	4b849af0-f989-4e95-9d79-60aeade80a4ca	creating	test-bck	10	None	volumebackups
a1821891-c7a1-4a31-a962-9f9fb254ebd6	4b849af0-f989-4e95-9d79-60aeade80a4ca	available	marica-test2-bck	10	206	volumebackups

```
root@wn-recas-uniba-30:~# swift stat volumebackups
  Account: AUTH_afb4978796f5422d9acdec64da6aaf5f
  Container: volumebackups
    Objects: 246
      Bytes: 13823875
      Read ACL: root@wn-
      Write ACL: volume_4
      Sync To: volume_4
      Sync Key: volume_4
Accept-Ranges: bytes
X-Storage-Policy: Policy-0
  Connection: close
  X-Timestamp: 1398779248.46843
  X-Trans-Id: tx4b892dee2660425aa764f-005488b98a
Content-Type: text/plain; charset=utf-8
```

Backup driver Ceph

Modificare il file cinder.conf - sezione DEFAULT

```
backup_driver=cinder.backup.drivers.ceph

# Ceph configuration file to use. (string value)
backup_ceph_conf=/etc/ceph/ceph.conf

# The Ceph user to connect with. Default here is to use the
# same user as for Cinder volumes. If not using cephx this
# should be set to None. (string value)
backup_ceph_user=cinder-backup

# The chunk size, in bytes, that a backup is broken into
# before transfer to the Ceph object store. (integer value)
#backup_ceph_chunk_size=134217728

# The Ceph pool where volume backups are stored. (string
# value)
backup_ceph_pool=backups

# RBD stripe unit to use when creating a backup image.
# (integer value)
#backup_ceph_stripe_unit=0

# RBD stripe count to use when creating a backup image.
# (integer value)
#backup_ceph_stripe_count=0

# If True, always discard excess bytes when restoring volumes
# i.e. pad with zeroes. (boolean value)
#restore_discard_excess_bytes=true
```

Backup di un volume RBD su Ceph

- Il driver è in grado di rilevare se il volume è un volume Ceph RBD
- in questo caso tenta di fare un backup incrementale e in caso di failure un backup full
- supporta il backup
 - ✓ all'interno dello stesso pool (not recommended)
 - ✓ tra pool diversi
 - ✓ tra cluster diversi

Ceph backup: under the hood

Workflow di creazione del primo backup di un volume

1. Create a base backup image (if it does not exists) used for storing differential exports
2. Snapshot source volume to create a new point-in-time
3. Perform differential transfer:

```
rbd export-diff --id cinder --conf /etc/ceph/ceph.conf --pool volumes volumes/volume-afa33905-0d87-42ff-ad36-9c75fdcf09ba@backup.4e50e949-3dcd-4ff1-89e0-a6a9c1beb5c1.snap.1418722200.64 -  
rbd import-diff --id cinder-backup --conf /etc/ceph/ceph.conf --pool backups - backups/volume-afa33905-0d87-42ff-ad36-9c75fdcf09ba.backup.base
```

Results in rbd:

```
# rbd -p volumes ls -l  
NAME SIZE PARENT FMT PROT  
LOCK  
volume-afa33905-0d87-42ff-ad36-9c75fdcf09ba 10240M 2  
volume-afa33905-0d87-42ff-ad36-9c75fdcf09ba@backup.4e50e949-3dcd-4ff1-89e0-a6a9c1beb5c1.snap.1418722200.64 10240M 2  
  
# rbd -p backups ls -l  
NAME SIZE PARENT FMT PROT  
LOCK  
volume-afa33905-0d87-42ff-ad36-9c75fdcf09ba.backup.base 10240M 2  
volume-afa33905-0d87-42ff-ad36-9c75fdcf09ba.backup.base@backup.4e50e949-3dcd-4ff1-89e0-a6a9c1beb5c1.snap.1418722200.64 10240M 2
```

Ceph backup: under the hood (2)

Workflow di creazione del backup successivo

1. Snapshot source volume to create a new point-in-time
2. Perform differential transfer using --from-snap:

```
rbd export-diff --id cinder --conf /etc/ceph/ceph.conf --pool volumes --from-snap backup.4e50e949-3dcd-4ff1-89e0-a6a9c1beb5c1.snap.1418722200.64 volumes/volume-afa33905-0d87-42ff-ad36-9c75fdcf09ba@backup.c255e3ca-f01b-4fe6-ad9f-af0524a7b531.snap.1418725945.25 -  
rbd import-diff --id cinder-backup --conf /etc/ceph/ceph.conf --pool backups - backups/volume-afa33905-0d87-42ff-ad36-9c75fdcf09ba.backup.base
```

Results in rbd:

```
# rbd -p volumes ls -l  
NAME SIZE PARENT FMT  
PROT LOCK  
volume-afa33905-0d87-42ff-ad36-9c75fdcf09ba 10240M 2  
volume-afa33905-0d87-42ff-ad36-9c75fdcf09ba@backup.c255e3ca-f01b-4fe6-ad9f-af0524a7b531.snap.1418725945.25 10240M 2  
  
# rbd -p backups ls -l  
NAME SIZE  
PARENT FMT PROT LOCK  
volume-afa33905-0d87-42ff-ad36-9c75fdcf09ba.backup.base 10240M  
2  
volume-afa33905-0d87-42ff-ad36-9c75fdcf09ba.backup.base@backup.4e50e949-3dcd-4ff1-89e0-a6a9c1beb5c1.snap.1418722200.64 10240M  
2  
volume-afa33905-0d87-42ff-ad36-9c75fdcf09ba.backup.base@backup.c255e3ca-f01b-4fe6-ad9f-af0524a7b531.snap.1418725945.25 10240M  
2
```

Verso il disaster-recovery

- La funzionalità di cinder **backup-restore** consente di ripristinare lo stato di un volume all'interno della stessa istanza di Openstack.
- A partire da Icehouse esiste un'estensione delle API di cinder backup:
 - **import/export** dei metadata
 - occorre patchare il client *Icehouse* (<https://review.openstack.org/#/c/72743>)
- In **Juno** le API di cinder sono state ulteriormente arricchite per supportare la replica dei volumi

Cinder backup export

```
# cinder --os-volume-api-version 2 backup-export 4e50e949-3dcd-4ff1-89e0-a6a9c1beb5c1
```

Property	Value
backup_service	cinder.backup.drivers.ceph
backup_url	eyJzdGF0dXMi0iAiYXZhaWxhYmxlIiwgIm9iamVjdF9jb3VudCI6IG51bGwsICJkZWxldGVkX2F0IjogbnVsbCwgInByb2p1Y3RfaWQi0iAiYWZiNDk30Dc5NmY1NDIyZDlhY2R1YzY0ZGE2YWFmNWYiLCAidXNlcl9pZCI6ICIxM2IzMDBmOTkwZW00DI2YT1zY2QyNTIw0D1lNmNhZCIsICJzZXJ2aWNlIjogImNpbmRlc5iYWNrdXAuZHJpdmVcy5jZXBoIiwgImF2YWlsYWJpbGl0eV96b25lIjogIm5vdmEiLC AiZGVsZXRLZCI6IGZhHN1LCAiY3J1YXR1ZF9hdCI6ICIyMDE0LTEyLTE2VDA50jMw0jAwLjAwMDAwMCIsICJ1cGRhdGVkX2F0IjogIjIwMTQtMTItMTZUMTE6Mjc6MTkuMDAwMDAwIiwgImRp c3BsYXlfZGVzY3JpcHRpb24i0iBudWxsLCAiaG9zdCI6ICJ3bi1yZWNhcy11bml iYS0zMCIsICJjb250YwluZXIi0iAiYmFja3VwcyIsICJ2b2x1bWVfaWQi0iAiYwZhMzM5MDUtMGQ4Ny00MmZmLWFk MzYt0WM3NWZkY2Yw0WJhIiwgImRpc3BsYXlfbmFtZSI6ICJ2b2wt dGVzdC1iY2siLC AiZmFpbF9yZWFzb24i0iBudWxsLCAic2VydmljZV9tZXRhZGF0YSI6IG51bGwsICJpZCI6ICI0ZTUwZTk00S0z ZGNkLTRmZjEt0D1lMC1hNmE5YzFiZWI1YzEiLC Aiic2l6ZSI6IDEwfQ==

base64 --decode

```
{"status": "available", "object_count": null, "deleted_at": null, "project_id": "afb4978796f5422d9acdec64da6aaf5f", "user_id": "13b300f990ec4826a23cd252089e6cad", "service": "cinder.backup.drivers.ceph", "availability_zone": "nova", "deleted": false, "created_at": "2014-12-16T09:30:00.000000", "updated_at": "2014-12-16T11:27:19.000000", "display_description": null, "host": "wn-recas-uniba-30", "container": "backups", "volume_id": "afa33905-0d87-42ff-ad36-9c75fdcf09ba", "display_name": "vol-test-bck", "fail_reason": null, "service_metadata": null, "id": "4e50e949-3dcd-4ff1-89e0-a6a9c1beb5c1", "size": 10}
```

Cinder backup import

```
# cinder --os-volume-api-version 2 backup-import cinder.backup.drivers.ceph  
eyJzdGF0dXMi0iAiYXZhaWxhYmxlIiwgIm9iamVjdF9jb3VudCI6IG51bGwsICJkZWxldGVkX2F0IjogbnVsbCwgInByb2plY3RfaWQi0  
iAiYWZiNDk30Dc5NmY1NDIyZDlhY2RlYzY0ZGE2YWFmNWYiLCAidXNlc19pZCI6ICIxM2IzMDBmOTkwZWM00DI2YTIzY2QyNTIwODl1Nm  
NhZCIsICJzZXJ2aWNlIjogImNpbmRlc15iYWNrdXAuZHJpdmVycy5jZXBoIiwgImF2YwlsYWJpbGl0eV96b25lIjogIm5vdmEiLCAiZGV  
sZXRLZCI6IGZhHNllCAiY3JlYXRLZF9hdCI6ICJyMDE0LTEyLTE2VDA50jMw0jAwLjAwMDAwMCIsICJ1cGRhdGVkX2F0IjogIjIwMTQt  
MTItMTZUMTE6Mjc6MTkuMDAwMDAwIiwgImRpc3BsYXlfZGVzY3JpcHRpb24i0iBudWxsLCAiaG9zdCI6ICJ3bi1yZWNhcyclbml1YS0zM  
CIIsICJjb250YWluZXIi0iAiYmFja3VwcyIsICJ2b2x1bWVfaWQi0iAiYwZhMzM5MDUtMGQ4Ny00MmZmLWFkMzYt0WM3NWZkY2Yw0WJhIi  
wgImRpc3BsYXlfbmFtZSI6ICJ2b2wtdGVzdC1iY2siLCAiZmFpbF9yZWFBz24i0iBudWxsLCAic2VydmIjZV9tZXRhZGF0YSI6IG51bGw  
sICJpZCI6ICJ0ZTUwZTk00S0zZGNkLTRmZjEt0D1lMC1hNmE5YzFiZWI1YzEiLCAic2l6ZSI6IDEwfQ==  
+-----+  
| Property | Value |  
+-----+  
| id | 9a3d360a-8cb2-42a3-9289-7279cf6b67cc |  
| name | None |  
+-----+
```

```
# cinder backup-show 9a3d360a-8cb2-42a3-9289-7279cf6b67cc  
+-----+  
| Property | Value |  
+-----+  
| availability_zone | nova |  
| container | backups |  
| created_at | 2014-12-16T13:52:42.000000 |  
| description | None |  
| fail_reason | None |  
| id | 9a3d360a-8cb2-42a3-9289-7279cf6b67cc |  
| name | vol-test-bck |  
| object_count | None |  
| size | 10 |  
| status | available |  
| volume_id | 0000-0000-0000-0000 |  
+-----+
```