



SUSE Enterprise Storage 7.1

Guide d'opérations et d'administration

Guide d'opérations et d'administration

SUSE Enterprise Storage 7.1


par Tomáš Bažant, Alexandra Settle, et Liam Proven

Date de publication : 20 mar 2025

<https://documentation.suse.com> 

Copyright © 2020–2025 SUSE LLC et contributeurs. Tous droits réservés.

Sauf indication contraire, ce document est concédé sous licence Creative Commons Attribution-ShareAlike 4.0 International (CC-BY-SA 4.0) : <https://creativecommons.org/licenses/by-sa/4.0/legalcode> .

Pour les marques commerciales SUSE, consultez le site Web <http://www.suse.com/company/legal/> . Toutes les marques commerciales de fabricants tiers appartiennent à leur propriétaire respectif. Les symboles de marque

commerciale (®, ™, etc.) désignent des marques de SUSE et de ses sociétés affiliées. Des astérisques (*) désignent des marques commerciales de fabricants tiers.

Toutes les informations de cet ouvrage ont été regroupées avec le plus grand soin. Cela ne garantit cependant pas sa complète exactitude. Ni SUSE LLC, ni les sociétés affiliées, ni les auteurs, ni les traducteurs ne peuvent être tenus responsables des erreurs possibles ou des conséquences qu'elles peuvent entraîner.

Table des matières

À propos de ce guide xviii

- 1 Documentation disponible xviii
- 2 Commentaires xix
- 3 Conventions relatives à la documentation xx
- 4 Support xxii
 - Déclaration de support pour SUSE Enterprise Storage xxii • Avant-premières technologiques xxiii
- 5 Contributeurs Ceph xxiv
- 6 Commandes et invites de commande utilisées dans ce guide xxv
 - Commandes associées à Salt xxv • Commandes associées à Ceph xxv • Commandes Linux générales xxvii • Informations supplémentaires xxvii

I CEPH DASHBOARD 1

1 À propos de Ceph Dashboard 2

2 Interface utilisateur Web du tableau de bord 3

- 2.1 Connexion 3
- 2.2 Menu utilitaire 5
- 2.3 Menu principal 6
- 2.4 Volet de contenu 7
- 2.5 Fonctionnalités courantes de l'interface utilisateur Web 7
- 2.6 Widgets du tableau de bord 7
 - Widgets de statut 8 • Widgets de capacité 8 • Widgets de performance 9

3 Gestion des utilisateurs et des rôles Ceph Dashboard 11

- 3.1 Liste des utilisateurs 11
- 3.2 Ajout de nouveaux utilisateurs 11
- 3.3 Modification des utilisateurs 12
- 3.4 Suppression d'utilisateurs 12
- 3.5 Liste des rôles des utilisateurs 13
- 3.6 Ajout de rôles personnalisés 13
- 3.7 Modification des rôles personnalisés 15
- 3.8 Suppression des rôles personnalisés 15

4 Affichage des éléments internes de la grappe 16

- 4.1 Affichage des noeuds de grappe 16
- 4.2 Accès à l'inventaire de la grappe 16
- 4.3 Affichage des instances Ceph Monitor 17
- 4.4 Affichage des services 18
- 4.5 Affichage des OSD Ceph 19
 - Ajout d'OSD 22
- 4.6 Affichage de la configuration de la grappe 26
- 4.7 Affichage de la carte CRUSH 26
- 4.8 Affichage des modules Manager 27
- 4.9 Affichage des journaux 28
- 4.10 Affichage de la surveillance 28

5 Gestion des réserves 29

- 5.1 Ajout d'une nouvelle réserve 30
- 5.2 Suppression de réserves 30

5.3 Modification des options d'une réserve 31

6 Gestion du périphérique de traitement par blocs RADOS (RBD) 32

6.1 Affichage des détails sur les RBD 33

6.2 Affichage de la configuration d'un RBD 34

6.3 Création de RBD 35

6.4 Suppression de RBD 36

6.5 Création d'instantanés de périphériques de traitement par blocs RADOS (RBD) 36

6.6 Mise en miroir de RBD 37

Configuration de grappes primaires et secondaires 38 • Activation du daemon `rbd-mirror` 39 • Désactivation de la mise en miroir 40 • Démarrage des homologues 41 • Suppression d'un homologue de grappe 42 • Configuration de la réplication de réserve dans Ceph Dashboard 42 • Vérification du fonctionnement de la réplication d'image RBD 46

6.7 Gestion des passerelles iSCSI 49

Ajout de cibles iSCSI 50 • Modification des cibles iSCSI 52 • Suppression de cibles iSCSI 52

6.8 Qualité de service (QoS) RBD 52

Configuration globale des options 53 • Configuration des options d'une nouvelle réserve 54 • Configuration des options d'une réserve existante 54 • Options de configuration 54 • Création d'options QoS RBD avec une nouvelle image RBD 55 • Modification des options QoS RBD sur les images existantes 55 • Modification des options de configuration lors de la copie ou du clonage d'images 55

7 Gestion de NFS Ganesha 56

7.1 Création d'exportations NFS 57

7.2 Suppression des exportations NFS 59

7.3 Modification d'exportations NFS 59

8 Gestion de CephFS 61

8.1 Affichage de l'aperçu CephFS 61

9 Gestion de la passerelle Object Gateway 63

9.1 Affichage des instances Object Gateway 63

9.2 Gestion des utilisateurs Object Gateway 64

Ajout d'un nouvel utilisateur de la passerelle 65 • Suppression d'utilisateurs de la passerelle 67 • Modification des détails des utilisateurs de la passerelle 67

9.3 Gestion des compartiments Object Gateway 67

Ajout d'un nouveau compartiment 67 • Affichage des détails d'un compartiment 68 • Modification du compartiment 69 • Suppression d'un compartiment 70

10 Configuration manuelle 71

10.1 Configuration de la prise en charge de TLS/SSL 71

Création de certificats auto-signés 72 • Utilisation de certificats signés par une autorité de certification 72

10.2 Modification du nom d'hôte et du numéro de port 73

10.3 Modification des noms d'utilisateur et des mots de passe 74

10.4 Activation de l'interface client de gestion d'Object Gateway 75

10.5 Activation de la gestion iSCSI 76

10.6 Activation de Single Sign-On 77

11 Gestion des utilisateurs et des rôles via la ligne de commande 79

11.1 Gestion de la stratégie de mot de passe 79

11.2 Gestion des comptes utilisateur 80

| | |
|------|--|
| 11.3 | Rôles et autorisations des utilisateurs 81 |
| | Définition des étendues de sécurité 81 • Spécification des rôles utilisateur 82 |
| 11.4 | Configuration du proxy 85 |
| | Accès au tableau de bord avec des proxys inverses 85 • Désactivation des réacheminements 85 • Configuration des codes de statut d'erreur 86 • Exemple de configuration HAProxy 86 |
| 11.5 | Audit des requêtes API 87 |
| 11.6 | Configuration de NFS Ganesha dans Ceph Dashboard 88 |
| | Configuration de plusieurs grappes NFS Ganesha 88 |
| 11.7 | Plug-ins de débogage 89 |
| | |
| II | OPÉRATION DE GRAPPE 90 |
| 12 | Détermination de l'état d'une grappe 91 |
| 12.1 | Vérification de l'état d'une grappe 91 |
| 12.2 | Vérification de l'état de santé de la grappe 93 |
| 12.3 | Vérification des statistiques d'utilisation d'une grappe 103 |
| 12.4 | Vérification de l'état des OSD 105 |
| 12.5 | Contrôle des OSD pleins 105 |
| 12.6 | Vérification de l'état des instances Monitor 106 |
| 12.7 | Vérification des états des groupes de placement 107 |
| 12.8 | Capacité de stockage 107 |
| 12.9 | Surveillance des OSD et des groupes de placement 110 |
| | Surveillance des OSD 111 • Assignation d'ensembles de groupes de placement 112 • Homologation 114 • Surveillance des états des groupes de placement 114 • Identification de l'emplacement d'un objet 120 |
| 13 | Tâches opérationnelles 122 |
| 13.1 | Modification de la configuration d'une grappe 122 |

- 13.2 Ajout de noeuds 122
- 13.3 Suppression de noeuds 123
- 13.4 Gestion des OSD 125
 - Liste des périphériques de disque 125 • Effacement de périphériques de disque 126 • Ajout d'OSD à l'aide de la spécification DriveGroups 126 • Suppression des OSD 136 • Remplacement d'OSD 137
- 13.5 Déplacement du Salt Master vers un nouveau noeud 139
- 13.6 Mise à jour des noeuds de grappe 141
 - Dépôts logiciels 141 • Préparation du dépôt 141 • Temps d'indisponibilité des services Ceph 141 • Exécution de la mise à jour 142
- 13.7 Mise à jour de Ceph 142
 - Démarrage de la mise à jour 142 • Surveillance de la mise à jour 143 • Annulation d'une mise à jour 143
- 13.8 Arrêt ou redémarrage de la grappe 143
- 13.9 Suppression d'une grappe Ceph entière 144

14 Exécution des services Ceph 145

- 14.1 Exécution de services individuels 145
- 14.2 Exécution de types de service 146
- 14.3 Exécution de services sur un seul noeud 146
 - Identification des services et des cibles 146 • Exécution de l'ensemble des services sur un noeud 147 • Exécution d'un service spécifique sur un noeud 147 • Vérification de l'état des services 148
- 14.4 Arrêt et redémarrage de l'ensemble de la grappe Ceph 148

15 Sauvegarde et restauration 150

- 15.1 Sauvegarde de la configuration et des données de grappe 150
 - Sauvegarde de la configuration de ceph-salt 150 • Sauvegarde de la configuration Ceph 150 • Sauvegarde de la configuration Salt 150 • Sauvegarde des configurations personnalisées 151

15.2 Restauration d'un noeud Ceph 151

16 Surveillance et alertes 153

16.1 Configuration d'images personnalisées ou locales 154

16.2 Mise à jour des services de surveillance 156

16.3 Désactivation de la surveillance 157

16.4 Configuration de Grafana 157

16.5 Configuration du module Prometheus Manager 158

Configuration de l'interface réseau 158 • Configuration du paramètre `scrape_interval` 159 • Configuration du cache 159 • Activation de la surveillance des images RBD 160

16.6 Modèle de sécurité de Prometheus 160

16.7 Passerelle SNMP de Prometheus Alertmanager 161

III STOCKAGE DE DONNÉES DANS UNE GRAPPE 162

17 Gestion des données stockées 163

17.1 Périphériques OSD 164

Classes de périphériques 164

17.2 Compartiments 172

17.3 Ensembles de règles 176

Itération de l'arborescence de noeuds 177 • `firstn` et `indep` 179

17.4 Groupes de placement 180

Utilisation de groupes de placement 180 • Détermination de la valeur de `PG_NUM` 182 • Définition du nombre de groupes de placement 184 • Détermination du nombre de groupes de placement 184 • Détermination des statistiques relatives aux groupes de placement d'une grappe 185 • Détermination des statistiques relatives aux groupes de placement bloqués 185 • Recherche d'une assignation de groupe de placement 185 • Récupération des statistiques d'un groupe de placement 186 • Nettoyage d'un

- groupe de placement 186 • Définition de priorités pour le renvoi et la récupération des groupes de placement 186 • Rétablissement des objets perdus 187 • Activation de la mise à l'échelle automatique des groupes de placement 188
- 17.5 Manipulation de la carte CRUSH 189
 - Modification d'une carte CRUSH 189 • Ajout ou déplacement d'un OSD 190 • Différence entre **ceph osd reweight** et **ceph osd crush reweight** 191 • Suppression d'un OSD 192 • Ajout d'un compartiment 192 • Déplacement d'un compartiment 192 • Suppression d'un compartiment 192
- 17.6 Nettoyage des groupes de placement 193
- 18 Gestion des réserves de stockage 196**
 - 18.1 Création d'une réserve 197
 - 18.2 Liste des réserves 198
 - 18.3 Modification du nom d'une réserve 199
 - 18.4 Suppression d'une réserve 199
 - 18.5 Autres opérations 200
 - Association de réserves à une application 200 • Définition de quotas de réserve 201 • Affichage des statistiques d'une réserve 201 • Obtention de valeurs d'une réserve 203 • Définition des valeurs d'une réserve 203 • Définition du nombre de répliques d'objets 207
 - 18.6 Migration d'une réserve 209
 - Limites 210 • Migration à l'aide du niveau de cache 210 • Migration d'images RBD 213
 - 18.7 Instantanés de réserve 213
 - Création d'un instantané d'une réserve 214 • Liste des instantanés d'une réserve 214 • Suppression d'un instantané d'une réserve 214
 - 18.8 Compression des données 214
 - Activation de la compression 215 • Options de compression de réserve 215 • Options de compression globales 217

19 Réserves codées à effacement 219

- 19.1 Conditions préalables pour les réserves codées à effacement 219
- 19.2 Création d'un exemple de réserve codée à effacement 219
- 19.3 Profils de code à effacement 220
 - Création d'un profil de code à effacement 223 • Suppression d'un profil de code à effacement 224 • Affichage des détails d'un profil de code à effacement 224 • Liste des profils de code à effacement 225
- 19.4 Marquage des réserves codées à effacement avec périphérique de bloc RADOS (RBD) 225

20 Périphérique de bloc RADOS 226

- 20.1 Commandes de périphériques de bloc 226
 - Création d'une image de périphérique de bloc dans une réserve répliquée 227 • Création d'une image de périphérique de bloc dans une réserve codée à effacement 227 • Liste des images de périphériques de bloc 228 • Récupération d'informations sur l'image 228 • Redimensionnement d'une image de périphérique de bloc 228 • Suppression d'une image de périphérique de bloc 228
- 20.2 Montage et démontage 229
 - Création d'un compte utilisateur Ceph 229 • Authentification des utilisateurs 230 • Préparation du périphérique de bloc RADOS à utiliser 230 • **rbdmap** : assignation de périphériques RBD au moment du démarrage 232 • Augmentation de la taille des périphériques RBD 233
- 20.3 Images instantanées 234
 - Activation et configuration de cephx 234 • Notions de base sur les instantanés 235 • Superposition d'instantanés 237
- 20.4 Miroirs d'image RBD 241
 - Configuration de la réserve 242 • Configuration de l'image RBD 246 • Vérification de l'état du miroir 251
- 20.5 Paramètres de cache 252
- 20.6 Paramètres QoS 253

- 20.7 Paramètres de la lecture anticipée 255
- 20.8 Fonctions avancées 256
- 20.9 Assignment RBD à l'aide d'anciens clients de kernel 258
- 20.10 Activation des périphériques de bloc et de Kubernetes 259
 - Utilisation de périphériques de bloc Ceph dans Kubernetes 262

IV ACCÈS AUX DONNÉES DE LA GRAPPE 266

21 Ceph Object Gateway 267

- 21.1 Restrictions d'Object Gateway et règles de dénomination 267
 - Limitations des compartiments 267 • Limitations des objets stockés 267 • Limitations d'en-tête HTTP 268
- 21.2 Déploiement de la passerelle Object Gateway 268
- 21.3 Exploitation du service Object Gateway 268
- 21.4 Options de configuration 268
- 21.5 Gestion des accès à la passerelle Object Gateway 269
 - Accès à Object Gateway 269 • Gestion des comptes S3 et Swift 271
- 21.6 Interfaces clients HTTP 275
- 21.7 Activation de HTTPS/SSL pour les passerelles Object Gateway 275
 - Création d'un certificat auto-signé 275 • Configuration d'Object Gateway avec SSL 276
- 21.8 Modules de synchronisation 277
 - Configuration des modules de synchronisation 277 • Synchronisation des zones 278 • Module de synchronisation Elasticsearch 280 • Module de synchronisation cloud 283 • Module de synchronisation de l'archivage 288
- 21.9 authentification LDAP 289
 - Mécanisme d'authentification 289 • Configuration requise 289 • Configuration de la passerelle Object Gateway en vue de l'utilisation de l'authentification LDAP 290 • Utilisation d'un filtre de

- recherche personnalisé pour limiter l'accès des utilisateurs 291 • Génération d'un jeton d'accès pour l'authentification LDAP 292
- 21.10 Partitionnement d'index de compartiment 293
 - Repartitionnement d'index de compartiment 293 • Partitionnement d'index des nouveaux compartiments 296
- 21.11 Intégration à OpenStack Keystone 297
 - Configuration d'OpenStack 297 • Configuration de la passerelle Ceph Object Gateway 298
- 21.12 Placement de réserve et classes de stockage 300
 - Affichage des cibles de placement 300 • Classes de stockage 301 • Configuration des groupes de zones et des zones 301 • Personnalisation du placement 303 • Utilisation des classes de stockage 305
- 21.13 Passerelles Object Gateway multisites 305
 - Exigences et hypothèses 306 • Configuration d'une zone maître 307 • Configuration des zones secondaires 313 • Maintenance générale d'Object Gateway 319 • Basculement et reprise après sinistre 321

22 Passerelle Ceph iSCSI 323

- 22.1 Cibles gérées par ceph-iscsi 323
 - Connexion à open-iscsi 323 • Connexion Microsoft Windows (initiateur iSCSI de Microsoft) 327 • Connexion de VMware 334
- 22.2 Conclusion 340

23 Système de fichiers en grappe 341

- 23.1 Montage de CephFS 341
 - Préparation du client 341 • Création d'un fichier de secret 342 • Montage de CephFS 342
- 23.2 Démontage de CephFS 344
- 23.3 Montage de CephFS dans /etc/fstab 344

- 23.4 Daemons MDS actifs multiples (MDS actif-actif) 344
 - Utilisation de MDS actif-actif 344 • Augmentation de la taille de la grappe active MDS 345 • Diminution du nombre de rangs 346 • Épinglage manuel d'arborescences de répertoires à un rang 346
- 23.5 Gestion du basculement 347
 - Configuration de daemons de secours avec relecture 347
- 23.6 Définition des quotas CephFS 348
 - Limites des quotas CephFS 348 • Configuration des quotas CephFS 349
- 23.7 Gestion des instantanés CephFS 350
 - Création d'instantanés 350 • Suppression d'instantanés 351
- 24 Exportation des données Ceph via Samba 352**
- 24.1 Exportation de CephFS via un partage Samba 352
 - Configuration et exportation de paquetages Samba 352 • Exemple de passerelle unique 353 • Configuration de la haute disponibilité 356
- 24.2 Jointure de la passerelle Samba et d'Active Directory 362
 - Préparation de l'installation de Samba 362 • Vérification de DNS 363 • Résolution des enregistrements SRV 363 • Configuration de Kerberos 364 • Résolution du nom de l'hôte local 364 • Configuration de Samba 365 • Jointure du domaine Active Directory 369 • Configuration de NSS 369 • Démarrage des services 369 • Test de la connectivité winbindd 370
- 25 NFS Ganesha 371**
- 25.1 Création d'un service NFS 372
- 25.2 Démarrage ou redémarrage de NFS Ganesha 373
- 25.3 Liste des objets dans la réserve de récupération NFS 373
- 25.4 Création d'une exportation NFS 373
- 25.5 Vérification de l'exportation NFS 374
- 25.6 Montage de l'exportation NFS 375
- 25.7 Plusieurs grappes NFS Ganesha 375

V INTÉGRATION DES OUTILS DE VIRTUALISATION 376

26 libvirt et Ceph 377

- 26.1 Configuration de Ceph avec libvirt 377
- 26.2 Préparation du gestionnaire de machines virtuelles 378
- 26.3 Création d'une machine virtuelle 379
- 26.4 Configuration de la machine virtuelle 379
- 26.5 Résumé 382

27 Ceph comme support de l'instance QEMU/KVM 383

- 27.1 Installation qemu-block-rbd 383
- 27.2 Utilisation de QEMU 383
- 27.3 Création d'images avec QEMU 384
- 27.4 Redimensionnement d'images avec QEMU 384
- 27.5 Récupération d'informations d'image avec QEMU 384
- 27.6 Exécution de QEMU avec RBD 385
- 27.7 Activation du rejet et de TRIM 385
- 27.8 Définition des options du cache QEMU 386

VI CONFIGURATION D'UNE GRAPPE 388

28 Configuration de la grappe Ceph 389

- 28.1 Configuration du fichier ceph.conf 389
 - Accès à ceph.conf dans des images de conteneur 389
- 28.2 Base de données de configuration 390
 - Configuration des sections et des masques 390 • Définition et lecture des options de configuration 391 • Configuration des daemons lors de l'exécution 391
- 28.3 config-key stocker 394
 - Passerelle iSCSI 395

- 28.4 Ceph OSD et BlueStore 395
 - Configuration du dimensionnement automatique du cache 395
- 28.5 Ceph Object Gateway 396
 - Paramètres généraux 397 • Configuration des interfaces clients HTTP 407

29 Modules Ceph Manager 410

- 29.1 Équilibreur 410
 - Mode « crush-compat » 411 • Planification et exécution de l'équilibrage des données 411
- 29.2 Activation du module de télémétrie 413

30 Authentification avec cephx 415

- 30.1 Architecture d'authentification 415
- 30.2 Les zones de gestion principales 418
 - Informations de base 419 • Gestion des utilisateurs 422 • Gestion des trousseaux 427 • Utilisation de la ligne de commande 430

A Mises à jour de la maintenance de Ceph basées sur les versions intermédiaires de « Pacific » en amont 431

Glossaire 432

À propos de ce guide

Ce guide est axé sur les tâches de routine que vous devez effectuer en tant qu'administrateur après le déploiement de la grappe Ceph de base (opérations au quotidien). Il décrit également toutes les méthodes prises en charge pour accéder aux données stockées dans une grappe Ceph. SUSE Enterprise Storage 7.1 est une extension de SUSE Linux Enterprise Server 15 SP3. Il réunit les fonctionnalités du projet de stockage Ceph (<http://ceph.com/>), l'ingénierie d'entreprise et le support de SUSE. SUSE Enterprise Storage 7.1 permet aux organisations informatiques de déployer une architecture de stockage distribuée capable de prendre en charge un certain nombre de cas d'utilisation à l'aide de plates-formes matérielles courantes.

1 Documentation disponible



Note : documentation en ligne et dernières mises à jour

La documentation relative à nos produits est disponible à la page <https://documentation.suse.com>, où vous pouvez également rechercher les dernières mises à jour et parcourir ou télécharger la documentation dans différents formats. Les dernières mises à jour de la documentation sont disponibles dans la version en anglais.

En outre, la documentation sur le produit est disponible dans votre système sous `/usr/share/doc/manual`. Elle est incluse dans un paquetage RPM nommé `ses-manual_LANG_CODE`. S'il ne se trouve pas déjà sur votre système, installez-le. Par exemple :

```
# zypper install ses-manual_en
```

La documentation suivante est disponible pour ce produit :

Guide de déploiement (<https://documentation.suse.com/ses/html/ses-all/book-storage-deployment.html>)

Ce guide est axé sur le déploiement d'une grappe Ceph de base et sur la manière de déployer des services supplémentaires. Il couvre également les étapes de mise à niveau vers SUSE Enterprise Storage 7.1 à partir de la version précédente du produit.

Guide d'opérations et d'administration (<https://documentation.suse.com/ses/html/ses-all/book-storage-admin.html>) ↗

Ce guide est axé sur les tâches de routine que vous devez effectuer en tant qu'administrateur après le déploiement de la grappe Ceph de base (opérations au quotidien). Il décrit également toutes les méthodes prises en charge pour accéder aux données stockées dans une grappe Ceph.

Guide de renforcement de la sécurité (<https://documentation.suse.com/ses/html/ses-all/book-storage-security.html>) ↗

Ce guide explique comment garantir la sécurité de votre grappe.

Guide de dépannage (<https://documentation.suse.com/ses/html/ses-all/book-storage-trouble-shooting.html>) ↗

Ce guide passe en revue divers problèmes courants lors de l'exécution de SUSE Enterprise Storage 7.1, ainsi que d'autres problèmes liés aux composants pertinents tels que Ceph ou Object Gateway.

Guide de SUSE Enterprise Storage pour Windows (<https://documentation.suse.com/ses/html/ses-all/book-storage-windows.html>) ↗

Ce guide décrit l'intégration, l'installation et la configuration des environnements Microsoft Windows et de SUSE Enterprise Storage à l'aide du pilote Windows.

2 Commentaires

Vos commentaires et contributions concernant cette documentation sont les bienvenus. Pour nous en faire part, plusieurs canaux sont à votre disposition :

Requêtes de service et support

Pour connaître les services et les options de support disponibles pour votre produit, visitez le site <http://www.suse.com/support/> ↗.

Pour ouvrir une requête de service, vous devez disposer d'un abonnement SUSE enregistré auprès du SUSE Customer Center. Accédez à <https://scc.suse.com/support/requests> ↗, connectez-vous, puis cliquez sur *Créer un(e) nouveau(elle)*.

Signalement d'erreurs

Signalez les problèmes liés à la documentation à l'adresse <https://bugzilla.suse.com/> ↗. Pour signaler des problèmes, vous avez besoin d'un compte Bugzilla.

Pour simplifier ce processus, vous pouvez utiliser les liens *Report Documentation Bug* (Signaler une erreur dans la documentation) en regard des titres dans la version HTML de ce document. Ces liens présélectionnent le produit et la catégorie appropriés dans Bugzilla et ajoutent un lien vers la section actuelle. Vous pouvez directement commencer à signaler le bogue.

Contributions

Pour contribuer à cette documentation, utilisez les liens *Edit Source* (Modifier la source), en regard des titres dans la version HTML de ce document. Ils vous permettent d'accéder au code source sur GitHub, où vous pouvez ouvrir une demande d'ajout (Pull request). Pour apporter votre contribution, vous avez besoin d'un compte GitHub.

Pour plus d'informations sur l'environnement de documentation utilisé pour cette documentation, reportez-vous au fichier lisezmoi de l'espace de stockage à l'adresse <https://github.com/SUSE/doc-ses>.





Messagerie

Vous pouvez également signaler des erreurs et envoyer vos commentaires concernant la documentation à l'adresse doc-team@suse.com. Veillez à inclure le titre du document, la version du produit et la date de publication du document. Mentionnez également le numéro et le titre de la section concernée (ou incluez l'URL), et décrivez brièvement le problème.

3 Conventions relatives à la documentation

Les conventions typographiques et mentions suivantes sont utilisées dans cette documentation :

- /etc/passwd : noms de répertoires et de fichiers
- MARQUE_RÉSERVATION : l'élément MARQUE_RÉSERVATION doit être remplacé par la valeur réelle
- CHEMIN : variable d'environnement
- ls, --help : commandes, options et paramètres
- user : nom de l'utilisateur ou du groupe
- package_name : nom d'un paquetage logiciel
- **Alt** , **Alt - F1** : touche ou combinaison de touches sur lesquelles appuyer. Les touches sont affichées en majuscules comme sur un clavier.

- *Fichier, Fichier > Enregistrer sous* : options de menu, boutons
-  Ce paragraphe n'est utile que pour les architectures AMD64/Intel 64. Les flèches marquent le début et la fin du bloc de texte. 
-  Ce paragraphe ne s'applique qu'aux architectures IBM Z et POWER. Les flèches marquent le début et la fin du bloc de texte. 
- *Chapitre 1, “Exemple de chapitre”* : renvoi à un autre chapitre de ce guide.
- Commandes à exécuter avec les privilèges root. Souvent, vous pouvez également leur ajouter en préfixe la commande sudo pour les exécuter sans privilèges.

```
# command
> sudo command
```

- Commandes pouvant être exécutées par des utilisateurs non privilégiés.

```
> command
```

- Avis



Warning: note d'avertissement

Information essentielle dont vous devez prendre connaissance avant de continuer. Met en garde contre des problèmes de sécurité ou des risques de perte de données, de détérioration matérielle ou de blessure physique.



Important: note importante

Information importante dont vous devez prendre connaissance avant de continuer.



Note: note de remarque

Information supplémentaire, par exemple sur les différences dans les versions des logiciels.



Tip: note indiquant une astuce

Information utile, telle qu'un conseil ou un renseignement pratique.

- Notes compactes



Information supplémentaire, par exemple sur les différences dans les versions des logiciels.



Information utile, telle qu'un conseil ou un renseignement pratique.

4 Support

Vous trouverez ci-dessous la déclaration de support pour SUSE Enterprise Storage et des informations générales sur les avant-premières technologiques. Pour plus d'informations sur le cycle de vie du produit, reportez-vous au site <https://www.suse.com/lifecycle>.

Si vous avez droit au support, vous trouverez des instructions détaillées sur la collecte d'informations pour un ticket de support sur le site <https://documentation.suse.com/sles-15/html/SLES-all/cha-adm-support.html>.

4.1 Déclaration de support pour SUSE Enterprise Storage

Pour bénéficier d'un support, vous devez disposer d'un abonnement adéquat auprès de SUSE. Pour connaître les offres de support spécifiques auxquelles vous pouvez accéder, rendez-vous sur la page <https://www.suse.com/support/> et sélectionnez votre produit.

Les niveaux de support sont définis comme suit :

N1

Identification du problème : support technique conçu pour fournir des informations de compatibilité, un support pour l'utilisation, une maintenance continue, la collecte d'informations et le dépannage de base à l'aide de la documentation disponible.

N2

Isolement du problème : support technique conçu pour analyser des données, reproduire des problèmes clients, isoler la zone problématique et fournir une solution aux problèmes qui ne sont pas résolus au niveau 1 ou préparer le niveau 3.

N3

Résolution des problèmes : support technique conçu pour résoudre les problèmes en impliquant des ingénieurs afin de corriger des défauts produit identifiés par le support de niveau 2.

Pour les clients et partenaires sous contrat, SUSE Enterprise Storage est fourni avec un support de niveau 3 pour tous les paquetages, excepté les suivants :

- Avant-premières technologiques.
- Son, graphiques, polices et illustrations.
- Paquetages nécessitant un contrat de client supplémentaire.
- Certains paquetages fournis avec le module *Workstation Extension* (Extension de poste de travail), qui bénéficient uniquement d'un support de niveau 2
- Paquetages dont les noms se terminent par `-devel` (contenant les fichiers d'en-tête et les ressources développeurs similaires) ne bénéficient d'un support que dans le cadre de leurs paquetages principaux.


SUSE offre un support uniquement pour l'utilisation de paquetages d'origine, autrement dit les paquetages qui ne sont pas modifiés ni recompilés.

4.2 Avant-premières technologiques


Les avant-premières technologiques sont des paquetages, des piles ou des fonctions fournis par SUSE pour donner un aperçu des innovations à venir. Ces avant-premières technologiques sont fournies pour vous permettre de tester de nouvelles technologies au sein de votre environnement. Tous vos commentaires sont les bienvenus ! Si vous testez une avant-première technologique, veuillez contacter votre représentant SUSE et l'informer de votre expérience et de vos cas d'utilisation. Vos remarques sont utiles pour un développement futur.

Les avant-premières technologiques présentent les limites suivantes :

- Les avant-premières technologiques sont toujours en cours de développement. Ainsi, elles peuvent être incomplètes ou instables, ou *non* adaptées à une utilisation en production.
- Les avant-premières technologiques ne bénéficient *pas* du support technique.
- Les avant-premières technologiques peuvent être disponibles uniquement pour des architectures matérielles spécifiques.
- Les détails et fonctionnalités des avant-premières technologiques sont susceptibles d'être modifiés. Il en résulte que la mise à niveau vers des versions ultérieures d'une avant-première technologique peut être impossible et nécessiter une nouvelle installation.
- SUSE peut découvrir qu'une avant-première ne répond pas aux besoins des clients ou du marché, ou n'est pas conforme aux normes de l'entreprise. Les avant-premières technologiques peuvent être supprimées d'un produit à tout moment. SUSE ne s'engage pas à fournir à l'avenir une version prise en charge de ces technologies.

Pour obtenir un aperçu des avant-premières technologiques fournies avec votre produit, reportez-vous aux notes de version à l'adresse https://www.suse.com/releasenotes/x86_64/SUSE-Enterprise-Storage/7.1 .

5 Contributeurs Ceph

Le projet Ceph et sa documentation sont le résultat du travail de centaines de contributeurs et d'organisations. Pour plus d'informations, consultez la page <https://ceph.com/contributors/> .

6 Commandes et invites de commande utilisées dans ce guide

En tant qu'administrateur de grappe Ceph, vous configurez et ajustez le comportement de la grappe en exécutant des commandes spécifiques. Il existe plusieurs types de commandes dont vous avez besoin :

6.1 Commandes associées à Salt

Ces commandes vous aident à déployer des noeuds de grappe Ceph, à exécuter des commandes simultanément sur plusieurs noeuds de la grappe (ou tous), ou à ajouter ou supprimer des noeuds de la grappe. Les commandes les plus fréquemment utilisées sont **ceph-salt** et **ceph-salt config**. Vous devez exécuter les commandes Salt sur le noeud Salt Master en tant qu'utilisateur root. Ces commandes sont introduites avec l'invite suivante :

```
root@master #
```

Par exemple :

```
root@master # ceph-salt config ls
```

6.2 Commandes associées à Ceph

Il s'agit de commandes de niveau inférieur permettant de configurer et d'affiner tous les aspects de la grappe et de ses passerelles sur la ligne de commande, par exemple **ceph**, **cephadm**, **rbd** ou **radosgw-admin**.

Pour exécuter les commandes associées à Ceph, vous devez disposer d'un accès en lecture à une clé Ceph. Les fonctionnalités de la clé définissent alors vos privilèges au sein de l'environnement Ceph. Une option consiste à exécuter les commandes Ceph en tant qu'utilisateur root (ou via **sudo**) et à employer le trousseau de clés par défaut sans restriction « ceph.client.admin.key ».

L'option plus sûre et recommandée consiste à créer une clé individuelle plus restrictive pour chaque administrateur et à la placer dans un répertoire où les utilisateurs peuvent la lire, par exemple :

```
~/ceph/ceph.client.USERNAME.keyring
```



Astuce : chemin des clés Ceph

Pour utiliser un administrateur et un trousseau de clés personnalisés, vous devez spécifier le nom d'utilisateur et le chemin d'accès à la clé chaque fois que vous exécutez la commande **ceph** à l'aide des options `-n client.NOM_UTILISATEUR` et `--keyring CHEMIN/VERS/TROUSSEAU`.

Pour éviter cela, incluez ces options dans la variable `CEPH_ARGS` des fichiers `~/ .bashrc` des différents utilisateurs.

Bien que vous puissiez exécuter les commandes associées à Ceph sur n'importe quel noeud de la grappe, nous vous recommandons de les lancer sur le noeud Admin. Dans cette documentation, nous employons l'utilisateur `cephuser` pour exécuter les commandes. Elles sont donc introduites avec l'invite suivante :

```
cephuser@adm >
```

Par exemple :

```
cephuser@adm > ceph auth list
```



Astuce : commandes de noeuds spécifiques

Si la documentation vous demande d'exécuter une commande sur un noeud de la grappe avec un rôle spécifique, cela sera réglé par l'invite. Par exemple :

```
cephuser@mon >
```

6.2.1 Exécution de **ceph-volume**

À partir de SUSE Enterprise Storage 7, les services Ceph s'exécutent de manière conteneurisée. Si vous devez exécuter **ceph-volume** sur un noeud OSD, vous devez l'ajouter au début de la commande **cephadm**, par exemple :

```
cephuser@adm > cephadm ceph-volume simple scan
```

6.3 Commandes Linux générales

Les commandes Linux non associées à Ceph, telles que `mount`, `cat` ou `openssl`, sont introduites avec les invites `cephuser@adm >` ou `#`, selon les privilèges que la commande concernée nécessite.

6.4 Informations supplémentaires

Pour plus d'informations sur la gestion des clés Ceph, reportez-vous à la [Section 30.2, « Les zones de gestion principales »](#).

I Ceph Dashboard

- 1 À propos de Ceph Dashboard 2
- 2 Interface utilisateur Web du tableau de bord 3
- 3 Gestion des utilisateurs et des rôles Ceph Dashboard 11
- 4 Affichage des éléments internes de la grappe 16
- 5 Gestion des réserves 29
- 6 Gestion du périphérique de traitement par blocs RADOS (RBD) 32
- 7 Gestion de NFS Ganesha 56
- 8 Gestion de CephFS 61
- 9 Gestion de la passerelle Object Gateway 63
- 10 Configuration manuelle 71
- 11 Gestion des utilisateurs et des rôles via la ligne de commande 79

1 À propos de Ceph Dashboard

Ceph Dashboard est une application Web intégrée de gestion et de surveillance Ceph qui permet de gérer divers aspects et objets de la grappe. Le tableau de bord est automatiquement activé après le déploiement de la grappe de base, comme décrit dans le *Manuel « Guide de déploiement », Chapitre 7 « Déploiement de la grappe Bootstrap à l'aide de ceph-salt »*.

Ceph Dashboard pour SUSE Enterprise Storage 7.1 ajoute à Ceph Manager davantage de fonctionnalités de gestion Web pour faciliter l'administration de Ceph, notamment la surveillance et l'administration des applications. Vous n'avez plus besoin de connaître les commandes complexes relatives à Ceph pour gérer et surveiller votre grappe Ceph. Vous pouvez utiliser l'interface intuitive de Ceph Dashboard ou son API REST intégrée.

Le module Ceph Dashboard permet de visualiser des informations et des statistiques relatives à la grappe Ceph à l'aide d'un serveur Web hébergé par [ceph-mgr](#). Pour plus d'informations sur Ceph Manager, reportez-vous au *Manuel « Guide de déploiement », Chapitre 1 « SES et Ceph », Section 1.2.3 « Noeuds et daemons Ceph »*.

2 Interface utilisateur Web du tableau de bord

2.1 Connexion

Pour vous connecter à Ceph Dashboard, faites pointer votre navigateur vers son URL, y compris le numéro de port. Exécutez la commande suivante pour rechercher l'adresse :

```
cephuser@adm > ceph mgr services | grep dashboard  
"dashboard": "https://host:port/",
```

La commande renvoie l'URL où se trouve Ceph Dashboard. Si vous rencontrez des problèmes avec cette commande, reportez-vous au *Manuel « Troubleshooting Guide », Chapitre 10 « Troubleshooting the Ceph Dashboard », Section 10.1 « Locating the Ceph Dashboard »*.

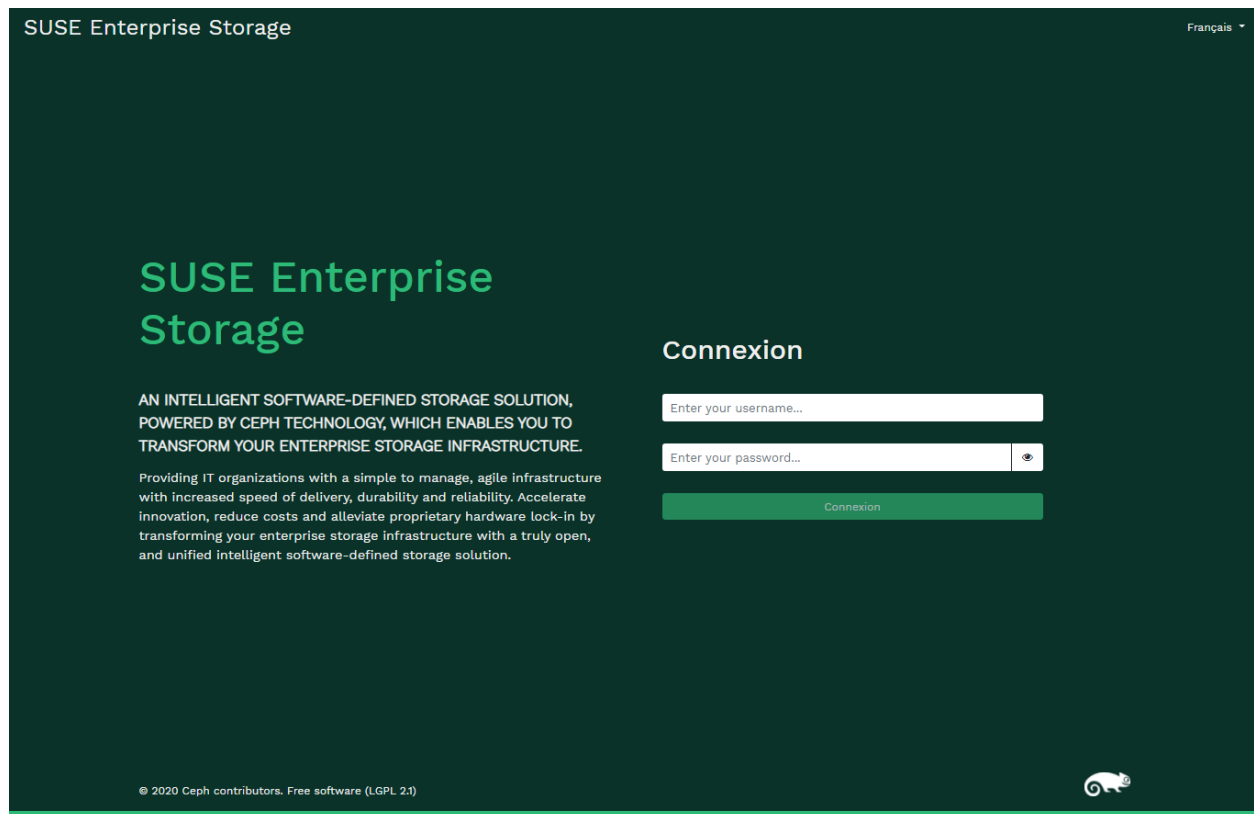


FIGURE 2.1 : ÉCRAN DE CONNEXION À CEPH DASHBOARD

Connectez-vous à l'aide des informations d'identification créées lors du déploiement de la grappe (voir *Manuel « Guide de déploiement », Chapitre 7 « Déploiement de la grappe Bootstrap à l'aide de ceph-salt », Section 7.2.9 « Configuration des informations d'identification de connexion de Ceph Dashboard »*).



Astuce : compte utilisateur personnalisé

Si vous ne souhaitez pas utiliser le compte *admin* par défaut pour accéder à Ceph Dashboard, créez un compte utilisateur personnalisé avec des privilèges d'administrateur. Pour plus d'informations, reportez-vous au [Chapitre 11, Gestion des utilisateurs et des rôles via la ligne de commande](#).



Important

Dès qu'une mise à niveau vers une nouvelle version majeure de Ceph (nom de code : Pacific) est disponible, le tableau de bord Ceph affiche un message approprié en haut de la zone de notification. Pour effectuer la mise à niveau, suivez les instructions du Manuel « Guide de déploiement », Chapitre 11 « Mise à niveau de SUSE Enterprise Storage 7 vers la version 7.1 ».

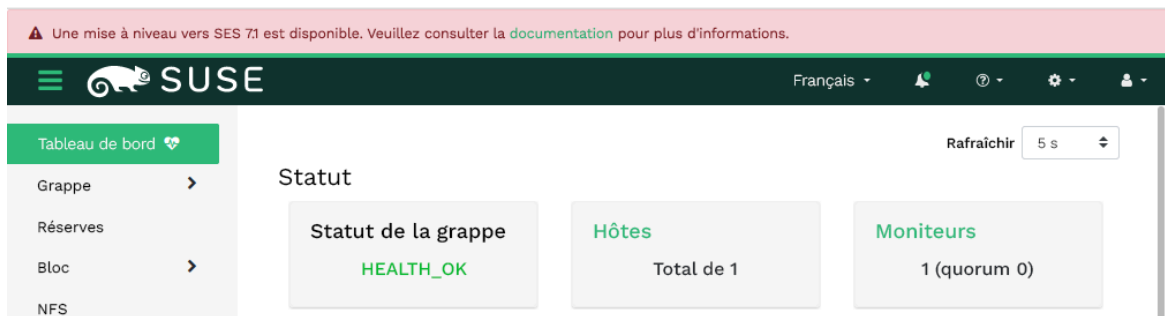


FIGURE 2.2 : NOTIFICATION CONCERNANT UNE NOUVELLE VERSION DE SUSE ENTERPRISE STORAGE

L'interface utilisateur du tableau de bord est divisée en plusieurs *blocs* : le *menu utilitaire* dans la partie supérieure droite de l'écran, le *menu principal* à gauche et le *volet de contenu* principal.

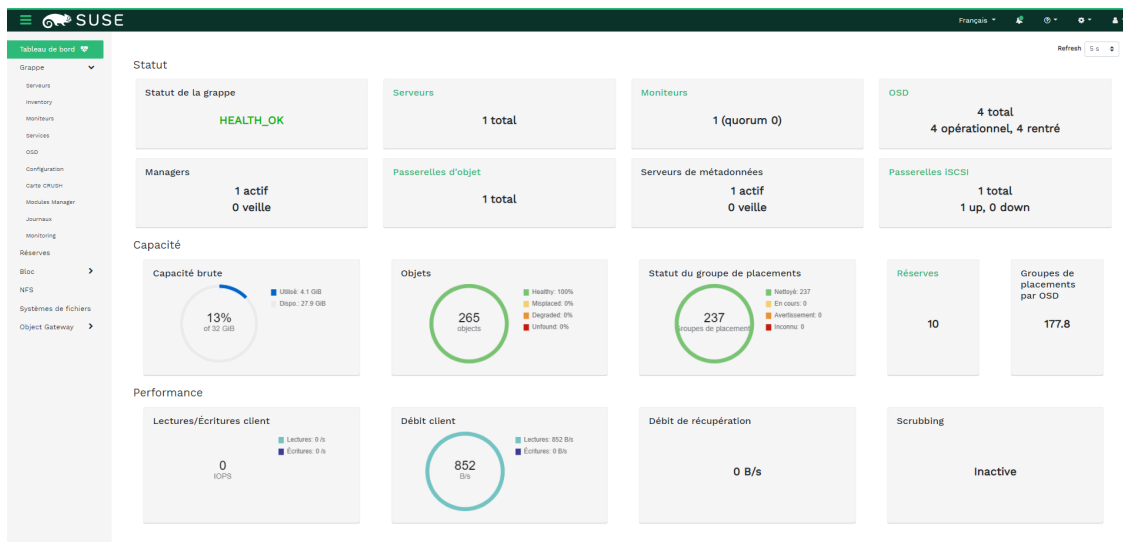


FIGURE 2.3 : PAGE D'ACCUEIL DE CEPH DASHBOARD

2.2 Menu utilitaire

La partie supérieure droite de l'écran contient un menu utilitaire. Il comprend des tâches générales liées davantage au tableau de bord qu'à la grappe Ceph. Vous pouvez cliquer sur ses options pour accéder aux éléments suivants :

- Modification de la langue de l'interface du tableau de bord en tchèque, allemand, anglais, espagnol, français, indonésien, italien, japonais, coréen, polonais, portugais (brésilien) et chinois.
- Tâches et notifications
- Documentation, informations sur l'API REST ou d'autres informations sur le tableau de bord.
- Gestion des utilisateurs et configuration de la télémétrie



Note

Pour obtenir une description plus détaillée de la ligne de commande pour les rôles utilisateur, reportez-vous au [Chapitre 11, Gestion des utilisateurs et des rôles via la ligne de commande](#).

- Configuration de la connexion, changement du mot de passe ou déconnexion

2.3 Menu principal

Le menu principal du tableau de bord occupe la partie gauche de l'écran. Il traite les aspects suivants :

Tableau de bord

Permet de retourner sur la page d'accueil de Ceph Dashboard.

Grappe

Permet d'afficher des informations détaillées sur les hôtes, l'inventaire, les instances Ceph Monitor, les services, les OSD Ceph, la configuration de grappe, la carte CRUSH, les modules Ceph Manager, les journaux et la surveillance.

Réserves

Permet d'afficher et de gérer les réserves de la grappe.

Bloc

Permet d'afficher des informations détaillées et de gérer les images de RBD, la mise en miroir et iSCSI.

NFS

Permet d'afficher et de gérer les déploiements NFS Ganesha.



Note

Si NFS Ganesha n'est pas déployé, une note d'information apparaît. Reportez-vous à la [Section 11.6, « Configuration de NFS Ganesha dans Ceph Dashboard »](#).

Systèmes de fichiers

Permet d'afficher et de gérer les CephFS.

Object Gateway

Permet d'afficher et de gérer les daemons, les utilisateurs et les compartiments d'Object Gateway.



Note


Si Object Gateway n'est pas déployé, une note d'information apparaît. Reportez-vous à la [Section 10.4, « Activation de l'interface client de gestion d'Object Gateway »](#).


2.4 Volet de contenu


Le volet de contenu occupe la partie principale de l'écran du tableau de bord. La page d'accueil du tableau de bord présente de nombreux widgets utiles pour vous informer brièvement sur l'état actuel de la grappe, la capacité et les performances.


2.5 Fonctionnalités courantes de l'interface utilisateur Web

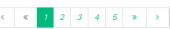
Dans Ceph Dashboard, la plupart des opérations s'effectuent à l'aide de *listes*, par exemple, des listes de réserves, de noeuds OSD ou de périphériques RBD. Par défaut, toutes les listes se rafraîchissent automatiquement toutes les cinq secondes. Les widgets communs suivants vous aident à gérer ou à ajuster ces listes :

Cliquez sur  pour déclencher un rafraîchissement manuel de la liste.

Cliquez sur  pour afficher ou masquer des colonnes de table individuelles.

Cliquez sur  et entrez (ou sélectionnez) le nombre de lignes à afficher sur une seule page.

Cliquez à l'intérieur de  et filtrez les lignes en saisissant la chaîne à rechercher.

Utilisation  pour modifier la page affichée si la liste s'étend sur plusieurs pages.

2.6 Widgets du tableau de bord

Chaque widget du tableau de bord affiche des informations d'état spécifiques liées à un aspect particulier d'une grappe Ceph en cours d'exécution. Certains widgets sont des liens actifs. Si vous cliquez sur ces derniers, ils vous redirigent vers une page détaillée associée au sujet qu'ils représentent.



Astuce : plus de détails grâce au pointeur de la souris

Certains widgets graphiques vous montrent plus de détails lorsque vous passez le pointeur de la souris sur eux.

2.6.1 Widgets de statut

Les widgets *Statut* vous donnent un bref aperçu du statut actuel de la grappe.



FIGURE 2.4 : WIDGETS DE STATUT

Statut de la grappe

Présente des informations de base sur l'état de santé de la grappe.

Hôtes

Affiche le nombre total de noeuds de la grappe.

Moniteurs

Affiche le nombre de moniteurs en cours d'exécution et leur quorum.

OSD

Affiche le nombre total d'OSD, ainsi que le nombre d'OSD *démarrés* et *rentrés*.

Gestionnaires

Affiche le nombre de daemons Ceph Manager actifs et en veille.

Passerelles d'objet

Affiche le nombre d'instances Object Gateway en cours d'exécution.

Serveur de métadonnées (MDS)

Affiche le nombre de serveurs de métadonnées.

Passerelles iSCSI

Affiche le nombre de passerelles iSCSI configurées.

2.6.2 Widgets de capacité

Les widgets *Capacité* affichent de brèves informations sur la capacité de stockage.

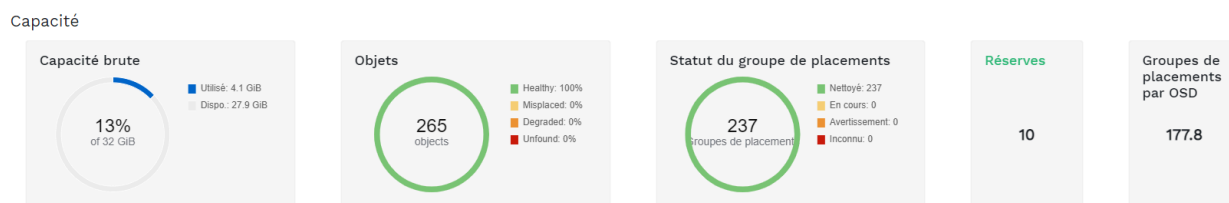


FIGURE 2.5 : WIDGETS DE CAPACITÉ

Capacité brute

Affiche le rapport entre la capacité de stockage brute utilisée et celle disponible.

Objets

Affiche le nombre d'objets de données stockés dans la grappe.

Statut du groupe de placements

Affiche une représentation graphique des groupes de placements en fonction de leur statut.

Réserves

Affiche le nombre de réserve dans la grappe.

Groupes de placements par OSD

Affiche le nombre moyen de groupes de placements par OSD.

2.6.3 Widgets de performance

Les widgets *Performance* font référence aux données de performance de base des clients Ceph.

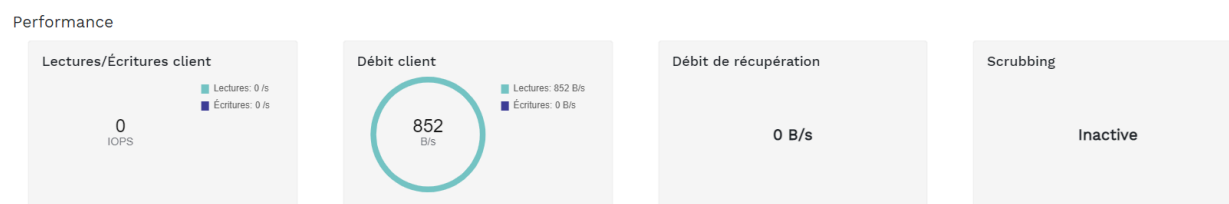


FIGURE 2.6 : WIDGETS DE PERFORMANCE

Lectures/Écritures client

Quantité d'opérations de lecture et d'écriture des clients par seconde.

Débit client

Quantité de données transférées vers et depuis les clients Ceph, en octets par seconde.

Débit de récupération

Débit de données récupérées par seconde.

Nettoyage


Affiche le statut du nettoyage (voir [Section 17.4.9, « Nettoyage d'un groupe de placement »](#)), à savoir Inactif, Activé ou Actif.

3 Gestion des utilisateurs et des rôles Ceph Dashboard

La gestion des utilisateurs du tableau de bord effectuée à l'aide de commandes Ceph sur la ligne de commande a déjà été présentée au [Chapitre 11, Gestion des utilisateurs et des rôles via la ligne de commande](#).

Cette section décrit comment gérer les comptes utilisateur à l'aide de l'interface utilisateur Web du tableau de bord.

3.1 Liste des utilisateurs

Cliquez sur  dans le menu de l'utilitaire, puis sélectionnez *Gestion des utilisateurs*.

La liste contient le nom d'utilisateur de chaque utilisateur, son nom complet, son adresse électronique, la liste des rôles assignés, l'état d'activation du rôle et la date d'expiration du mot de passe.



| Nom d'utilisateur | Nom | Adresse électronique | Rôles | Activé | Password expiration date |
|-------------------|------------------|----------------------|-------------------------------|--------|--------------------------|
| admin | | | administrator | ✓ | |
| Alex | Alexandra Settle | tux@example.com | cluster-manager, pool-manager | ✓ | |
| dashboard user 1 | Dashboard User1 | du1@example.com | | ✓ | |
| rgw user | RGW User | rgw@example.com | pool-manager, rgw-manager | ✓ | |

0 sélectionné(s) / 4 total

FIGURE 3.1 : GESTION DES UTILISATEURS

3.2 Ajout de nouveaux utilisateurs

Cliquez sur *Créer* en haut à gauche de l'en-tête de tableau pour ajouter un nouvel utilisateur. Entrez son nom d'utilisateur, son mot de passe et éventuellement un nom complet et une adresse électronique.

Créer Utilisateur

Nom d'utilisateur *

potato

✓

Mot de passe ?

.....

✓

👁

Confirmer le mot de passe

.....

✓

👁

Password expiration date ?

Password expiration date...

✕

Nom complet

Mr Potato

✓

Adresse électronique

potato@example.com

✓

Rôles

✎ There are no roles.

☒ Activé

☒ User must change password at next login

Créer Utilisateur

Annuler

FIGURE 3.2 : AJOUT D'UN UTILISATEUR

Cliquez sur la petite icône de crayon pour assigner des rôles prédéfinis à l'utilisateur. Confirmez en cliquant sur le bouton *Créer un utilisateur*.


3.3 Modification des utilisateurs

Dans le tableau, cliquez sur la ligne d'un utilisateur pour la mettre en surbrillance, puis sélectionnez *Modifier* pour éditer les détails de l'utilisateur. Confirmez en cliquant sur le bouton *Modifier l'utilisateur*.

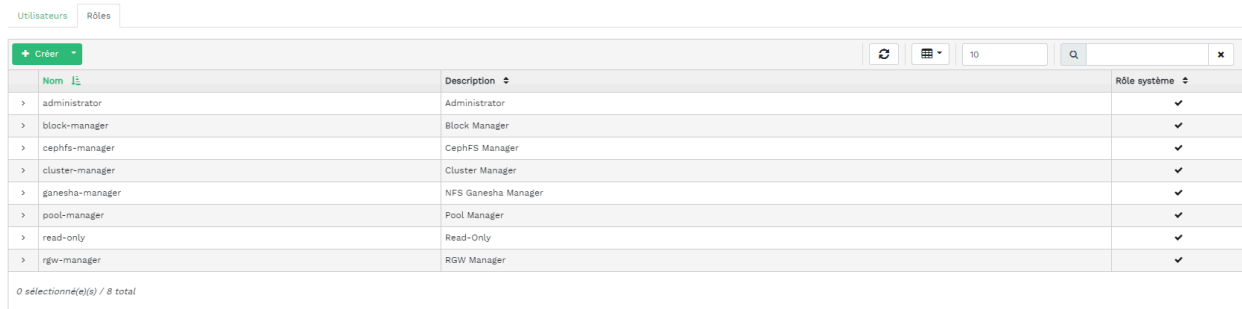
3.4 Suppression d'utilisateurs

Cliquez sur la ligne d'un utilisateur dans le tableau pour la mettre en surbrillance, puis sélectionnez la zone de liste déroulante en regard de *Modifier* et sélectionnez *Supprimer* dans la liste pour supprimer le compte utilisateur. Cochez la case *Oui* et confirmez avec le bouton *Supprimer l'utilisateur*.

3.5 Liste des rôles des utilisateurs

Cliquez sur  dans le menu de l'utilitaire, puis sélectionnez *Gestion des utilisateurs*. Cliquez ensuite sur l'onglet *Rôles*.

La liste contient le nom et la description de chaque rôle, et indique s'il s'agit d'un rôle système.



| Nom | Description | Rôle système |
|-----------------|---------------------|--------------|
| administrator | Administrator | ✓ |
| block-manager | Block Manager | ✓ |
| cephfs-manager | CephFS Manager | ✓ |
| cluster-manager | Cluster Manager | ✓ |
| ganesha-manager | NFS Ganesha Manager | ✓ |
| pool-manager | Pool Manager | ✓ |
| read-only | Read-Only | ✓ |
| rgw-manager | RGW Manager | ✓ |

0 sélectionné(s) / 8 total

FIGURE 3.3 : RÔLES UTILISATEUR

3.6 Ajout de rôles personnalisés

Cliquez sur *Créer* en haut à gauche de l'en-tête de tableau pour ajouter un nouveau rôle personnalisé. Entrez le *Nom* et la *Description*, puis en regard de *Autorisations*, sélectionnez les autorisations appropriées.



Astuce : purge des rôles personnalisés

Si vous créez des rôles utilisateur personnalisés et avez l'intention de supprimer ultérieurement la grappe Ceph à l'aide de la commande **ceph-salt purge**, vous devez d'abord purger les rôles personnalisés. Pour plus de détails, reportez-vous à la [Section 13.9](#), « *Suppression d'une grappe Ceph entière* ».

Créer un rôle

Nom *

ganesha pool user ✓

Description

a user that can only manage ganesha and pools ✓

Autorisations

| <input type="checkbox"/> Tout | <input type="checkbox"/> Lire | <input type="checkbox"/> Créer | <input type="checkbox"/> Mettre à jour | <input type="checkbox"/> Supprimer |
|---|-------------------------------------|-------------------------------------|--|-------------------------------------|
| <input type="checkbox"/> cephfs | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> config-opt | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> dashboard-settings | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> grafana | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> hosts | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> iscsi | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> log | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> manager | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> monitor | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input checked="" type="checkbox"/> nfs-ganesha | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> |
| <input type="checkbox"/> osd | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> pool | <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> prometheus | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> rbd-image | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> rbd-mirroring | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> rgw | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> user | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

Créer un rôle

Retour

FIGURE 3.4 : AJOUT D'UN RÔLE



Astuce : activation multiple

Si vous cochez la case qui précède le nom du périmètre, vous activez toutes les autorisations pour ce périmètre. Si vous cochez la case *Tout*, vous activez toutes les autorisations pour tous les périmètres.

Confirmez en cliquant sur le bouton *Créer un rôle*.

3.7 Modification des rôles personnalisés

Cliquez sur la ligne d'un utilisateur dans le tableau pour la mettre en surbrillance, puis sélectionnez *Modifier* en haut à gauche de l'en-tête du tableau pour éditer la description et les autorisations du rôle personnalisé. Confirmez en cliquant sur le bouton *Edit Role* (Modifier le rôle).

3.8 Suppression des rôles personnalisés

Cliquez sur la ligne d'un rôle dans le tableau pour la mettre en surbrillance, puis sélectionnez la zone de liste déroulante en regard de *Modifier* et sélectionnez *Supprimer* dans la liste pour supprimer le rôle. Cochez la case *Oui* et confirmez avec le bouton *Supprimer le rôle*.

4 Affichage des éléments internes de la grappe

L'élément de menu *Grappe* permet d'afficher des informations détaillées sur les hôtes de la grappe Ceph, l'inventaire, les instances Ceph Monitor, les services, les OSD, la configuration, la carte CRUSH, Ceph Manager, les journaux et les fichiers de surveillance.

4.1 Affichage des noeuds de grappe

Cliquez sur *Grappe* > *Hôtes* pour afficher la liste des noeuds de la grappe.



| Liste d'hôtes | | Performance globale | |
|-----------------------------|---|---------------------|------------------------|
| Créer | | 10 | Q |
| Nom d'hôte | Services | Labels | Version |
| > master | | | |
| > node1 | mgr,node1.wbmqsa, mon,node1, osd,0, osd,3 | | 15.2.4-557-g4ac763f0b3 |
| > node2 | mgr,node2.qcwalx, mon,node2, osd,1, osd,4 | | 15.2.4-557-g4ac763f0b3 |
| > node3 | mgr,node3.rhkzzy, mon,node3, osd,2, osd,5 | | 15.2.4-557-g4ac763f0b3 |
| 0 sélectionné(e)s / 4 total | | | |

FIGURE 4.1 : HÔTES

Cliquez sur la flèche de liste déroulante en regard d'un nom de noeud dans la colonne *Nom d'hôte* pour afficher les détails des performances du noeud.




La colonne *Services* répertorie tous les daemons qui s'exécutent sur chaque noeud associé. Cliquez sur un nom de daemon pour afficher sa configuration détaillée.

4.2 Accès à l'inventaire de la grappe

Cliquez sur *Grappe* > *Inventaire* pour afficher la liste des périphériques. La liste inclut le chemin, le type, la disponibilité, le fournisseur, le modèle, la taille et les OSD du périphérique.

Cliquez pour sélectionner un nom de noeud dans la colonne *Nom d'hôte*. Une fois le nom sélectionné, cliquez sur *Identifier* pour identifier le périphérique sur lequel l'hôte s'exécute. Cela indique au périphérique de faire clignoter ses voyants. Sélectionnez la durée de cette opération entre 1, 2, 5, 10 et 15 minutes. Cliquez sur *Exécuter*.

Devices

| Identify | | | | | | |  |  | 10 |  Norm d'hôte | Any |
|------------|-------------|------|-----------|--------|-------|--------|---|---|----|---|-----|
| Nom d'hôte | Device path | Type | Available | Vendor | Model | Taille | OSD | | | | |
| master | /dev/vda | HDD | | 0x1af4 | | 42 GiB | | | | | |
| node1 | /dev/vda | HDD | | 0x1af4 | | 42 GiB | | | | | |
| node1 | /dev/vdb | HDD | | 0x1af4 | | 8 GiB | | osd.0 | | | |
| node1 | /dev/vdc | HDD | | 0x1af4 | | 8 GiB | | osd.3 | | | |
| node2 | /dev/vda | HDD | | 0x1af4 | | 42 GiB | | | | | |
| node2 | /dev/vdb | HDD | | 0x1af4 | | 8 GiB | | osd.1 | | | |
| node2 | /dev/vdc | HDD | | 0x1af4 | | 8 GiB | | osd.4 | | | |
| node3 | /dev/vda | HDD | | 0x1af4 | | 42 GiB | | | | | |
| node3 | /dev/vdb | HDD | | 0x1af4 | | 8 GiB | | osd.2 | | | |
| node3 | /dev/vdc | HDD | | 0x1af4 | | 8 GiB | | osd.5 | | | |

0 sélectionné(s) / 10 total

FIGURE 4.2 : SERVICES

4.3 Affichage des instances Ceph Monitor

Cliquez sur *Grappe* > *Moniteurs* pour afficher la liste des noeuds de la grappe hébergeant des moniteurs Ceph en cours d'exécution. Le volet de contenu est divisé en deux vues : Statut et Dans le quorum ou Hors quorum.

Le tableau *Statut* affiche des statistiques générales sur les instances Ceph Monitor en cours d'exécution, notamment les informations suivantes :

- ID de grappe
- monmap modifié
- monmap epoch
- quorum con
- quorum mon
- required con
- required mon

Les volets Dans le quorum et Hors quorum incluent le nom de chaque moniteur, son numéro de rang, son adresse IP publique et le nombre de sessions ouvertes.

Cliquez sur un nom de noeud dans la colonne *Nom* pour afficher la configuration Ceph Monitor associée.

| | |
|----------------|---|
| Statut | |
| ID de grappe | 05766fa4-a9a7-11eb-9e46-525400b22828 |
| monmap modifié | 2021-04-30T11:27:20.46562Z |
| monmap epoch | 1 |
| quorum con | 4540138292840890367 |
| quorum mon | kraken,luminous,mimic,osdmap-prune,nautilus,octopus |
| required con | 2449958747315978244 |
| required mon | kraken,luminous,mimic,osdmap-prune,nautilus,octopus |

Dans le quorum

| Nom | Rang | Adresse publique | Sessions ouvertes |
|---------|------|----------------------|-------------------|
| node1 | 0 | 10.20.156.201:6789/0 | |
| node2 | 2 | 10.20.156.202:6789/0 | |
| node3 | 1 | 10.20.156.203:6789/0 | |
| 3 total | | | |

Hors quorum

| Nom | Rang | Adresse publique |
|--------------------|------|------------------|
| No data to display | | |
| 0 total | | |

FIGURE 4.3 : CEPH MONITOR

4.4 Affichage des services

Cliquez sur *Grappe* > *Services* pour afficher les détails de chacun des services disponibles : *crash*, Ceph Manager et les instances Ceph Monitor. La liste inclut le nom et l'ID de l'image du conteneur, le statut de ce qui est en cours d'exécution, la taille et la date du dernier rafraîchissement. Cliquez sur la flèche de liste déroulante en regard d'un nom de service dans la colonne *Service* pour afficher les détails du daemon. La liste détaillée inclut le nom d'hôte, le type et l'ID du daemon, l'ID du conteneur, le nom et l'ID de l'image du conteneur, le numéro de version, le statut et la date du dernier rafraîchissement.

| Cluster > Services | | | | | | | |
|-----------------------------|--|--------------------|----------------------|--|--------------------|-----------------------------|---------|
| ✖ Supprimer | | | | | | | |
| Service | Container image name | Container image ID | Placement | Running | Taille | Last Refreshed | |
| ▼ crash | registry.suse.de/dev/storag/7.0/containers/sep/ceph/cephlatest | 6549871c3f67 | | | 4 | 2020-08-14T13:37:34.148847 | |
| Daemons | | | | | | | |
| Nom d'hôte | Daemon type | Daemon ID | Container ID | Container image name | Container image ID | Version | Statut |
| master | crash | master | 3acfc11b607e | registry.suse.de/dev/storag/7.0/containers/sep/ceph/cephlatest | 6549871c3f67 | 15.2.4.557 | running |
| node1 | crash | node1 | 3d56e2a421eb | registry.suse.de/dev/storag/7.0/containers/sep/ceph/cephlatest | 6549871c3f67 | 15.2.4.557 | running |
| node2 | crash | node2 | 8fa9790b9a51 | registry.suse.de/dev/storag/7.0/containers/sep/ceph/cephlatest | 6549871c3f67 | 15.2.4.557 | running |
| node3 | crash | node3 | b047531bbf2a | registry.suse.de/dev/storag/7.0/containers/sep/ceph/cephlatest | 6549871c3f67 | 15.2.4.557 | running |
| 4 total | | | | | | | |
| ➤ mgr | registry.suse.de/dev/storag/7.0/containers/sep/ceph/cephlatest | dcfacef0831b | master | 1 | 1 | 2021-05-05T13:37:46.693795Z | |
| ➤ mon | registry.suse.de/dev/storag/7.0/containers/sep/ceph/cephlatest | dcfacef0831b | master:10.20.165.200 | 1 | 1 | 2021-05-05T13:37:46.694300Z | |
| ➤ node-exporter | registry.suse.com/caasp/v4.5/prometheus-node-exporter:0.18.1 | a149a78bcd37 | master | 1 | 1 | 2021-05-05T13:37:46.695817Z | |
| ➤ osd.sesdev_osd_deployer | registry.suse.de/dev/storag/7.0/containers/sep/ceph/cephlatest | dcfacef0831b | master | 4 | 4 | 2021-05-05T13:37:46.694390Z | |
| 1 sélectionné(e)s / 5 total | | | | | | | |

FIGURE 4.4 : SERVICES

4.5 Affichage des OSD Ceph

Cliquez sur *Grappe* > *OSD* pour afficher la liste des noeuds hébergeant des daemons OSD en cours d'exécution. La liste comprend, pour chaque noeud, le nom, l'ID, le statut, la classe de périphérique, le nombre de groupes de placements, la taille, l'utilisation, un graphique des octets de lecture/d'écriture dans le temps, ainsi que le taux d'opérations de lecture/écriture par seconde.

Liste des OSD Performance globale

+ Créer Cluster-wide configuration 10 Q x

| | Hôte | ID | Statut | Device class | Groupes de placements | Taille | Drapeaux | Utilisation | Octets de lecture | Octets d'écriture | Opérations de lecture | Opérations d'écriture |
|--------------------------|---------|----|---------------------------------|--------------|-----------------------|--------|----------|----------------------------|-------------------|-------------------|-----------------------|-----------------------|
| <input type="checkbox"/> | > node1 | 0 | in up | hdd | 0 | 8 GiB | | <div><div></div></div> 12% | | | 0 /s | 0 /s |
| <input type="checkbox"/> | > node2 | 1 | in up | hdd | 1 | 8 GiB | | <div><div></div></div> 13% | | | 0 /s | 0 /s |
| <input type="checkbox"/> | > node3 | 2 | in up | hdd | 1 | 8 GiB | | <div><div></div></div> 13% | | | 0 /s | 0 /s |
| <input type="checkbox"/> | > node1 | 3 | in up | hdd | 1 | 8 GiB | | <div><div></div></div> 13% | | | 0 /s | 0 /s |
| <input type="checkbox"/> | > node2 | 4 | in up | hdd | 1 | 8 GiB | | <div><div></div></div> 12% | | | 0 /s | 0 /s |
| <input type="checkbox"/> | > node3 | 5 | in up | hdd | 0 | 8 GiB | | <div><div></div></div> 12% | | | 0 /s | 0 /s |

0 sélectionné(e)(s) / 6 total

FIGURE 4.5 : INSTANCES CEPH OSD

Sélectionnez *Drapeaux* dans le menu déroulant de *configuration de la grappe* dans l'en-tête du tableau pour ouvrir une fenêtre contextuelle. Elle contient une liste des drapeaux qui s'appliquent à l'ensemble de la grappe. Vous pouvez activer ou désactiver des drapeaux individuels et confirmer avec le bouton *Soumettre*.

Drapeaux OSD à l'échelle de la grappe

☐ Pas rentrés

Les OSD qui ont été marqués comme sortis ne seront pas marqués comme rentrés à leur démarrage.

☐ Pas sortis

Les OSD ne seront pas marqués automatiquement comme sortis après l'intervalle configuré.

☐ Pas démarrés

Les OSD ne sont pas autorisés à démarrer.

☐ Pas arrêtés

Les rapports de défaillance des OSD sont ignorés, de sorte que les moniteurs ne marquent pas les OSD comme étant arrêtés.

☐ Pause

Met en pause les lectures et écritures

☐ Pas de nettoyage

Nettoyage désactivé

☐ Pas de nettoyage en profondeur

Le nettoyage en profondeur est désactivé

☐ Pas de renvoi

Le renvoi des groupes de placements est suspendu

☐ No Rebalance

OSD will choose not to backfill unless PG is also degraded

☐ Pas de récupération

La récupération des groupes de placements est suspendue

☒ Tri au niveau du bit

Utiliser le tri au niveau du bit

☒ Variables snapdir purgées

Soumettre

Annuler

FIGURE 4.6 : DRAPEAUX OSD

Sélectionnez *Priorité de récupération* dans le menu déroulant de *configuration de la grappe* dans l'en-tête du tableau pour ouvrir une fenêtre contextuelle. Elle contient une liste des priorité de récupération des OSD qui s'appliquent à l'ensemble de la grappe. Vous pouvez activer le profil de priorité préféré et affiner les différentes valeurs ci-dessous. Confirmez en cliquant sur *Soumettre*.

Priorité de récupération des OSD

Priorité *

Personnalisé

☒ Personnaliser les valeurs de priorité

Nbre max. de renvois * ?

1

Nombre maximum de récupérations actives * ?

0

Nombre maximum de démarrages uniques de récupérations *

1

Mise en veille de la récupération * ?

0

Soumettre

Annuler

FIGURE 4.7 : PRIORITÉ DE RÉCUPÉRATION DES OSD

Cliquez sur la flèche de liste déroulante en regard d'un nom de noeud dans la colonne *Hôte* pour afficher un tableau étendu avec des détails sur les paramètres et les performances des périphériques. En parcourant plusieurs onglets, vous pouvez voir les listes *Attributs*, *Métadonnées*, *Santé du périphérique*, *Compteur de performance*, ainsi qu'un *Histogramme* graphique des opérations de lecture et d'écriture et les *Détails des performances*.

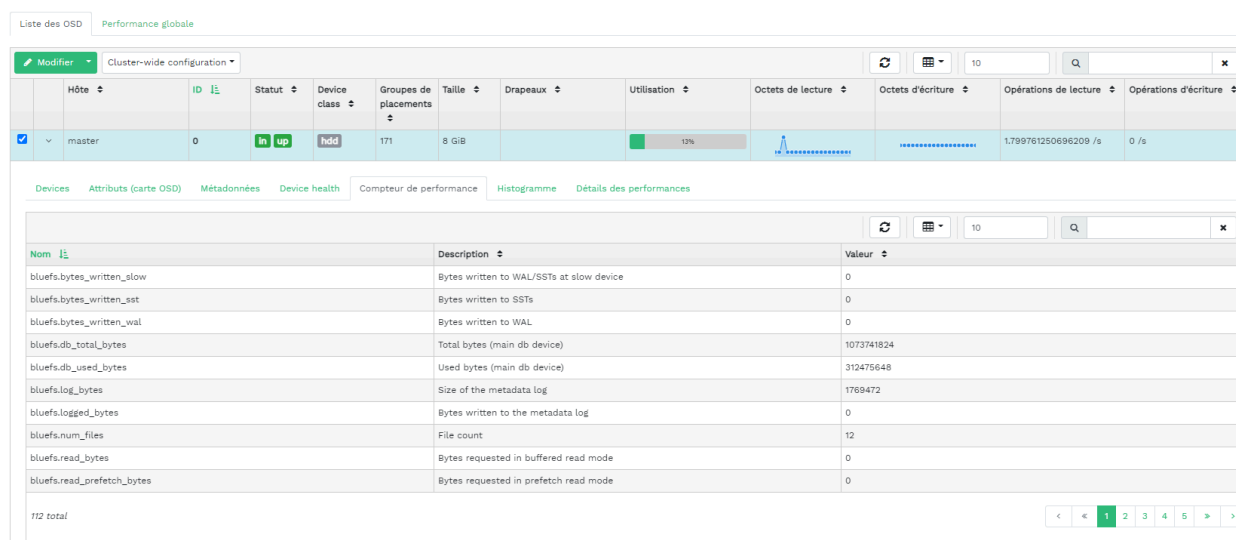


FIGURE 4.8 : DÉTAILS DES OSD



Astuce : exécution de tâches spécifiques sur les OSD

Une fois que vous avez cliqué sur un nom de noeud OSD, la ligne du tableau est mise en surbrillance. Cela signifie que vous pouvez désormais effectuer une tâche sur le noeud. Vous pouvez choisir d'effectuer l'une des opérations suivantes : *Modifier*, *Créer*, *Nettoyer*, *Nettoyage en profondeur*, *Réévaluer*, *Marquer comme sorti*, *Marquer comme rentré*, *Marquer comme arrêté*, *Marquer comme perdu*, *Purger*, *Détruire* ou *Supprimer*.

Cliquez sur la flèche vers le bas en haut à gauche de l'en-tête du tableau à côté du bouton *Créer* et sélectionnez la tâche à effectuer.

4.5.1 Ajout d'OSD

Pour ajouter de nouveaux OSD, procédez comme suit :

1. Vérifiez que certains noeuds de la grappe disposent de périphériques de stockage dont le statut est Disponible. Cliquez ensuite sur la flèche vers le bas en haut à gauche de l'en-tête du tableau et sélectionnez *Créer*. Cela ouvre la fenêtre *Créer des OSD*.

Créer OSD

Primary devices [?](#) [+ Ajouter](#)

Shared devices

WAL devices [?](#) [+ Ajouter](#)

DB devices [?](#) [+ Ajouter](#)

Configuration

Fonctionnalités ☐ Encryption

[Aperçu](#) [Annuler](#)

FIGURE 4.9 : CRÉER DES OSD

2. Pour ajouter des périphériques de stockage principaux pour les OSD, cliquez sur *Ajouter*. Avant de pouvoir ajouter des périphériques de stockage, vous devez spécifier des critères de filtrage en haut à droite du tableau *Périphériques principaux*, par exemple *Type disque dur*. Cliquez sur *Ajouter* pour confirmer.

Périphériques primaires

10 [Type](#) [disque dur](#) [Type : disque dur](#) [Effacer les filtres](#)

| Nom d'hôte | Chemin d'accès au périphérique | Type | Fournisseur | Modèle | Taille |
|--------------|--------------------------------|------------|-------------|--------|--------|
| doc-ses-min1 | /dev/vdb | Disque dur | Ox1af4 | | 12 Gio |
| doc-ses-min1 | /dev/vdc | Disque dur | Ox1af4 | | 12 Gio |

Total de 2

Nombre de périphériques : 2. Capacité brute : 24 Gio.

[Ajouter](#) [Annuler](#)

FIGURE 4.10 : AJOUT DE PÉRIPHÉRIQUES PRINCIPAUX

3. Dans la fenêtre *Créer des OSD* mise à jour, ajoutez éventuellement des périphériques WAL et BD partagés, ou activez le chiffrement des périphériques.

Créer des OSD

Périphériques primaires ? Type : disque dur Effacer

| Nom d'hôte | Chemin d'accès au périphérique | Type | Fournisseur | Modèle | Taille |
|--------------|--------------------------------|------------|-------------|--------|--------|
| doc-ses-min1 | /dev/vdb | Disque dur | Ox1af4 | | 12 Gio |
| doc-ses-min1 | /dev/vdc | Disque dur | Ox1af4 | | 12 Gio |

Total de 2

Capacité brute : 24 Gio

Périphériques partagés

Périphériques WAL ? + Ajouter

Périphériques DB ? + Ajouter

Configuration

Fonctions
☐ Chiffrement

Aperçu
Annuler

FIGURE 4.11 : CRÉATION D'OSD AVEC DES PÉRIPHÉRIQUES PRINCIPAUX AJOUTÉS

4. Cliquez sur *Aperçu* pour afficher l'aperçu des spécifications DriveGroups pour les périphériques ajoutés précédemment. Confirmez en cliquant sur *Créer*.

Aperçu de la création des OSD

Groupes d'unités

```
[
  {
    "service_type": "osd",
    "service_id": "dashboard-admin-1600784434446",
    "host_pattern": "*",
    "data_devices": {
      "rotational": true
    }
  }
]
```

Créer

Annuler

FIGURE 4.12 :

- Les nouveaux périphériques seront ajoutés à la liste des OSD.

| | Hôte | ID | Statut | Device class | Groupes de placement | Taille | Drapeaux | Utilisation | Octets de lecture | Octets d'écriture | Opérations de lecture | Opérations d'écriture |
|--------------------------|----------------|----|---------------------------------|--------------|----------------------|--------|----------|-------------|-------------------|-------------------|-----------------------|-----------------------|
| <input type="checkbox"/> | > doc-ses-min2 | 0 | in up | hdd | 119 | 10 GiB | | 11% | | | 0.7999105934891158 /s | 0 /s |
| <input type="checkbox"/> | > doc-ses-min3 | 1 | in up | hdd | 108 | 10 GiB | | 11% | | | 1.5998816768416986 /s | 0 /s |
| <input type="checkbox"/> | > doc-ses-min4 | 2 | in up | hdd | 126 | 10 GiB | | 11% | | | 0 /s | 0 /s |
| <input type="checkbox"/> | > doc-ses-min1 | 3 | in up | hdd | 96 | 12 GiB | | 9% | | | 0.3999455526088382 /s | 0 /s |
| <input type="checkbox"/> | > doc-ses-min1 | 4 | in up | hdd | 76 | 12 GiB | | 9% | | | 1.9995708432976873 /s | 0 /s |

FIGURE 4.13 : OSD NOUVELLEMENT AJOUTÉS

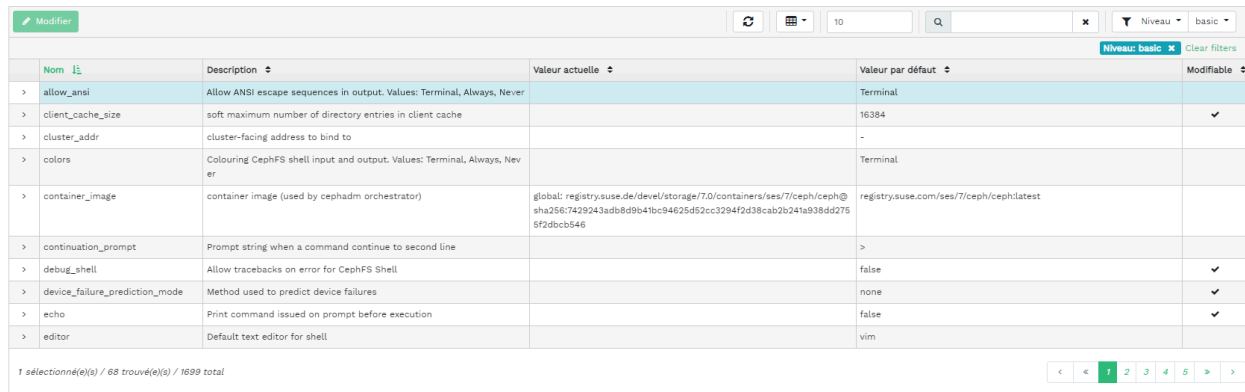


Note

Il n'y a pas de visualisation de la progression du processus de création des OSD. Leur création prend un certain temps. Les OSD apparaissent dans la liste lorsqu'ils ont été déployés. Si vous souhaitez vérifier le statut du déploiement, affichez les journaux en cliquant sur *Grappe* > *Journaux*.

4.6 Affichage de la configuration de la grappe

Cliquez sur *Grappe* > *Configuration* pour afficher une liste complète des options de configuration de la grappe Ceph. La liste contient le nom de l'option, sa brève description et ses valeurs actuelles et par défaut, et indique si l'option est modifiable.



The screenshot shows a web interface for Ceph configuration. At the top, there is a 'Modifier' button and a search bar. Below the search bar, there is a table with columns: 'Nom', 'Description', 'Valeur actuelle', 'Valeur par défaut', and 'Modifiable'. The table lists various configuration options like 'allow_ansi', 'client_cache_size', 'cluster_addr', 'colors', 'container_image', 'continuation_prompt', 'debug_shell', 'device_failure_prediction_mode', 'echo', and 'editor'. Each row has a dropdown arrow next to the 'Nom' column. At the bottom of the table, there is a status bar indicating '1 sélectionné(e) / 68 trouvé(e) / 1699 total'.

| Nom | Description | Valeur actuelle | Valeur par défaut | Modifiable |
|--------------------------------|--|---|---|------------|
| allow_ansi | Allow ANSI escape sequences in output. Values: Terminal, Always, Never | | Terminal | |
| client_cache_size | soft maximum number of directory entries in client cache | | 16384 | ✓ |
| cluster_addr | cluster-facing address to bind to | | - | |
| colors | Colouring CephFS shell input and output. Values: Terminal, Always, Never | | Terminal | |
| container_image | container image (used by cephadm orchestrator) | global: registry.suse.de/dev/storage/7.0/containers/ses/7/ceph/ceph@sha256:7429243adb8d9b41bc94625d52cc3294f2d38cab2b241a938dd2755f2dbcb546 | registry.suse.com/ses/7/ceph/cephlatest | |
| continuation_prompt | Prompt string when a command continue to second line | | > | |
| debug_shell | Allow tracebacks on error for CephFS Shell | | false | ✓ |
| device_failure_prediction_mode | Method used to predict device failures | | none | ✓ |
| echo | Print command issued on prompt before execution | | false | ✓ |
| editor | Default text editor for shell | | vim | |

FIGURE 4.14 : CONFIGURATION DE LA GRAPPE

Cliquez sur la flèche de liste déroulante en regard d'une option de configuration dans la colonne *Nom* pour afficher un tableau étendu avec des informations détaillées sur l'option, telles que son type de valeur, les valeurs minimales et maximales autorisées, si elle peut être mise à jour au moment de l'exécution, etc.

Lorsqu'une option spécifique est mise en surbrillance, vous pouvez éditer sa ou ses valeurs en cliquant sur le bouton *Modifier* en haut à gauche de l'en-tête du tableau. Confirmez les modifications en cliquant sur *Enregistrer*.

4.7 Affichage de la carte CRUSH

Cliquez sur *Grappe* > *Carte CRUSH* pour afficher une carte CRUSH de la grappe. Pour plus d'informations générales sur les cartes CRUSH, reportez-vous à la [Section 17.5, « Manipulation de la carte CRUSH »](#).

Cliquez sur la racine, les noeuds ou des OSD spécifiques pour afficher des informations plus détaillées, telles que la pondération CRUSH, la profondeur dans l'arborescence de la carte, la classe de périphérique de l'OSD, etc.



FIGURE 4.15 : CARTE CRUSH

4.8 Affichage des modules Manager

Cliquez sur *Graphe* > *Modules Manager* pour afficher une liste des modules Ceph Manager disponibles. Chaque ligne reprend un nom de module et indique s'il est actuellement activé ou non.

| <div> <div>Modifier</div> <div> <div></div> <div></div> </div> <div>10</div> <div> <div></div> <div></div> </div> </div> | |
|--|--------|
| Nom | Activé |
| ansible | |
| balancer | |
| crash | |
| dashboard | ✓ |
| deepsea | |
| devicehealth | |
| diskprediction_local | |
| influx | |
| insights | |
| iostat | ✓ |
| <div>1 sélectionné/Total de 24</div> <div> <div><</div> <div><<</div> <div>1</div> <div>2</div> <div>3</div> <div>>></div> <div>></div> </div> | |

FIGURE 4.16 : MODULES MANAGER

Cliquez sur la flèche de liste déroulante en regard d'un module dans la colonne *Nom* pour afficher un tableau étendu avec des paramètres détaillés dans le tableau *Détails* ci-dessous. Pour les éditer, cliquez sur *Modifier* en haut à gauche de l'en-tête du tableau. Confirmez les modification en cliquant sur *Mise à jour*.

Cliquez sur la flèche de la liste déroulante en regard du bouton *Modifier* en haut à gauche de l'en-tête du tableau pour *Activer* ou *Désactiver* un module.

4.9 Affichage des journaux

Cliquez sur *Grappe* > *Journaux* pour afficher une liste des entrées de journal récentes de la grappe. Chaque ligne comprend un tampon horaire, le type de l'entrée de journal et le message consigné. Cliquez sur l'onglet *Journaux d'audit* pour afficher les entrées de journal du sous-système d'audit. Reportez-vous à la [Section 11.5, « Audit des requêtes API »](#) pour connaître les commandes permettant d'activer ou de désactiver les audits.



FIGURE 4.17 : JOURNAUX

4.10 Affichage de la surveillance

Cliquez sur *Grappe* > *Surveillance* pour gérer et afficher les détails des alertes Prometheus.

Si Prometheus est actif, dans ce volet de contenu, vous pouvez afficher des informations détaillées sur les *Alertes actives*, *Toutes les alertes* ou les *Silences*.



Note

Si Prometheus n'est pas déployé, une bannière d'information apparaît avec un lien vers la documentation appropriée.

5 Gestion des réserves



Astuce : complément d'informations sur les réserves

Pour plus d'informations générales sur les réserves Ceph, reportez-vous au [Chapitre 18, Gestion des réserves de stockage](#). Pour des informations spécifiques sur les réserves codées à effacement, reportez-vous au [Chapitre 19, Réserves codées à effacement](#).

Pour répertorier toutes les réserves disponibles, cliquez sur *Pools* (Réserves) dans le menu principal.

La liste affiche, pour chaque réserve, le nom, le type, l'application associée, le statut du groupe de placements, la taille des répliques, la dernière modification, le profil de code à effacement, l'ensemble de règles crush, l'utilisation et les statistiques de lecture/écriture.

Liste des réserves

Performance globale

| <div>+ Ajouter</div> | | | | | | | | | | | | <div><div></div></div> | | <div>10</div> | <div><div></div></div> | <div></div> | <div></div> |
|---------------------------|----------|--------------|-------------------------------|--------------|------------|-----------------------------|---------------------|-------------|-------------------|--------------------|--------------|------------------------|--|---------------|------------------------|-------------|-------------|
| Nom | Type | Applications | Statut du groupe de placement | Taille de la | Derni chan | Profil de code à effacement | Jeu de règles Crush | Utilisation | Octets en lecture | Octets en écriture | Opéra de lec | Opéra d'écrit | | | | | |
| .rgw.root | répliqué | rgw | 8 actifs+nettoyés | 3 | 22 | | replicated_rule | 0 % | <div></div> | <div></div> | 0/s | 0/s | | | | | |
| cephfs_data | répliqué | cephfs | 256 actifs+nettoyés | 3 | 209 | | replicated_rule | 0 % | <div></div> | <div></div> | 0/s | 0/s | | | | | |
| cephfs_metadata | répliqué | cephfs | 64 actifs+nettoyés | 3 | 210 | | replicated_rule | 0 % | <div></div> | <div></div> | 0/s | 0/s | | | | | |
| default.rgw.buckets.index | répliqué | rgw | 8 actifs+nettoyés | 3 | 75 | | replicated_rule | 0 % | <div></div> | <div></div> | 0/s | 0/s | | | | | |
| default.rgw.control | répliqué | rgw | 8 actifs+nettoyés | 3 | 25 | | replicated_rule | 0 % | <div></div> | <div></div> | 0/s | 0/s | | | | | |
| default.rgw.log | répliqué | rgw | 8 actifs+nettoyés | 3 | 30 | | replicated_rule | 0 % | <div></div> | <div></div> | 0/s | 0/s | | | | | |
| default.rgw.meta | répliqué | rgw | 8 actifs+nettoyés | 3 | 28 | | replicated_rule | 0 % | <div></div> | <div></div> | 0/s | 0/s | | | | | |
| family_photos | répliqué | cephfs | 128 actifs+nettoyés | 3 | 226 | | replicated_rule | 0 % | <div></div> | <div></div> | 0/s | 0/s | | | | | |
| testing_rbd_pool | répliqué | cephfs,rbd | 128 actifs+nettoyés | 3 | 76 | | replicated_rule | 0 % | <div></div> | <div></div> | 0,8/s | 0/s | | | | | |
| 0 sélectionné/Total de 9 | | | | | | | | | | | | | | | | | |

FIGURE 5.1 : LISTE DES RÉSERVES

Cliquez sur la flèche de liste déroulante en regard d'un nom de réserve dans la colonne *Nom* pour afficher une table étendue contenant des informations détaillées sur la réserve, telles que les détails généraux, les détails de performances et la configuration.

5.1 Ajout d'une nouvelle réserve

Pour ajouter une nouvelle réserve, cliquez sur *Créer* en haut à gauche du tableau des réserves. Dans le formulaire de la réserve, vous pouvez entrer le nom de la réserve, son type, ses applications, son mode de compression et ses quotas, y compris le nombre maximal d'octets et d'objets. Le formulaire de la réserve calcule lui-même le nombre de groupes de placements qui convient le mieux à cette réserve. Le calcul est basé sur la quantité d'OSD dans la grappe et le type de réserve sélectionné avec ses paramètres spécifiques. Dès qu'un nombre de groupes de placements est défini manuellement, il est remplacé par un nombre calculé. Confirmez en cliquant sur le bouton *Create pool* (Créer une réserve).

Créer Pool

Nom * potato-pool ✓

Type de réserve * replicated ✓ ▾

PG Autoscale on ✓ ▾

Taille de réplication * 3

Applications cephfs ✕

CRUSH

Jeu de règles Crush replicated_rule ▾ ⌂ + 🗑

Compression

Mode none ✓ ▾

Quotas

Max bytes ⓘ par exemple, 10 Gio

Max objects ⓘ 0

Créer Pool Annuler

FIGURE 5.2 : AJOUT D'UNE NOUVELLE RÉSERVE

5.2 Suppression de réserves

Pour supprimer une réserve, sélectionnez-la dans la ligne du tableau. Cliquez sur la flèche de liste déroulante en regard du bouton *Créer*, puis sur *Supprimer*.

5.3 Modification des options d'une réserve

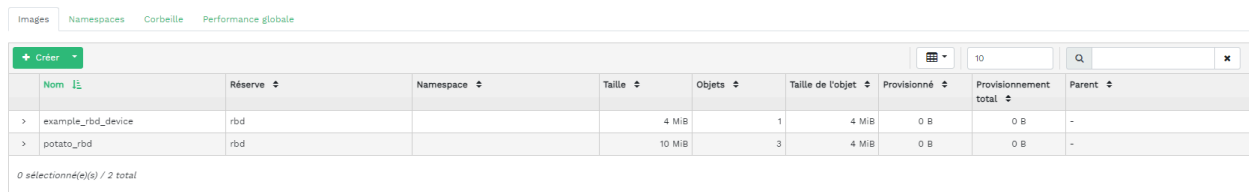
Pour modifier les options d'une réserve, sélectionnez la réserve dans la ligne du tableau, puis cliquez sur *Éditer* en haut à gauche du tableau des réserves.

Vous pouvez modifier le nom de la réserve, augmenter le nombre de groupes de placements et éditer la liste des applications et les paramètres de compression. Confirmez en cliquant sur le bouton *Edit pool* (Modifier une réserve).

6 Gestion du périphérique de traitement par blocs RADOS (RBD)

Pour répertorier tous les périphériques de bloc RADOS (RADOS Block Device, RBD) disponibles, cliquez sur *Bloc > Images* dans le menu principal.

La liste affiche de brèves informations sur le périphérique, telles que son nom, le nom de la réserve associée, l'espace de noms, la taille du périphérique, le nombre et la taille des objets sur le périphérique, des informations sur le provisioning des détails et le parent.



The screenshot shows a web interface with a top navigation bar containing 'Images', 'Namespaces', 'Corbeille', and 'Performance globale'. Below this is a table titled 'Images' with a '+ Créer' button and a search bar. The table has the following columns: Nom, Réserve, Namespace, Taille, Objets, Taille de l'objet, Provisionné, Provisionnement total, and Parent. Two rows are visible: 'example_rbd_device' and 'potato_rbd'.

| Nom | Réserve | Namespace | Taille | Objets | Taille de l'objet | Provisionné | Provisionnement total | Parent |
|----------------------|---------|-----------|--------|--------|-------------------|-------------|-----------------------|--------|
| > example_rbd_device | rbd | | 4 MiB | 1 | 4 MiB | 0 B | 0 B | - |
| > potato_rbd | rbd | | 10 MiB | 3 | 4 MiB | 0 B | 0 B | - |

0 sélectionné(e)s / 2 total

FIGURE 6.1 : LISTE DES IMAGES RBD

6.1 Affichage des détails sur les RBD

Pour afficher des informations plus détaillées sur un périphérique, cliquez sur sa ligne dans le tableau :

| | | | |
|-------------------------|---|---------------|-------------|
| Détails | Instantanés | Configuration | Performance |
| Nom | example_rbd_device | | |
| Réserve | rbd | | |
| Réserve de données | - | | |
| Créé | 30/04/2021 15:02:38 | | |
| Taille | 4 MiB | | |
| Objets | 1 | | |
| Taille de l'objet | 4 MiB | | |
| Fonctionnalités | deep-flatten exclusive-lock fast-diff layering object-map | | |
| Provisionné | 0 B | | |
| Provisionnement total | 0 B | | |
| Unité de segmentation | 4 MiB | | |
| Nombre de segmentations | 1 | | |
| Parent | - | | |
| Préfixe du nom de bloc | rbd_data.37ff2b17a0d1 | | |
| Tri | 22 | | |
| Format Version | 2 | | |

FIGURE 6.2 : DÉTAILS DES RBD

6.2 Affichage de la configuration d'un RBD

Pour afficher la configuration détaillée d'un périphérique, cliquez sur sa ligne dans le tableau, puis sur l'onglet *Configuration* dans le tableau inférieur :

| Détails | Instantanés | Configuration | Performance |
|------------------------------------|---|-------------------------|-------------|
| Nom ↕ | Description ↕ | Clef ↕ | Source ↕ |
| Rafale de bits/s | Limite de rafale d'octets en E/S souhaitée. | rbd_qos_bps_burst | Image |
| Limite de bits/s | Limite souhaitée d'octets E/S par seconde. | rbd_qos_bps_limit | Image |
| Rafale E/S par seconde | Limite de rafale d'opérations E/S souhaitée. | rbd_qos_iops_burst | Image |
| Limite E/S par seconde | Limite souhaitée d'opérations E/S par seconde. | rbd_qos_iops_limit | Image |
| Rafale E/S par seconde en lecture | Limite de rafale d'octets lus par seconde. | rbd_qos_read_bps_burst | Image |
| Limite de bits/s en lecture | Limite souhaitée de lecture d'octets par seconde. | rbd_qos_read_bps_limit | Image |
| Rafale E/S par seconde en lecture | Limite de rafale d'opérations de lecture | rbd_qos_read_iops_burst | Image |
| Limite E/S par seconde en lecture | Limite souhaitée d'opérations de lecture par seconde. | rbd_qos_read_iops_limit | Image |
| Rafale E/S par seconde en écriture | Limite de rafale de lectures d'octets souhaitée. | rbd_qos_write_bps_burst | Image |
| Limite de bits/s en écriture | Limite souhaitée d'écriture d'octets par seconde. | rbd_qos_write_bps_limit | Image |

FIGURE 6.3 : CONFIGURATION D'UN RBD

6.3 Création de RBD

Pour ajouter un nouveau périphérique, cliquez sur *Créer* en haut à gauche de l'en-tête du tableau, puis procédez comme suit dans l'écran *Créer RBD* :

FIGURE 6.4 : AJOUT D'UN NOUVEAU RBD

1. Entrez le nom du nouveau périphérique. Reportez-vous au *Manuel* « *Guide de déploiement* », Chapitre 2 « *Configuration matérielle requise et recommandations* », Section 2.11 « *Limitations concernant les noms* » pour plus d'informations sur les règles de dénomination.
2. Sélectionnez la réserve à laquelle est assignée l'application `rbd` à partir de laquelle le nouveau périphérique RBD va être créé.

3. Spécifiez la taille du nouveau périphérique.
4. Spécifiez des options supplémentaires pour le périphérique. Pour affiner les paramètres de l'appareil, cliquez sur *Avancé* et entrez des valeurs pour la taille d'objet, l'unité de segmentation ou le nombre de segments. Pour spécifier les limites de la qualité de service (Quality of Service, QoS), cliquez sur *Qualité de service* et indiquez les valeurs.
5. Confirmez en cliquant sur le bouton *Créer RBD*.

6.4 Suppression de RBD

Pour supprimer un périphérique, sélectionnez-le dans la ligne du tableau. Cliquez sur la flèche de liste déroulante en regard du bouton *Créer*, puis sur *Supprimer*. Confirmez la suppression en cliquant sur *Delete RBD* (Supprimer le RBD).



Astuce : déplacement de RBD vers la corbeille

La suppression d'un RBD est une opération irréversible. En revanche, *Déplacer vers la corbeille* vous permet de restaurer ultérieurement le périphérique en le sélectionnant sur l'onglet *Corbeille* du principal tableau et en cliquant sur *Restaurer* en haut à gauche de l'en-tête du tableau.

6.5 Création d'instantanés de périphériques de traitement par blocs RADOS (RBD)

Pour créer un instantané de périphérique de traitement par blocs RADOS, sélectionnez le périphérique dans la ligne du tableau. Le volet de contenu de configuration détaillé apparaît. Sélectionnez l'onglet *Instantanés* et cliquez sur *Créer* en haut à gauche de l'en-tête du tableau. Entrez le nom de l'instantané et confirmez en cliquant sur *Create RBD Snapshot* (Créer un instantané de RBD).

Après avoir sélectionné un instantané, vous pouvez effectuer des opérations supplémentaires sur le périphérique, telles que renommer, protéger, cloner, copier ou supprimer. *Rollback* restaure l'état de l'appareil à partir de l'instantané actuel.

| | | |
|---------|-------------|---------------|
| Détails | Instantanés | Configuration |
|---------|-------------|---------------|

| + Créer ▾ | | <div> <div></div> <div>10</div> </div> | <div> <div></div> <div></div> </div> | × |
|------------------------------|--------|--|--------------------------------------|---------------------|
| Nom | Taille | Provisionné | État | Créé |
| testing_rbd-20190215T095402Z | 10 Mio | 0 o | NON PROTÉGÉ | 15/02/2019 10:54:08 |
| testing_rbd-20190405T074138Z | 10 Mio | 0 o | NON PROTÉGÉ | 15/04/2019 09:41:42 |
| 0 sélectionné/Total de 2 | | | | |

FIGURE 6.5 : INSTANTANÉS DE RBD

6.6 Mise en miroir de RBD

Les images RBD peuvent être mises en miroir de manière asynchrone entre deux grappes Ceph. Vous pouvez utiliser Ceph Dashboard pour configurer la réplication d'images RBD entre plusieurs grappes. Cette fonctionnalité est disponible en deux modes :

Mode basé sur un journal

Ce mode utilise la fonctionnalité de journalisation de l'image RBD afin de garantir une réplication ponctuelle, cohérente entre les grappes en cas de panne.

Mode basé sur des instantanés

Ce mode utilise des instantanés en miroir d'image RBD planifiés régulièrement ou créés manuellement pour répliquer des images RBD cohérentes entre les grappes en cas de panne.

La mise en miroir est configurée réserve par réserve au sein des grappes homologues et peut être configurée sur un sous-ensemble spécifique d'images dans la réserve ou configurée pour mettre en miroir automatiquement toutes les images d'une réserve lorsque vous utilisez la mise en miroir basée sur le journal uniquement.

La mise en miroir est configurée à l'aide de la commande `rbd`, qui est installée par défaut dans SUSE Enterprise Storage 7.1. Le daemon `rbd-mirror` est chargé d'extraire les mises à jour d'image de la grappe homologue distante et de les appliquer à l'image au sein de la grappe locale. Reportez-vous à la [Section 6.6.2, « Activation du daemon `rbd-mirror` »](#) pour plus d'informations sur l'activation du daemon `rbd-mirror`.

Selon les besoins de réplication, la mise en miroir RBD peut être configurée pour une réplication unidirectionnelle ou bidirectionnelle :

Réplication unidirectionnelle

Lorsque les données sont mises en miroir uniquement à partir d'une grappe primaire vers une grappe secondaire, le daemon `rbd-mirror` s'exécute uniquement sur la grappe secondaire.

Réplication bidirectionnelle

Lorsque les données sont mises en miroir à partir des images primaires sur une grappe vers des images non primaires sur une autre grappe (et inversement), le daemon `rbd-mirror` s'exécute sur les deux grappes.



Important

Chaque instance du daemon `rbd-mirror` doit pouvoir se connecter simultanément aux grappes Ceph locales et distantes, par exemple tous les hôtes de moniteur et OSD. En outre, le réseau doit disposer de suffisamment de bande passante entre les deux centres de données pour gérer le workload en miroir.



Astuce : informations générales

Pour des informations générales et l'approche de ligne de commande pour la mise en miroir de périphérique de bloc RADOS, reportez-vous à la [Section 20.4, « Miroirs d'image RBD »](#).

6.6.1 Configuration de grappes primaires et secondaires

Une grappe *primaire* est celle sur laquelle la réserve d'origine avec les images est créée. Une grappe *secondaire* est celle sur laquelle la réserve ou les images sont répliquées à partir de la grappe *primaire*.



Note : assignation d'un nom relatif

Les termes *primaire* et *secondaire* peuvent être relatifs dans le contexte de la réplication, étant donné qu'ils se rapportent davantage aux réserves individuelles qu'aux grappes. Par exemple, dans la réplication bidirectionnelle, une réserve peut être mise en miroir de la grappe *primaire* vers la grappe *secondaire*, tandis qu'une autre réserve peut être mise en miroir de la grappe *secondaire* vers la grappe *primaire*.

6.6.2 Activation du daemon `rbd-mirror`

Les procédures suivantes montrent comment effectuer les tâches d'administration de base pour configurer la mise en miroir à l'aide de la commande `rbd`. La mise en miroir est configurée réserve par réserve au sein des grappes Ceph.

Les étapes de configuration d'une réserve doivent être effectuées sur les deux grappes homologues. Ces procédures supposent que deux grappes, nommées « *primary* » (primaire) et « *secondary* » (secondaire), sont accessibles depuis un seul hôte pour plus de clarté.

Le daemon `rbd-mirror` effectue la réplication réelle des données de grappe.

1. Renommez `ceph.conf` et les fichiers de trousseau de clés et copiez-les de l'hôte primaire vers l'hôte secondaire :

```
cephuser@secondary > cp /etc/ceph/ceph.conf /etc/ceph/primary.conf
cephuser@secondary > cp /etc/ceph/ceph.admin.client.keyring \
/etc/ceph/primary.client.admin.keyring
cephuser@secondary > scp PRIMARY_HOST:/etc/ceph/ceph.conf \
/etc/ceph/secondary.conf
cephuser@secondary > scp PRIMARY_HOST:/etc/ceph/ceph.client.admin.keyring \
/etc/ceph/secondary.client.admin.keyring
```

2. Pour activer la mise en miroir sur une réserve avec `rbd`, indiquez `mirror pool enable`, le nom de la réserve et le mode de mise en miroir :

```
cephuser@adm > rbd mirror pool enable POOL_NAME MODE
```



Note

Le mode de mise en miroir peut être image ou pool. Par exemple :

```
cephuser@secondary > rbd --cluster primary mirror pool enable image-pool
image
cephuser@secondary > rbd --cluster secondary mirror pool enable image-pool
image
```

3. Dans Ceph Dashboard, accédez à *Bloc* > *Mise en miroir*. Le tableau *Daemons* à gauche montre les daemons rbd-mirror en cours d'exécution et leur état de santé.

Daemons

| <div><div><div></div></div><div><div></div></div><div>10</div><div><div></div></div><div><div></div></div><div><div></div></div></div> | | | | |
|--|------|--------------|------------------------|---------|
| Instance ↕ | ID ↕ | Nom d'hôte ↕ | Version ↕ | Santé ↕ |
| 292255 | test | doc-ses-min4 | 14.2.2-354-g8878cf2360 | OK |
| Total de 1 | | | | |

FIGURE 6.6 : EXÉCUTION DE DAEMONS rbd-mirror

6.6.3 Désactivation de la mise en miroir

Pour désactiver la mise en miroir sur une réserve avec rbd, spécifiez la commande mirror pool disable et le nom de la réserve :

```
cephuser@adm > rbd mirror pool disable POOL_NAME
```

Lorsque la mise en miroir est désactivée sur une réserve de cette manière, la mise en miroir est également désactivée sur toutes les images (dans la réserve) pour lesquelles la mise en miroir a été explicitement activée.

6.6.4 Démarrage des homologues

Pour que `rd-mirror` découvre sa grappe homologue, l'homologue doit être enregistré dans la réserve et un compte utilisateur doit être créé. Ce processus peut être automatisé avec `rd` et les commandes `mirror pool peer bootstrap create` ainsi que `mirror pool peer bootstrap import`.

Pour créer manuellement un nouveau jeton Bootstrap avec `rd`, spécifiez la commande `mirror pool peer bootstrap create`, un nom de réserve, ainsi qu'un nom de site facultatif pour décrire la grappe locale :

```
cephuser@adm > rbd mirror pool peer bootstrap create [--site-name local-site-name] pool-name
```

La sortie de la commande `mirror pool peer bootstrap create` sera un jeton qui doit être fourni à la commande `mirror pool peer bootstrap import`. Par exemple, sur la grappe primaire :

```
cephuser@adm > rbd --cluster primary mirror pool peer bootstrap create --site-name primary
image-pool
eyJmc2lkIjojOWY1MjgyZGI0Yjg5ODU0NTk2LTgwOTgtMzIwYzFmYzY5MmYzIiwia2xpZW50X2lkIjoicmJkLWlpcnJvcilwZWVya2V5IjojQVFBUnczOWQwdkhvQmhBQVlMM1I4RmR5dHNJQU50bkFTZ0l0TVE9PSIsIm1vbl9ob3N0I
WlpcnJvcilwZWVya2V5IjojQVFBUnczOWQwdkhvQmhBQVlMM1I4RmR5dHNJQU50bkFTZ0l0TVE9PSIsIm1vbl9ob3N0I
joiW3YyOjE5Mi4xNjguMS4zOjY4MjAsdjE6MTkyLjE2OC4xLjM6NjgyMV0ifQ==
```

Pour importer manuellement le jeton Bootstrap créé par une autre grappe à l'aide de la commande `rd`, spécifiez la commande `mirror pool peer bootstrap import`, le nom de la réserve, un chemin d'accès au jeton créé (ou « - » pour lire à partir de l'entrée standard), ainsi qu'un nom de site facultatif pour décrire la grappe locale et une direction de mise en miroir (par défaut, `rx-tx` pour la mise en miroir bidirectionnelle, mais cela peut également être défini sur `rx-only` pour la mise en miroir unidirectionnelle) :

```
cephuser@adm > rbd mirror pool peer bootstrap import [--site-name local-site-name] [--direction rx-only or rx-tx] pool-name token-path
```

Par exemple, sur la grappe secondaire :

```
cephuser@adm > cat >>E0F < token
eyJmc2lkIjojOWY1MjgyZGI0Yjg5ODU0NTk2LTgwOTgtMzIwYzFmYzY5MmYzIiwia2xpZW50X2lkIjoicmJkLWlpcn
JvcilwZWVya2V5IjojQVFBUnczOWQwdkhvQmhBQVlMM1I4RmR5dHNJQU50bkFTZ0l0TVE9PSIsIm1vbl9ob3N0I
```

```
joiW3Yy0jE5Mi4xNjguMS4z0jY4MjAsdjE6MTkyLjE2OC4xLjM6NjgyMV0ifQ==  
EOF  
cephuser@adm > rbd --cluster secondary mirror pool peer bootstrap import --site-name  
secondary image-pool token
```

6.6.5 Suppression d'un homologue de grappe

Pour supprimer une grappe Ceph homologue de mise en miroir avec la commande `rbd`, indiquez la commande `mirror pool peer remove`, le nom de la réserve et l'UUID de l'homologue (disponible dans le résultat de la commande `rbd mirror pool info`) :

```
cephuser@adm > rbd mirror pool peer remove pool-name peer-uuid
```

6.6.6 Configuration de la réplication de réserve dans Ceph Dashboard

Le daemon `rbd-mirror` doit avoir accès à la grappe primaire pour pouvoir mettre en miroir des images RBD. Veillez à avoir suivi les étapes de la [Section 6.6.4, « Démarrage des homologues »](#) avant de poursuivre.

1. Sur les grappes *primaire* et *secondaire*, créez une réserve portant le même nom et assignez-lui l'application `rbd`. Reportez-vous à la [Section 5.1, « Ajout d'une nouvelle réserve »](#) pour plus de détails sur la création d'une réserve.

Créer Pool

Nom *

mirrored-pool

✓

Type de réserve *

replicated

✓

PG Autoscale

off

✓

Groupe de placements *

4

✓

Taille de réplication *

3

Applications

✎

rbd

✕

CRUSH

Jeu de règles Crush

replicated_rule

⌵

🔍

+

🗑

Compression

Mode

none

⌵

Quotas

Max bytes 📄

par exemple, 10 Gio

Max objects 📄

0

Configuration RBD

Qualité de service ➕

Créer Pool

Annuler

2. Sur les tableaux de bord des grappes *primaire* et *secondaire*, accédez à *Bloc* > *Mise en miroir*. Dans le tableau *Réserves* à droite, cliquez sur le nom de la réserve à répliquer et, après avoir cliqué sur *Mode d'édition*, sélectionnez le mode de réplication. Dans cet exemple, nous travaillerons avec un mode de réplication *réserve*, ce qui signifie que toutes les images d'une réserve donnée seront répliquées. Confirmez en cliquant sur *Mise à jour*.



Modifier le mode de mise en miroir de la réserve ×

Pour modifier le mode de mise en miroir de la réserve `mirrored-pool`, sélectionnez un nouveau mode dans la liste, puis cliquez sur `Mettre à jour`.

Mode

Réserve ✓

Mise à jour Annuler

FIGURE 6.8 : CONFIGURATION DU MODE DE RÉPLICATION



Important : erreur ou avertissement sur la grappe primaire

Après la mise à jour du mode de réplication, un drapeau d'erreur ou d'avertissement apparaît dans la colonne de droite correspondante. Cela est dû au fait que la réserve n'a encore aucun utilisateur homologue qui lui est assigné pour la réplication. Ignorez ce drapeau pour la grappe *primaire* étant donné que nous assignons un utilisateur homologue à la grappe *secondaire* uniquement.

3. Sur le tableau de bord de la grappe *secondaire*, accédez à *Bloc > Mise en miroir*. Ajoutez l'homologue de miroir de réserve en sélectionnant *Ajouter un homologue*. Spécifiez les détails de la grappe *primaire* :



The screenshot shows a web form titled "Ajouter un homologue de mise en miroir de la réserve" with a close button (X) in the top right corner. Below the title, there is a text instruction: "Ajoutez les attributs d'homologue de mise en miroir de la réserve `mirrored-pool`, puis cliquez sur **Soumettre**." The form contains four input fields: "Nom de la grappe *" with the value "primary", "ID CephX *" with the value "rbd-mirror-peer", "Adresses du moniteur" with the value "10.100.24.60,10.100.24.61,10.100.24.62", and "Clef CephX" with the value "AQAlr5Vd4y/UMRAATF8ee/wnPF2x3P9DtmEP2Q==". At the bottom right, there are two buttons: "Soumettre" (green) and "Annuler" (white).

FIGURE 6.9 : AJOUT DES INFORMATIONS D'IDENTIFICATION DE L'HOMOLOGUE

Nom de la grappe

Chaîne unique arbitraire qui identifie la grappe primaire (par exemple « primaire »). Le nom de la grappe doit être différent de celui de la grappe secondaire réel.

ID CephX

ID utilisateur Ceph que vous avez créé en tant qu'homologue de mise en miroir. Dans cet exemple, il s'agit de « rbd-mirror-peer ».

Adresses du moniteur

Liste des adresses IP des noeuds Ceph Monitor de la grappe principale, séparées par des virgules.

Clé CephX

Clé liée à l'ID utilisateur homologue. Vous pouvez la récupérer en exécutant l'exemple de commande suivant sur la grappe primaire :

```
cephuser@adm > ceph auth print_key pool-mirror-peer-name
```

Confirmez en cliquant sur *Soumettre*.

Réserves

Modifier le mode

10

| Nom | Mode | Leader | nb en local | nb à distance | Santé |
|------------------|-------|--------|-------------|---------------|-------|
| example_rbd_pool | pool | 292255 | 2 | 2 | OK |
| mirrored-pool | pool | 292255 | 0 | 0 | OK |
| pool3 | image | 292255 | 2 | 2 | OK |
| pool4 | pool | 292255 | 1 | 1 | OK |

1 sélectionné/Total de 4

FIGURE 6.10 : LISTE DES RÉSERVES RÉPLIQUÉES

6.6.7 Vérification du fonctionnement de la réplication d'image RBD

Lorsque le daemon `rbd-mirror` est en cours d'exécution et que la réplication d'image RBD est configurée sur Ceph Dashboard, il est temps de vérifier si la réplication fonctionne réellement :

1. Sur l'instance Ceph Dashboard de la grappe *primaire*, créez une image RBD de sorte que sa réserve parent soit la réserve que vous avez déjà créée à des fins de réplication. Activez les fonctionnalités Verrou exclusif et Journalisation pour l'image. Reportez-vous à la [Section 6.3, « Création de RBD »](#) pour plus de détails sur la création d'images RBD.

Créer un RBD

Nom *

mirrored-image1

Réserve *

mirrored-pool

☐

Utiliser une réserve de données dédiée

Taille *

60 Gio

Fonctionnalités

☐

Aplatissement en profondeur

☒

Superposition

☒

Verrou exclusif

☐

Assignation d'objet (nécessite exclusive-lock)

☒

Journalisation (nécessite exclusive-lock)

☐

Diff. rapide (nécessite object-map)

[Avancé...](#)

Créer un RBD

Annuler

- Après avoir créé l'image que vous souhaitez répliquer, ouvrez l'instance Ceph Dashboard de la grappe *secondaire* et accédez à *Bloc > Mise en miroir*. Le tableau *Réserves* sur la droite rend compte du changement du *nombre d'images à distance* et synchronise le *nombre d'images en local*.

Réserves

Modifier le mode

10

| Nom | Mode | Leader | nb en local | nb à distance | Santé |
|------------------|-------|--------|-------------|---------------|-------|
| example_rbd_pool | pool | 292255 | 2 | 2 | OK |
| mirrored-pool | pool | 292255 | 1 | 1 | OK |
| pool3 | image | 292255 | 2 | 2 | OK |
| pool4 | pool | 292255 | 1 | 1 | OK |

1 sélectionné/Total de 4

FIGURE 6.12 : NOUVELLE IMAGE RBD SYNCHRONISÉE



Astuce : progression de la réplication

Le tableau *Images* en bas de la page affiche le statut de la réplication des images RBD. L'onglet *Problèmes* inclut les éventuels problèmes, l'onglet *Synchronisation en cours* affiche la progression de la réplication des images et l'onglet *Prêt* répertorie toutes les images dont la réplication a réussi.

Images

Problèmes Synchronisation en cours Prêt

| | | | |
|---------------|-----------------|--|-----------|
| | | | |
| Réserve | Image | Description | État |
| mirrored-pool | mirrored-image1 | replaying, master_position=[object_number=3, tag_tid=1, entry_tid=3], mirror_position=[object_number=3, tag_tid=1, entry_tid=3], entries_behind_master=0 | Relecture |
| pool3 | img1 | replaying, master_position=[object_number=1, tag_tid=2, entry_tid=6401], mirror_position=[object_number=1, tag_tid=2, entry_tid=6401], entries_behind_master=0 | Relecture |
| pool3 | new_image1 | replaying, master_position=[object_number=1, tag_tid=3, entry_tid=641], mirror_position=[object_number=1, tag_tid=3, entry_tid=641], entries_behind_master=0 | Relecture |
| pool4 | img4 | replaying, master_position=[object_number=3, tag_tid=1, entry_tid=3], mirror_position=[object_number=3, tag_tid=1, entry_tid=3], entries_behind_master=0 | Relecture |
| Total de 6 | | | |

FIGURE 6.13 : STATUT DE LA RÉPLICATION DES IMAGES RBD

3. Sur la grappe *primaire*, écrivez des données dans l'image RBD. Sur l'instance Ceph Dashboard de la grappe *secondaire*, accédez à *Bloc > Images* et surveillez si la taille de l'image correspondante augmente au fur et à mesure que des données sont inscrites sur la grappe primaire.

6.7 Gestion des passerelles iSCSI



Astuce : plus d'informations sur les passerelles iSCSI

Pour plus d'informations sur les passerelles iSCSI, reportez-vous au [Chapitre 22, Passerelle Ceph iSCSI](#).

Pour répertorier toutes les passerelles disponibles et les images assignées, cliquez sur *Bloc > iSCSI* dans le menu principal. Un onglet *Présentation* s'ouvre et répertorie les passerelles iSCSI actuellement configurées et les images RBD assignées.

Le tableau *Passerelles* répertorie l'état de chaque passerelle, le nombre de cibles iSCSI et le nombre de sessions. Le tableau *Images* répertorie le nom de chaque image assignée, le type de backstore du nom de la réserve associée et d'autres détails statistiques.

L'onglet *Cibles* répertorie les cibles iSCSI actuellement configurées.



| Cible | Portails | Images | # Sessions |
|--------------------------------------|---------------------------------------|-------------------------------|------------|
| > iqn.2001-07.com.ceph:1619785904397 | master.ses7-mini.test:10.20.165.200 | rbd/example_rbd_device_potato | 0 |
| > iqn.2001-07.com.ceph:1619785974221 | master.ses7-mini.test:192.168.121.185 | rbd/potato-rbd | 0 |

0 sélectionné(s) / 2 total

FIGURE 6.14 : LISTE DES CIBLES iSCSI

Pour afficher des informations plus détaillées sur une cible, cliquez sur la flèche de liste déroulante sur la ligne cible du tableau. Un schéma structuré en arborescence s'ouvre et liste les disques, les portails, les initiateurs et les groupes. Cliquez sur un élément pour le développer et afficher son contenu détaillé, éventuellement avec une configuration associée dans le tableau sur la droite.

Présentation

Cibles

Créer

Authentification de la découverte

10

Q

x

| Cible | Portails | Images | # Sessions |
|------------------------------------|---|--------------------------------|------------|
| iqn.2001-07.com.ceph:1597683071527 | node1.asettle-dashboards.test:10.20.164.201 | rbid/example_rbd_device_potato | 0 |

Topologie iSCSI

iqn.2001-07.com.ceph:1597683071527

Disks

rbid/example_rbd_device_potato

Portals

node1.asettle-dashboards.test:10.20.164.201

Initiators

Groups

rbid/example_rbd_device_potato

| Nom | Actuel | Valeur par défaut |
|------------------|--------------------------------------|-------------------|
| backstore | utilisateur:rbid (tcmu-runner) | rbid |
| hw_max_sectors | 1024 | 1024 |
| lun | 0 | |
| max_data_area_mb | 8 | 8 |
| osd_op_timeout | 30 | 30 |
| qfull_timeout | 5 | 5 |
| wwn | bf60abfd-9159-4098-bc9b-2be4daaefa5c | |
| 7 total | | |

| > | iqn.2001-07.com.ceph:1597683089358 | node1.asettle-dashboards.test:10.20.164.201 | rbid/potato-rbid |
|---|------------------------------------|---|------------------|
| | | | 0 |

1 sélectionné(e)(s) / 2 total

FIGURE 6.15 : DÉTAILS D'UNE CIBLE iSCSI

6.7.1 Ajout de cibles iSCSI

Pour ajouter une nouvelle cible iSCSI, cliquez sur *Créer* en haut à gauche dans le tableau *Cibles*, puis entrez les informations requises.

Créer Target

IQN cible *

iqn.2001-07.com.ceph:1620221227294

+ Ajouter un portail

Portails *

master.ses7-mini.test:10.20.165.200

+ Ajouter un portail

Images

rbd/potato_rbd

lun: 0

Backstore: rbd.

+ Ajouter une image

☐ Authentification ACL

Utilisateur

Mot de passe

Utilisateur commun

Mot de passe commun

Créer Target

Annuler

FIGURE 6.16 : AJOUT D'UNE NOUVELLE CIBLE

51

Ajout de cibles iSCSI | SES 7.1

1. Entrez l'adresse cible de la nouvelle passerelle.
2. Cliquez sur *Add portal* (Ajouter un portail) et sélectionnez un ou plusieurs portails iSCSI dans la liste.
3. Cliquez sur *Add image* (Ajouter une image) et sélectionnez une ou plusieurs images RBD pour la passerelle.
4. Si vous devez utiliser l'authentification pour accéder à la passerelle, sélectionnez la case à cocher *Authentication ACL*, puis entrez les informations d'identification. Vous pouvez trouver des options d'authentification plus avancées après avoir activé *l'authentification mutuelle* et *l'authentification par détection*.
5. Confirmez en cliquant sur le bouton *Create Target* (Créer une cible).

6.7.2 Modification des cibles iSCSI

Pour modifier une cible iSCSI existante, cliquez sur sa ligne dans le tableau *Cibles*, puis cliquez sur *Modifier* en haut à gauche du tableau.

Vous pouvez ensuite modifier la cible iSCSI, ajouter ou supprimer des portails et ajouter ou supprimer des images RBD associées. Vous pouvez également ajuster les informations d'authentification pour la passerelle.

6.7.3 Suppression de cibles iSCSI

Pour supprimer une cible iSCSI, sélectionnez la ligne du tableau, puis cliquez sur la flèche déroulante en regard du bouton *Modifier*, puis sélectionnez *Supprimer*. Cochez la case *Oui* et confirmez avec le bouton *Delete iSCSI* (Supprimer la cible iSCSI).

6.8 Qualité de service (QoS) RBD



Astuce : informations supplémentaires

Pour plus d'informations générales et une description des options de configuration QoS RBD, reportez-vous à la [Section 20.6, « Paramètres QoS »](#).

Les options QoS peuvent être configurées à différents niveaux.

- Globalement
- Par réserve
- Par image

La configuration *globale* est en haut de la liste et sera utilisée pour toutes les images RBD récemment créées et pour les images qui ne remplacent pas ces valeurs au niveau de la couche image RBD ou réserve. Une valeur d'option spécifiée globalement peut être remplacée par réserve ou par image. Les options spécifiées sur une réserve seront appliquées à toutes les images RBD de cette réserve, excepté si elles sont remplacées par une option de configuration définie sur une image. Les options spécifiées sur une image remplaceront les options spécifiées sur une réserve et celles définies globalement.

De cette façon, il est possible de définir des valeurs par défaut de manière globale, de les adapter pour toutes les images RBD d'une réserve spécifique et de remplacer la configuration de la réserve pour des images RBD individuelles.

6.8.1 Configuration globale des options

Pour configurer globalement les options de périphérique de traitement par blocs RADOS (RBD), sélectionnez *Grappe* > *Configuration* dans le menu principal.

1. Pour lister toutes les options de configuration globale disponibles, en regard de *Niveau*, choisissez *Avancé* dans le menu déroulant.
2. Filtrez ensuite les résultats du tableau en entrant rbd_qos dans le champ de recherche. Cela liste toutes les options de configuration disponibles pour QoS.
3. Pour modifier une valeur, cliquez sur sa ligne dans le tableau, puis sélectionnez *Modifier* en haut à gauche du tableau. La boîte de dialogue *Modifier* contient six champs différents pour spécifier les valeurs. Les valeurs d'option de configuration RBD sont requises dans la zone de texte *mgr*.



Note

Contrairement aux autres boîtes de dialogue, celle-ci ne vous permet pas de spécifier la valeur dans des unités pratiques. Vous devez définir ces valeurs en octets ou en entrées/sorties par seconde (IOPS), en fonction de l'option que vous modifiez.

6.8.2 Configuration des options d'une nouvelle réserve

Pour créer une réserve et configurer ses options de configuration RBD, cliquez sur *Réserve* > *Créer*. Sélectionnez *replicated* (répliqué) comme type de réserve. Vous devez ensuite ajouter la balise d'application `rbd` à la réserve pour être en mesure de configurer les options QoS RBD.



Note

Il n'est pas possible de définir les options de configuration RBD QoS sur une réserve codée à effacement. Pour configurer les options QoS RBD pour les réserves codées à effacement, vous devez modifier la réserve de métadonnées répliquée d'une image RBD. La configuration sera ensuite appliquée à la réserve de données codée à effacement de cette image.

6.8.3 Configuration des options d'une réserve existante

Pour configurer les options QoS RBD sur une réserve existante, cliquez sur *Réserve*, puis sur la ligne de la réserve dans le tableau, et sélectionnez *Modifier* en haut à gauche du tableau.

Dans la boîte de dialogue, vous devriez voir la section *Configuration RBD* suivie d'une section *Qualité de service*.



Note

Si aucune d'elles n'apparaît, vous êtes probablement occupé à modifier une réserve *codée à effacement*, qui ne permet pas de définir des options de configuration RBD, ou la réserve n'est pas configurée pour être utilisée par des images RBD. Dans ce dernier cas, assignez la balise d'application `rbd` à la réserve et les sections de configuration correspondantes s'afficheront.

6.8.4 Options de configuration

Cliquez sur *Qualité de service* + pour développer les options de configuration. Une liste de toutes les options disponibles s'affiche. Les unités des options de configuration sont déjà affichées dans les zones de texte. En cas d'option d'octets par seconde, vous pouvez utiliser des raccourcis tels que « 1M » ou « 5G ». Ils seront automatiquement convertis en « 1 Mo/s » et « 5 Go/s » respectivement.

Si vous cliquez sur le bouton de réinitialisation à droite de chaque zone de texte, toute valeur éventuelle définie pour la réserve est supprimée. En revanche, cela ne supprime pas les valeurs de configuration des options configurées globalement ou au niveau d'une image RBD.

6.8.5 Création d'options QoS RBD avec une nouvelle image RBD

Pour créer une image RBD avec des options QoS RBD définies sur cette image, sélectionnez *Bloc > Images*, puis cliquez sur *Créer*. Cliquez sur *Avancé...* pour développer la section de configuration avancée. Cliquez sur *Qualité de service* + pour ouvrir toutes les options de configuration disponibles.

6.8.6 Modification des options QoS RBD sur les images existantes

Pour modifier les options QoS RBD pour une image existante, sélectionnez *Bloc > Images* et cliquez sur la ligne du tableau de la réserve, puis sur *Modifier*. La boîte de dialogue de modification s'affiche. Cliquez sur *Avancé...* pour développer la section de configuration avancée. Cliquez sur *Qualité de service* + pour ouvrir toutes les options de configuration disponibles.

6.8.7 Modification des options de configuration lors de la copie ou du clonage d'images

Si une image RBD est clonée ou copiée, par défaut, les valeurs définies sur cette image particulière sont également copiées. Si vous le souhaitez, vous pouvez les modifier lors de la copie ou du clonage en spécifiant les valeurs de configuration mises à jour dans la boîte de dialogue de copie/clonage, de la même manière que lors de la création ou de l'édition d'une image RBD. De cette façon, seules les valeurs de l'image RBD copiée ou clonée sont définies (ou réinitialisées). Cette opération ne modifie ni la configuration d'image RBD source, ni la configuration globale.

Si vous choisissez de réinitialiser la valeur d'option lors de la copie/du clonage, aucune valeur ne sera définie pour cette option sur cette image. Cela signifie que toute valeur de cette option spécifiée pour la réserve parent sera utilisée si elle est configurée pour la réserve parent. Dans le cas contraire, le système appliquera la valeur par défaut globale.

7 Gestion de NFS Ganesha



Important

NFS Ganesha prend en charge les versions 4.1 et ultérieures de NFS. Il ne prend pas en charge la version 3 de NFS.



Astuce : plus d'informations sur NFS Ganesha

Pour plus d'informations générales sur NFS Ganesha, reportez-vous au [Chapitre 25, NFS Ganesha](#).

Pour répertorier toutes les exportations NFS disponibles, cliquez sur *NFS* dans le menu principal. La liste affiche le répertoire de chaque exportation, le nom d'hôte du daemon, le type d'interface dorsale de stockage et son type d'accès.

| + Créer | | | | | | |
|--------------------------|----------------|---------------------|--------------------|---------|---------------------|--------------|
| | | | | | | |
| | Chemin | Pseudo | Grappe | Daemons | Backend de stockage | Type d'accès |
| > | /potato/potato | /exportimus-maximus | ganesha-sesdev_nfs | | CephFS | MDONLY_RO |
| > | /root | /exportcephfs | ganesha-sesdev_nfs | | CephFS | RW |
| > | /root/potato | /exportpotato | ganesha-sesdev_nfs | | CephFS | MDONLY |
| 0 sélectionné/Total de 3 | | | | | | |

FIGURE 7.1 : LISTE DES EXPORTATIONS NFS

Pour afficher des informations plus détaillées sur une exportation NFS, cliquez sur sa ligne dans le tableau.

| | | |
|----------------------------|--------------------|-------------|
| Détails | | Clients (0) |
| Type d'accès | RW | |
| Système de fichiers CephFS | sesdev_fs | |
| Utilisateur CephFS | admin | |
| Grappe | ganesha-sesdev_nfs | |
| Daemons | | |
| Protocole NFS | NFSv3, NFSv4 | |
| Chemin | /root | |
| Pseudo | /exportcephfs | |
| Squash | no_root_squash | |
| Backend de stockage | CephFS | |
| Transport | TCP, UDP | |

FIGURE 7.2 : DÉTAILS DE L'EXPORTATION NFS

7.1 Création d'exportations NFS

Pour ajouter une nouvelle exportation NFS, cliquez sur *Créer* en haut à gauche du tableau des exportations, puis entrez les informations requises.

Créer une exportation NFS

Graphe *

ganesha-sesdev_nfs

Daemons

Aucun élément sélectionné.

+ Ajouter un daemon

Backend de stockage *

CephFS

✓

ID utilisateur CephFS *

admin

✓

Nom CephFS *

sesdev_fs

✓

Libellé de sécurité

☐ Activer le libellé de sécurité

Chemin CephFS *

/root

✓

Un nouveau répertoire sera créé

Protocole NFS *

☒ NFSv3
 ☒ NFSv4

Balise NFS ?

Pseudo * ?

/exportcephfs

✓

Type d'accès *

RW

✓

Autorise toutes les opérations

Squash *

no_root_squash

✓

Protocole de transport *

☒ UDP
 ☒ TCP

Clients

Accès possible par tous les clients

+ Ajouter des clients

Créer une exportation NFS

Annuler

FIGURE 7.3 : AJOUT D'UNE NOUVELLE EXPORTATION NFS

1. Sélectionnez un ou plusieurs daemons NFS Ganesha qui exécuteront l'exportation.
2. Sélectionnez une interface dorsale de stockage.



Important

À l'heure actuelle, seules les exportations NFS soutenues par CephFS sont prises en charge.

3. Sélectionnez un ID utilisateur et d'autres options associées à l'interface dorsale.
4. Entrez le chemin de répertoire pour l'exportation NFS. Si le répertoire n'existe pas sur le serveur, il est alors créé.
5. Indiquez d'autres options liées à NFS, telles que la version du protocole NFS prise en charge, le pseudo, le type d'accès, l'action squash ou le protocole de transport.
6. Si vous devez limiter l'accès à des clients spécifiques, cliquez sur *Add clients* (Ajouter des clients) et ajoutez leurs adresses IP, ainsi que le type d'accès et les options d'action squash.
7. Confirmez en cliquant sur *Créer une exportation NFS*.

7.2 Suppression des exportations NFS

Pour supprimer une exportation, sélectionnez-la et mettez-la en surbrillance dans la ligne du tableau. Cliquez sur la flèche de liste déroulante en regard du bouton *Modifier* et sélectionnez *Supprimer*. Cochez la case *Oui* et confirmez avec le bouton *Supprimer l'exportation NFS*.

7.3 Modification d'exportations NFS

Pour modifier une exportation existante, sélectionnez-la et mettez-la en surbrillance dans la ligne du tableau, puis cliquez sur *Modifier* en haut à gauche du tableau des exportations.

Vous pouvez ensuite régler tous les détails de l'exportation NFS.

Modifier l'exportation NFS

Graphe *

ganesha-sesdev_nfs

Daemons

Aucun élément sélectionné.

+ Ajouter un daemon

Backend de stockage *

CephFS

ID utilisateur CephFS *

admin

Nom CephFS *

sesdev_fs

Libellé de sécurité

☐ Activer le libellé de sécurité

Chemin CephFS *

/root

Protocole NFS *

☒ NFSv3
☒ NFSv4

Balise NFS ?

Pseudo * ?

/exportcephfs

Type d'accès *

RW

Autorise toutes les opérations

Squash *

no_root_squash

Protocole de transport *

☒ UDP
☒ TCP

Clients

Accès possible par tous les clients

+ Ajouter des clients

Modifier l'exportation NFS

Annuler

FIGURE 7.4 : MODIFICATION D'UNE EXPORTATION NFS

8 Gestion de CephFS



Astuce : informations supplémentaires

Pour trouver des informations détaillées sur CephFS, reportez-vous au [Chapitre 23, Système de fichiers en grappe](#).

8.1 Affichage de l'aperçu CephFS

Cliquez sur *Systèmes de fichiers* dans le menu principal pour afficher la vue d'ensemble des systèmes de fichiers configurés. Le tableau principal affiche le nom de chaque système de fichiers, sa date de création et s'il est activé ou non.

Vous pouvez cliquer sur la ligne d'un système de fichiers dans le tableau afin de découvrir des détails sur son rang et sur les réserves ajoutées au système de fichiers.

| Nom | Créé | Activé |
|-----------|---------------------|--------|
| sesdev_fs | 30/04/2021 12:30:19 | ✓ |

Détails

Clients 2

Directories

Détails des performances

Rangs

| Rang | État | Daemon | Activité | Dentries | Inodes |
|------|--------|------------------------|------------|----------|--------|
| 0 | active | sesdev_fs.master.Ingtr | Reqs: 0 /s | 10 | 13 |

1 total

Réserves

| Réserve | Type | Taille |
|----------------------------|------|----------|
| cephfs.sesdev_fs. data | | 13.1 GiB |
| cephfs.sesdev_fs. metadata | | 13.1 GiB |

2 total

Standbys

| Daemons en veille |
|------------------------|
| sesdev_fs.node3.jszwzi |

FIGURE 8.1 : DÉTAILS CEPHFS

Au bas de l'écran, vous pouvez voir des statistiques comptant le nombre d'inodes MDS connexes et les demandes des clients, recueillies en temps réel.

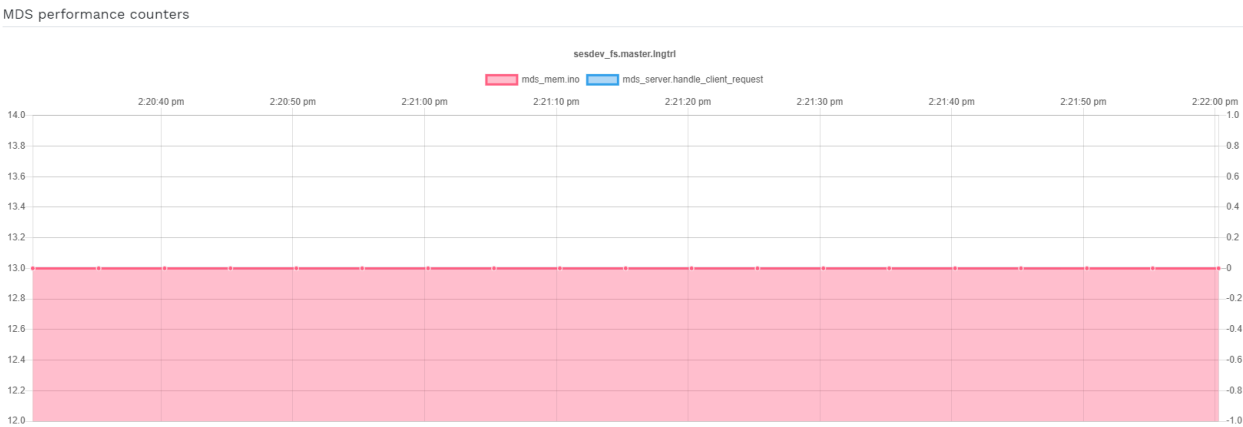


FIGURE 8.2 : DÉTAILS CEPHFS

9 Gestion de la passerelle Object Gateway



Important

Avant de commencer, il se peut que vous rencontriez la notification suivante lorsque vous tentez d'accéder à l'interface client Object Gateway sur Ceph Dashboard :

Information

No RGW credentials found, please consult the documentation on how to enable RGW for the dashboard.

Please consult the documentation on how to configure and enable the Object Gateway management functionality.

Cela est dû au fait que la passerelle Object Gateway n'a pas été configurée automatiquement par `cephadm` pour Ceph Dashboard. Si vous rencontrez cette notification, suivez les instructions de la [Section 10.4, « Activation de l'interface client de gestion d'Object Gateway »](#) pour activer manuellement l'interface client Object Gateway pour Ceph Dashboard.



Astuce : plus d'informations sur Object Gateway

Pour plus d'informations générales sur Object Gateway, reportez-vous au [Chapitre 21, Ceph Object Gateway](#).

9.1 Affichage des instances Object Gateway

Pour afficher une liste des passerelles Object Gateway configurées, cliquez sur *Object Gateway* > *Daemons*. La liste comprend l'ID de la passerelle, le nom de l'hôte du noeud de grappe sur lequel le daemon de la passerelle est en cours d'exécution et le numéro de version de la passerelle.

Cliquez sur la flèche de liste déroulante en regard du nom de la passerelle pour afficher des informations détaillées sur cette dernière. L'onglet *Compteurs de performance* affiche les détails des opérations de lecture/d'écriture et les statistiques de cache.

| | | |
|--------------------|---|------------------------|
| Détails | Compteurs de performance | Détails de performance |
| arch | x86_64 | |
| ceph_release | octopus | |
| ceph_version | ceph version 15.2.4-557-g4ac763f0b3 (4ac763f0b3864d9168bc4a46fef26d7fa759545e) octopus (stable) | |
| ceph_version_short | 15.2.4-557-g4ac763f0b3 | |
| container_hostname | node1 | |
| container_image | registry.suse.de/devel/storage/7.0/containers/ses/7/ceph/ceph | |
| cpu | Processeur Intel Core (Haswell, pas TSX) | |
| distro | sles | |
| distro_description | SUSE Linux Enterprise Server 15 SP2 | |
| distro_version | 15.2 | |
| frontend_config#0 | beast port=80 | |
| frontend_type#0 | beast | |
| hostname | node1 | |
| kernel_description | N° 1 SMP Mer 29 jul 18:54:11 UTC 2020 (dbe0add) | |
| kernel_version | 5.3.18-24.9-default | |
| mem_swap_kb | 0 | |
| mem_total_kb | 4020668 | |
| num_handles | 1 | |
| os | Linux | |
| pid | 1 | |
| zone_id | 2a664005-94ad-432a-b873-d563fed68496 | |
| zone_name | valeur par défaut | |
| zonegroup_id | cc4ec3c6-c611-4bfd-a155-e3e05552d5cd | |
| zonegroup_name | valeur par défaut | |

FIGURE 9.1 : DÉTAILS DE LA PASSERELLE

9.2 Gestion des utilisateurs Object Gateway

Cliquez sur *Object Gateway* > *Utilisateurs* pour afficher une liste des utilisateurs Object Gateway existants.

Cliquez sur la flèche de liste déroulante en regard du nom d'utilisateur pour afficher des détails sur le compte utilisateur, tels que les informations de statut ou les détails de quota utilisateur et de compartiments.

Détails
Clés

| | |
|---------------------------------|-----------|
| Nom d'utilisateur | rgw-admin |
| Nom complet | admin |
| Suspendu | Non |
| Système | Oui |
| Nombre maximal de compartiments | 1000 |

Quota utilisateur

| | |
|-------------------------|-----|
| Activé | Non |
| Taille maximale | - |
| Nombre maximal d'objets | - |

Quota de compartiments

| | |
|-------------------------|-----|
| Activé | Non |
| Taille maximale | - |
| Nombre maximal d'objets | - |

FIGURE 9.2 : UTILISATEURS DE LA PASSERELLE

9.2.1 Ajout d'un nouvel utilisateur de la passerelle

Pour ajouter un nouvel utilisateur de la passerelle, cliquez sur *Créer* en haut à gauche de l'en-tête du tableau. Remplissez les informations d'identification, les détails sur la clé S3 et les quotas utilisateur et de compartiments, puis confirmez en cliquant sur *Créer un utilisateur*.

Créer un utilisateur

Nom d'utilisateur *

exemple_utilisateur_rgw

✓

Nom complet *

Exemple d'utilisateur

✓

Adresse électronique

exemple@utilisateur.com

✓

Nombre max. compartiments

Personnalisé

✓

1000

☐ Suspendu

Clé S3

☒ Générer automatiquement la clé

Quota utilisateur

☐ Activé

Quota de compartiments

☒ Activé

☒ Taille illimitée

☒ Nombre illimité d'objets

Créer un utilisateur

Annuler

FIGURE 9.3 : AJOUT D'UN NOUVEL UTILISATEUR DE LA PASSERELLE

9.2.2 Suppression d'utilisateurs de la passerelle

Pour supprimer un utilisateur de la passerelle, sélectionnez-le et mettez-le en surbrillance. Cliquez sur le bouton de liste déroulante en regard de *Modifier* et sélectionnez *Supprimer* dans la liste pour supprimer le compte utilisateur. Cochez la case *Oui* et confirmez avec le bouton *Supprimer l'utilisateur*.

9.2.3 Modification des détails des utilisateurs de la passerelle

Pour modifier les détails d'un utilisateur de la passerelle, sélectionnez-le et mettez-le en surbrillance. Cliquez sur *Modifier* en haut à gauche de l'en-tête du tableau.

Modifiez les informations utilisateur de base ou supplémentaires, telles que les fonctions, clés, utilisateurs secondaires et les détails de quota. Confirmez en cliquant sur le bouton *Modifier l'utilisateur*.

L'onglet *Clés* comprend une liste en lecture seule des utilisateurs de la passerelle et de leurs clés secrète et d'accès. Pour afficher les clés, cliquez sur un nom d'utilisateur dans la liste, puis sélectionnez *Afficher* en haut à gauche de l'en-tête du tableau. Dans la boîte de dialogue *Clef S3*, cliquez sur l'icône en forme d'oeil pour dévoiler les clés, ou cliquez sur l'icône du presse-papiers pour copier la clé associée dans le presse-papiers.

9.3 Gestion des compartiments Object Gateway

Les compartiments Object Gateway (OGW) implémentent la fonctionnalité des conteneurs OpenStack Swift. Ils servent de conteneurs pour le stockage des objets de données.

Cliquez sur *Object Gateway > Compartiments* pour afficher une liste des compartiments Object Gateway.

9.3.1 Ajout d'un nouveau compartiment

Pour ajouter un nouvel compartiment Object Gateway, cliquez sur *Créer* en haut à gauche de l'en-tête du tableau. Entrez le nom du compartiment, sélectionnez le propriétaire et définissez la cible de placement. Confirmez en cliquant sur le bouton *Create Bucket* (Créer un compartiment).



Note

À ce stade, vous pouvez également activer le verrouillage en sélectionnant *Activé*. Cela peut toutefois être configuré après la création. Pour plus d'informations, reportez-vous à la [Section 9.3.3, « Modification du compartiment »](#).

9.3.2 Affichage des détails d'un compartiment

Pour afficher des informations détaillées sur un compartiment Object Gateway, cliquez sur la flèche de liste déroulante en regard du nom du compartiment.

| | |
|-------------------------|--|
| Détails | |
| Nom | export |
| ID | 2a664005-94ad-432a-b873-d563fed68496.14523.1 |
| Propriétaire | rgw-admin |
| Type d'index | Normal |
| Règle de placement | default-placement |
| Marqueur | 2a664005-94ad-432a-b873-d563fed68496.14523.1 |
| Marqueur maximal | 0#,1#,2#,3#,4#,5#,6#,7#,8#,9#,10# |
| Version | 0#1,1#1,2#1,3#1,4#1,5#1,6#1,7#1,8#1,9#1,10#1 |
| Version principale | 0#0,1#0,2#0,3#0,4#0,5#0,6#0,7#0,8#0,9#0,10#0 |
| Heure de modification | 24/08/2020 13:24:34 |
| Groupe de zones | cc4ec3c6-c611-4bfd-a155-e3e05552d5cd |
| Contrôle des versions | Suspendu |
| Suppression MFA | Désactivé |
| Quota de compartiments | |
| Activé | Non |
| Taille maximale | Illimité |
| Nombre maximal d'objets | Illimité |
| Verrouillage | |
| Activé | Non |

FIGURE 9.4 : DÉTAILS D'UN COMPARTIMENT DE PASSERELLE



Astuce : quota de compartiments

Sous le tableau *Détails*, vous pouvez trouver des détails sur les paramètres de quota de compartiments et de verrouillage.

9.3.3 Modification du compartiment

Sélectionnez et mettez en surbrillance un compartiment, puis cliquez sur *Modifier* en haut à gauche de l'en-tête du tableau.

Vous pouvez mettre à jour le propriétaire du compartiment ou activer le contrôle des versions, l'authentification multi-critères ou le verrouillage. Cliquez sur *Modifier le compartiment* pour confirmer les modifications.

Modifier le compartiment


Id

eaf156f1-e787-4c5c-8e86-06cba6481d65.44187.1

Nom

root


Propriétaire *

asettle 


Cible de placement

default-placement


Contrôle des versions

☐ Activé 

Authentification multi-critères (MFA)

☐ Suppression activée 

Verrouillage

☐ Activé 

Modifier le compartiment

Annuler

FIGURE 9.5 : MODIFICATION DES DÉTAILS DU COMPARTIMENT

9.3.4 Suppression d'un compartiment

Pour supprimer un compartiment Object Gateway, sélectionnez-le et mettez-le en surbrillance. Cliquez sur le bouton de liste déroulante en regard de *Modifier* et sélectionnez *Supprimer* dans la liste pour supprimer le compartiment. Cochez la case *Oui* et confirmez avec le bouton *Delete bucket* (Supprimer le compartiment).

10 Configuration manuelle

Cette section présente des informations avancées pour les utilisateurs qui préfèrent configurer manuellement les paramètres du tableau de bord via la ligne de commande.

10.1 Configuration de la prise en charge de TLS/SSL

Par défaut, toutes les connexions HTTP vers le tableau de bord sont sécurisées via TLS/SSL. Une connexion sécurisée nécessite un certificat SSL. Vous pouvez soit utiliser un certificat auto-signé, soit générer un certificat à faire signer par une autorité de certification bien connue.



Astuce : désactivation de SSL

Vous pouvez souhaiter désactiver le support SSL pour une raison spécifique, par exemple, si le tableau de bord s'exécute derrière un proxy qui ne prend pas en charge SSL.

Soyez toutefois prudent lorsque vous désactivez SSL, car **les noms d'utilisateur et les mots de passe** seront envoyés au tableau de bord **non chiffrés**.

Pour désactiver SSL, exécutez la commande suivante :

```
cephuser@adm > ceph config set mgr mgr/dashboard/ssl false
```



Astuce : redémarrage des processus Ceph Manager

Vous devez redémarrer manuellement les processus Ceph Manager après avoir modifié le certificat et la clé SSL. Pour ce faire, vous pouvez exécuter la commande

```
cephuser@adm > ceph mgr fail ACTIVE-MANAGER-NAME
```

ou désactiver, puis réactiver le module de tableau de bord, ce qui déclenche une régénération du gestionnaire :

```
cephuser@adm > ceph mgr module disable dashboard  
cephuser@adm > ceph mgr module enable dashboard
```

10.1.1 Création de certificats auto-signés

La création d'un certificat auto-signé pour une communication sécurisée est simple. Elle vous permet d'exécuter rapidement le tableau de bord.



Note : exigence des navigateurs Web

En cas de certificat auto-signé, la plupart des navigateurs Web exigent une confirmation explicite avant d'établir une connexion sécurisée au tableau de bord.

Pour générer et installer un certificat auto-signé, utilisez la commande intégrée suivante :

```
cephuser@adm > ceph dashboard create-self-signed-cert
```

10.1.2 Utilisation de certificats signés par une autorité de certification

Pour sécuriser correctement la connexion au tableau de bord et pour éviter l'exigence de confirmation explicite imposée par les navigateurs Web en cas de certificat auto-signé, nous vous recommandons d'utiliser un certificat signé par une autorité de certification.

Vous pouvez générer une paire de clés de certificat avec une commande similaire à la suivante :

```
# openssl req -new -nodes -x509 \  
-subj "/O=IT/CN=ceph-mgr-dashboard" -days 3650 \  
-keyout dashboard.key -out dashboard.crt -extensions v3_ca
```

La commande ci-dessus produit les fichiers `dashboard.key` et `dashboard.crt`. Une fois le fichier `dashboard.crt` signé par une autorité de certification, activez-le pour toutes les instances Ceph Manager en exécutant les commandes suivantes :

```
cephuser@adm > ceph dashboard set-ssl-certificate -i dashboard.crt  
cephuser@adm > ceph dashboard set-ssl-certificate-key -i dashboard.key
```



Astuce : des certificats différents pour chaque instance du gestionnaire

Si vous avez besoin de certificats différents pour chaque instance Ceph Manager, modifiez les commandes et incluez le nom de l'instance comme suit. Remplacez *NAME* par le nom de l'instance Ceph Manager (généralement le nom d'hôte associé) :

```
cephuser@adm > ceph dashboard set-ssl-certificate NAME -i dashboard.crt  
cephuser@adm > ceph dashboard set-ssl-certificate-key NAME -i dashboard.key
```

10.2 Modification du nom d'hôte et du numéro de port

Ceph Dashboard se lie à une adresse TCP/IP et à un port TCP spécifiques. Par défaut, l'instance Ceph Manager actuellement active qui héberge le tableau de bord se lie au port TCP 8443 (ou 8080 lorsque SSL est désactivé).



Note

Si un pare-feu est activé sur les hôtes qui exécutent Ceph Manager (et donc Ceph Dashboard), vous devrez peut-être modifier la configuration pour autoriser l'accès à ces ports. Pour plus d'informations sur les paramètres de pare-feu pour Ceph, reportez-vous au *Manuel « Troubleshooting Guide », Chapitre 13 « Hints and tips », Section 13.7 « Firewall settings for Ceph »*.

Ceph Dashboard se lie par défaut à « :: », ce qui correspond à toutes les adresses IPv4 et IPv6 disponibles. Vous pouvez modifier l'adresse IP et le numéro de port de l'application Web afin qu'ils s'appliquent à toutes les instances Ceph Manager en utilisant les commandes suivantes :

```
cephuser@adm > ceph config set mgr mgr/dashboard/server_addr IP_ADDRESS  
cephuser@adm > ceph config set mgr mgr/dashboard/server_port PORT_NUMBER
```



Astuce : configuration séparée des instances Ceph Manager

Étant donné que chaque daemon `ceph-mgr` héberge sa propre instance du tableau de bord, vous devrez peut-être les configurer séparément. Pour changer l'adresse IP et le numéro de port pour une instance spécifique du gestionnaire, utilisez les commandes suivantes (remplacez `NAME` par l'ID de l'instance `ceph-mgr`) :

```
cephuser@adm > ceph config set mgr mgr/dashboard/NAME/server_addr IP_ADDRESS
cephuser@adm > ceph config set mgr mgr/dashboard/NAME/server_port PORT_NUMBER
```



Astuce : liste des noeuds d'extrémité configurés

La commande `ceph mgr services` affiche tous les noeuds d'extrémité actuellement configurés. Recherchez la clé `dashboard` pour obtenir l'URL d'accès au tableau de bord.

10.3 Modification des noms d'utilisateur et des mots de passe

Si vous ne souhaitez pas utiliser le compte d'administrateur par défaut, créez un compte utilisateur différent et associez-le à au moins un rôle. Nous fournissons un ensemble de rôles système prédéfinis que vous pouvez utiliser. Pour plus d'informations, reportez-vous au [Chapitre 11, Gestion des utilisateurs et des rôles via la ligne de commande](#).

Pour créer un utilisateur avec des privilèges d'administrateur, utilisez la commande suivante :

```
cephuser@adm > ceph dashboard ac-user-create USER_NAME PASSWORD administrator
```

10.4 Activation de l'interface client de gestion d'Object Gateway

Pour utiliser la fonctionnalité de gestion d'Object Gateway du tableau de bord, vous devez fournir les informations d'identification de connexion d'un utilisateur avec le drapeau `system` activé :

1. Si vous n'avez pas d'utilisateur avec le drapeau `system`, créez-en un :

```
cephuser@adm > radosgw-admin user create --uid=USER_ID --display-name=DISPLAY_NAME --system
```

Prenez note des clés `clé_accès` et `clé_secrète` dans la sortie de la commande.

2. Vous pouvez également obtenir les informations d'identification d'un utilisateur existant à l'aide de la commande `radosgw-admin` :

```
cephuser@adm > radosgw-admin user info --uid=USER_ID
```

3. Fournissez les informations d'identification reçues au tableau de bord dans des fichiers distincts :

```
cephuser@adm > ceph dashboard set-rgw-api-access-key ACCESS_KEY_FILE
cephuser@adm > ceph dashboard set-rgw-api-secret-key SECRET_KEY_FILE
```



Note

Par défaut, le pare-feu est activé dans SUSE Linux Enterprise Server 15 SP3. Pour plus d'informations sur la configuration du pare-feu, reportez-vous au *Manuel « Troubleshooting Guide », Chapitre 13 « Hints and tips », Section 13.7 « Firewall settings for Ceph »*.

Plusieurs points sont à prendre en considération :

- Le nom de l'hôte et le numéro de port d'Object Gateway sont déterminés automatiquement.
- Si plusieurs zones sont utilisées, le système détermine automatiquement l'hôte dans le groupe de zones maître et la zone maître. Cela est suffisant pour la plupart des configurations, mais dans certaines circonstances, vous pouvez souhaiter définir le nom de l'hôte et le port manuellement :

```
cephuser@adm > ceph dashboard set-rgw-api-host HOST
```

```
cephuser@adm > ceph dashboard set-rgw-api-port PORT
```

- Voici d'autres paramètres dont vous pouvez avoir besoin :

```
cephuser@adm > ceph dashboard set-rgw-api-scheme SCHEME # http or https
cephuser@adm > ceph dashboard set-rgw-api-admin-resource ADMIN_RESOURCE
cephuser@adm > ceph dashboard set-rgw-api-user-id USER_ID
```

- Si vous utilisez un certificat auto-signé ([Section 10.1, « Configuration de la prise en charge de TLS/SSL »](#)) dans votre configuration Object Gateway, désactivez la vérification du certificat dans le tableau de bord pour éviter les refus de connexions dus à des certificats signés par une autorité de certification inconnue ou ne correspondant pas au nom d'hôte :

```
cephuser@adm > ceph dashboard set-rgw-api-ssl-verify False
```

- Si Object Gateway prend trop de temps pour traiter les requêtes et que le tableau de bord s'exécute selon des timeouts, la valeur de ces derniers peut être ajustée (elle est par défaut de 45 secondes) :

```
cephuser@adm > ceph dashboard set-rest-requests-timeout SECONDS
```

10.5 Activation de la gestion iSCSI

Ceph Dashboard gère les cibles iSCSI à l'aide de l'API REST fournie par le service `rgw-target-api` de la passerelle Ceph iSCSI. Assurez-vous qu'elle est installée et activée sur les passerelles iSCSI.



Note

La fonctionnalité de gestion iSCSI de Ceph Dashboard dépend de la dernière version 3 du projet `ceph-iscsi`. Vérifiez que votre système d'exploitation fournit la version correcte, sinon Ceph Dashboard n'activera pas les fonctions de gestion.

Si l'API REST `ceph-iscsi` est configurée en mode HTTPS et qu'elle utilise un certificat auto-signé, configurez le tableau de bord pour éviter la vérification du certificat SSL lors de l'accès à l'API `ceph-iscsi`.

Désactivez la vérification SSL de l'API :

```
cephuser@adm > ceph dashboard set-iscsi-api-ssl-verification false
```

Définissez les passerelles iSCSI disponibles :

```
cephuser@adm > ceph dashboard iscsi-gateway-list
cephuser@adm > ceph dashboard iscsi-gateway-add scheme://username:password@host[:port]
cephuser@adm > ceph dashboard iscsi-gateway-rm gateway_name
```

10.6 Activation de Single Sign-On

Single Sign-on (SSO) est une méthode de contrôle d'accès qui permet aux utilisateurs de se connecter avec un seul ID et mot de passe à plusieurs applications simultanément.

Ceph Dashboard prend en charge l'authentification externe des utilisateurs via le protocole SAML 2.0. Étant donné que l'*autorisation* passe toujours par le tableau de bord, vous devez d'abord créer des comptes utilisateur et les associer aux rôles souhaités. En revanche, le processus d'*authentification* peut être exécuté par un *fournisseur d'identité* (IdP) existant.

Pour configurer Single Sign-on, utilisez la commande suivante :

```
cephuser@adm > ceph dashboard sso setup saml2 CEPH_DASHBOARD_BASE_URL \
  IDP_METADATA IDP_USERNAME_ATTRIBUTE \
  IDP_ENTITY_ID SP_X_509_CERT \
  SP_PRIVATE_KEY
```

Paramètres :

CEPH_DASHBOARD_BASE_URL

URL de base à laquelle Ceph Dashboard est accessible (par exemple, « <https://cephdashboard.local> »).

IDP_METADATA

URL, chemin ou contenu du fichier XML de métadonnées IdP (par exemple, « <https://myidp/metadata> »).

IDP_USERNAME_ATTRIBUTE

Facultatif. Attribut qui sera utilisé pour obtenir le nom d'utilisateur à partir de la réponse d'authentification. Sa valeur par défaut est « uid ».

IDP_ENTITY_ID

Facultatif. Utilisez ce paramètre lorsqu'il existe plusieurs ID d'entité pour les métadonnées IdP.

SP_X_509_CERT / SP_PRIVATE_KEY

Facultatif. Chemin du fichier ou contenu du certificat qui sera utilisé par Ceph Dashboard (Fournisseur de service) pour la signature et le codage. Ces chemins de fichiers doivent être accessibles à partir de l'instance Ceph Manager active.



Note : requêtes SAML

La valeur de l'émetteur des requêtes SAML suivra ce modèle :

```
CEPH_DASHBOARD_BASE_URL/auth/saml2/metadata
```

Pour afficher la configuration SAML 2.0 actuelle, exécutez :

```
cephuser@adm > ceph dashboard sso show saml2
```

Pour désactiver Single Sign-on, exécutez :

```
cephuser@adm > ceph dashboard sso disable
```

Pour vérifier si SSO est activé, exécutez :

```
cephuser@adm > ceph dashboard sso status
```

Pour activer SSO, exécutez :

```
cephuser@adm > ceph dashboard sso enable saml2
```

11 Gestion des utilisateurs et des rôles via la ligne de commande

Cette section explique comment gérer les comptes utilisateur employés par Ceph Dashboard. Elle vous aide à créer ou modifier des comptes utilisateur, ainsi qu'à définir des rôles et autorisations appropriés pour les utilisateurs.

11.1 Gestion de la stratégie de mot de passe

Par défaut, la fonction de stratégie de mot de passe est activée, y compris les vérifications suivantes :

- Le mot de passe comporte-t-il plus de N caractères ?
- L'ancien et le nouveau mots de passe sont-ils identiques ?

La fonction de stratégie de mot de passe peut être complètement activée ou désactivée :

```
cephuser@adm > ceph dashboard set-pwd-policy-enabled true|false
```

Les vérifications individuelles suivantes peuvent être activées ou désactivées :

```
cephuser@adm > ceph dashboard set-pwd-policy-check-length-enabled true|false
cephuser@adm > ceph dashboard set-pwd-policy-check-oldpwd-enabled true|false
cephuser@adm > ceph dashboard set-pwd-policy-check-username-enabled true|false
cephuser@adm > ceph dashboard set-pwd-policy-check-exclusion-list-enabled true|false
cephuser@adm > ceph dashboard set-pwd-policy-check-complexity-enabled true|false
cephuser@adm > ceph dashboard set-pwd-policy-check-sequential-chars-enabled true|false
cephuser@adm > ceph dashboard set-pwd-policy-check-repetitive-chars-enabled true|false
```

En outre, les options suivantes sont disponibles pour configurer le comportement de la stratégie de mot de passe.

- Longueur minimale du mot de passe (8 par défaut) :

```
cephuser@adm > ceph dashboard set-pwd-policy-min-length N
```

- Complexité minimale du mot de passe (10 par défaut) :

```
cephuser@adm > ceph dashboard set-pwd-policy-min-complexity N
```

La complexité du mot de passe est calculée en classant chaque caractère du mot de passe.

- Liste de mots, séparés par des virgules, qui ne peuvent pas être utilisés dans un mot de passe :

```
cephuser@adm > ceph dashboard set-pwd-policy-exclusion-list word[,...]
```

11.2 Gestion des comptes utilisateur

Ceph Dashboard prend en charge la gestion de plusieurs comptes utilisateur. Chaque compte utilisateur se compose d'un nom d'utilisateur, d'un mot de passe (stocké sous forme codée à l'aide de `bcrypt`), d'un nom facultatif et d'une adresse électronique facultative.

Les comptes utilisateur sont stockés dans la base de données de configuration de Ceph Monitor et sont partagés globalement par toutes les instances de Ceph Manager.

Utilisez les commandes suivantes pour gérer les comptes utilisateur :

Afficher les utilisateurs existants :

```
cephuser@adm > ceph dashboard ac-user-show [USERNAME]
```

Créer un utilisateur :

```
cephuser@adm > ceph dashboard ac-user-create USERNAME -i [PASSWORD_FILE] [ROLENAME]  
[NAME] [EMAIL]
```

Supprimer un utilisateur :

```
cephuser@adm > ceph dashboard ac-user-delete USERNAME
```

Changer le mot de passe d'un utilisateur :

```
cephuser@adm > ceph dashboard ac-user-set-password USERNAME -i PASSWORD_FILE
```

Changer le nom et l'adresse électronique d'un utilisateur :

```
cephuser@adm > ceph dashboard ac-user-set-info USERNAME NAME EMAIL
```

Désactiver l'utilisateur

```
cephuser@adm > ceph dashboard ac-user-disable USERNAME
```

```
cephuser@adm > ceph dashboard ac-user-enable USERNAME
```

11.3 Rôles et autorisations des utilisateurs

Cette section décrit les étendues de sécurité que vous pouvez assigner à un rôle utilisateur, comment gérer les rôles des utilisateurs et comment les affecter aux comptes utilisateur.

11.3.1 Définition des étendues de sécurité

Les comptes utilisateur sont associés à un ensemble de rôles qui définissent les parties du tableau de bord auxquelles l'utilisateur peut accéder. Les parties du tableau de bord sont regroupées au sein d'une étendue de *sécurité*. Les étendues de sécurité sont prédéfinies et statiques. Les périmètres de sécurité suivants sont actuellement disponibles :

hosts

Inclut toutes les fonctionnalités liées à l'entrée de menu *Hôtes*.

config-opt

Inclut toutes les fonctionnalités liées à la gestion des options de configuration Ceph.

pool

Inclut toutes les fonctionnalités liées à la gestion des réserves.

osd

Inclut toutes les fonctionnalités liées à la gestion Ceph OSD.

monitor

Inclut toutes les fonctionnalités liées à la gestion Ceph Monitor.

rbd-image

Inclut toutes les fonctionnalités liées à la gestion des images de périphérique de bloc RADOS (RBD).

rbd-mirroring

Inclut toutes les fonctionnalités liées à la gestion de la mise en miroir de périphériques de bloc RADOS.

iscsi

Inclut toutes les fonctionnalités liées à la gestion iSCSI.

rgw

Inclut toutes les fonctionnalités liées à la gestion Object Gateway.

cephfs

Inclut toutes les fonctionnalités liées à la gestion CephFS.

manager

Inclut toutes les fonctionnalités liées à la gestion Ceph Manager.

log

Inclut toutes les fonctionnalités liées à la gestion des journaux Ceph.

grafana

Inclut toutes les fonctionnalités liées au proxy Grafana.

prometheus

Inclut toutes les fonctions liées à la gestion des alertes Prometheus.

dashboard-settings

Permet de modifier les paramètres du tableau de bord.

11.3.2 Spécification des rôles utilisateur

Un *rôle* spécifie un ensemble d'assignations entre une *étendue de sécurité* et un ensemble d'*autorisations*. Il existe quatre types d'autorisations : « read » (lire), « create » (créer), « update » (mettre à jour) et « delete » (supprimer).

L'exemple suivant spécifie un rôle pour lequel un utilisateur dispose des autorisations de lecture et de création concernant les fonctionnalités liées à la gestion des réserves, et possède toutes les autorisations pour les fonctionnalités liées à la gestion des images RBD :

```
{
  'role': 'my_new_role',
  'description': 'My new role',
  'scopes_permissions': {
    'pool': ['read', 'create'],
    'rbd-image': ['read', 'create', 'update', 'delete']
  }
}
```

```
}
```

Le tableau de bord fournit déjà un ensemble de rôles prédéfinis que nous appelons les *rôles système*. Vous pouvez les utiliser instantanément après une nouvelle installation de Ceph Dashboard :

administrateur

Fournit des autorisations complètes pour toutes les étendues de sécurité.

read-only

Fournit une autorisation de lecture pour toutes les étendues de sécurité, à l'exception des paramètres du tableau de bord.

block-manager

Fournit des autorisations complètes pour les étendues « rbd-image », « rbd-mirroring » et « iscsi ».

rgw-manager

Fournit des autorisations complètes pour l'étendue « rgw ».

cluster-manager

Fournit des autorisations complètes pour les étendues « hosts », « osd », « monitor », « manager » et « config-opt ».

pool-manager

Fournit des autorisations complètes pour l'étendue « pool ».

cephfs-manager

Fournit des autorisations complètes pour l'étendue « cephfs ».

11.3.2.1 Gestion des rôles personnalisés

Vous pouvez créer des rôles utilisateur à l'aide des commandes suivantes :

Créer un rôle :

```
cephuser@adm > ceph dashboard ac-role-create ROLENAME [DESCRIPTION]
```

Supprimer un rôle :

```
cephuser@adm > ceph dashboard ac-role-delete ROLENAME
```

Ajouter des autorisations d'étendue à un rôle :

```
cephuser@adm > ceph dashboard ac-role-add-scope-perms ROLENAME SCOPENAME PERMISSION
[PERMISSION...]
```

Supprimer des autorisations d'étendue d'un rôle :

```
cephuser@adm > ceph dashboard ac-role-del-perms ROLENAME SCOPENAME
```

11.3.2.2 Assignment de rôles à des comptes utilisateur

Utilisez les commandes suivantes pour assigner des rôles aux utilisateurs :

Définir des rôles utilisateur :

```
cephuser@adm > ceph dashboard ac-user-set-roles USERNAME ROLENAME [ROLENAME ...]
```

Ajouter des rôles supplémentaires à un utilisateur :

```
cephuser@adm > ceph dashboard ac-user-add-roles USERNAME ROLENAME [ROLENAME ...]
```

Supprimer des rôles d'un utilisateur :

```
cephuser@adm > ceph dashboard ac-user-del-roles USERNAME ROLENAME [ROLENAME ...]
```



Astuce : purge des rôles personnalisés

Si vous créez des rôles utilisateur personnalisés et avez l'intention de supprimer ultérieurement la grappe Ceph à l'aide de l'exécuteur **ceph.purge**, vous devez d'abord purger les rôles personnalisés. Pour plus de détails, reportez-vous à la [Section 13.9, « Suppression d'une grappe Ceph entière »](#).

11.3.2.3 Exemple : création d'un utilisateur et d'un rôle personnalisé

Cette section illustre une procédure de création d'un compte utilisateur pouvant gérer les images RBD, afficher et créer des réserves Ceph, et accéder en lecture seule à toutes les autres étendues.

1. Créez un utilisateur nommé tux :

```
cephuser@adm > ceph dashboard ac-user-create tux PASSWORD
```

2. Créez un rôle et spécifiez les autorisations d'étendue :

```
cephuser@adm > ceph dashboard ac-role-create rbd/pool-manager
cephuser@adm > ceph dashboard ac-role-add-scope-perms rbd/pool-manager \
  rbd-image read create update delete
cephuser@adm > ceph dashboard ac-role-add-scope-perms rbd/pool-manager pool read
create
```

3. Associez les rôles à l'utilisateur tux :

```
cephuser@adm > ceph dashboard ac-user-set-roles tux rbd/pool-manager read-only
```

11.4 Configuration du proxy

Si vous souhaitez établir une URL fixe pour accéder à Ceph Dashboard ou si vous ne souhaitez pas autoriser les connexions directes aux noeuds du gestionnaire, vous pouvez configurer un proxy qui transmet automatiquement les requêtes entrantes à l'instance ceph-mgr active.

11.4.1 Accès au tableau de bord avec des proxys inverses

Si vous accédez au tableau de bord via une configuration de proxy inverse, vous devrez peut-être le desservir sous un préfixe d'URL. Pour que le tableau de bord utilise des hyperliens qui incluent votre préfixe, vous pouvez définir le paramètre url_prefix :

```
cephuser@adm > ceph config set mgr mgr/dashboard/url_prefix URL_PREFIX
```

Ensuite, vous pouvez accéder au tableau de bord à l'adresse http://NOM_HÔTE:NUMÉRO_PORT/PRÉFIXE_URL/.

11.4.2 Désactivation des réacheminements

Si Ceph Dashboard se trouve derrière un proxy d'équilibrage de la charge tel que HAProxy, désactivez le comportement de redirection pour éviter les situations dans lesquelles les URL internes (impossibles à résoudre) sont publiées sur le client frontal. Utilisez la commande suivante pour que le tableau de bord réponde avec une erreur HTTP (500 par défaut) au lieu de rediriger vers le tableau de bord actif :

```
cephuser@adm > ceph config set mgr mgr/dashboard/standby_behaviour "error"
```


Pour rétablir le paramètre sur le comportement de redirection par défaut, utilisez la commande suivante :

```
cephuser@adm > ceph config set mgr mgr/dashboard/standby_behaviour "redirect"
```

11.4.3 Configuration des codes de statut d'erreur

Si le comportement de redirection est désactivé, vous devez personnaliser le code de statut HTTP des tableaux de bord de secours. Pour ce faire, exécutez la commande suivante :

```
cephuser@adm > ceph config set mgr mgr/dashboard/standby_error_status_code 503
```

11.4.4 Exemple de configuration HAProxy

L'exemple de configuration suivant concerne le passage TLS/SSL à l'aide de HAProxy.



Note

La configuration fonctionne dans les conditions suivantes : en cas de basculement du tableau de bord, le client frontal peut recevoir une réponse de redirection HTTP (303) et être redirigé vers un hôte impossible à résoudre.

Cela se produit lorsque le basculement intervient au cours de deux vérifications de l'état de santé HAProxy. Dans cette situation, le noeud de tableau de bord précédemment actif répond alors par un code 303 qui pointe vers le nouveau noeud actif. Pour éviter cela, vous devez envisager de désactiver le comportement de redirection sur les noeuds de secours.

```
defaults
    log global
    option log-health-checks
    timeout connect 5s
    timeout client 50s
    timeout server 450s

frontend dashboard_front
    mode http
    bind *:80
    option httplog
    redirect scheme https code 301 if !{ ssl_fc }
```

```
frontend dashboard_front_ssl
  mode tcp
  bind *:443
  option tcplog
  default_backend dashboard_back_ssl

backend dashboard_back_ssl
  mode tcp
  option httpchk GET /
  http-check expect status 200
  server x HOST:PORT ssl check verify none
  server y HOST:PORT ssl check verify none
  server z HOST:PORT ssl check verify none
```

11.5 Audit des requêtes API

L'API REST de Ceph Dashboard peut consigner les requêtes PUT, POST et DELETE dans le journal d'audit Ceph. La consignation est désactivée par défaut, mais vous pouvez l'activer avec la commande suivante :

```
cephuser@adm > ceph dashboard set-audit-api-enabled true
```

Si elle est activée, les paramètres suivants sont consignés pour chaque requête :

from

Origine de la requête, par exemple « https://[:1]:44410 ».

path

Chemin de l'API REST, par exemple /api/auth.

method

« PUT », « POST » ou « DELETE ».

user

Nom de l'utilisateur (ou « Aucun »).

Un exemple d'entrée de journal ressemble à ceci :

```
2019-02-06 10:33:01.302514 mgr.x [INF] [DASHBOARD] \
from='https://[:ffff:127.0.0.1]:37022' path='/api/rgw/user/exu' method='PUT' \
user='admin' params='{ "max_buckets": "1000", "display_name": "Example User", "uid":
"exu", "suspended": "0", "email": "user@example.com" }'
```



Astuce : désactivation de la consignation de la charge utile de requête

La consignation de la charge utile de requête (la liste des arguments et leurs valeurs) est activée par défaut. Vous pouvez la désactiver comme suit :

```
cephuser@adm > ceph dashboard set-audit-api-log-payload false
```

11.6 Configuration de NFS Ganesha dans Ceph Dashboard

Ceph Dashboard peut gérer les exportations NFS Ganesha qui utilisent CephFS ou Object Gateway comme backstore. Le tableau de bord gère les fichiers de configuration NFS Ganesha stockés dans des objets RADOS sur la grappe CephFS. NFS Ganesha doit stocker une partie de leur configuration dans la grappe Ceph.

Exécutez la commande suivante pour configurer l'emplacement de l'objet de configuration NFS Ganesha :

```
cephuser@adm > ceph dashboard set-ganesha-clusters-rados-pool-namespace pool_name[/namespace]
```

Vous pouvez désormais gérer les exportations NFS Ganesha à l'aide de Ceph Dashboard.

11.6.1 Configuration de plusieurs grappes NFS Ganesha

Ceph Dashboard prend en charge la gestion des exportations NFS Ganesha appartenant à différentes grappes NFS Ganesha. Il est recommandé que chaque grappe NFS Ganesha stocke ses objets de configuration dans une réserve/un espace de noms RADOS distinct pour isoler les configurations les unes des autres.

Utilisez la commande suivante pour spécifier les emplacements de la configuration de chaque grappe NFS Ganesha :

```
cephuser@adm > ceph dashboard set-ganesha-clusters-rados-pool-namespace cluster_id:pool_name[/namespace](,cluster_id:pool_name[/namespace]))*
```

cluster_id est une chaîne arbitraire qui identifie de manière unique la grappe NFS Ganesha.

Lors de la configuration de Ceph Dashboard avec plusieurs grappes NFS Ganesha, l'interface utilisateur Web vous permet automatiquement de choisir la grappe à laquelle une exportation appartient.

11.7 Plug-ins de débogage

Les plug-ins Ceph Dashboard étendent les fonctionnalités du tableau de bord. Le plug-in de débogage permet de personnaliser le comportement du tableau de bord en fonction du mode de débogage. Il peut être activé, désactivé ou vérifié à l'aide de la commande suivante :

```
cephuser@adm > ceph dashboard debug status
Debug: 'disabled'
cephuser@adm > ceph dashboard debug enable
Debug: 'enabled'
cephuser@adm > dashboard debug disable
Debug: 'disabled'
```

Par défaut, il est désactivé. Il s'agit du paramètre recommandé pour les déploiements de production. Si nécessaire, le mode débogage peut être activé sans redémarrage.

II Opération de grappe

- 12 Détermination de l'état d'une grappe **91**
- 13 Tâches opérationnelles **122**
- 14 Exécution des services Ceph **145**
- 15 Sauvegarde et restauration **150**
- 16 Surveillance et alertes **153**

12 Détermination de l'état d'une grappe

Lorsque vous disposez d'une grappe en cours d'exécution, vous pouvez utiliser l'outil **ceph** pour la surveiller. Pour déterminer l'état de la grappe, il faut généralement vérifier le statut des OSD Ceph, des moniteurs Ceph, des groupes de placement et des serveurs de métadonnées.



Astuce : mode interactif

Pour exécuter l'outil **ceph** en mode interactif, tapez **ceph** sans argument sur la ligne de commande. Le mode interactif est plus pratique si vous voulez entrer plusieurs commandes **ceph** consécutives. Par exemple :

```
cephuser@adm > ceph
ceph> health
ceph> status
ceph> quorum_status
ceph> mon stat
```

12.1 Vérification de l'état d'une grappe

La commande **ceph status** ou **ceph -s** permet de connaître l'état immédiat de la grappe :

```
cephuser@adm > ceph -s
cluster:
  id:      b4b30c6e-9681-11ea-ac39-525400d7702d
  health: HEALTH_OK

services:
  mon: 5 daemons, quorum ses-min1,ses-master,ses-min2,ses-min4,ses-min3 (age 2m)
  mgr: ses-min1.gpijpm(active, since 3d), standbys: ses-min2.oopvyh
  mds: my_cephfs:1 {0=my_cephfs.ses-min1.oterul=up:active}
  osd: 3 osds: 3 up (since 3d), 3 in (since 11d)
  rgw: 2 daemons active (myrealm.myzone.ses-min1.kwwazo, myrealm.myzone.ses-
min2.jngabw)

task status:
  scrub status:
    mds.my_cephfs.ses-min1.oterul: idle

data:
  pools: 7 pools, 169 pgs
```

```
objects: 250 objects, 10 KiB
usage:   3.1 GiB used, 27 GiB / 30 GiB avail
pgs:     169 active+clean
```

La sortie fournit les informations suivantes :

- ID de grappe
- État d'intégrité de la grappe
- Époque d'assignation du moniteur et état du quorum du moniteur
- Époque d'assignation des OSD et état des OSD
- Statut des instances Ceph Manager
- Statut des instances Object Gateway
- Version d'assignation des groupes de placement
- Nombre de groupes de placement et de réserves
- Quantité *théorique* de données stockées et nombre d'objets stockés
- Quantité totale de données stockées



Astuce : méthode utilisée par Ceph pour calculer l'utilisation des données

La valeur de used reflète la quantité réelle de stockage brut utilisée. La valeur de xxx Go/xxx Go désigne la quantité disponible de la capacité de stockage globale de la grappe (la quantité disponible correspond à la valeur inférieure). Le nombre théorique reflète la taille des données stockées avant qu'elles soient répliquées ou clonées ou qu'elles fassent l'objet d'un instantané. Par conséquent, la quantité de données réellement stockée dépasse généralement la quantité théorique stockée, car Ceph crée des répliques des données et peut également utiliser la capacité de stockage pour le clonage et la création d'instantanés.

Les autres commandes affichant des informations d'état immédiat sont les suivantes :

- ceph pg stat
- ceph osd pool stats

- ceph df
- ceph df detail

Pour obtenir les informations mises à jour en temps réel, utilisez l'une de ces commandes (y compris ceph -s) comme argument de la commande watch :

```
# watch -n 10 'ceph -s'
```

Appuyez sur `ctrl - c` pour refermer la sortie de la commande.

12.2 Vérification de l'état de santé de la grappe

Après avoir démarré votre grappe et avant de commencer à lire et/ou à écrire des données, vérifiez l'état d'intégrité de votre grappe :

```
cephuser@adm > ceph health
HEALTH_WARN 10 pgs degraded; 100 pgs stuck unclean; 1 mons down, quorum 0,2 \
node-1,node-2,node-3
```



Astuce

Si vous avez choisi des emplacements autres que ceux par défaut pour votre configuration ou votre trousseau de clés, vous pouvez les indiquer ici :

```
cephuser@adm > ceph -c /path/to/conf -k /path/to/keyring health
```

La grappe Ceph renvoie l'un des codes d'intégrité suivants :

OSD_DOWN

Un ou plusieurs OSD sont marqués comme étant arrêtés. Le daemon OSD peut avoir été arrêté ou les OSD homologues peuvent ne pas être en mesure d'accéder à l'OSD via le réseau. Un daemon arrêté ou bloqué, un hôte en panne ou une panne réseau font partie des causes les plus courantes de ce problème.

Vérifiez que l'hôte est intègre, que le daemon a été démarré et que le réseau fonctionne. Si le daemon est tombé en panne, le fichier journal du daemon (/var/log/ceph/ceph-osd.*) peut contenir des informations de débogage.

OSD_type *crush_DOWN*, par exemple, *OSD_HOST_DOWN*

Tous les OSD d'une sous-arborescence CRUSH particulière sont marqués comme étant arrêtés, par exemple tous les OSD d'un hôte.

OSD_ORPHAN

Un OSD est référencé dans la hiérarchie des cartes CRUSH, mais n'existe pas. L'OSD peut être retiré de la hiérarchie CRUSH avec :

```
cephuser@adm > ceph osd crush rm osd.ID
```

OSD_OUT_OF_ORDER_FULL

Les seuils d'utilisation pour *backfillfull* (par défaut : 0,90), *nearfull* (par défaut : 0,85), *full* (par défaut : 0,95) et/ou *failsafe_full* ne sont pas croissants. *backfillfull* < *nearfull*, *nearfull* < *full* et *full* < *failsafe_full* sont agencés dans cet ordre.

Pour lire les valeurs actuelles, exécutez la commande suivante :

```
cephuser@adm > ceph health detail
HEALTH_ERR 1 full osd(s); 1 backfillfull osd(s); 1 nearfull osd(s)
osd.3 is full at 97%
osd.4 is backfill full at 91%
osd.2 is near full at 87%
```

Les seuils peuvent être ajustés avec les commandes suivantes :

```
cephuser@adm > ceph osd set-backfillfull-ratio ratio
cephuser@adm > ceph osd set-nearfull-ratio ratio
cephuser@adm > ceph osd set-full-ratio ratio
```

OSD_FULL

Un ou plusieurs OSD ont dépassé le seuil *full* et empêchent la grappe de gérer les écritures. L'utilisation par réserve peut être vérifiée comme suit :

```
cephuser@adm > ceph df
```

Le ratio *full* actuellement défini peut être vu avec :

```
cephuser@adm > ceph osd dump | grep full_ratio
```

Une solution à court terme de restauration de la disponibilité en écriture consiste à augmenter légèrement le seuil *full* :

```
cephuser@adm > ceph osd set-full-ratio ratio
```

Ajoutez un nouveau stockage à la grappe en déployant plus d'OSD ou supprimez les données existantes afin de libérer de l'espace.

OSD_BACKFILLFULL

Un ou plusieurs OSD ont dépassé le seuil *backfillfull*, ce qui empêche le rééquilibrage des données sur ce périphérique. Cet avertissement anticipé indique que le rééquilibrage peut ne pas aboutir et que la grappe approche de sa pleine capacité. L'utilisation par réserve peut être vérifiée comme suit :

```
cephuser@adm > ceph df
```

OSD_NEARFULL

Un ou plusieurs OSD ont dépassé le seuil *nearfull*. Cet avertissement anticipé indique que la grappe approche de sa pleine capacité. L'utilisation par réserve peut être vérifiée comme suit :

```
cephuser@adm > ceph df
```

OSDMAP_FLAGS

Un ou plusieurs indicateurs de grappe présentant un intérêt ont été définis. À l'exception de *full*, ces indicateurs peuvent être définis ou effacés comme suit :

```
cephuser@adm > ceph osd set flag  
cephuser@adm > ceph osd unset flag
```

Ces indicateurs comprennent :

full

La grappe est marquée comme pleine et ne peut donc pas traiter les écritures.

pauserd, pausewr

Les lectures et les écritures sont mises en pause.

noup

Les OSD ne sont pas autorisés à démarrer.

nodown

Les rapports d'échec OSD sont ignorés de sorte que les moniteurs ne marquent pas les OSD comme *down* (arrêtés).

noin

Les OSD précédemment marqués comme *out* (hors service) ne sont pas marqués comme *in* (en service) lors de leur démarrage.

noout

Les OSD *down* (arrêtés) ne seront pas automatiquement marqués comme *out* (hors service) à l'issue de l'intervalle configuré.

nobackfill, norecover, norebalance

La récupération ou le rééquilibrage des données est suspendu.

noscrub, nodeep_scrub

Le nettoyage (reportez-vous à la [Section 17.6, « Nettoyage des groupes de placement »](#)) est désactivé.

notieragent

L'activité de hiérarchisation du cache est suspendue.

OSD_FLAGS

Un ou plusieurs OSD possèdent chacun un indicateur OSD. Ces indicateurs comprennent :

noup

L'OSD n'est pas autorisé à démarrer.

nodown

Les rapports d'échec sont ignorés pour cet OSD.

noin

Si cet OSD était auparavant marqué comme *out* automatiquement après un échec, il ne sera pas marqué comme *in* lors du démarrage.

noout

Si cet OSD est arrêté, il n'est pas automatiquement marqué comme *out* à l'issue de l'intervalle configuré.

Les indicateurs OSD peuvent être définis et effacés comme suit :

```
cephuser@adm > ceph osd add-flag osd-ID
cephuser@adm > ceph osd rm-flag osd-ID
```

OLD_CRUSH_TUNABLES

La carte CRUSH doit être mise à jour, car elle utilise des paramètres très anciens. Les paramètres les plus anciens qui peuvent être utilisés (c'est-à-dire la version client la plus ancienne pouvant se connecter à la grappe) sans déclencher cet avertissement d'intégrité sont déterminés par l'option de configuration mon_crush_min_required_version.

OLD_CRUSH_STRAW_CALC_VERSION

La carte CRUSH utilise une méthode non optimale plus ancienne afin de calculer les valeurs de pondération intermédiaires des compartiments straw. La carte CRUSH doit être mise à jour pour utiliser la nouvelle méthode (straw_calc_version=1).

CACHE_POOL_NO_HIT_SET

Une ou plusieurs réserves de cache ne sont pas configurées avec un jeu d'accès pour le suivi de l'utilisation, ce qui empêche l'agent de hiérarchisation d'identifier les objets inactifs à évincer du cache. Les jeux d'accès peuvent être configurés sur la réserve de cache comme suit :

```
cephuser@adm > ceph osd pool set poolname hit_set_type type
cephuser@adm > ceph osd pool set poolname hit_set_period period-in-seconds
cephuser@adm > ceph osd pool set poolname hit_set_count number-of-hitsets
cephuser@adm > ceph osd pool set poolname hit_set_fpp target-false-positive-rate
```

OSD_NO_SORTBITWISE

Aucun OSD d'une version antérieure à la version 12 « Luminous » ne s'exécute actuellement, mais l'indicateur `sortbitwise` n'est pas défini. Vous devez définir l'indicateur `sortbitwise` pour permettre le démarrage des OSD antérieurs à la version 12 « Luminous » ou plus récents :

```
cephuser@adm > ceph osd set sortbitwise
```

POOL_FULL

Une ou plusieurs réserves ont atteint leur quota et n'autorisent plus les écritures. Vous pouvez définir des quotas de réserve et leur utilisation comme suit :

```
cephuser@adm > ceph df detail
```

Vous pouvez augmenter le quota de réserve comme suit :

```
cephuser@adm > ceph osd pool set-quota poolname max_objects num-objects
cephuser@adm > ceph osd pool set-quota poolname max_bytes num-bytes
```

ou supprimer des données existantes pour réduire l'utilisation.

PG_AVAILABILITY

La disponibilité des données est réduite, ce qui signifie que la grappe ne peut pas traiter les requêtes de lecture ou d'écriture potentielles pour certaines de ses données. Plus précisément, un ou plusieurs groupes de placement se trouvent dans un état qui ne permet pas de traiter les requêtes d'E/S. Les états de groupe de placement posant problème sont les suivants : *peering* (homologation), *stale* (périmé), *incomplete* (incomplet) et l'absence d'état *active* (actif) (si ces conditions ne disparaissent pas rapidement). Pour plus de détails sur les groupes de placement affectés, exécutez la commande suivante :

```
cephuser@adm > ceph health detail
```

Dans la plupart des cas, la cause première est due au fait qu'un ou plusieurs OSD sont actuellement arrêtés. Pour connaître l'état des groupes de placement incriminés, exécutez la commande suivante :

```
cephuser@adm > ceph tell pgid query
```

PG_DEGRADED

La redondance des données est réduite pour certaines données, ce qui signifie que la grappe ne dispose pas du nombre de répliques souhaité pour toutes les données (pour les réserves répliquées) ou pour les fragments de code d'effacement (pour les réserves codées à effacement). Plus précisément, l'indicateur *degraded* (altéré) ou *undersized* (de taille insuffisante) est associé à un ou plusieurs groupes de placement (le nombre d'instances de ce groupe de placement est insuffisant dans la grappe) ou l'indicateur *clean* ne leur est pas associé depuis un certain temps. Pour plus de détails sur les groupes de placement affectés, exécutez la commande suivante :

```
cephuser@adm > ceph health detail
```

Dans la plupart des cas, la cause première est due au fait qu'un ou plusieurs OSD sont actuellement arrêtés. Pour connaître l'état des groupes de placement incriminés, exécutez la commande suivante :

```
cephuser@adm > ceph tell pgid query
```

PG_DEGRADED_FULL

La redondance des données peut être réduite ou menacée pour certaines données en raison d'un manque d'espace libre dans la grappe. Plus précisément, l'indicateur *backfill_toofull* ou *recovery_toofull* est associé à un ou plusieurs groupes de placement, ce qui signifie que la grappe ne parvient pas à migrer ou à récupérer les données, car un ou plusieurs OSD ont dépassé le seuil *backfillfull*.

PG_DAMAGED

Le nettoyage des données (voir [Section 17.6, « Nettoyage des groupes de placement »](#)) a détecté des problèmes de cohérence des données dans la grappe. Plus précisément, l'indicateur *inconsistent* ou *snaptrim_error* est associé à un ou plusieurs groupes de placement, ce qui indique qu'une opération de nettoyage antérieure a détecté un problème ou que l'indicateur *repair* est défini, car une réparation est actuellement en cours pour une telle incohérence.

OSD_SCRUB_ERRORS

Les nettoyages récents d'OSD ont révélé des incohérences.

CACHE_POOL_NEAR_FULL

Une réserve de niveau de cache est presque pleine. Dans ce contexte, l'état Full est déterminé par les propriétés *target_max_bytes* et *target_max_objects* de la réserve de cache. Lorsque la réserve atteint le seuil cible, les requêtes d'écriture dans la réserve peuvent être bloquées pendant que les données sont vidées et évincées du cache, un état qui entraîne généralement des latences très élevées et des performances médiocres. La taille cible de la réserve de cache peut être définie ainsi :

```
cephuser@adm > ceph osd pool set cache-pool-name target_max_bytes bytes  
cephuser@adm > ceph osd pool set cache-pool-name target_max_objects objects
```

L'activité normale de vidage et d'éviction du cache peut également être entravée en raison de la disponibilité ou des performances réduites du niveau de base ou de la charge globale de la grappe.

TOO_FEW_PGS

Le nombre de groupes de placement utilisés est inférieur au seuil configurable de *mon_pg_warn_min_per_osd* groupes de placement par OSD. Cela peut entraîner une distribution et un équilibrage sous-optimaux des données entre les OSD de la grappe, ce qui fait baisser les performances globales.

TOO_MANY_PGS

Le nombre de groupes de placement utilisés est supérieur au seuil configurable de *mon_pg_warn_max_per_osd* groupes de placement par OSD. Cela peut conduire à une utilisation plus importante de la mémoire pour les daemons OSD, à une homologation plus lente après des changements d'état de grappe (redémarrages, ajouts ou suppressions d'OSD, par exemple) et à une charge plus élevée des instances Ceph Manager et Ceph Monitor. Contrairement à la valeur *pg_num*, la valeur *pgp_num* peut être réduite pour des réserves existantes. Cela permet de regrouper efficacement certains groupes de placement sur les mêmes ensembles d'OSD, atténuant ainsi quelques-uns des effets négatifs décrits ci-dessus. Il est possible d'ajuster la valeur de *pgp_num* comme suit :

```
cephuser@adm > ceph osd pool set pool pgp_num value
```

SMALLER_PGP_NUM

La valeur de `pgp_num` est inférieure à celle de `pg_num` pour une ou plusieurs réserves. Ceci indique normalement que le nombre de groupes de placement a été augmenté indépendamment du comportement de placement. Ce problème est résolu en définissant `pgp_num` et `pg_num` sur la même valeur, ce qui déclenche la migration de données, comme suit :

```
cephuser@adm > ceph osd pool set pool pgp_num pg_num_value
```

MANY_OBJECTS_PER_PG

Pour une ou plusieurs réserves, le nombre moyen d'objets par groupe de placement est sensiblement supérieur à la moyenne globale de la grappe. Le seuil spécifique est contrôlé par la valeur de configuration de `mon_pg_warn_max_object_skew`. Cela indique généralement que la ou les réserves contenant la plupart des données de la grappe possèdent un nombre trop faible de groupes de placement et/ou que les autres réserves ne contenant pas autant de données possèdent un nombre excessif de groupes de placement. Pour ne plus afficher l'avertissement d'intégrité, vous pouvez relever le seuil en ajustant l'option de configuration `mon_pg_warn_max_object_skew` sur les moniteurs.

POOL_APP_NOT_ENABLED

Il existe une réserve qui contient un ou plusieurs objets, mais qui n'a pas été marquée pour une utilisation par une application particulière. Pour que cet avertissement ne s'affiche plus, étiquetez la réserve pour qu'elle soit utilisée par une application. Par exemple, si la réserve est utilisée par RBD :

```
cephuser@adm > rbd pool init pool_name
```

Si la réserve est utilisée par une application personnalisée « foo », vous pouvez également l'étiqueter à l'aide de la commande de bas niveau :

```
cephuser@adm > ceph osd pool application enable foo
```

POOL_FULL

Une ou plusieurs réserves ont atteint leur quota (ou sont proches des 100 %). Le seuil de déclenchement de cette condition d'erreur est contrôlé par l'option de configuration `mon_pool_quota_crit_threshold`. Les quotas de réserve peuvent être ajustés à la hausse ou à la baisse (voire supprimés) comme suit :

```
cephuser@adm > ceph osd pool set-quota pool max_bytes bytes  
cephuser@adm > ceph osd pool set-quota pool max_objects objects
```

Définir la valeur de quota sur 0 désactive le quota.

POOL_NEAR_FULL

Une ou plusieurs réserves se rapprochent de leur quota. Le seuil de déclenchement de cette condition d'avertissement est contrôlé par l'option de configuration `mon_pool_quota_warn_threshold`. Les quotas de réserve peuvent être ajustés à la hausse ou à la baisse (voire supprimés) comme suit :

```
cephuser@adm > ceph osd osd pool set-quota pool max_bytes bytes
cephuser@adm > ceph osd osd pool set-quota pool max_objects objects
```

Définir la valeur de quota sur 0 désactive le quota.

OBJECT_MISPLACED

Un ou plusieurs objets de la grappe ne sont pas stockés sur le noeud prévu par la grappe. Cela indique que la migration de données due à une modification récente de la grappe n'a pas encore abouti. Un mauvais placement des données n'est pas dangereux en soi. La cohérence des données n'est jamais compromise et les anciennes copies d'objets ne sont jamais supprimées tant que le nombre de nouvelles copies souhaité n'est pas atteint (dans les emplacements souhaités).

OBJECT_UNFOUND

Un ou plusieurs objets de la grappe sont introuvables. Plus précisément, les OSD savent qu'une copie nouvelle ou mise à jour d'un objet doit exister, mais la copie de cette version de l'objet est introuvable sur les OSD actuellement opérationnels. Les requêtes de lecture ou d'écriture sur les objets « introuvables » seront bloquées. Idéalement, l'OSD hors service qui héberge la copie la plus récente de l'objet introuvable peut être remis en service. Les OSD candidats peuvent être identifiés à partir de l'état d'homologation du ou des groupes de placement associés à l'objet introuvable :

```
cephuser@adm > ceph tell pgid query
```

REQUEST_SLOW

Une ou plusieurs requêtes OSD sont longues à traiter. Cela peut indiquer une charge extrême, un périphérique de stockage lent ou un bogue logiciel. Vous pouvez interroger la file d'attente des requêtes sur le ou les OSD en question, exécutez la commande suivante à partir de l'hôte OSD :

```
cephuser@adm > cephadm enter --name osd.ID -- ceph daemon osd.ID ops
```

Vous pouvez afficher le résumé des requêtes récentes les plus lentes :

```
cephuser@adm > cephadm enter --name osd.ID -- ceph daemon osd.ID dump_historic_ops
```


Vous pouvez trouver l'emplacement d'un OSD avec :

```
cephuser@adm > ceph osd find osd.id
```

REQUEST_STUCK

Une ou plusieurs requêtes d'OSD ont été bloquées pendant une période relativement longue, par exemple 4 096 secondes. Cela indique que l'état de santé de la grappe n'est pas bon depuis un certain temps (par exemple, en raison du faible nombre d'OSD actifs ou de groupes de placement inactifs), ou que l'OSD concernée présente un problème interne.

PG_NOT_SCRUBBED

Un ou plusieurs groupes de placement n'ont pas été nettoyés récemment (reportez-vous à la [Section 17.6, « Nettoyage des groupes de placement »](#)). Les groupes de placement sont normalement nettoyés toutes les `mon_scrub_interval` secondes ; cet avertissement se déclenche lorsque `mon_warn_not_scrubbed` intervalles se sont écoulés sans nettoyage. Les groupes de placement ne sont pas nettoyés s'ils ne sont pas marqués comme propres, ce qui peut arriver s'ils sont mal placés ou altérés (voir PG_AVAILABILITY et PG_DEGRADED ci-dessus). Vous pouvez lancer manuellement le nettoyage d'un groupe de placement :

```
cephuser@adm > ceph pg scrub pgid
```

PG_NOT_DEEP_SCRUBBED

Un ou plusieurs groupes de placement n'ont pas été nettoyés en profondeur récemment (reportez-vous à la [Section 17.6, « Nettoyage des groupes de placement »](#)). Les groupes de placement sont normalement nettoyés toutes les `osd_deep_mon_scrub_interval` secondes ; cet avertissement se déclenche lorsque `mon_warn_not_deep_scrubbed` secondes se sont écoulées sans nettoyage. Les groupes de placement n'ont pas été nettoyés (en profondeur) s'ils ne sont pas marqués comme propres, ce qui peut arriver s'ils sont mal placés ou altérés (voir PG_AVAILABILITY et PG_DEGRADED ci-dessus). Vous pouvez lancer manuellement le nettoyage d'un groupe de placement :

```
cephuser@adm > ceph pg deep-scrub pgid
```



Astuce

Si vous avez choisi des emplacements autres que ceux par défaut pour votre configuration ou votre trousseau de clés, vous pouvez les indiquer ici :

```
# ceph -c /path/to/conf -k /path/to/keyring health
```

12.3 Vérification des statistiques d'utilisation d'une grappe

Pour vérifier l'utilisation des données d'une grappe et leur distribution entre les réserves, utilisez la commande `ceph df`. Pour obtenir plus de détails, utilisez `ceph df detail`.

```
cephuser@adm > ceph df
--- RAW STORAGE ---
CLASS  SIZE      AVAIL    USED      RAW USED  %RAW USED
hdd    30 GiB    27 GiB   121 MiB   3.1 GiB    10.40
TOTAL  30 GiB    27 GiB   121 MiB   3.1 GiB    10.40

--- POOLS ---
POOL                                ID  STORED  OBJECTS  USED      %USED  MAX AVAIL
device_health_metrics              1      0 B        0      0 B        0    8.5 GiB
cephfs.my_cephfs.meta              2    1.0 MiB     22    4.5 MiB    0.02    8.5 GiB
cephfs.my_cephfs.data              3      0 B        0      0 B        0    8.5 GiB
.rgw.root                          4    1.9 KiB     13    2.2 MiB    0       8.5 GiB
myzone.rgw.log                     5    3.4 KiB    207     6 MiB    0.02    8.5 GiB
myzone.rgw.control                 6      0 B        8      0 B        0    8.5 GiB
myzone.rgw.meta                    7      0 B        0      0 B        0    8.5 GiB
```

La section `RAW STORAGE` de la sortie donne un aperçu de la quantité de stockage utilisée pour vos données par la grappe.

- `CLASS` : classe de stockage du périphérique. Reportez-vous à la [Section 17.1.1, « Classes de périphériques »](#) pour plus d'informations sur les classes de périphériques.
- `SIZE` : capacité de stockage globale de la grappe.
- `AVAIL` : quantité d'espace disponible dans la grappe.
- `USED` : espace (accumulé sur tous les OSD) alloué uniquement pour les objets de données conservés sur le périphérique de bloc.
- `RAW USED` : somme de l'espace « `USED` » et de l'espace alloué/réservé au niveau du périphérique de bloc pour Ceph, par exemple la partie blueFS pour BlueStore.
- `% RAW USED` : pourcentage de stockage brut utilisé. Utilisez ce nombre avec le ratio `full` et le ratio `near full` pour vous assurer que vous n'atteignez pas la capacité de votre grappe. Reportez-vous à la [Section 12.8, « Capacité de stockage »](#) pour plus de détails.



Note : niveau de remplissage de grappe

Lorsqu'un niveau de remplissage de stockage brut se rapproche de 100 %, vous devez ajouter un nouveau stockage à la grappe. Une utilisation plus élevée peut conduire à la saturation de certains OSD et à des problèmes d'intégrité de la grappe.

Utilisez la commande `ceph osd df tree` pour établir la liste de niveau de remplissage de tous les OSD.

La section POOLS de la sortie fournit la liste des réserves et l'utilisation théorique de chaque réserve. La sortie de cette section ne reflète *pas* les répliques, les clones ou les instantanés existants. Par exemple, si vous stockez un objet de 1 Mo de données, l'utilisation théorique est de 1 Mo, mais l'utilisation réelle peut être de 2 Mo ou plus selon le nombre de répliques, de clones et d'instantanés.

- P00L : nom de la réserve.
- ID : identifiant de la réserve.
- STORED : quantité de données stockées par l'utilisateur.
- OBJECTS : nombre théorique d'objets stockés par réserve.
- USED : quantité d'espace allouée exclusivement aux données par tous les noeuds OSD (en ko).
- % USED : pourcentage de stockage théorique utilisé par réserve.
- MAX AVAIL : espace maximal disponible dans la réserve indiquée.



Note

Les nombres figurant dans la section POOLS sont théoriques. Ils n'incluent pas le nombre de répliques, d'instantanés ou de clones. Par conséquent, la somme des montants USED et %USED ne correspond pas aux montants RAW USED et %RAW USED dans la section RAW STORAGE de la sortie.

12.4 Vérification de l'état des OSD

Vous pouvez vérifier les OSD pour vous assurer qu'ils sont opérationnels et activés à l'aide de la commande suivante :

```
cephuser@adm > ceph osd stat
```

ou

```
cephuser@adm > ceph osd dump
```

Vous pouvez également afficher les OSD en fonction de leur position dans la carte CRUSH.

ceph osd tree permet d'afficher une arborescence CRUSH avec un hôte, ses OSD, leur état et leur pondération :

```
cephuser@adm > ceph osd tree
```

| ID | CLASS | WEIGHT | TYPE NAME | STATUS | REWEIGHT | PRI-AFF |
|----|-------|---------|---------------|--------|----------|---------|
| -1 | 3 | 0.02939 | root default | | | |
| -3 | 3 | 0.00980 | rack mainrack | | | |
| -2 | 3 | 0.00980 | host osd-host | | | |
| 0 | 1 | 0.00980 | osd.0 | up | 1.00000 | 1.00000 |
| 1 | 1 | 0.00980 | osd.1 | up | 1.00000 | 1.00000 |
| 2 | 1 | 0.00980 | osd.2 | up | 1.00000 | 1.00000 |

12.5 Contrôle des OSD pleins

Ceph vous empêche d'écrire sur un OSD plein afin de vous éviter de perdre des données. Pour une grappe opérationnelle, un message d'avertissement doit s'afficher lorsque celle-ci est sur le point d'atteindre son ratio complet. La valeur par défaut de **monosd full ratio** est 0.95, c'est-à-dire 95 % de la capacité au-delà de laquelle les clients ne peuvent plus écrire de données dans la grappe. La valeur par défaut de **monosd nearfull ratio** est de 0.85, c'est-à-dire 85 % de la capacité à partir de laquelle un message d'avertissement d'intégrité est émis.

Les noeuds OSD pleins sont signalés par la commande **ceph health** :

```
cephuser@adm > ceph health
HEALTH_WARN 1 nearfull osds
osd.2 is near full at 85%
```

ou

```
cephuser@adm > ceph health
```

```
HEALTH_ERR 1 nearfull osds, 1 full osds
osd.2 is near full at 85%
osd.3 is full at 97%
```

La meilleure façon de gérer une grappe pleine consiste à ajouter de nouveaux hôtes/disques OSD permettant à la grappe de redistribuer les données à l'espace de stockage récemment disponible.



Astuce : prévention de la saturation des OSD

Un OSD plein utilise 100 % de son espace disque. Lorsqu'il atteint ce taux de remplissage, l'OSD se bloque sans avertissement. Voici quelques conseils à retenir lors de l'administration de noeuds OSD.

- L'espace disque de chaque OSD (généralement monté sous `/var/lib/ceph/osd/osd-{1,2...}`) doit être placé sur une partition ou un disque sous-jacent dédié.
- Vérifiez les fichiers de configuration Ceph et assurez-vous que Ceph ne stocke pas son fichier journal sur les partitions/disques dédiés aux OSD.
- Assurez-vous qu'aucun autre processus n'écrit sur les partitions/disques dédiés aux OSD.

12.6 Vérification de l'état des instances Monitor

Après avoir démarré la grappe et avant la première lecture et/ou écriture de données, vérifiez le statut du quorum des instances Ceph Monitor. Lorsque la grappe dessert déjà des requêtes, vérifiez périodiquement le statut des instances Ceph Monitor pour vous assurer qu'elles sont en cours d'exécution.

Pour afficher l'assignation des moniteurs, exécutez la commande suivante :

```
cephuser@adm > ceph mon stat
```

ou

```
cephuser@adm > ceph mon dump
```

Pour contrôler l'état du quorum de la grappe de moniteurs, exécutez la commande suivante :

```
cephuser@adm > ceph quorum_status
```

Ceph renvoie l'état du quorum. Par exemple, une grappe Ceph composée de trois moniteurs peut renvoyer les éléments suivants :

```
{ "election_epoch": 10,
  "quorum": [
    0,
    1,
    2],
  "monmap": { "epoch": 1,
    "fsid": "444b489c-4f16-4b75-83f0-cb8097468898",
    "modified": "2011-12-12 13:28:27.505520",
    "created": "2011-12-12 13:28:27.505520",
    "mons": [
      { "rank": 0,
        "name": "a",
        "addr": "192.168.1.10:6789\0"},
      { "rank": 1,
        "name": "b",
        "addr": "192.168.1.11:6789\0"},
      { "rank": 2,
        "name": "c",
        "addr": "192.168.1.12:6789\0"}
    ]
  }
}
```

12.7 Vérification des états des groupes de placement

Les groupes de placement assignent des objets aux OSD. Lorsque vous surveillez vos groupes de placement, vous voulez qu'ils soient actifs (active) et propres (clean). Pour une explication détaillée, reportez-vous à la [Section 12.9, « Surveillance des OSD et des groupes de placement »](#).

12.8 Capacité de stockage

Lorsqu'une grappe de stockage Ceph se rapproche de sa capacité maximale, Ceph vous empêche d'écrire ou de lire à partir d'OSD Ceph afin d'éviter toute perte de données. Par conséquent, laisser une grappe de production s'approcher de son ration « full » n'est pas une bonne pratique, car cela nuit au principe de haute disponibilité. Le ratio « full » par défaut est défini sur 0,95, soit 95 % de la capacité. C'est une valeur très agressive pour une grappe de test avec un petit nombre d'OSD.



Astuce : augmentation de la capacité de stockage

Lorsque vous surveillez votre grappe, soyez attentif aux avertissements liés au ratio near-full. Cela signifie qu'une défaillance de certains OSD pourrait entraîner une interruption de service temporaire en cas d'échec d'un ou de plusieurs OSD. Pensez à ajouter d'autres OSD pour augmenter la capacité de stockage.

Un scénario courant pour les grappes de test implique qu'un administrateur système supprime un OSD Ceph de la grappe de stockage Ceph pour examiner le rééquilibrage de la grappe. Il supprime ensuite un autre OSD Ceph, puis encore un autre, et ainsi de suite jusqu'à ce que la grappe atteigne le ratio « full » et se verrouille. Nous recommandons un minimum de planification de la capacité, même avec une grappe de test. La planification vous permet d'estimer la capacité de réserve dont vous avez besoin pour maintenir une haute disponibilité. Idéalement, vous souhaitez envisager une série de défaillances Ceph OSD où la grappe peut récupérer un état actif et propre (active + clean) sans remplacer ces Ceph OSD immédiatement. Vous pouvez exécuter une grappe dans un état actif et altéré (active + degraded), mais ce n'est pas idéal pour des conditions de fonctionnement normales.

Le diagramme suivant représente une grappe de stockage Ceph simpliste contenant 33 noeuds Ceph avec un Ceph OSD par hôte, chacun d'eux disposant d'un accès en lecture-écriture pour une unité de 3 To. Cette grappe présente une capacité réelle maximale de 99 To. L'option mon osd full ratio est définie sur 0,95. Si la grappe arrive à 5 To de la capacité restante, elle ne permettra plus aux clients de lire et d'écrire des données. Par conséquent, la capacité d'exploitation de la grappe de stockage est de 95 To, et pas de 99.

| Rack 1 | Rack 2 | Rack 3 | Rack 4 | Rack 5 | Rack 6 |
|--------|--------|--------|--------|--------|---------|
| OSD 1 | OSD 7 | OSD 13 | OSD 19 | OSD 25 | OSD 31 |
| OSD 2 | OSD 8 | OSD 14 | OSD 20 | OSD 26 | OSD 32 |
| OSD 3 | OSD 9 | OSD 15 | OSD 21 | OSD 27 | OSD 33 |
| OSD 4 | OSD 10 | OSD 16 | OSD 22 | OSD 28 | Réserve |
| OSD 5 | OSD 11 | OSD 17 | OSD 23 | OSD 29 | Réserve |
| OSD 6 | OSD 12 | OSD 18 | OSD 24 | OSD 30 | Réserve |

FIGURE 12.1 : GRAPPE CEPH

Il est normal dans une telle grappe qu'un ou deux OSD échouent. Un scénario moins fréquent, mais plausible, implique une défaillance du routeur ou de l'alimentation d'un rack, ce qui met hors service plusieurs OSD simultanément (par exemple, les OSD 7 à 12). Dans un tel scénario, vous devez toujours viser à disposer d'une grappe qui peut rester opérationnelle et atteindre un état `active + clean`, même si cela implique d'ajouter sans délai plusieurs hôtes avec des OSD supplémentaires. Si l'utilisation de votre capacité est trop élevée, vous ne pouvez pas risquer de perdre des données. Cela dit, vous pourriez encore sacrifier la disponibilité des données pendant la résolution d'une panne au sein d'un domaine défaillant si l'utilisation de la capacité de la grappe dépasse le ratio « full ». C'est pour cette raison que nous recommandons au moins une certaine planification approximative de la capacité.

Identifiez deux nombres pour votre grappe :

1. Le nombre d'OSD.
2. La capacité totale de la grappe.

Si vous divisez la capacité totale de votre grappe par le nombre d'OSD que celle-ci contient, vous obtenez la capacité moyenne d'un OSD au sein de votre grappe. Pensez à multiplier cette valeur par le nombre d'OSD qui, selon vous, pourraient échouer simultanément lors d'opérations normales (un nombre relativement faible). Enfin, multipliez la capacité de la grappe par le ratio « full » pour arriver à une capacité d'exploitation maximale. Ensuite, soustrayez la quantité de données des OSD susceptibles d'échouer pour obtenir un ratio « full » raisonnable. Répétez ce processus avec un nombre plus élevé de défaillances OSD (un rack d'OSD) afin d'obtenir un nombre raisonnable pour un ratio « nearfull ».

Les paramètres suivants ne s'appliquent qu'à la création d'une grappe et sont ensuite stockés dans la carte OSD :

```
[global]
mon osd full ratio = .80
mon osd backfillfull ratio = .75
mon osd nearfull ratio = .70
```



Astuce

Ces paramètres ne s'appliquent que lors de la création d'une grappe. Par la suite, ils doivent être modifiés dans la carte OSD à l'aide des commandes `ceph osd set-near-full-ratio` et `ceph osd set-full-ratio`.

mon osd full ratio

Pourcentage d'espace disque utilisé avant qu'un OSD soit considéré comme complet (full). La valeur par défaut est 0,95.

mon osd backfillfull ratio

Pourcentage d'espace disque utilisé avant qu'un OSD soit considéré comme trop rempli (full) pour un renvoi. La valeur par défaut est 0,90.

mon osd nearfull ratio

Pourcentage d'espace disque utilisé avant qu'un OSD soit considéré comme presque complet (nearfull). La valeur par défaut est 0,85.



Astuce : vérification de la pondération des OSD

Si certains OSD sont presque complets (nearfull), mais que d'autres ont beaucoup de capacité, la pondération CRUSH peut être problématique pour les OSD nearfull.

12.9 Surveillance des OSD et des groupes de placement

Les principes de haute disponibilité et de haute fiabilité exigent une approche de tolérance aux pannes dans le cadre de la gestion des problèmes matériels et logiciels. Ceph n'a pas de point d'échec unique et peut desservir les requêtes de données en mode altéré (« degraded »). Le placement de données de Ceph introduit une couche d'indirection pour s'assurer que les données ne se lient pas directement à des adresses OSD particulières. Cela signifie que le suivi des pannes du système nécessite de trouver le groupe de placement et les OSD sous-jacents à l'origine du problème.



Astuce : accès en cas de défaillance

Une panne dans une partie de la grappe peut vous empêcher d'accéder à un objet particulier. Cela ne signifie pas que vous ne pouvez pas accéder à d'autres objets. Lorsque vous rencontrez une panne, suivez les étapes de surveillance de vos OSD et groupes de placement. Commencez ensuite le dépannage.

Ceph peut généralement se réparer lui-même. Toutefois, lorsque des problèmes persistent, la surveillance des OSD et des groupes de placement vous aide à identifier ce qui ne va pas.

12.9.1 Surveillance des OSD

Le statut d'un OSD est soit *dans la grappe* (« rentré », c'est-à-dire « in » en anglais), soit *hors de la grappe* (« sorti » - « out »). Par ailleurs, il est soit *opérationnel et en cours d'exécution* (« opérationnel » - « up »), soit *arrêté et pas en cours d'exécution* (« arrêté » - « down »). Si un OSD est « opérationnel », il peut être dans la grappe (vous pouvez lire et écrire des données) ou hors de celle-ci. S'il était dans la grappe et a été sorti récemment de cette dernière, Ceph migre les groupes de placement vers d'autres OSD. Si un OSD est hors de la grappe, CRUSH ne lui assigne pas de groupes de placement. Si un OSD est « arrêté », il doit également être « sorti ».



Note : état altéré

Si un OSD est « arrêté » et « rentré », il y a un problème et la grappe présente un état altéré.

Si vous exécutez une commande telle que `ceph health`, `ceph -s` ou `ceph -w`, vous pouvez remarquer que la grappe ne renvoie pas toujours la valeur `HEALTH OK` (État de santé OK). En ce qui concerne les OSD, vous devez vous attendre à ce que la grappe ne renvoie *pas* la valeur `HEALTH OK` dans les circonstances suivantes :

- Vous n'avez pas encore démarré la grappe (elle ne répondra pas).
- Vous avez démarré ou redémarré la grappe et elle n'est pas encore prête, car les groupes de placement sont en cours de création et les OSD, en cours d'homologation.
- Vous avez ajouté ou supprimé un OSD.
- Vous avez modifié votre assignation de grappe.

Un aspect important de la surveillance des OSD consiste à s'assurer que lorsque la grappe est opérationnelle et en cours d'exécution, tous ses OSD le sont aussi. Pour vérifier si tous les OSD sont en cours d'exécution, utilisez la commande suivante :

```
# ceph osd stat
x osds: y up, z in; epoch: eNNNN
```

Le résultat devrait vous indiquer le nombre total d'OSD (x), combien sont « opérationnels » (y), combien sont « rentrés » (z), et l'époque d'assignation (eNNNN). Si le nombre d'OSD « rentrés » dans la grappe est supérieur au nombre d'OSD qui sont « opérationnels », exécutez la commande suivante pour identifier les daemons `ceph-osd` qui ne sont pas en cours d'exécution :

```
# ceph osd tree
#ID CLASS WEIGHT  TYPE NAME                STATUS REWEIGHT PRI-AFF
-1          2.000000 pool openstack
-3          2.000000 rack dell-2950-rack-A
-2          2.000000 host dell-2950-A1
0  ssd 1.000000    osd.0                up  1.000000 1.000000
1  ssd 1.000000    osd.1                down 1.000000 1.000000
```

Par exemple, si un OSD avec l'ID 1 est arrêté, démarrez-le :

```
cephuser@osd > sudo systemctl start ceph-CLUSTER_ID@osd.0.service
```

Reportez-vous au *Manuel « Troubleshooting Guide », Chapitre 4 « Troubleshooting OSDs », Section 4.3 « OSDs not running »* pour obtenir des informations sur les problèmes associés aux OSD qui se sont arrêtés ou qui ne redémarrent pas.

12.9.2 Assignation d'ensembles de groupes de placement

Lorsque CRUSH assigne des groupes de placement aux OSD, il examine le nombre de répliques pour la réserve et attribue le groupe de placement aux OSD de telle sorte que chaque réplique du groupe de placement soit assignée à un OSD différent. Par exemple, si la réserve nécessite trois répliques d'un groupe de placement, CRUSH peut les assigner à `osd.1`, `osd.2` et `osd.3` respectivement. CRUSH vise en fait un placement pseudo-aléatoire qui tiendra compte des domaines de défaillance que vous avez définis dans votre carte CRUSH, de sorte que vous verrez rarement des groupes de placement assignés à des OSD voisins les plus proches dans une vaste grappe. Nous nous référons à l'ensemble d'OSD qui devraient contenir les répliques d'un groupe de placement

particulier en tant qu'*ensemble agissant*. Dans certains cas, un OSD de l'ensemble agissant est arrêté ou n'est pas en mesure de desservir les requêtes d'objets dans le groupe de placement pour l'une ou l'autre raison. Ces situations peuvent se présenter dans l'un des scénarios suivants :

- Vous avez ajouté ou supprimé un OSD. Ensuite, CRUSH a réassigné le groupe de placement à d'autres OSD et a donc modifié la composition de l'*ensemble agissant*, provoquant la migration de données avec un processus de renvoi (« backfill »).
- Un OSD était « arrêté », a été redémarré et est maintenant en cours de récupération.
- Un OSD de l'*ensemble agissant* est « arrêté » ou n'est pas en mesure de desservir les requêtes, et un autre OSD a assumé temporairement ses fonctions.

Ceph traite une requête client à l'aide de l'*ensemble opérationnel*, qui est l'ensemble d'OSD qui traitera réellement les requêtes. Dans la plupart des cas, l'*ensemble opérationnel* et l'*ensemble agissant* sont pratiquement identiques. Lorsqu'ils ne le sont pas, cela peut indiquer que Ceph migre des données, qu'un OSD est en cours de récupération ou qu'il existe un problème (par exemple, dans de tels scénarios, Ceph renvoie habituellement un état HEALTH WARN avec un message « stuck stale » [obsolète bloqué]).

Pour récupérer une liste des groupes de placement, exécutez la commande suivante :

```
cephuser@adm > ceph pg dump
```

Pour voir quels OSD se trouvent dans l'*ensemble agissant* ou dans l'*ensemble opérationnel* pour un groupe de placement donné, exécutez la commande suivante :

```
cephuser@adm > ceph pg map PG_NUM  
osdmap eNNN pg RAW_PG_NUM (PG_NUM) -> up [0,1,2] acting [0,1,2]
```

Le résultat devrait vous indiquer l'époque osdmap (eNNN), le nombre de groupes de placement (PG_NUM), les OSD dans l'*ensemble opérationnel* (« up ») et les OSD dans l'*ensemble agissant* (« acting ») :



Astuce : indicateur de problème de grappe

Si l'*ensemble opérationnel* et l'*ensemble agissant* ne correspondent pas, cela peut indiquer que la grappe est occupée à se rééquilibrer ou qu'elle rencontre un problème potentiel.

12.9.3 Homologation

Avant de pouvoir écrire des données dans un groupe de placement, celui-ci doit présenter un état actif (« active ») et propre (« clean »). Pour que Ceph puisse déterminer l'état actuel d'un groupe de placement, l'OSD primaire du groupe de placement (le premier OSD dans l'*ensemble agissant*) effectue une homologation avec les OSD secondaires et tertiaires pour établir un accord sur l'état actuel du groupe de placement (dans le cas d'une réserve avec trois répliques du groupe de placement).

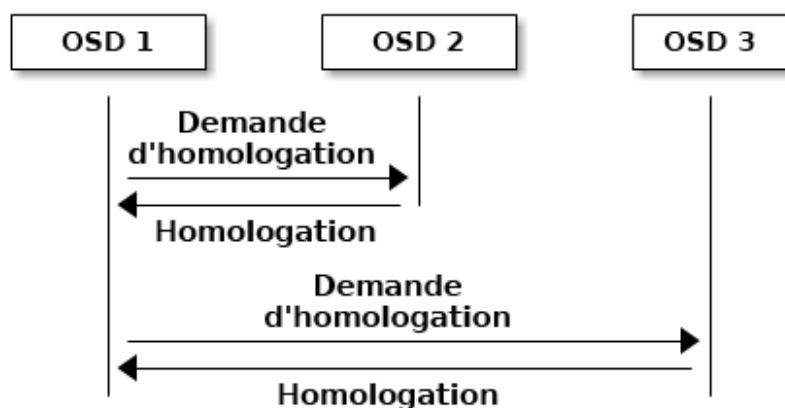


FIGURE 12.2 : SCHÉMA D'HOMOLOGATION

12.9.4 Surveillance des états des groupes de placement

Si vous exécutez une commande telle que `ceph health`, `ceph -s` ou `ceph -w`, vous remarquerez peut-être que la grappe ne renvoie pas toujours un message `HEALTH OK` (État de santé OK). Après avoir vérifié si les OSD sont en cours d'exécution, vous devez également vérifier l'état des groupes de placement.

Attendez-vous à ce que la grappe ne renvoie **pas** un message `HEALTH OK` dans un certain nombre de circonstances liées aux homologations des groupes de placement :

- Vous avez créé une réserve et les groupes de placement n'ont pas encore effectué l'homologation.
- Les groupes de placement sont en cours de récupération.
- Vous avez ajouté ou supprimé un OSD au niveau de la grappe.

- Vous avez modifié votre carte CRUSH et vos groupes de placement sont en cours de migration.
- Les données sont incohérentes entre différentes répliques d'un groupe de placement.
- Ceph est en train de nettoyer les répliques d'un groupe de placement.
- Ceph ne dispose pas de suffisamment de capacité de stockage pour effectuer des opérations de renvoi.

Si en raison de l'une des circonstances mentionnées ci-dessus, Ceph renvoie un message `HEALTH WARN`, ne paniquez pas. Dans de nombreux cas, la grappe se rétablit d'elle-même. Parfois, il peut cependant arriver que vous deviez intervenir. Un aspect important de la surveillance des groupes de placement consiste à vous assurer que lorsque la grappe est opérationnelle et en cours d'exécution, tous les groupes de placement sont actifs et, de préférence, dans un état propre. Pour voir l'état de tous les groupes de placement, exécutez la commande suivante :

```
cephuser@adm > ceph pg stat
x pgs: y active+clean; z bytes data, aa MB used, bb GB / cc GB avail
```

Le résultat devrait vous indiquer le nombre total de groupes de placement (x), le nombre de groupes de placement dans un état particulier tel que le « active + clean » (y) et la quantité de données stockées (z).

En plus des états des groupes de placement, Ceph renverra également la quantité de capacité de stockage utilisée (aa), la quantité de capacité de stockage restante (bb) et la capacité de stockage totale pour le groupe de placement. Ces chiffres peuvent être importants dans certains cas :

- Vous vous approchez de votre ratio `nearfull` (presque complet) ou `full` (complet).
- Vos données ne sont pas distribuées dans l'ensemble de la grappe en raison d'une erreur dans votre configuration CRUSH.



Astuce : ID de groupe de placement

Les ID de groupe de placement se composent du numéro de la réserve (pas de son nom) suivi d'un point (.) et de l'ID du groupe de placement (un nombre hexadécimal). Vous pouvez consulter les numéros des réserves et leur nom dans la sortie de la commande `ceph osd lspools`. Par exemple, la réserve par défaut `rbd` correspond au numéro de réserve 0. Un ID de groupe de placement complet se présente sous la forme suivante :

```
POOL_NUM.PG_ID
```

Il ressemble généralement à ceci :

```
0.1f
```

Pour récupérer une liste des groupes de placement, exécutez la commande suivante :

```
cephuser@adm > ceph pg dump
```

Vous pouvez également mettre la sortie au format JSON et l'enregistrer dans un fichier :

```
cephuser@adm > ceph pg dump -o FILE_NAME --format=json
```

Pour interroger un groupe de placement particulier, exécutez la commande suivante :

```
cephuser@adm > ceph pg POOL_NUM.PG_ID query
```

La liste suivante décrit en détail les statuts courants des groupes de placement.

CREATING

Lorsque vous créez une réserve, celle-ci crée le nombre de groupes de placement que vous avez spécifié. Ceph renvoie le statut « creating » (en cours de création) lorsqu'un ou plusieurs groupes de placement sont en cours de création. Une fois les groupes créés, les OSD qui font partie de l'*ensemble agissant* du groupe de placement effectuent l'homologation. Lorsque l'homologation est terminée, le statut du groupe de placement doit être « active + clean », ce qui signifie qu'un client Ceph peut commencer à écrire dans le groupe de placement.

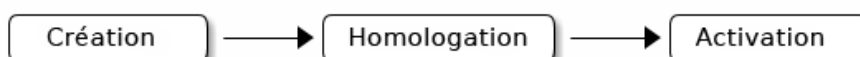


FIGURE 12.3 : ÉTAT DES GROUPES DE PLACEMENT

PEERING

Lorsque Ceph effectue l'homologation (« peering ») d'un groupe de placement, il amène les OSD qui stockent les répliques du groupe de placement à un accord concernant l'état des objets et des métadonnées que ce dernier contient. Lorsque Ceph termine l'homologation,

cela signifie que les OSD qui stockent le groupe de placement s'accordent sur l'état actuel du groupe de placement. Toutefois, l'achèvement du processus d'homologation ne signifie **pas** que chaque réplique dispose du contenu le plus récent.



Note : historique faisant autorité

Ceph ne reconnaîtra une opération d'écriture pour un client **que** lorsque tous les OSD de l'*ensemble agissant* conserveront l'opération d'écriture. Cette pratique garantit qu'au moins un membre de l'*ensemble agissant* aura un enregistrement de chaque opération d'écriture reconnue depuis la dernière opération d'homologation réussie. Avec un enregistrement précis de chaque opération d'écriture reconnue, Ceph peut construire et développer un nouvel historique faisant autorité pour le groupe de placement, à savoir un ensemble complet et organisé d'opérations qui, si elles sont effectuées, permettent de mettre à jour la copie d'un OSD de groupe de placement.

ACTIVE

Une fois que Ceph a terminé le processus d'homologation, un groupe de placement peut devenir actif (« active »). L'état « active » signifie que les données du groupe de placement sont généralement disponibles dans le groupe de placement primaire et les répliques pour les opérations de lecture et d'écriture.

CLEAN

Lorsqu'un groupe de placement est dans l'état propre (« clean »), cela signifie que l'OSD primaire et les OSD des répliques ont été homologués, et qu'il n'y a pas de répliques errantes pour le groupe de placement. Ceph a effectué le bon nombre de répliques de tous les objets du groupe de placement.

DEGRADED

Lorsqu'un client écrit un objet sur l'OSD primaire, ce dernier est responsable de l'écriture des répliques sur les OSD des répliques. Une fois que l'OSD primaire a écrit l'objet dans le stockage, le groupe de placement reste dans un état altéré (« degraded ») jusqu'à ce que l'OSD primaire ait reçu une confirmation des OSD des répliques indiquant que Ceph a bien créé les objets des répliques.

La raison pour laquelle un groupe de placement peut être à la fois actif et altéré (« active + degraded ») est qu'un OSD peut être actif même s'il ne détient pas encore tous les objets. Si un OSD s'arrête, Ceph marque chaque groupe de placement assigné à l'OSD comme

altéré. Lorsque l'OSD en question redevient opérationnel, tous les OSD doivent à nouveau effectuer l'homologation. Toutefois, un client peut toujours écrire un nouvel objet sur un groupe de placement altéré s'il est actif.

Si un OSD est arrêté et que la condition d'altération persiste, Ceph peut marquer l'OSD arrêté comme étant sorti (« out ») de la grappe et réassigne les données de l'OSD arrêté à un autre OSD. Le délai entre le moment où un OSD est marqué comme arrêté et celui où il est considéré comme sorti de la grappe est déterminé par l'option `mon osd down out interval`, qui est définie par défaut sur 600 secondes.

Un groupe de placement peut également être marqué comme altéré parce que Ceph ne parvient pas à trouver un ou plusieurs objets qui devraient être dans le groupe de placement. Bien que vous ne puissiez pas lire ou écrire sur des objets introuvables, vous pouvez toujours accéder à tous les autres objets du groupe de placement altéré.

RECOVERING

Ceph a été conçu pour permettre une tolérance aux pannes dans un environnement où les problèmes matériels et logiciels sont permanents. Lorsqu'un OSD est arrêté, son contenu peut devenir obsolète par rapport à l'état actuel d'autres répliques dans les groupes de placement. Lorsque l'OSD redevient opérationnel, le contenu des groupes de placement doit alors être mis à jour pour refléter l'état actuel. Durant cette opération, l'état de l'OSD peut indiquer qu'il est en cours de récupération (« recovering »).

La récupération n'est pas toujours simple, car une défaillance matérielle peut provoquer une défaillance en cascade de plusieurs OSD. Par exemple, un commutateur réseau pour un rack ou une armoire peut tomber en panne, de sorte les OSD de plusieurs machines hôtes se retrouvent dans un état obsolète par rapport à celui de la grappe. Chacun de ces OSD doit alors récupérer une fois la panne résolue.

Ceph fournit un certain nombre de paramètres pour équilibrer les conflits de ressources entre les nouvelles requêtes de service et la nécessité de récupérer les objets de données et de restaurer les groupes de placement vers l'état actuel. Le paramètre `osd recovery delay start` permet à un OSD de redémarrer, d'effectuer un nouveau processus d'homologation et même de traiter certaines demandes de relecture avant d'entamer sa récupération. Le paramètre `osd recovery thread timeout` définit un timeout de thread, car plusieurs OSD peuvent échouer, redémarrer et effectuer une nouvelle homologation à des moments différents. Le paramètre `osd recovery max active` limite le nombre de demandes de récupération qu'un OSD traite simultanément afin d'éviter qu'il échoue. Le paramètre `osd recovery max chunk` limite la taille des tranches de données récupérées pour éviter la congestion du réseau.

BACK FILLING

Lorsqu'un nouvel OSD rejoint la grappe, CRUSH réassigne des groupes de placement des OSD de la grappe à l'OSD récemment ajouté. Obliger le nouvel OSD à accepter immédiatement les groupes de placement réassignés peut lui imposer une charge excessive. Le principe de renvoi (« backfilling ») de l'OSD avec les groupes de placement permet à ce processus de commencer en arrière-plan. Une fois le renvoi terminé, le nouvel OSD commence à desservir les requêtes lorsqu'il est prêt.

Pendant les opérations de renvoi, vous pouvez voir l'un des états suivants : « `backfill_wait` » indique qu'une opération de renvoi est en attente, mais pas encore en cours ; « `backfill` » indique qu'une opération de renvoi est en cours ; « `backfill_too_full` » indique qu'une opération de renvoi a été demandée, mais n'a pas pu être effectuée en raison d'une capacité de stockage insuffisante. Lorsqu'un groupe de placement ne peut pas être renvoyé, il peut être considéré comme incomplet (« incomplete »).

Ceph fournit plusieurs paramètres pour gérer la charge associée à la réassignation de groupes de placement à un OSD (en particulier un nouvel OSD). Par défaut, le paramètre `osd_max_backfills` définit le nombre maximal de renvois simultanés vers un OSD ou à partir de celui-ci à 10. Le paramètre `backfill_full_ratio` permet à un OSD de refuser une demande de renvoi si l'OSD se rapproche de son ratio « full » (90 % par défaut) et de changer avec la commande `ceph osd set-backfillfull-ratio`. Si un OSD refuse une demande de renvoi, le paramètre `osd_backfill_retry_interval` permet à un OSD de réessayer la demande (après 10 secondes, par défaut). Les OSD peuvent également définir les paramètres `osd_backfill_scan_min` et `osd_backfill_scan_max` pour gérer les intervalles d'analyse (64 et 512, par défaut).

REMAPPED

Lorsque l'*ensemble agissant* qui dessert un groupe de placement change, les données migrent de l'ancien *ensemble agissant* vers le nouvel *ensemble agissant*. Un nouvel OSD primaire peut nécessiter un certain temps pour desservir des requêtes. C'est pourquoi il peut demander à l'ancien OSD primaire de continuer à traiter les requêtes jusqu'à ce que la migration du groupe de placement soit terminée. Une fois la migration des données terminée, l'assignation utilise l'OSD primaire du nouvel *ensemble agissant*.

STALE

Tandis que Ceph utilise des pulsations pour s'assurer que les hôtes et les daemons sont en cours d'exécution, les daemons `ceph-osd` peuvent également se retrouver dans un état bloqué (« stuck ») dans lequel ils ne rendent pas compte des statistiques en temps opportun (par exemple, en cas de défaillance temporaire du réseau). Par défaut, les daemons OSD

signalent leurs statistiques de groupe de placement, de démarrage et d'échec toutes les demi-secondes (0,5), ce qui est plus fréquent que les seuils de pulsation. Si l'OSD primaire de l'ensemble agissant d'un groupe de placement ne parvient pas à rendre compte au moniteur ou si d'autres OSD ont signalé l'OSD primaire comme étant arrêté, les moniteurs marquent le groupe de placement comme obsolète (« stale »).

Lorsque vous démarrez votre grappe, il est courant de voir l'état « stale » tant que le processus d'homologation n'est pas terminé. En revanche, lorsque votre grappe est en cours d'exécution depuis un certain temps, si des groupes de placement sont dans l'état « stale », cela indique que l'OSD primaire pour ces groupes de placement est arrêté ou ne rend pas compte de ses statistiques de groupe de placement au moniteur.

12.9.5 Identification de l'emplacement d'un objet

Pour stocker les données d'objet dans le magasin d'objets Ceph, un client Ceph doit définir un nom d'objet et spécifier une réserve associée. Le client Ceph récupère la dernière assignation de grappe et l'algorithme CRUSH calcule comment assigner l'objet à un groupe de placement, puis calcule comment assigner le groupe de placement à un OSD de façon dynamique. Pour trouver l'emplacement de l'objet, tout ce dont vous avez besoin est le nom de l'objet et le nom de la réserve. Par exemple :

```
cephuser@adm > ceph osd map POOL_NAME OBJECT_NAME [NAMESPACE]
```

EXEMPLE 12.1 : LOCALISATION D'UN OBJET

À titre d'exemple, créons un objet. Spécifiez un nom d'objet « test-object-1 », un chemin vers un fichier d'exemple « testfile.txt » contenant certaines données d'objet et un nom de réserve « data » à l'aide de la commande **rados put** sur la ligne de commande :

```
cephuser@adm > rados put test-object-1 testfile.txt --pool=data
```

Pour vérifier que le magasin d'objets Ceph a stocké l'objet, exécutez la commande suivante :

```
cephuser@adm > rados -p data ls
```

Maintenant, identifiez l'emplacement de l'objet. Ceph renverra l'emplacement de l'objet :

```
cephuser@adm > ceph osd map data test-object-1
osdmap e537 pool 'data' (0) object 'test-object-1' -> pg 0.d1743484 \
(0.4) -> up ([1,0], p0) acting ([1,0], p0)
```

Pour supprimer l'exemple d'objet, il suffit de le supprimer à l'aide de la commande **rados** **rm** :

```
cephuser@adm > rados rm test-object-1 --pool=data
```

13 Tâches opérationnelles

13.1 Modification de la configuration d'une grappe

Pour modifier la configuration d'une grappe Ceph existante, procédez comme suit :

1. Exportez la configuration actuelle de la grappe dans un fichier :

```
cephuser@adm > ceph orch ls --export --format yaml > cluster.yaml
```

2. Modifiez le fichier de configuration et mettez à jour les lignes appropriées. Vous trouverez des exemples de spécification dans le *Manuel « Guide de déploiement », Chapitre 8 « Déploiement des services essentiels restants à l'aide de cephadm »* et à la [Section 13.4.3, « Ajout d'OSD à l'aide de la spécification DriveGroups »](#).

3. Appliquez la nouvelle configuration :

```
cephuser@adm > ceph orch apply -i cluster.yaml
```

13.2 Ajout de noeuds

Pour ajouter un noeud à une grappe Ceph, procédez comme suit :

1. Installez SUSE Linux Enterprise Server et SUSE Enterprise Storage sur le nouvel hôte. Reportez-vous au *Manuel « Guide de déploiement », Chapitre 5 « Installation et configuration de SUSE Linux Enterprise Server »* (Guide de sécurité, Chapitre 17 « Masquage et pare-feu ») pour plus d'informations.
2. Configurez l'hôte en tant que minion Salt d'un Salt Master préexistant. Reportez-vous au *Manuel « Guide de déploiement », Chapitre 6 « Déploiement de Salt »* (Guide de sécurité, Chapitre 17 « Masquage et pare-feu ») pour plus d'informations.
3. Ajoutez le nouvel hôte à `ceph-salt` et informez-en cephadm, par exemple :

```
root@master # ceph-salt config /ceph_cluster/minions add ses-min5.example.com
root@master # ceph-salt config /ceph_cluster/roles/cephadm add ses-min5.example.com
```

Reportez-vous au Manuel « Guide de déploiement », Chapitre 7 « Déploiement de la grappe Bootstrap à l'aide de ceph-salt », Section 7.2.2 « Ajout de minions Salt » (Guide de sécurité, Chapitre 17 « Masquage et pare-feu ») pour plus d'informations.

4. Vérifiez que le noeud a bien été ajouté à `ceph-salt` :

```
root@master # ceph-salt config /ceph_cluster/minions ls
o- minions ..... [Minions: 5]
[...]
o- ses-min5.example.com ..... [no roles]
```

5. Appliquez la configuration au nouvel hôte de la grappe :

```
root@master # ceph-salt apply ses-min5.example.com
```

6. Vérifiez que l'hôte qui vient d'être ajouté appartient désormais à l'environnement `cephadm` :

```
cephuser@adm > ceph orch host ls
HOST                ADDR                LABELS    STATUS
[...]
ses-min5.example.com  ses-min5.example.com
```

13.3 Suppression de noeuds



Astuce : suppression des OSD

Si le noeud que vous souhaitez supprimer exécute des OSD, commencez par supprimer ces derniers et vérifiez qu'aucun OSD ne s'exécute sur ce noeud. Pour plus d'informations sur la suppression des OSD, reportez-vous à la [Section 13.4.4, « Suppression des OSD »](#).

Pour supprimer un noeud d'une grappe, procédez comme suit :

1. Pour tous les types de service Ceph, à l'exception de `node-exporter` et `crash`, supprimez le nom d'hôte du noeud du fichier de spécification de placement de la grappe (par exemple, `cluster.yml`). Pour plus d'informations, reportez-vous au Manuel « Guide de déploiement », Chapitre 8 « Déploiement des services essentiels restants à l'aide de cephadm », Section 8.2 « Spéci-

figuration de service et de placement ». Par exemple, si vous souhaitez supprimer l'hôte nommé `ses-min2`, supprimez toutes les occurrences de `- ses-min2` de toutes les sections `placement` :

Remplacez

```
service_type: rgw
service_id: EXAMPLE_NFS
placement:
  hosts:
    - ses-min2
    - ses-min3
```

par

```
service_type: rgw
service_id: EXAMPLE_NFS
placement:
  hosts:
    - ses-min3
```

Appliquez vos modifications au fichier de configuration :

```
cephuser@adm > ceph orch apply -i rgw-example.yaml
```

2. Supprimez le noeud de l'environnement de cephadm :

```
cephuser@adm > ceph orch host rm ses-min2
```

3. Si le noeud exécute les services `crash.osd.1` et `crash.osd.2`, supprimez-les en exécutant la commande suivante sur l'hôte :

```
root@minion > cephadm rm-daemon --fsid CLUSTER_ID --name SERVICE_NAME
```

Par exemple :

```
root@minion > cephadm rm-daemon --fsid b4b30c6e... --name crash.osd.1
root@minion > cephadm rm-daemon --fsid b4b30c6e... --name crash.osd.2
```

4. Supprimez tous les rôles du minion à supprimer :

```
cephuser@adm > ceph-salt config /ceph_cluster/roles/tuned/throughput remove ses-min2
cephuser@adm > ceph-salt config /ceph_cluster/roles/tuned/latency remove ses-min2
cephuser@adm > ceph-salt config /ceph_cluster/roles/cephadm remove ses-min2
cephuser@adm > ceph-salt config /ceph_cluster/roles/admin remove ses-min2
```

Si le minion que vous souhaitez supprimer est le minion Bootstrap, vous devez également supprimer le rôle Bootstrap :

```
cephuser@adm > ceph-salt config /ceph_cluster/roles/bootstrap reset
```

- Après avoir supprimé tous les OSD sur un hôte unique, supprimez ce dernier de la carte CRUSH :

```
cephuser@adm > ceph osd crush remove bucket-name
```



Note

Le nom du compartiment doit être identique au nom d'hôte.

- Vous pouvez à présent supprimer le minion de la grappe :

```
cephuser@adm > ceph-salt config /ceph_cluster/minions remove ses-min2
```



Important

En cas d'échec et si le minion que vous essayez de supprimer se trouve dans un état de désactivation permanente, vous devrez supprimer le noeud du Salt Master :

```
root@master # salt-key -d minion_id
```

Supprimez ensuite manuellement le noeud du fichier *pillar_root/ceph-salt.sls*. Celui-ci se trouve généralement dans */srv/pillar/ceph-salt.sls*.

13.4 Gestion des OSD

Cette section explique comment ajouter, effacer ou supprimer des OSD dans une grappe Ceph.

13.4.1 Liste des périphériques de disque

Pour identifier les périphériques de disque utilisés et inutilisés sur tous les noeuds de la grappe, listez-les en exécutant la commande suivante :

```
cephuser@adm > ceph orch device ls
```


| HOST | PATH | TYPE | SIZE | DEVICE | AVAIL | REJECT | REASONS |
|------------|----------|------|-------|--------|-------|--|---------|
| ses-master | /dev/vda | hdd | 42.0G | | False | locked | |
| ses-min1 | /dev/vda | hdd | 42.0G | | False | locked | |
| ses-min1 | /dev/vdb | hdd | 8192M | 387836 | False | locked, LVM detected, Insufficient space (<5GB) on vgs | |
| ses-min2 | /dev/vdc | hdd | 8192M | 450575 | True | | |

13.4.2 Effacement de périphériques de disque

Pour réutiliser un périphérique de disque, vous devez d'abord l'effacer :

```
ceph orch device zap HOST_NAME DISK_DEVICE
```

Par exemple :

```
cephuser@adm > ceph orch device zap ses-min2 /dev/vdc
```



Note

Si vous avez déjà déployé des OSD à l'aide de DriveGroups ou de l'option `--all-available-devices` alors que l'indicateur `unmanaged` n'était pas défini, `cephadm` déploiera automatiquement ces OSD une fois que vous les aurez effacés.

13.4.3 Ajout d'OSD à l'aide de la spécification DriveGroups

Les *groupes d'unités* (« DriveGroups ») spécifient les dispositions des OSD dans la grappe Ceph. Ils sont définis dans un fichier YAML unique. Dans cette section, nous utiliserons le fichier `drive_groups.yml` comme exemple.

Un administrateur doit spécifier manuellement un groupe d'OSD interdépendants (OSD hybrides déployés sur un mélange de disques durs et SSD) ou qui partagent des options de déploiement identiques (par exemple, même magasin d'objets, même option de chiffrement, OSD autonomes). Pour éviter de lister explicitement les périphériques, les groupes d'unités utilisent une liste d'éléments de filtre qui correspondent à quelques champs sélectionnés de rapports d'inventaire de **ceph-volume**. `cephadm` fournit le code qui traduit ces DriveGroups en listes de périphériques réelles pour inspection par l'utilisateur.

La commande permettant d'appliquer la spécification d'OSD à la grappe est la suivante :

```
cephuser@adm > ceph orch apply osd -i drive_groups.yml
```

Pour afficher un aperçu des opérations et tester votre application, vous pouvez utiliser l'option `--dry-run` en combinaison avec la commande `ceph orch apply osd`. Par exemple :

```
cephuser@adm > ceph orch apply osd -i drive_groups.yml --dry-run
...
+-----+-----+-----+-----+-----+
|SERVICE|NAME  |HOST  |DATA      |DB  |WAL  |
+-----+-----+-----+-----+-----+
|osd      |test  |mgr0  |/dev/sda  |-   |-   |
|osd      |test  |mgr0  |/dev/sdb  |-   |-   |
+-----+-----+-----+-----+-----+
```

Si la sortie de l'option `--dry-run` correspond à vos attentes, réexécutez simplement la commande sans l'option `--dry-run`.

13.4.3.1 OSD non gérés

Tous les périphériques de disque propres disponibles correspondant à la spécification Drive-Groups sont automatiquement utilisés comme OSD une fois que vous les avez ajoutés à la grappe. Ce comportement est appelé mode *managed* (géré).

Pour désactiver le mode *managed*, ajoutez la ligne `unmanaged: true` aux spécifications appropriées, par exemple :

```
service_type: osd
service_id: example_drvgrp_name
placement:
  hosts:
    - ses-min2
    - ses-min3
encrypted: true
unmanaged: true
```



Astuce

Pour faire passer des OSD déjà déployés du mode *managed* au mode *unmanaged*, ajoutez les lignes `unmanaged: true`, le cas échéant, au cours de la procédure décrite à la [Section 13.1](#), « *Modification de la configuration d'une grappe* ».

13.4.3.2 Spécification DriveGroups

Voici un exemple de fichier de spécification DriveGroups :

```
service_type: osd
service_id: example_drvgrp_name
placement:
  host_pattern: '*'
data_devices:
  drive_spec: DEVICE_SPECIFICATION
db_devices:
  drive_spec: DEVICE_SPECIFICATION
wal_devices:
  drive_spec: DEVICE_SPECIFICATION
block_wal_size: '5G' # (optional, unit suffixes permitted)
block_db_size: '5G' # (optional, unit suffixes permitted)
encrypted: true      # 'True' or 'False' (defaults to 'False')
```



Note

L'option précédemment appelée « encryption » (chiffrement) dans DeepSea a été renommée « encrypted » (chiffré). Lorsque vous appliquez des DriveGroups dans SUSE Enterprise Storage 7, veuillez à utiliser cette nouvelle terminologie dans votre spécification de services, sinon l'opération **ceph orch apply** échouera.

13.4.3.3 Périphériques de disque correspondants

Vous pouvez décrire les spécifications à l'aide des filtres suivants :

- Par modèle de disque :

```
model: DISK_MODEL_STRING
```

- Par fournisseur de disque :

```
vendor: DISK_VENDOR_STRING
```



Astuce

Saisissez toujours DISK_VENDOR_STRING en minuscules.

Pour obtenir des informations sur le modèle et le fournisseur du disque, examinez la sortie de la commande suivante :

```
cephuser@adm > ceph orch device ls
HOST      PATH      TYPE  SIZE DEVICE_ID          MODEL      VENDOR
ses-min1  /dev/sdb  ssd   29.8G SATA_SSD_AF34075704240015  SATA SSD   ATA
ses-min2  /dev/sda  ssd   223G Micron_5200_MTFDDAK240TDN  Micron_5200_MTFD ATA
[...]
```

- Selon qu'un disque est rotatif ou non. Les disques SSD et NVMe ne sont pas rotatifs.

```
rotational: 0
```

- Déployez un noeud à l'aide de *tous* les disques disponibles pour les OSD :

```
data_devices:
  all: true
```

- En outre, vous pouvez limiter le nombre de disques correspondants :

```
limit: 10
```

13.4.3.4 Filtrage des périphériques par taille

Vous pouvez filtrer les périphériques de disque par leur taille, soit en fonction d'une taille précise, soit selon une plage de tailles. Le paramètre `size:` (taille :) accepte les arguments sous la forme suivante :

- « 10G » : inclut les disques d'une taille exacte.
- « 10G:40G » : inclut les disques dont la taille est dans la plage.
- « :10G » : inclut les disques dont la taille est inférieure ou égale à 10 Go.
- « 40G: » : inclut les disques dont la taille est égale ou supérieure à 40 Go.

EXEMPLE 13.1 : CORRESPONDANCE PAR TAILLE DE DISQUE

```
service_type: osd
service_id: example_drvgrp_name
placement:
  host_pattern: '*'
data_devices:
  size: '40TB:'
```

```
db_devices:
  size: ':2TB'
```



Note : guillemets requis

Lorsque vous utilisez le délimiteur « : », vous devez entourer la taille par des guillemets simples, faute de quoi le signe deux-points est interprété comme un nouveau hachage de configuration.



Astuce : abréviations des unités

Au lieu d'indiquer les tailles en gigaoctets (G), vous pouvez les spécifier en mégaoctets (M) ou téraoctets (T).

13.4.3.5 Exemples de DriveGroups

Cette section comprend des exemples de différentes configurations OSD.

EXEMPLE 13.2 : CONFIGURATION SIMPLE

Cet exemple décrit deux noeuds avec la même configuration :

- 20 HDD
 - Fournisseur : Intel
 - Modèle : SSD-123-foo
 - Taille : 4 To
- 2 SSD
 - Fournisseur : Micron
 - Modèle : MC-55-44-ZX
 - Taille : 512 Go

Le fichier drive_groups.yml correspondant se présentera comme suit :

```
service_type: osd
service_id: example_drvgrp_name
placement:
  host_pattern: '*'
```

```
data_devices:
  model: SSD-123-foo
db_devices:
  model: MC-55-44-XZ
```

Une telle configuration est simple et valide. Le problème est qu'un administrateur peut ajouter des disques de fournisseurs différents par la suite et ceux-ci ne seront pas inclus. Vous pouvez améliorer cela en limitant les filtres aux propriétés de base des unités :

```
service_type: osd
service_id: example_drvgrp_name
placement:
  host_pattern: '*'
data_devices:
  rotational: 1
db_devices:
  rotational: 0
```

Dans l'exemple précédent, nous imposons de déclarer tous les périphériques rotatifs comme « périphériques de données » et tous les périphériques non rotatifs seront utilisés comme « périphériques partagés » (wal, db).

Si vous savez que les unités de plus de 2 To seront toujours les périphériques de données plus lents, vous pouvez filtrer par taille :

```
service_type: osd
service_id: example_drvgrp_name
placement:
  host_pattern: '*'
data_devices:
  size: '2TB:'
db_devices:
  size: ':2TB'
```

EXEMPLE 13.3 : CONFIGURATION AVANCÉE

Cet exemple décrit deux configurations distinctes : 20 HDD devraient partager 2 SSD, tandis que 10 SSD devraient partager 2 NVMe.

- 20 HDD
 - Fournisseur : Intel
 - Modèle : SSD-123-foo
 - Taille : 4 To
- 12 SSD

- Fournisseur : Micron
- Modèle : MC-55-44-ZX
- Taille : 512 Go
- 2 NVMe
 - Fournisseur : Samsung
 - Modèle : NVME-QQQQQ-987
 - Taille : 256 Go

Une telle configuration peut être définie avec deux dispositions comme suit :

```
service_type: osd
service_id: example_drvgrp_name
placement:
  host_pattern: '*'
data_devices:
  rotational: 0
db_devices:
  model: MC-55-44-XZ
```

```
service_type: osd
service_id: example_drvgrp_name2
placement:
  host_pattern: '*'
data_devices:
  model: MC-55-44-XZ
db_devices:
  vendor: samsung
  size: 256GB
```

EXEMPLE 13.4 : CONFIGURATION AVANCÉE AVEC DES NOEUDS NON UNIFORMES

Les exemples précédents supposaient que tous les noeuds avaient les mêmes unités. Cependant, ce n'est pas toujours le cas :

Noeuds 1 à 5 :

- 20 HDD

- Fournisseur : Intel
- Modèle : SSD-123-foo
- Taille : 4 To
- 2 SSD
 - Fournisseur : Micron
 - Modèle : MC-55-44-ZX
 - Taille : 512 Go

Noeuds 6 à 10 :

- 5 NVMe
 - Fournisseur : Intel
 - Modèle : SSD-123-foo
 - Taille : 4 To
- 20 SSD
 - Fournisseur : Micron
 - Modèle : MC-55-44-ZX
 - Taille : 512 Go

Vous pouvez utiliser la clé « target » dans la disposition pour cibler des noeuds spécifiques. La notation de cible Salt aide à garder les choses simples :

```
service_type: osd
service_id: example_drvgrp_one2five
placement:
  host_pattern: 'node[1-5]'
data_devices:
  rotational: 1
db_devices:
  rotational: 0
```

suivi de


```

service_type: osd
service_id: example_drvgrp_rest
placement:
  host_pattern: 'node[6-10]'
data_devices:
  model: MC-55-44-XZ
db_devices:
  model: SSD-123-foo

```

EXEMPLE 13.5 : CONFIGURATION EXPERTE

Tous les cas précédents supposaient que les WAL et les DB utilisaient le même périphérique. Il est cependant possible également de déployer le WAL sur un périphérique dédié :

- 20 HDD
 - Fournisseur : Intel
 - Modèle : SSD-123-foo
 - Taille : 4 To
- 2 SSD
 - Fournisseur : Micron
 - Modèle : MC-55-44-ZX
 - Taille : 512 Go
- 2 NVMe
 - Fournisseur : Samsung
 - Modèle : NVME-QQQQQ-987
 - Taille : 256 Go

```

service_type: osd
service_id: example_drvgrp_name
placement:
  host_pattern: '*'
data_devices:
  model: MC-55-44-XZ
db_devices:
  model: SSD-123-foo
wal_devices:

```

```
model: NVME-QQQQ-987
```

EXEMPLE 13.6 : CONFIGURATION COMPLEXE (ET PEU PROBABLE)

Dans la configuration suivante, nous essayons de définir :

- 20 HDD soutenus par 1 NVMe
- 2 HDD soutenus par 1 SSD (db) et 1 NVMe (wal)
- 8 SSD soutenus par 1 NVMe
- 2 SSD autonomes (chiffrés)
- 1 HDD est de rechange et ne doit pas être déployé

Le résumé des unités utilisées est le suivant :

- 23 HDD
 - Fournisseur : Intel
 - Modèle : SSD-123-foo
 - Taille : 4 To
- 10 SSD
 - Fournisseur : Micron
 - Modèle : MC-55-44-ZX
 - Taille : 512 Go
- 1 NVMe
 - Fournisseur : Samsung
 - Modèle : NVME-QQQQQ-987
 - Taille : 256 Go

La définition des groupes d'unités sera la suivante :

```
service_type: osd
service_id: example_drvgrp_hdd_nvme
placement:
  host_pattern: '*'
data_devices:
```

```
rotational: 0
db_devices:
  model: NVME-QQQQ-987
```

```
service_type: osd
service_id: example_drvgrp_hdd_ssd_nvme
placement:
  host_pattern: '*'
data_devices:
  rotational: 0
db_devices:
  model: MC-55-44-XZ
wal_devices:
  model: NVME-QQQQ-987
```

```
service_type: osd
service_id: example_drvgrp_ssd_nvme
placement:
  host_pattern: '*'
data_devices:
  model: SSD-123-foo
db_devices:
  model: NVME-QQQQ-987
```

```
service_type: osd
service_id: example_drvgrp_standalone_encrypted
placement:
  host_pattern: '*'
data_devices:
  model: SSD-123-foo
encrypted: True
```

Il reste un HDD dans la mesure où le fichier est en cours d'analyse du haut vers le bas.

13.4.4 Suppression des OSD

Avant de supprimer un noeud OSD de la grappe, vérifiez que cette dernière dispose de plus d'espace disque disponible que le disque OSD que vous allez supprimer. Gardez à l'esprit que la suppression d'un OSD entraîne un rééquilibrage de l'ensemble de la grappe.

1. Identifiez l'OSD à supprimer en obtenant son ID :

```
cephuser@adm > ceph orch ps --daemon_type osd
NAME    HOST          STATUS      REFRESHED  AGE  VERSION
osd.0   target-ses-090 running (3h) 7m ago    3h   15.2.7.689 ...
```

```
osd.1 target-ses-090 running (3h) 7m ago 3h 15.2.7.689 ...
osd.2 target-ses-090 running (3h) 7m ago 3h 15.2.7.689 ...
osd.3 target-ses-090 running (3h) 7m ago 3h 15.2.7.689 ...
```

2. Supprimez un ou plusieurs OSD de la grappe :

```
cephuser@adm > ceph orch osd rm OSD1_ID OSD2_ID ...
```

Par exemple :

```
cephuser@adm > ceph orch osd rm 1 2
```

3. Vous pouvez exécuter une requête pour connaître l'état de l'opération de suppression :

```
cephuser@adm > ceph orch osd rm status
```

| OSD_ID | HOST | STATE | PG_COUNT | REPLACE | FORCE | STARTED_AT |
|--------|-------------|-------------------------|----------|---------|-------|----------------------------|
| 2 | cephadm-dev | done, waiting for purge | 0 | True | False | 2020-07-17 13:01:43.147684 |
| 3 | cephadm-dev | draining | 17 | False | True | 2020-07-17 13:01:45.162158 |
| 4 | cephadm-dev | started | 42 | False | True | 2020-07-17 13:01:45.162158 |

13.4.4.1 Arrêt de la suppression d'un OSD

Après avoir planifié la suppression d'un OSD, vous pouvez interrompre l'opération si nécessaire. La commande suivante permet de rétablir l'état initial de l'OSD et de le supprimer de la file d'attente :

```
cephuser@adm > ceph orch osd rm stop OSD_SERVICE_ID
```

13.4.5 Remplacement d'OSD

Plusieurs raisons peuvent vous pousser à remplacer un disque OSD. Par exemple :

- Le disque OSD est défaillant ou, d'après les informations SMART, sera bientôt défaillant et ne peut plus être utilisé pour stocker des données en toute sécurité.
- Vous devez mettre à niveau le disque OSD, par exemple pour augmenter sa taille.
- Vous devez modifier la disposition du disque OSD.
- Vous envisagez de passer d'une disposition non-LVM à une disposition LVM.

Pour remplacer un OSD tout en conservant son ID, exécutez la commande suivante :

```
cephuser@adm > ceph orch osd rm OSD_SERVICE_ID --replace
```

Par exemple :

```
cephuser@adm > ceph orch osd rm 4 --replace
```

Le remplacement d'un OSD est identique à la suppression d'un OSD (pour plus de détails, reportez-vous à la [Section 13.4.4, « Suppression des OSD »](#)) à la différence près que l'OSD n'est pas supprimé de façon permanente de la hiérarchie CRUSH et se voit assigner un indicateur destroyed (détruit).

L'indicateur destroyed (détruit) sert à déterminer les ID d'OSD qui seront réutilisés lors du prochain déploiement d'OSD. Les nouveaux disques ajoutés qui correspondent à la spécification `DriveGroups` (pour plus de détails, reportez-vous à la [Section 13.4.3, « Ajout d'OSD à l'aide de la spécification DriveGroups »](#)) se verront assigner les ID d'OSD de leur homologue remplacé.



Astuce

L'ajout de l'option `--dry-run` ne permet pas d'exécuter le remplacement réel, mais d'afficher un aperçu des étapes qui se produiraient normalement.



Note

En cas de remplacement d'un OSD à la suite d'un échec, nous vous recommandons vivement de déclencher un nettoyage en profondeur des groupes de placement. Pour plus d'informations, reportez-vous au [Section 17.6, « Nettoyage des groupes de placement »](#).

Exécutez la commande suivante pour lancer un nettoyage en profondeur :

```
cephuser@adm > ceph osd deep-scrub osd.OSD_NUMBER
```



Important : échec du périphérique partagé

En cas d'échec d'un périphérique partagé pour DB/WAL, vous devez effectuer la procédure de remplacement pour tous les OSD qui partagent le périphérique ayant échoué.

13.5 Déplacement du Salt Master vers un nouveau noeud

Si vous devez remplacer l'hôte Salt Master par un autre, procédez comme suit :

1. Exportez la configuration de la grappe et sauvegardez le fichier JSON exporté. Pour plus de détails, reportez-vous au *Manuel « Guide de déploiement », Chapitre 7 « Déploiement de la grappe Bootstrap à l'aide de ceph-salt », Section 7.2.14 « Exportation des configurations de grappe »*.
2. Si l'ancien Salt Master est également le seul noeud d'administration de la grappe, déplacez manuellement les fichiers `/etc/ceph/ceph.client.admin.keyring` et `/etc/ceph/ceph.conf` vers le nouveau Salt Master.
3. Arrêtez et désactivez le service `systemd` Salt Master sur l'ancien noeud Salt Master :

```
root@master # systemctl stop salt-master.service
root@master # systemctl disable salt-master.service
```

4. Si l'ancien noeud Salt Master ne se trouve plus dans la grappe, arrêtez et désactivez également le service `systemd` du minion Salt :

```
root@master # systemctl stop salt-minion.service
root@master # systemctl disable salt-minion.service
```



Avertissement

N'arrêtez ou ne désactivez pas `salt-minion.service` si des daemons Ceph (MON, MGR, OSD, MDS, passerelle, surveillance) s'exécutent sur l'ancien noeud Salt Master.

5. Installez SUSE Linux Enterprise Server 15 SP3 sur le nouveau Salt Master en suivant la procédure décrite dans le *Manuel « Guide de déploiement », Chapitre 5 « Installation et configuration de SUSE Linux Enterprise Server »*.



Astuce : transition des minions Salt

Pour simplifier la transition des minions Salt vers le nouveau Salt Master, retirez la clé publique Salt Master d'origine de chacun d'eux :

```
root@minion > rm /etc/salt/pki/minion/minion_master.pub
root@minion > systemctl restart salt-minion.service
```

6. Installez le paquetage `salt-master` et, le cas échéant, le paquetage `salt-minion` sur le nouveau Salt Master.

7. Installez `ceph-salt` sur le nouveau noeud Salt Master :

```
root@master # zypper install ceph-salt
root@master # systemctl restart salt-master.service
root@master # salt '*' saltutil.sync_all
```



Important

Veillez à exécuter les trois commandes avant de continuer. Les commandes sont idempotentes ; peu importe si elles sont exécutées à plusieurs reprises.

8. Incluez le nouveau Salt Master dans la grappe comme décrit dans le *Manuel « Guide de déploiement », Chapitre 7 « Déploiement de la grappe Bootstrap à l'aide de ceph-salt », Section 7.1 « Installation ceph-salt », le Manuel « Guide de déploiement », Chapitre 7 « Déploiement de la grappe Bootstrap à l'aide de ceph-salt », Section 7.2.2 « Ajout de minions Salt » et le Manuel « Guide de déploiement », Chapitre 7 « Déploiement de la grappe Bootstrap à l'aide de ceph-salt », Section 7.2.4 « Spécification du noeud Admin ».*
9. Importez la configuration de grappe sauvegardée et appliquez-la :

```
root@master # ceph-salt import CLUSTER_CONFIG.json
root@master # ceph-salt apply
```



Important

Renommez le `minion id` du Salt Master dans le fichier `CLUSTER_CONFIG.json` exporté avant de l'importer.

13.6 Mise à jour des noeuds de grappe

Gardez les noeuds de grappe Ceph à jour en appliquant régulièrement des mises à jour progressives.

13.6.1 Dépôts logiciels

Avant d'appliquer des correctifs à la grappe avec les paquetages les plus récents, vérifiez que tous les noeuds de la grappe ont accès aux dépôts pertinents. Pour obtenir la liste complète des dépôts requis, reportez-vous au *Manuel « Guide de déploiement », Chapitre 10 « Mise à niveau de SUSE Enterprise Storage 6 vers la version 7.1 », Section 10.1.5.1 « Dépôts logiciels »*.

13.6.2 Préparation du dépôt

Si vous utilisez un outil de préparation (SUSE Manager, Subscription Management Tool ou RMT, par exemple) qui met à disposition des dépôts logiciels pour les noeuds de la grappe, vérifiez que les phases pour les dépôts de mise à jour de SUSE Linux Enterprise Server et de SUSE Enterprise Storage sont créées au même moment.

Il est vivement recommandé d'utiliser un outil de préparation pour appliquer des correctifs de niveau `frozen` ou `staged`. Cela garantit le même niveau de correctif aux noeuds qui rejoignent la grappe et à ceux qui y sont déjà en cours d'exécution. Vous évitez ainsi de devoir appliquer les correctifs les plus récents à tous les noeuds de la grappe avant que de nouveaux noeuds puissent la rejoindre.

13.6.3 Temps d'indisponibilité des services Ceph

Selon la configuration, les noeuds de grappe peuvent être redémarrés pendant la mise à jour. S'il existe un point d'échec unique pour des services tels qu'Object Gateway, Samba Gateway, NFS Ganesha ou iSCSI, les machines clientes peuvent être temporairement déconnectées des services dont les noeuds sont redémarrés.

13.6.4 Exécution de la mise à jour

Pour mettre à jour les paquetages logiciels sur tous les noeuds de grappe vers la dernière version, exécutez la commande suivante :

```
root@master # ceph-salt update
```

13.7 Mise à jour de Ceph

Vous pouvez demander à cephadm de mettre à jour Ceph d'une version de correctifs vers une autre. La mise à jour automatique des services Ceph respecte l'ordre recommandé : elle commence par les instances Ceph Manager, Ceph Monitor, puis continue avec d'autres services tels que les OSD Ceph et les instances Metadata Server et Object Gateway. Chaque daemon est redémarré uniquement après que Ceph indique que la grappe restera disponible.



Note

La procédure de mise à jour ci-dessous utilise la commande **ceph orch upgrade**. Gardez à l'esprit que les instructions suivantes expliquent comment mettre à jour votre grappe Ceph avec une version de produit (par exemple, une mise à jour de maintenance), et *non* comment mettre à niveau votre grappe d'une version de produit à une autre.

13.7.1 Démarrage de la mise à jour

Avant de démarrer la mise à jour, vérifiez que tous les noeuds sont en ligne et que votre grappe est saine :

```
cephuser@adm > cephadm shell -- ceph -s
```

Pour effectuer une mise à jour vers une version spécifique de Ceph :

```
cephuser@adm > ceph orch upgrade start --image REGISTRY_URL
```

Par exemple :

```
cephuser@adm > ceph orch upgrade start --image registry.suse.com/ses/7.1/ceph/ceph:latest
```

Mettez à niveau les paquetages sur les hôtes :

```
cephuser@adm > ceph-salt update
```

13.7.2 Surveillance de la mise à jour

Exécutez la commande suivante pour déterminer si une mise à jour est en cours :

```
cephuser@adm > ceph orch upgrade status
```

Pendant l'exécution de la mise à jour, vous verrez une barre de progression dans la sortie d'état de Ceph :

```
cephuser@adm > ceph -s
[...]
progress:
  Upgrade to registry.suse.com/ses/7.1/ceph/ceph:latest (00h 20m 12s)
  [=====.....] (time remaining: 01h 43m 31s)
```

Vous pouvez également consulter le journal cephadm :

```
cephuser@adm > ceph -W cephadm
```

13.7.3 Annulation d'une mise à jour

Vous pouvez arrêter le processus de mise à jour à tout moment :

```
cephuser@adm > ceph orch upgrade stop
```

13.8 Arrêt ou redémarrage de la grappe

Dans certains cas, il faudra peut-être arrêter ou redémarrer l'ensemble de la grappe. Nous vous recommandons de contrôler soigneusement les dépendances des services en cours d'exécution. Les étapes suivantes fournissent un aperçu de l'arrêt et du démarrage de la grappe :

1. Ordonnez à la grappe Ceph de ne pas marquer les OSD comme étant hors service :

```
cephuser@adm > ceph osd set noout
```

2. Arrêtez les daemons et les noeuds dans l'ordre suivant :

1. Clients de stockage
2. Passerelles, par exemple NFS Ganesha ou Object Gateway

3. Serveur de métadonnées
 4. Ceph OSD
 5. Ceph Manager
 6. Ceph Monitor
3. Si nécessaire, effectuez des tâches de maintenance.
4. Démarrez les noeuds et les serveurs dans l'ordre inverse du processus d'arrêt :
1. Ceph Monitor
 2. Ceph Manager
 3. Ceph OSD
 4. Serveur de métadonnées
 5. Passerelles, par exemple NFS Ganesha ou Object Gateway
 6. Clients de stockage
5. Supprimez l'indicateur noout :

```
cephuser@adm > ceph osd unset noout
```

13.9 Suppression d'une grappe Ceph entière

La commande **ceph-salt purge** permet de supprimer l'intégralité de la grappe Ceph. Si d'autres grappes Ceph sont déployées, celle spécifiée par **ceph -s** est purgée. De cette façon, vous pouvez nettoyer l'environnement de grappe lors du test de différentes configurations.

Pour éviter toute suppression accidentelle, l'orchestration vérifie si la sécurité est désengagée. Vous pouvez désengager les mesures de sécurité et supprimer la grappe Ceph en exécutant les commandes suivantes :

```
root@master # ceph-salt disengage-safety  
root@master # ceph-salt purge
```

14 Exécution des services Ceph

Vous pouvez exécuter les services Ceph au niveau d'un daemon, d'un noeud ou d'une grappe. Selon l'approche dont vous avez besoin, utilisez `cephadm` ou la commande **`systemctl`**.

14.1 Exécution de services individuels

Si vous devez exécuter un service spécifique, commencez par l'identifier :

```
cephuser@adm > ceph orch ps
```

| NAME | HOST | STATUS | REFRESHED | [...] |
|------------------------------------|----------|---------------|-----------|-------|
| mds.my_cephfs.ses-min1.oterul | ses-min1 | running (5d) | 8m ago | |
| mgr.ses-min1.gpijpm | ses-min1 | running (5d) | 8m ago | |
| mgr.ses-min2.oopvyh | ses-min2 | running (5d) | 8m ago | |
| mon.ses-min1 | ses-min1 | running (5d) | 8m ago | |
| mon.ses-min2 | ses-min2 | running (5d) | 8m ago | |
| mon.ses-min4 | ses-min4 | running (5d) | 7m ago | |
| osd.0 | ses-min2 | running (61m) | 8m ago | |
| osd.1 | ses-min3 | running (61m) | 7m ago | |
| osd.2 | ses-min4 | running (61m) | 7m ago | |
| rgw.myrealm.myzone.ses-min1.kwazo | ses-min1 | running (5d) | 8m ago | |
| rgw.myrealm.myzone.ses-min2.jngabw | ses-min2 | error | 8m ago | |

Pour identifier un service sur un noeud spécifique, exécutez :

```
ceph orch ps NODE_HOST_NAME
```

Par exemple :

```
cephuser@adm > ceph orch ps ses-min2
```

| NAME | HOST | STATUS | REFRESHED |
|---------------------|----------|---------------|-----------|
| mgr.ses-min2.oopvyh | ses-min2 | running (5d) | 3m ago |
| mon.ses-min2 | ses-min2 | running (5d) | 3m ago |
| osd.0 | ses-min2 | running (67m) | 3m ago |



Astuce

La commande **`ceph orch ps`** prend en charge plusieurs formats de sortie. Pour changer de format, ajoutez l'option `--format FORMAT`, *FORMAT* correspondant au format `json`, `json-pretty` ou `yaml`. Par exemple :

```
cephuser@adm > ceph orch ps --format yaml
```

Une fois que vous connaissez le nom du service, vous pouvez le démarrer, le redémarrer ou l'arrêter :

```
ceph orch daemon COMMAND SERVICE_NAME
```

Par exemple, pour redémarrer le service OSD avec l'ID 0, exécutez :

```
cephuser@adm > ceph orch daemon restart osd.0
```

14.2 Exécution de types de service

Si vous devez exécuter un type de service spécifique sur l'ensemble de la grappe Ceph, utilisez la commande suivante :

```
ceph orch COMMAND SERVICE_TYPE
```

Remplacez *COMMAND* par *start*, *stop* ou *restart*.

Par exemple, la commande suivante permet de redémarrer toutes les instances MON de la grappe, quels que soient les noeuds sur lesquels elles s'exécutent :

```
cephuser@adm > ceph orch restart mon
```

14.3 Exécution de services sur un seul noeud

La commande **systemctl** permet d'exécuter des cibles et des services *systemd* associés à Ceph sur un seul noeud.

14.3.1 Identification des services et des cibles

Avant d'exécuter des cibles et des services *systemd* associés à Ceph, vous devez identifier les noms de leurs fichiers unité. Les noms de fichier des services se présentent comme suit :

```
ceph-FSID@SERVICE_TYPE.ID.service
```

Par exemple :

```
ceph-b4b30c6e-9681-11ea-ac39-525400d7702d@mon.doc-ses-min1.service
```

```
ceph-b4b30c6e-9681-11ea-ac39-525400d7702d@rgw.myrealm.myzone.doc-ses-min1.kwwazo.service
```

FSID

ID unique de la grappe Ceph. Celui-ci figure dans la sortie de la commande **`ceph fsid`**.

SERVICE_TYPE

Type du service, par exemple `osd`, `mon` ou `rgw`.

ID

Chaîne d'identification du service. Pour les OSD, il s'agit du numéro d'ID du service. Pour les autres services, il peut s'agir d'un nom d'hôte du noeud ou d'autres chaînes pertinentes pour le type de service.



Astuce

La partie `SERVICE_TYPE.ID` est identique au contenu de la colonne `NAME` dans la sortie de la commande **`ceph orch ps`**.

14.3.2 Exécution de l'ensemble des services sur un noeud

Les cibles `systemd` de Ceph permettent d'exécuter simultanément *tous* les services sur un noeud ou tous les services *appartenant à une grappe* identifiée par son `FSID`.

Par exemple, pour arrêter tous les services Ceph sur un noeud, quelle que soit la grappe à laquelle ils appartiennent, exécutez :

```
root@minion > systemctl stop ceph.target
```

Pour redémarrer tous les services appartenant à une grappe Ceph avec l'ID `b4b30c6e-9681-11ea-ac39-525400d7702d`, exécutez :

```
root@minion > systemctl restart ceph-b4b30c6e-9681-11ea-ac39-525400d7702d.target
```

14.3.3 Exécution d'un service spécifique sur un noeud

Après avoir identifié le nom d'un service spécifique, exécutez-le comme suit :

```
systemctl COMMAND SERVICE_NAME
```

Par exemple, pour redémarrer un seul service OSD avec l'ID 1 sur une grappe avec l'ID `b4b30c6e-9681-11ea-ac39-525400d7702d`, exécutez :

```
# systemctl restart ceph-b4b30c6e-9681-11ea-ac39-525400d7702d@osd.1.service
```

14.3.4 Vérification de l'état des services

Vous pouvez interroger `systemd` pour connaître l'état des services. Par exemple :

```
# systemctl status ceph-b4b30c6e-9681-11ea-ac39-525400d7702d@osd.0.service
```

14.4 Arrêt et redémarrage de l'ensemble de la grappe Ceph

Une panne de courant programmée peut nécessiter l'arrêt et le redémarrage de la grappe. Pour arrêter tous les services associés à Ceph et redémarrer sans problème, suivez les étapes ci-dessous.

PROCÉDURE 14.1 : ARRÊT DE L'ENSEMBLE DE LA GRAPPE CEPH

1. Arrêtez ou déconnectez tous les clients qui accèdent à la grappe.
2. Pour empêcher CRUSH de rééquilibrer automatiquement la grappe, définissez la grappe sur `noout` :

```
cephuser@adm > ceph osd set noout
```

3. Arrêtez tous les services Ceph sur tous les noeuds de la grappe :

```
root@master # ceph-salt stop
```

4. Mettez tous les noeuds de grappe hors tension :

```
root@master # salt -G 'ceph-salt:member' cmd.run "shutdown -h"
```

PROCÉDURE 14.2 : DÉMARRAGE DE L'ENSEMBLE DE LA GRAPPE CEPH

1. Mettez le noeud Admin sous tension.
2. Mettez les noeuds Ceph Monitor sous tension.

3. Mettez les noeuds Ceph OSD sous tension.
4. Désélectionnez l'option noout préalablement sélectionnée :

```
root@master # ceph osd unset noout
```

5. Mettez toutes les passerelles configurées sous tension.
6. Mettez sous tension ou connectez les clients de la grappe.

15 Sauvegarde et restauration

Ce chapitre explique quelles parties de la grappe Ceph vous devriez sauvegarder afin d'être en mesure de restaurer sa fonctionnalité.

15.1 Sauvegarde de la configuration et des données de grappe

15.1.1 Sauvegarde de la configuration de ceph-salt

Exportez la configuration de la grappe. Pour plus d'informations, reportez-vous au *Manuel « Guide de déploiement »*, Chapitre 7 « Déploiement de la grappe Bootstrap à l'aide de ceph-salt », Section 7.2.14 « Exportation des configurations de grappe ».

15.1.2 Sauvegarde de la configuration Ceph

Sauvegardez l'annuaire `/etc/ceph`. Il contient des informations de configuration de grappe cruciales. Par exemple, vous avez besoin de la sauvegarde de `/etc/ceph` lorsque vous devez remplacer le noeud Admin.

15.1.3 Sauvegarde de la configuration Salt

Vous devez sauvegarder le répertoire `/etc/salt/`. Il contient les fichiers de configuration Salt, par exemple la clé de Salt Master et les clés clients acceptées.

Les fichiers Salt ne sont pas strictement requis pour la sauvegarde du noeud Admin, mais facilitent le redéploiement de la grappe Salt. S'il n'existe pas de sauvegarde de ces fichiers, les minions Salt doivent être enregistrés à nouveau au niveau du nouveau noeud Admin.



Note : sécurité de la clé privée de Salt Master

Assurez-vous que la sauvegarde de la clé privée de Salt Master est stockée à un emplacement sûr. La clé de Salt Master peut être utilisée pour manipuler tous les noeuds de la grappe.

15.1.4 Sauvegarde des configurations personnalisées

- Données et personnalisation de Prometheus.
- Personnalisation de Grafana.
- Modifications manuelles de la configuration iSCSI.
- Clés Ceph.
- Assignation et règles CRUSH. Enregistrez la carte CRUSH décompilée, y compris les règles CRUSH dans le fichier `crushmap-backup.txt` en exécutant la commande suivante :

```
cephuser@adm > ceph osd getcrushmap | crushtool -d - -o crushmap-backup.txt
```

- Configuration de la passerelle Samba. Si vous utilisez une seule passerelle, sauvegardez `/etc/samba/smb.conf`. Si vous utilisez une configuration haute disponibilité, sauvegardez également les fichiers de configuration CTDB et Pacemaker. Pour plus de détails sur la configuration utilisée par les passerelles Samba, reportez-vous au [Chapitre 24, Exportation des données Ceph via Samba](#).
- Configuration de NFS Ganesha. Uniquement nécessaire en cas d'utilisation de la configuration HA. Pour plus de détails sur la configuration utilisée par NFS Ganesha, reportez-vous au [Chapitre 25, NFS Ganesha](#).

15.2 Restauration d'un noeud Ceph

La procédure de récupération d'un noeud à partir d'une sauvegarde consiste à réinstaller le noeud, à remplacer ses fichiers de configuration, puis à réorchestrer la grappe afin que le noeud de remplacement soit de nouveau ajouté.

Si vous devez redéployer le noeud Admin, reportez-vous à la [Section 13.5, « Déplacement du Salt Master vers un nouveau noeud »](#).

Pour les minions, il est généralement plus simple de reconstruire et de redéployer.

1. Réinstallez le noeud. Pour plus d'informations, reportez-vous au *Manuel « Guide de déploiement », Chapitre 5 « Installation et configuration de SUSE Linux Enterprise Server »*
2. Installez Salt. Pour plus d'informations, reportez-vous au *Manuel « Guide de déploiement », Chapitre 6 « Déploiement de Salt »*

3. Après avoir restauré le répertoire `/etc/salt` à partir d'une sauvegarde, activez et redémarrez les services Salt applicables, par exemple :

```
root@master # systemctl enable salt-master
root@master # systemctl start salt-master
root@master # systemctl enable salt-minion
root@master # systemctl start salt-minion
```

4. Supprimez la clé publique principale de l'ancien noeud Salt Master de tous les minions.

```
root@master # rm /etc/salt/pki/minion/minion_master.pub
root@master # systemctl restart salt-minion
```

5. Restaurez tout ce qui était local sur le noeud Admin.
6. Importez la configuration de la grappe à partir du fichier JSON exporté précédemment. Pour plus d'informations, reportez-vous au *Manuel « Guide de déploiement », Chapitre 7 « Déploiement de la grappe Bootstrap à l'aide de ceph-salt », Section 7.2.14 « Exportation des configurations de grappe »*.
7. Appliquez la configuration de grappe importée :

```
root@master # ceph-salt apply
```

16 Surveillance et alertes

Dans SUSE Enterprise Storage 7.1, cephadm déploie une pile de surveillance et d'alerte. Les utilisateurs doivent soit définir les services (p. ex., Prometheus, Alertmanager et Grafana) qu'ils souhaitent déployer à l'aide de cephadm dans un fichier de configuration YAML, soit utiliser l'interface de ligne de commande pour les déployer. En cas de déploiement de plusieurs services du même type, une configuration hautement disponible est déployée. Le service Node exporter fait figure d'exception à cette règle.

Les services de surveillance suivants peuvent être déployés à l'aide de cephadm :

- **Prometheus** est le toolkit de surveillance et d'alerte. Il collecte les données fournies par les exportateurs Prometheus et déclenche les alertes préconfigurées lorsque les seuils prédéfinis sont atteints.
- **Alertmanager** traite les alertes envoyées par le serveur Prometheus. Il déduplique les alertes, les regroupe et les achemine vers le récepteur approprié. Par défaut, Ceph Dashboard est automatiquement configuré en tant que récepteur.
- **Grafana** est le logiciel de visualisation et d'alerte. La fonctionnalité d'alerte de Grafana n'est pas utilisée par cette pile de surveillance. Pour les alertes, le service Alertmanager est utilisé.
- **Node exporter** est un exportateur pour Prometheus qui fournit des données à propos du noeud sur lequel il est installé. Il est recommandé d'installer Node exporter sur tous les noeuds.

Le module Prometheus Manager fournit un exportateur Prometheus pour transmettre les compteurs de performance Ceph à partir du point de collecte dans `ceph-mgr`.

La configuration de Prometheus, y compris les cibles de *récupération* (mesures fournissant des daemons), est définie automatiquement par cephadm. cephadm déploie également une liste d'alertes par défaut, par exemple `health error` (erreur de santé), `10% OSDs down` (10 % OSD arrêtés) ou `pgs inactive` (GP inactifs).

Par défaut, le trafic vers Grafana est chiffré par TLS. Vous pouvez fournir votre propre certificat TLS ou utiliser un certificat auto-signé. Si aucun certificat personnalisé n'a été configuré avant le déploiement de Grafana, un certificat auto-signé est automatiquement créé et configuré pour Grafana.

Vous pouvez configurer des certificats personnalisés pour Grafana en procédant comme suit :

1. Configurez les fichiers de certificat :

```
cephuser@adm > ceph config-key set mgr/cephadm/grafana_key -i $PWD/key.pem
cephuser@adm > ceph config-key set mgr/cephadm/grafana_cert -i $PWD/certificate.pem
```

2. Redémarrez le service Ceph Manager :

```
cephuser@adm > ceph orch restart mgr
```

3. Reconfigurez le service Grafana pour refléter les nouveaux chemins de certificat et définissez l'URL correcte pour Ceph Dashboard :

```
cephuser@adm > ceph orch reconfig grafana
```

Alertmanager gère les alertes envoyées par le serveur Prometheus. Il s'occupe de les déduplicer, de les regrouper et de les acheminer vers le bon récepteur. Les alertes peuvent être désactivées à l'aide d'Alertmanager, mais les silences peuvent également être gérés au moyen de Ceph Dashboard.

Il est recommandé de déployer `Node exporter` sur tous les nœuds. Pour ce faire, vous pouvez utiliser le fichier `monitoring.yaml` avec le type de service `node-exporter`. Pour plus d'informations sur le déploiement des services, reportez-vous au Manuel « Guide de déploiement », Chapitre 8 « Déploiement des services essentiels restants à l'aide de `cephadm` », Section 8.3.8 « Déploiement de la pile de surveillance ».

16.1 Configuration d'images personnalisées ou locales



Astuce

Cette section explique comment modifier la configuration des images de conteneur utilisées lors du déploiement ou de la mise à jour de services. Elle n'inclut pas les commandes nécessaires au déploiement ou au redéploiement des services.

La méthode recommandée pour déployer la pile de surveillance consiste à appliquer sa spécification comme décrit dans le Manuel « *Guide de déploiement* », Chapitre 8 « *Déploiement des services essentiels restants à l'aide de cephadm* », Section 8.3.8 « *Déploiement de la pile de surveillance* ».

Pour déployer des images de conteneur personnalisées ou locales, celles-ci doivent être définies dans cephadm. Pour ce faire, vous devez exécuter la commande suivante :

```
cephuser@adm > ceph config set mgr mgr/cephadm/OPTION_NAME VALUE
```

Dans laquelle *OPTION_NAME* correspond à l'un des noms suivants :

- container_image_prometheus
- container_image_node_exporter
- container_image_alertmanager
- container_image_grafana

Si aucune option n'est définie ou si le paramètre a été supprimé, les images suivantes sont utilisées comme *VALUE* :

- registry.suse.com/ses/7.1/ceph/prometheus-server:2.32.1
- registry.suse.com/ses/7.1/ceph/prometheus-node-exporter:1.1.2
- registry.suse.com/ses/7.1/ceph/prometheus-alertmanager:0.21.0
- registry.suse.com/ses/7.1/ceph/grafana:7.5.12

Par exemple :

```
cephuser@adm > ceph config set mgr mgr/cephadm/container_image_prometheus prom/  
prometheus:v1.4.1
```



Note

Si vous définissez une image personnalisée, la valeur par défaut sera remplacée (mais pas écrasée). La valeur par défaut change lorsque des mises à jour sont disponibles. Si vous définissez une image personnalisée, vous ne pourrez pas mettre à jour automatiquement

le composant pour lequel vous avez configuré l'image personnalisée. Vous devez mettre à jour manuellement la configuration (nom de l'image et balise) pour pouvoir installer des mises à jour.

Si vous choisissez plutôt de suivre les recommandations, vous pouvez réinitialiser l'image personnalisée que vous avez définie auparavant, après quoi la valeur par défaut sera de nouveau utilisée. Utilisez **ceph config rm** pour réinitialiser l'option de configuration :

```
cephuser@adm > ceph config rm mgr mgr/cephadm/OPTION_NAME
```

Par exemple :

```
cephuser@adm > ceph config rm mgr mgr/cephadm/container_image_prometheus
```

16.2 Mise à jour des services de surveillance

Comme indiqué dans la [Section 16.1, « Configuration d'images personnalisées ou locales »](#), cephadm est fourni avec les URL des images de conteneur recommandées et testées, lesquelles sont utilisées par défaut.

Si vous mettez à jour les paquetages Ceph, de nouvelles versions de ces URL peuvent être fournies. Cette opération met simplement à jour l'emplacement à partir duquel les images de conteneur sont extraites, mais ne met à jour aucun service.

Une fois les URL des nouvelles images de conteneur mises à jour, que ce soit manuellement comme décrit dans la [Section 16.1, « Configuration d'images personnalisées ou locales »](#) ou automatiquement via une mise à jour du paquetage Ceph, les services de surveillance peuvent être mis à jour.

Pour ce faire, utilisez **ceph orch reconfig** comme suit :

```
cephuser@adm > ceph orch reconfig node-exporter
cephuser@adm > ceph orch reconfig prometheus
cephuser@adm > ceph orch reconfig alertmanager
cephuser@adm > ceph orch reconfig grafana
```

Il n'existe actuellement aucune commande unique permettant de mettre à jour tous les services de surveillance. L'ordre dans lequel ces services sont mis à jour n'a pas d'importance.



Note

Si vous utilisez des images de conteneur personnalisées, les URL spécifiées pour les services de surveillance ne seront pas modifiées automatiquement en cas de mise à jour des paquetages Ceph. Si vous avez spécifié des images de conteneur personnalisées, vous devrez indiquer manuellement les URL des nouvelles images de conteneur, notamment si vous utilisez un registre local de conteneurs.

Les URL des images de conteneur recommandées à utiliser sont indiquées dans la [Section 16.1, « Configuration d'images personnalisées ou locales »](#).

16.3 Désactivation de la surveillance

Pour désactiver la pile de surveillance, exécutez les commandes suivantes :

```
cephuser@adm > ceph orch rm grafana
cephuser@adm > ceph orch rm prometheus --force # this will delete metrics data
collected so far
cephuser@adm > ceph orch rm node-exporter
cephuser@adm > ceph orch rm alertmanager
cephuser@adm > ceph mgr module disable prometheus
```

16.4 Configuration de Grafana

Le back-end Ceph Dashboard a besoin de l'URL de Grafana pour pouvoir vérifier l'existence des tableaux de bord Grafana avant même que le front-end ne les charge. En raison de la nature de l'implémentation de Grafana dans Ceph Dashboard, cela signifie que deux connexions fonctionnelles sont nécessaires pour pouvoir afficher les graphiques Grafana dans Ceph Dashboard :

- Le back-end (module Ceph MGR) doit vérifier l'existence du graphique demandé. Si cette requête aboutit, il informe le front-end qu'il peut accéder à Grafana en toute sécurité.
- Le front-end demande alors les graphiques Grafana directement à partir du navigateur de l'utilisateur à l'aide d'une [iframe](#). L'instance Grafana est accessible directement, sans détour, via Ceph Dashboard.

Cela dit, il se peut qu'en raison de votre environnement, le navigateur de l'utilisateur puisse difficilement accéder directement à l'URL configurée dans Ceph Dashboard. Pour résoudre ce problème, vous pouvez configurer une URL distincte qui sera uniquement utilisée pour indiquer au front-end (le navigateur de l'utilisateur) l'URL à utiliser pour accéder à Grafana.

Pour modifier l'URL renvoyée au front-end, exécutez la commande suivante :

```
cephuser@adm > ceph dashboard set-grafana-frontend-api-url GRAFANA-SERVER-URL
```

Si aucune valeur n'est définie pour cette option, elle sera simplement redéfinie sur la valeur de l'option `Grafana_API_URL`, laquelle est définie automatiquement et mise à jour périodiquement par cephadm. Si elle est définie, elle indiquera au navigateur d'utiliser cette URL pour accéder à Grafana.

16.5 Configuration du module Prometheus Manager

Le module Prometheus Manager est un module intégré à Ceph qui étend les fonctionnalités de Ceph. Le module lit les (méta)données de Ceph concernant son état et sa santé, et fournit à Prometheus les données (récupérées) dans un format exploitable.



Note

Le module Prometheus Manager doit être redémarré pour que les modifications apportées à la configuration soient appliquées.

16.5.1 Configuration de l'interface réseau

Par défaut, le module Prometheus Manager accepte les requêtes HTTP sur le port 9283 sur toutes les adresses IPv4 et IPv6 de l'hôte. Le port et l'adresse d'écoute peuvent être configurés avec `ceph config-key set`, avec les clés `mgr/prometheus/server_addr` et `mgr/prometheus/server_port`. Ce port est inscrit dans le registre de Prometheus.

Pour mettre à jour le paramètre `server_addr`, exécutez la commande suivante :

```
cephuser@adm > ceph config set mgr mgr/prometheus/server_addr 0.0.0.0
```

Pour mettre à jour le paramètre `server_port`, exécutez la commande suivante :

```
cephuser@adm > ceph config set mgr mgr/prometheus/server_port 9283
```

16.5.2 Configuration du paramètre `scrape_interval`

Par défaut, le module Prometheus Manager est configuré avec un intervalle de récupération de 15 secondes. Nous vous déconseillons d'utiliser un intervalle de récupération inférieur à 10 secondes. Pour configurer un intervalle de récupération différent dans le module Prometheus, définissez le paramètre `scrape_interval` sur la valeur souhaitée :



Important

Pour fonctionner correctement et ne causer aucun problème, le paramètre `scrape_interval` de ce module doit toujours être défini de manière à correspondre à l'intervalle de récupération de Prometheus.

```
cephuser@adm > ceph config set mgr mgr/prometheus/scrape_interval 15
```

16.5.3 Configuration du cache

Sur les grandes grappes (plus de 1 000 OSD), le temps nécessaire à la récupération des mesures peut devenir considérable. Sans le cache, le module Prometheus Manager peut surcharger le gestionnaire et causer l'arrêt ou l'absence de réponse des instances de Ceph Manager. Le cache est donc activé par défaut et ne peut pas être désactivé, mais cela signifie qu'il peut devenir obsolète. Le cache est considéré comme obsolète lorsque le temps nécessaire à la récupération des mesures à partir de Ceph dépasse l'intervalle de récupération (`scrape_interval`) configuré.

Dans ce cas-là, un avertissement est consigné et le module :

- Répond avec un code d'état HTTP 503 (service non disponible).
- Renvoie le contenu du cache, même s'il est obsolète.

Ce comportement peut être configuré à l'aide des commandes `ceph config set`.

Pour indiquer au module de répondre avec des données potentiellement obsolètes, configurez-le pour qu'il renvoie :

```
cephuser@adm > ceph config set mgr mgr/prometheus/stale_cache_strategy return
```

Pour indiquer au module de répondre avec un message de service non disponible, configurez-le pour échouer :

```
cephuser@adm > ceph config set mgr mgr/prometheus/stale_cache_strategy fail
```

16.5.4 Activation de la surveillance des images RBD

Le module Prometheus Manager peut éventuellement collecter des statistiques d'E/S par image RBD en activant des compteurs de performances d'OSD dynamiques. Les statistiques sont recueillies pour toutes les images figurant dans les réserves spécifiées dans le paramètre de configuration `mgr/prometheus/rbd_stats_pools`.

Le paramètre est une liste d'entrées `pool[/namespace]` séparées par des virgules ou des espaces. Si l'espace de noms n'est pas spécifié, les statistiques sont collectées pour tous les espaces de noms de la réserve.

Par exemple :

```
cephuser@adm > ceph config set mgr mgr/prometheus/rbd_stats_pools "pool1,pool2,poolN"
```

Le module analyse les réserves et espaces de noms spécifiés, établit une liste de toutes les images disponibles et la rafraîchit périodiquement. L'intervalle peut être configuré à l'aide du paramètre `mgr/prometheus/rbd_stats_pools_refresh_interval` (en secondes) et est de 300 secondes (cinq minutes) par défaut.

Par exemple, si vous avez modifié et défini l'intervalle de synchronisation sur 10 minutes :

```
cephuser@adm > ceph config set mgr mgr/prometheus/rbd_stats_pools_refresh_interval 600
```

16.6 Modèle de sécurité de Prometheus

Le modèle de sécurité de Prometheus part du principe que les utilisateurs non approuvés ont accès au noeud d'extrémité HTTP et aux journaux de Prometheus. Les utilisateurs non approuvés ont accès à l'ensemble des (méta)données collectées par Prometheus et contenues dans la base de données, ainsi qu'à diverses informations opérationnelles et de débogage.

Cependant, l'API HTTP de Prometheus est limitée aux opérations en lecture seule. Les configurations ne peuvent pas être modifiées à l'aide de l'API et les secrets ne sont pas exposés. Prometheus dispose par ailleurs de mesures intégrées pour limiter l'impact des attaques par déni de service.

16.7 Passerelle SNMP de Prometheus Alertmanager

Si vous voulez être informé des alertes Prometheus via des trappes SNMP, vous pouvez installer la passerelle SNMP de Prometheus Alertmanager via cephadm ou Ceph Dashboard. Pour effectuer cette opération pour SNMPv2c, par exemple, vous devez créer un fichier de spécification de service et de placement incluant le contenu suivant :



Note

Pour plus d'informations sur les fichiers de service et de placement, reportez-vous au Manuel « Guide de déploiement », Chapitre 8 « Déploiement des services essentiels restants à l'aide de cephadm », Section 8.2 « Spécification de service et de placement ».

```
service_type: snmp-gateway
service_name: snmp-gateway
placement:
  ADD_PLACEMENT_HERE
spec:
  credentials:
    snmp_community: ADD_COMMUNITY_STRING_HERE
    snmp_destination: ADD_FQDN_HERE:ADD_PORT_HERE
    snmp_version: V2c
```

Vous pouvez également utiliser Ceph Dashboard pour déployer le service de passerelle SNMP pour SNMPv2c et SNMPv3. Pour plus d'informations, reportez-vous à la [Section 4.4, « Affichage des services »](#).

III Stockage de données dans une grappe

- 17 Gestion des données stockées **163**
- 18 Gestion des réserves de stockage **196**
- 19 Réserves codées à effacement **219**
- 20 Périphérique de bloc RADOS **226**

17 Gestion des données stockées

L'algorithme CRUSH détermine comment stocker et récupérer des données en calculant les emplacements de stockage de données. CRUSH donne aux clients Ceph les moyens de communiquer directement avec les OSD plutôt que via un serveur ou un courtier centralisé. Grâce à une méthode algorithmique de stockage et de récupération des données, Ceph évite que son évolutivité soit entravée par un point de défaillance unique, un goulot d'étranglement des performances ou une limite physique.

CRUSH requiert une assignation de votre grappe et l'utilise pour stocker et récupérer de façon pseudo-aléatoire des données sur les OSD avec une distribution uniforme des données sur l'ensemble de la grappe.

Les cartes CRUSH contiennent une liste d'OSD, une liste de compartiments (« buckets ») pour l'agrégation des périphériques à des emplacements physiques et une liste de règles indiquant à CRUSH comment répliquer les données dans les réserves d'une grappe Ceph. En reflétant l'organisation physique sous-jacente de l'installation, CRUSH peut modéliser (et ainsi corriger) les sources potentielles de défaillances de périphériques corrélés. Les sources courantes incluent la proximité physique, une source d'alimentation partagée et un réseau partagé. En codant ces informations dans l'assignation de grappe, les stratégies de placement CRUSH peuvent séparer les répliques d'objet entre différents domaines de défaillance, tout en conservant la distribution souhaitée. Par exemple, pour prévoir le traitement de défaillances simultanées, il peut être souhaitable de s'assurer que les répliques de données se trouvent sur des périphériques utilisant des étagères, des racks, des alimentations électriques, des contrôleurs et/ou des emplacements physiques différents.

Une fois que vous avez déployé une grappe Ceph, une carte CRUSH par défaut est générée, ce qui est parfait pour votre environnement de sandbox Ceph. Cependant, lorsque vous déployez une grappe de données à grande échelle, vous devez envisager sérieusement de développer une carte CRUSH personnalisée, car elle vous aidera à gérer votre grappe Ceph, à améliorer les performances et à garantir la sécurité des données.

Par exemple, si un OSD tombe en panne, une carte CRUSH peut vous aider à localiser le centre de données physique, la salle, la rangée et le rack de l'hôte avec l'OSD défaillant dans le cas où vous auriez besoin d'une intervention sur site ou de remplacer le matériel.

De même, CRUSH peut vous aider à identifier les défaillances plus rapidement. Par exemple, si tous les OSD d'un rack particulier tombent en panne simultanément, la défaillance peut provenir d'un commutateur réseau ou de l'alimentation du rack, plutôt que des OSD eux-mêmes.

Une carte CRUSH personnalisée vous aide également à identifier les emplacements physiques où Ceph stocke des copies redondantes de données lorsque le ou les groupes de placement (voir [Section 17.4, « Groupes de placement »](#)) associés à un hôte défaillant se trouvent dans un état altéré.

Une carte CRUSH comporte trois sections principales.

- *Périphériques OSD* : comprend tous les périphériques de stockage d'objets correspondant à un daemon `ceph-osd`.
- *Compartiments* est une agrégation hiérarchique constituée d'emplacements de stockage (par exemple, des rangées, des racks, des hôtes, etc.) et de leurs pondérations assignées.
- *Ensembles de règles* : définit la manière de sélectionner les compartiments.

17.1 Périphériques OSD

Pour assigner des groupes de placement aux OSD, une carte CRUSH nécessite une liste de périphériques OSD (nom du daemon OSD). La liste des périphériques apparaît en premier dans la carte CRUSH.

```
#devices
device NUM osd.OSD_NAME class CLASS_NAME
```

Par exemple :

```
#devices
device 0 osd.0 class hdd
device 1 osd.1 class ssd
device 2 osd.2 class nvme
device 3 osd.3 class ssd
```

En règle générale, un daemon OSD est assigné à un seul disque.

17.1.1 Classes de périphériques

La flexibilité de la carte CRUSH pour le contrôle du placement de données est l'une des forces de Ceph. C'est aussi l'une des parties les plus difficiles à gérer de la grappe. Les *classes de périphériques* automatisent les modifications les plus courantes apportées aux cartes CRUSH que l'administrateur devait effectuer manuellement auparavant.

17.1.1.1 Problème de gestion de CRUSH

Les grappes Ceph sont souvent créées avec plusieurs types de périphériques de stockage : HDD, SSD, NVMe ou même des classes mixtes de ce qui précède. Nous appelons ces différents types de périphériques de stockage *classes de périphériques* pour éviter toute confusion entre la propriété *type* des compartiments CRUSH (par exemple, hôte, rack ou ligne ; voir [Section 17.2, « Compartiments »](#) pour plus de détails). Les Ceph OSD soutenus par des disques SSD sont beaucoup plus rapides que ceux s'appuyant sur des disques rotatifs, ce qui les rend plus appropriés pour certains workloads. Ceph facilite la création de réserves RADOS pour différents ensembles de données ou workloads, et l'assignation de règles CRUSH distinctes pour contrôler le placement de données pour ces réserves.

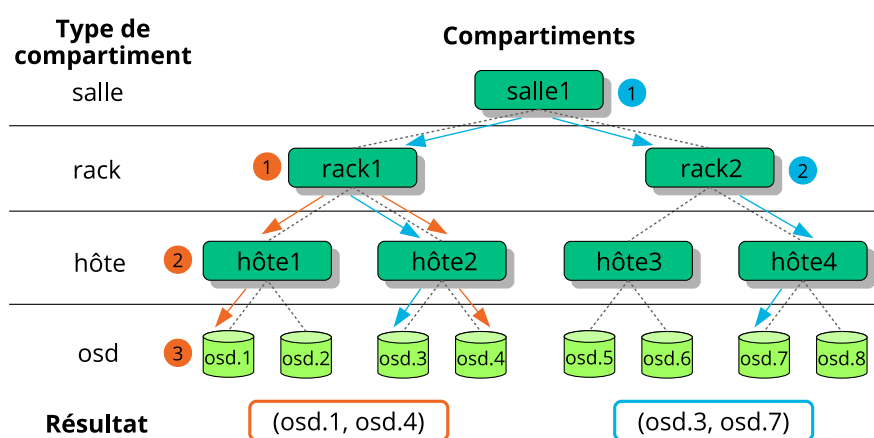


FIGURE 17.1 : OSD AVEC CLASSES DE PÉRIPHÉRIQUES MIXTES

Toutefois, la configuration de règles CRUSH pour placer les données uniquement sur une certaine classe de périphériques est fastidieuse. Les règles fonctionnent en termes de hiérarchie CRUSH, mais si les périphériques sont mélangés sur des hôtes ou racks identiques (comme dans l'exemple de hiérarchie ci-dessus), ils seront (par défaut) mélangés et apparaîtront dans les mêmes sous-arborescences de la hiérarchie. Les séparer manuellement dans des arborescences distinctes impliquait la création de plusieurs versions de chaque noeud intermédiaire pour chaque classe de périphériques dans les versions précédentes de SUSE Enterprise Storage.

17.1.1.2 Classes de périphériques

Une solution élégante proposée par Ceph consiste à ajouter une propriété appelée *device class* (classe de périphériques) à chaque OSD. Par défaut, les OSD définissent automatiquement leur classe de périphériques sur « hdd », « ssd » ou « nvme » en fonction des propriétés matérielles exposées par le kernel Linux. Ces classes de périphériques sont répertoriées dans une nouvelle colonne de la sortie de la commande `ceph osd tree` :

```
cephuser@adm > ceph osd tree
ID CLASS WEIGHT  TYPE NAME        STATUS REWEIGHT PRI-AFF
-1      83.17899 root default
-4      23.86200   host cpach
 2  hdd  1.81898     osd.2      up  1.00000 1.00000
 3  hdd  1.81898     osd.3      up  1.00000 1.00000
 4  hdd  1.81898     osd.4      up  1.00000 1.00000
 5  hdd  1.81898     osd.5      up  1.00000 1.00000
 6  hdd  1.81898     osd.6      up  1.00000 1.00000
 7  hdd  1.81898     osd.7      up  1.00000 1.00000
 8  hdd  1.81898     osd.8      up  1.00000 1.00000
15  hdd  1.81898     osd.15     up  1.00000 1.00000
10  nvme 0.93100     osd.10     up  1.00000 1.00000
 0  ssd  0.93100     osd.0      up  1.00000 1.00000
 9  ssd  0.93100     osd.9      up  1.00000 1.00000
```

Si la détection automatique de la classe de périphériques échoue, par exemple parce que le pilote du périphérique n'expose pas correctement les informations sur ce dernier via `/sys/block`, vous pouvez ajuster les classes de périphériques à partir de la ligne de commande :

```
cephuser@adm > ceph osd crush rm-device-class osd.2 osd.3
done removing class of osd(s): 2,3
cephuser@adm > ceph osd crush set-device-class ssd osd.2 osd.3
set osd(s) 2,3 to class 'ssd'
```

17.1.1.3 Définition des règles de placement CRUSH

Les règles CRUSH peuvent limiter le placement à une classe de périphériques spécifique. Par exemple, vous pouvez créer une réserve **répliquée** « fast » qui distribue des données uniquement sur les disques SSD en exécutant la commande suivante :

```
cephuser@adm > ceph osd crush rule create-
replicated RULE_NAME ROOT FAILURE_DOMAIN_TYPE DEVICE_CLASS
```

Par exemple :

```
cephuser@adm > ceph osd crush rule create-replicated fast default host ssd
```

Créez une réserve nommée « fast_pool » et assignez-la à la règle « fast » :

```
cephuser@adm > ceph osd pool create fast_pool 128 128 replicated fast
```

Le processus de création des règles de **code à effacement** est légèrement différent. Tout d'abord, vous créez un profil de code à effacement qui inclut une propriété pour votre classe de périphériques souhaitée. Ensuite, utilisez ce profil lors de la création de la réserve codée à effacement :

```
cephuser@adm > ceph osd erasure-code-profile set myprofile \
k=4 m=2 crush-device-class=ssd crush-failure-domain=host
cephuser@adm > ceph osd pool create mypool 64 erasure myprofile
```

Si vous devez modifier manuellement la carte CRUSH pour personnaliser votre règle, la syntaxe a été étendue de sorte à permettre de spécifier la classe de périphériques. Par exemple, la règle CRUSH générée par les commandes ci-dessus ressemble à ceci :

```
rule ecpool {
  id 2
  type erasure
  min_size 3
  max_size 6
  step set_chooseleaf_tries 5
  step set_choose_tries 100
  step take default class ssd
  step chooseleaf indep 0 type host
  step emit
}
```

La différence importante ici est que la commande « take » inclut le suffixe supplémentaire de « classe CLASS_NAME ».

17.1.1.4 Commandes supplémentaires

Pour répertorier les classes de périphériques utilisées dans une carte CRUSH, exécutez :

```
cephuser@adm > ceph osd crush class ls
[
  "hdd",
```

```
"ssd"  
]
```

Pour répertorier les règles CRUSH existantes, exécutez :

```
cephuser@adm > ceph osd crush rule ls  
replicated_rule  
fast
```

Pour afficher les détails de la règle CRUSH nommée « fast », exécutez :

```
cephuser@adm > ceph osd crush rule dump fast  
{  
  "rule_id": 1,  
  "rule_name": "fast",  
  "ruleset": 1,  
  "type": 1,  
  "min_size": 1,  
  "max_size": 10,  
  "steps": [  
    {  
      "op": "take",  
      "item": -21,  
      "item_name": "default~ssd"  
    },  
    {  
      "op": "chooseleaf_firstn",  
      "num": 0,  
      "type": "host"  
    },  
    {  
      "op": "emit"  
    }  
  ]  
}
```

Pour répertorier les OSD appartenant à une classe « ssd », exécutez :

```
cephuser@adm > ceph osd crush class ls-osd ssd  
0  
1
```

17.1.1.5 Migration d'une règle SSD héritée vers des classes de périphériques

Dans une version SUSE Enterprise Storage antérieure à la version 5, vous deviez modifier manuellement la carte CRUSH et maintenir une hiérarchie parallèle pour chaque type de périphérique spécialisé (comme SSD) afin d'écrire des règles qui s'appliquent à ces appareils. Depuis SUSE Enterprise Storage 5, la fonction de classe de périphériques permet d'effectuer cette opération en toute transparence.

Vous pouvez transformer une règle et une hiérarchie héritées en nouvelles règles basées sur la classe à l'aide de la commande **crushtool**. Plusieurs types de transformation sont possibles :

crushtool --reclassify-root *ROOT_NAME* *DEVICE_CLASS*

Cette commande prend tout ce qui se trouve dans la hiérarchie sous *ROOT_NAME* et ajuste toutes les règles qui font référence à cette racine via

```
take ROOT_NAME
```

vers

```
take ROOT_NAME class DEVICE_CLASS
```

Elle réattribue des numéros aux compartiments de sorte que les anciens ID sont utilisés pour l'arborescence fantôme (« shadow tree ») de la classe spécifiée. Par conséquent, aucun mouvement de données ne se produit.

EXEMPLE 17.1 : **crushtool --reclassify-root**

Considérez la règle existante suivante :

```
rule replicated_ruleset {
  id 0
  type replicated
  min_size 1
  max_size 10
  step take default
  step chooseleaf firstn 0 type rack
  step emit
}
```

Si vous reclassez la racine « default » en tant que classe « hdd », la règle devient la suivante :

```
rule replicated_ruleset {
  id 0
  type replicated
```

```

    min_size 1
    max_size 10
    step take default class hdd
    step chooseleaf firstn 0 type rack
    step emit
}

```

crushtool --set-subtree-class *BUCKET_NAME* *DEVICE_CLASS*

Cette méthode marque chaque périphérique de la sous-arborescence associée à la racine *BUCKET_NAME* avec la classe de périphériques spécifiée.

`--set-subtree-class` est normalement utilisé avec l'option `--reclassify-root` pour garantir que tous les périphériques de cette racine sont étiquetés avec la bonne classe. Cependant, certains de ces périphériques peuvent volontairement avoir une classe différente et vous ne souhaitez donc pas changer leur étiquette. Dans de tels cas, excluez l'option `--set-subtree-class`. Gardez à l'esprit que ce type de réaffectation n'est pas parfait, car la règle précédente est distribuée entre des périphériques de différentes classes, tandis que les règles ajustées seront uniquement assignées aux périphériques de la classe spécifiée.

crushtool --reclassify-bucket *MATCH_PATTERN* *DEVICE_CLASS* *DEFAULT_PATTERN*

Cette méthode permet de fusionner une hiérarchie spécifique à un type parallèle avec la hiérarchie normale. Par exemple, de nombreux utilisateurs possèdent des cartes CRUSH similaires à la suivante :

EXEMPLE 17.2 : **crushtool --reclassify-bucket**

```

host nodel {
    id -2          # do not change unnecessarily
    # weight 109.152
    alg straw
    hash 0 # rjenkins1
    item osd.0 weight 9.096
    item osd.1 weight 9.096
    item osd.2 weight 9.096
    item osd.3 weight 9.096
    item osd.4 weight 9.096
    item osd.5 weight 9.096
    [...]
}

host nodel-ssd {
    id -10         # do not change unnecessarily
    # weight 2.000
    alg straw
    hash 0 # rjenkins1
}

```

```

    item osd.80 weight 2.000
    [...]
}

root default {
    id -1          # do not change unnecessarily
    alg straw
    hash 0 # rjenkins1
    item node1 weight 110.967
    [...]
}

root ssd {
    id -18        # do not change unnecessarily
    # weight 16.000
    alg straw
    hash 0 # rjenkins1
    item node1-ssd weight 2.000
    [...]
}

```

Cette fonction reclasse chaque compartiment qui correspond à un modèle donné. Le modèle peut ressembler à `%suffix` ou `prefix%`. Dans l'exemple ci-dessus, vous utiliseriez le modèle `%-ssd`. Pour chaque compartiment correspondant, la partie restante du nom qui est représentée par le caractère joker « % » spécifie le compartiment de base. Tous les périphériques du compartiment correspondant sont étiquetés avec la classe de périphériques spécifiée, puis déplacés vers le compartiment de base. Si le compartiment de base n'existe pas (par exemple, si « node12-ssd » existe, mais pas « node12 »), il est créé et lié sous le compartiment parent par défaut spécifié. Les anciens ID de compartiment sont conservés pour les nouveaux compartiments fantômes afin d'empêcher le mouvement de données. Les règles avec des étapes `take` qui font référence aux anciens compartiments sont ajustées.

`crushtool --reclassify-bucket` *BUCKET_NAME* *DEVICE_CLASS* *BASE_BUCKET*

Vous pouvez utiliser l'option `--reclassify-bucket` sans caractère joker pour assigner un compartiment unique. Par exemple, dans l'exemple précédent, nous voulons que le compartiment « ssd » soit assigné au compartiment par défaut.

La commande finale pour convertir l'assignation composée des fragments ci-dessus serait la suivante :

```

cephuser@adm > ceph osd getcrushmap -o original
cephuser@adm > crushtool -i original --reclassify \
    --set-subtree-class default hdd \
    --reclassify-root default hdd \

```

```
--reclassify-bucket %-ssd ssd default \  
--reclassify-bucket ssd ssd default \  
-o adjusted
```

Afin de vérifier que la conversion est correcte, il existe une option `--compare` qui teste un grand échantillon d'entrées dans la carte CRUSH et compare si le même résultat revient. Ces entrées sont contrôlées par les mêmes options que celles qui s'appliquent à `--test`. Pour l'exemple ci-dessus, la commande se présenterait comme suit :

```
cephuser@adm > crushtool -i original --compare adjusted  
rule 0 had 0/10240 mismatched mappings (0)  
rule 1 had 0/10240 mismatched mappings (0)  
maps appear equivalent
```



Astuce

S'il existait des différences, vous verriez le taux d'entrées réassignées dans les parenthèses.

Si vous êtes satisfait de la carte CRUSH ajustée, vous pouvez l'appliquer à la grappe :

```
cephuser@adm > ceph osd setcrushmap -i adjusted
```

17.1.1.6 Informations supplémentaires

Pour plus de détails sur les cartes CRUSH, reportez-vous à la [Section 17.5, « Manipulation de la carte CRUSH »](#).

Pour plus de détails sur les réserves Ceph en général, reportez-vous au [Chapitre 18, Gestion des réserves de stockage](#).

Pour plus de détails sur les réserves codées à effacement, reportez-vous au [Chapitre 19, Réserves codées à effacement](#).

17.2 Compartiments

Les cartes CRUSH contiennent une liste d'OSD pouvant être organisée en une arborescence de « compartiments » afin d'agréger les périphériques dans des emplacements physiques. Les OSD individuels comprennent les feuilles de l'arborescence.

| | | |
|----|------------|--|
| 0 | osd | Périphérique ou OSD spécifique (<u>osd . 1</u> , <u>osd . 2</u> , etc). |
| 1 | host | Nom d'un hôte contenant un ou plusieurs OSD. |
| 2 | chassis | Identificateur du châssis du rack contenant l' <u>hôte</u> . |
| 3 | rack | Rack d'un ordinateur. La valeur par défaut est <u>unknown - rack</u> . |
| 4 | row | Rangée dans une série de racks. |
| 5 | pdu | Abréviation de « Power Distribution Unit » (unité de distribution de l'alimentation). |
| 6 | pod | Abréviation de « Point of Delivery » (point de livraison) : dans ce contexte, groupe de PDU ou groupe de rangées de racks. |
| 7 | room | Salle contenant des rangées de racks. |
| 8 | datacenter | Centre de données physique comprenant une ou plusieurs salles. |
| 9 | region | Région géographique du monde (par exemple, NAM, LAM, EMEA, APAC, etc) |
| 10 | root | Noeud racine de l'arborescence des compartiments OSD (normalement défini sur la valeur <u>default</u>). |



Astuce

Vous pouvez modifier les types existants et créer vos propres types de compartiment.

Les outils de déploiement de Ceph génèrent une carte CRUSH contenant un compartiment pour chaque hôte et une racine nommée « default », qui est utile pour la réserve rbd par défaut. Les types de compartiment restants permettent de stocker des informations sur l'emplacement physique des noeuds/compartiments, ce qui facilite grandement l'administration des grappes lorsque des OSD, des hôtes ou le matériel réseau sont défectueux et que l'administrateur doit accéder au matériel physique.

Chaque compartiment possède un type, un nom unique (chaîne), un identifiant unique exprimé en tant que nombre entier négatif, une pondération par rapport à la capacité totale de son ou ses éléments, l'algorithme de compartiment (straw2 par défaut) et le hachage (0 par défaut, reflet du hachage CRUSH rjenkins1). Un compartiment peut contenir un ou plusieurs éléments. Les éléments peuvent être constitués d'autres compartiments ou OSD. Les éléments peuvent posséder une pondération relative les uns par rapport aux autres.

```
[bucket-type] [bucket-name] {  
  id [a unique negative numeric ID]  
  weight [the relative capacity/capability of the item(s)]  
  alg [the bucket type: uniform | list | tree | straw2 | straw ]  
  hash [the hash type: 0 by default]  
  item [item-name] weight [weight]  
}
```

L'exemple suivant illustre la façon dont vous pouvez utiliser des compartiments pour agréger une réserve et des emplacements physiques, tels qu'un centre de données, une salle, un rack et une rangée.

```
host ceph-osd-server-1 {  
  id -17  
  alg straw2  
  hash 0  
  item osd.0 weight 0.546  
  item osd.1 weight 0.546  
}  
  
row rack-1-row-1 {  
  id -16  
  alg straw2  
  hash 0  
  item ceph-osd-server-1 weight 2.00  
}  
  
rack rack-3 {  
  id -15  
  alg straw2  
  hash 0  
  item rack-3-row-1 weight 2.00  
  item rack-3-row-2 weight 2.00  
  item rack-3-row-3 weight 2.00  
  item rack-3-row-4 weight 2.00  
  item rack-3-row-5 weight 2.00  
}
```

```

rack rack-2 {
    id -14
    alg straw2
    hash 0
    item rack-2-row-1 weight 2.00
    item rack-2-row-2 weight 2.00
    item rack-2-row-3 weight 2.00
    item rack-2-row-4 weight 2.00
    item rack-2-row-5 weight 2.00
}

rack rack-1 {
    id -13
    alg straw2
    hash 0
    item rack-1-row-1 weight 2.00
    item rack-1-row-2 weight 2.00
    item rack-1-row-3 weight 2.00
    item rack-1-row-4 weight 2.00
    item rack-1-row-5 weight 2.00
}

room server-room-1 {
    id -12
    alg straw2
    hash 0
    item rack-1 weight 10.00
    item rack-2 weight 10.00
    item rack-3 weight 10.00
}

datacenter dc-1 {
    id -11
    alg straw2
    hash 0
    item server-room-1 weight 30.00
    item server-room-2 weight 30.00
}

root data {
    id -10
    alg straw2
    hash 0
    item dc-1 weight 60.00
    item dc-2 weight 60.00
}

```

17.3 Ensembles de règles

Les cartes CRUSH prennent en charge la notion de « règles CRUSH », lesquelles déterminent le placement des données dans une réserve. Pour les grappes vastes, vous pouvez créer un grand nombre de réserves dans lesquelles chaque réserve peut avoir son propre ensemble de règles ou ses propres règles CRUSH. La carte CRUSH par défaut a une règle pour la racine par défaut. Si vous voulez plus de racines et plus de règles, vous devez les créer plus tard ou elles seront créées automatiquement lors de la création de réserves.



Note

Dans la plupart des cas, vous n'avez pas besoin de modifier les règles par défaut. Lorsque vous créez une réserve, son ensemble de règles par défaut est 0.

Une règle est définie selon le format suivant :

```
rule rulename {  
  
    ruleset ruleset  
    type type  
    min_size min-size  
    max_size max-size  
    step step  
  
}
```

ruleset

Nombre entier. Classifie une règle en tant que membre d'un ensemble de règles. Option activée en définissant l'ensemble de règles dans une réserve. Elle est obligatoire. La valeur par défaut est 0.

type

Chaîne. Décrit une règle pour une réserve codée « replicated » (répliqué) ou « erasure » (effacement). Cette option est obligatoire. La valeur par défaut est replicated.

min_size

Nombre entier. Si un groupe de réserves produit moins de répliques que ce nombre, CRUSH ne sélectionne PAS cette règle. Cette option est obligatoire. La valeur par défaut est 2.

max_size

Nombre entier. Si un groupe de réserves produit plus de répliques que ce nombre, CRUSH ne sélectionne PAS cette règle. Cette option est obligatoire. La valeur par défaut est 10.

step take *compartiment*

Récupère un compartiment spécifié par un nom, puis commence à effectuer une itération en profondeur dans l'arborescence. Cette option est obligatoire. Pour en savoir plus sur l'itération dans l'arborescence, consultez la [Section 17.3.1, « Itération de l'arborescence de noeuds »](#).

step *cible* *modenum* **type** *type-compartiment*

cible peut être choose ou chooseleaf. Lorsque la valeur est définie sur choose, un nombre de compartiments est sélectionné. chooseleaf sélectionne directement les OSD (noeuds feuilles) dans la sous-arborescence de chaque compartiment dans l'ensemble des compartiments.

mode peut être firstn ou indep. Reportez-vous à la [Section 17.3.2, « firstn et indep »](#).

Sélectionne le nombre de compartiments du type donné. Où N correspond au nombre d'options disponible, si *num* > 0 && < N, choisissez autant de compartiments ; si *num* < 0, cela signifie N - *num* ; et si *num* == 0, choisissez N compartiments (tous disponibles). Suit step take ou step choose.

step emit

Affiche la valeur actuelle et vide la pile. Figure généralement à la fin d'une règle, mais permet également de former des arborescences différentes dans la même règle. Suit step choose.

17.3.1 Itération de l'arborescence de noeuds

La structure des compartiments peut être considérée comme une arborescence de noeuds. Les compartiments sont des noeuds et les OSD sont les feuilles de cette arborescence.

Les règles de la carte CRUSH définissent la façon dont les OSD sont sélectionnés dans cette arborescence. Une règle commence par un noeud, puis réalise une itération dans l'arborescence pour renvoyer un ensemble d'OSD. Il n'est pas possible de définir quelle branche doit être sélectionnée. Au lieu de cela, l'algorithme CRUSH garantit que l'ensemble des OSD remplit les conditions de réplication et répartit équitablement les données.

Avec `step take compartiment`, l'itération dans l'arborescence des noeuds commence à partir du compartiment donné (et non pas du type de compartiment). Pour que les OSD de toutes les branches de l'arborescence puissent être renvoyés, le compartiment doit être le compartiment racine. Dans le cas contraire, l'itération se poursuit simplement dans une sous-arborescence.

Après `step take`, une ou plusieurs entrées `step choose` figurent dans la définition de la règle. Chaque `step choose` choisit un nombre défini de noeuds (ou de branches) dans le noeud supérieur précédemment sélectionné.

À la fin de l'itération, les OSD sélectionnés sont renvoyés avec `step emit`.

`step chooseleaf` est une fonction pratique qui sélectionne les OSD directement dans les branches du compartiment donné.

La [Figure 17.2, « Exemple d'arborescence »](#) illustre la façon dont `step` permet d'effectuer un traitement itératif dans une arborescence. Les flèches et les chiffres orange correspondent à `exemple1a` et `exemple1b`, tandis que la couleur bleue est associée à `exemple2` dans les définitions de règles suivantes.

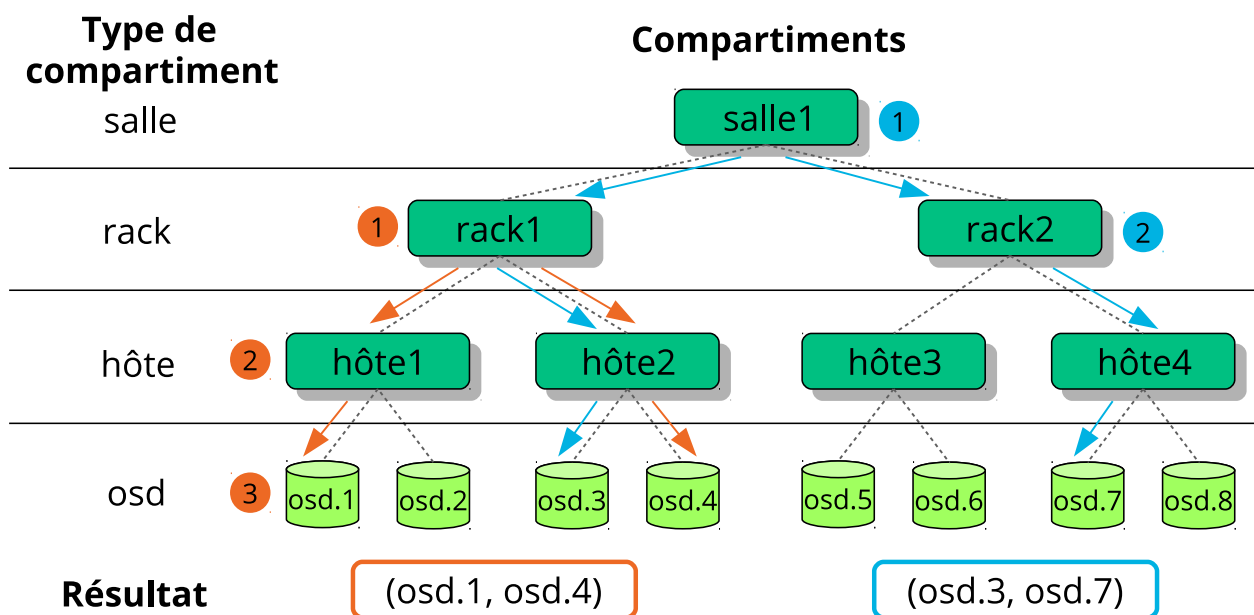


FIGURE 17.2 : EXEMPLE D'ARBORESCENCE

```
# orange arrows
rule exemple1a {
  ruleset 0
  type replicated
  min_size 2
  max_size 10
  # orange (1)
```

```

    step take rack1
    # orange (2)
    step choose firstn 0 host
    # orange (3)
    step choose firstn 1 osd
    step emit
}

rule example1b {
    ruleset 0
    type replicated
    min_size 2
    max_size 10
    # orange (1)
    step take rack1
    # orange (2) + (3)
    step chooseleaf firstn 0 host
    step emit
}

# blue arrows
rule example2 {
    ruleset 0
    type replicated
    min_size 2
    max_size 10
    # blue (1)
    step take room1
    # blue (2)
    step chooseleaf firstn 0 rack
    step emit
}

```

17.3.2 firstn et indep

Une règle CRUSH définit les remplacements des noeuds ou des OSD défaillants (voir [Section 17.3, « Ensembles de règles »](#)). Le mot clé `step` nécessite le paramètre `firstn` ou le paramètre `indep`. La [Figure 17.3, « Méthodes de remplacement de noeud »](#) fournit un exemple.

`firstn` ajoute des noeuds de remplacement à la fin de la liste des noeuds actifs. Dans le cas d'un noeud défaillant, les noeuds sains suivants sont décalés vers la gauche afin de combler l'espace laissé vacant par le noeud défaillant. Il s'agit de la méthode par défaut souhaitée pour les *réserves répliquées*, car un noeud secondaire possède déjà toutes les données et peut donc prendre immédiatement en charge les tâches du noeud principal.

indep sélectionne des noeuds de remplacement fixes pour chaque noeud actif. Le remplacement d'un noeud défaillant ne modifie pas l'ordre des noeuds restants. Cette approche est souhaitée pour les *réserves codées à effacement*. Dans les réserves codées à effacement, les données stockées sur un noeud dépendent de la position de celui-ci dans la sélection des noeuds. En cas de modification de l'ordre des noeuds, toutes les données des noeuds affectés doivent être déplacées.

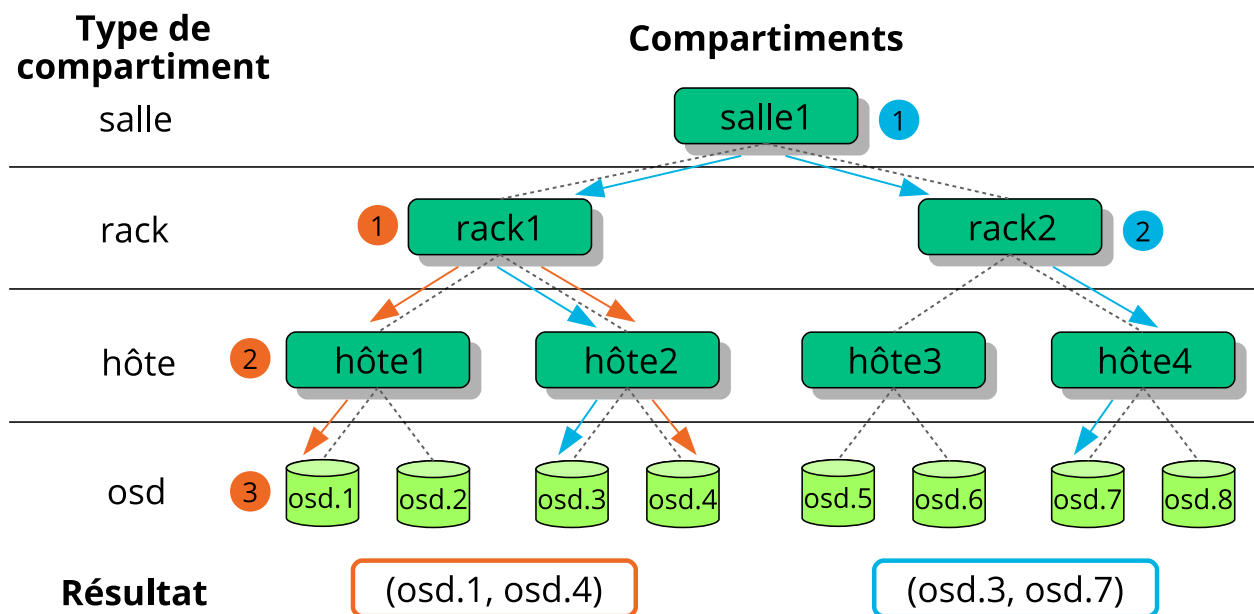


FIGURE 17.3 : MÉTHODES DE REMPLACEMENT DE NOEUD

17.4 Groupes de placement

Ceph assigne les objets aux groupes de placement (PG). Les groupes de placement sont des partitions ou des fragments d'une réserve d'objets logique qui placent les objets en tant que groupe dans des OSD. Les groupes de placement réduisent la quantité de métadonnées par objet lorsque Ceph stocke les données dans les OSD. Un plus grand nombre de groupes de placement, par exemple, 100 par OSD, permet un meilleur équilibrage.

17.4.1 Utilisation de groupes de placement

Un groupe de placement (PG) regroupe des objets au sein d'une réserve. La principale raison est que le fait que le suivi du placement des objets et des métadonnées sur une base « par objet » est coûteux en termes de calcul. Par exemple, un système avec des millions d'objets ne peut pas suivre directement le placement de chacun de ses objets.

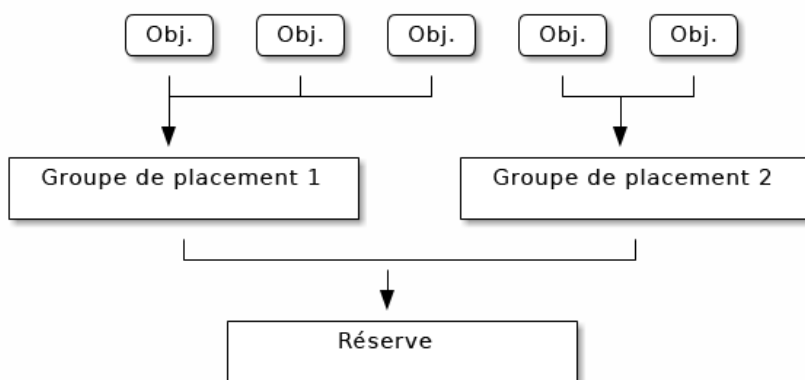


FIGURE 17.4 : GROUPES DE PLACEMENT D'UNE RÉSERVE

Le client Ceph calcule à quel groupe de placement un objet appartient. Pour ce faire, il hache l'ID d'objet et applique une opération basée sur le nombre de groupes de placement dans la réserve définie et l'ID de cette dernière.

Le contenu de l'objet au sein d'un groupe de placement est stocké dans un ensemble d'OSD. Par exemple, dans une réserve répliquée de taille deux, chaque groupe de placement stocke des objets sur deux OSD :

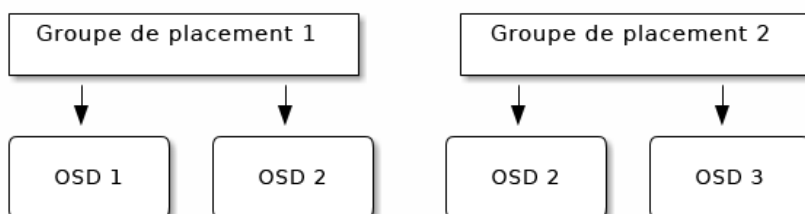


FIGURE 17.5 : GROUPES DE PLACEMENT ET OSD

Si l'OSD 2 échoue, un autre OSD est assigné au groupe de placement 1 et est rempli avec des copies de tous les objets de l'OSD 1. Si la taille de la réserve est modifiée de deux à trois, un OSD supplémentaire est assigné au groupe de placement et reçoit des copies de tous les objets du groupe de placement.

Les groupes de placement ne sont pas propriétaires de l'OSD, ils le partagent avec d'autres groupes de placement de la même réserve, voire avec d'autres réserves. Si l'OSD 2 échoue, le groupe de placement 2 devra également restaurer des copies d'objets à l'aide de l'OSD 3.

Lorsque le nombre de groupes de placement augmente, les nouveaux groupes de placement se voient assigner des OSD. Le résultat de la fonction CRUSH change également et certains objets des anciens groupes de placement sont copiés dans les nouveaux groupes de placement et retirés des anciens.

17.4.2 Détermination de la valeur de `PG_NUM`



Note

Depuis Ceph Nautilus (v14.x), vous pouvez utiliser le module `pg_autoscaler` de Ceph Manager pour mettre à l'échelle automatiquement les groupes de placement si nécessaire. Si vous souhaitez activer cette fonction, reportez-vous au *Manuel « Deploying and Administering SUSE Enterprise Storage with Rook »*, Chapitre 8 « Configuration », Section 8.1.1.1 « Default PG and PGP counts ».

Lorsque vous créez une nouvelle réserve, vous pouvez toujours choisir la valeur de `PG_NUM` manuellement :

```
# ceph osd pool create POOL_NAME PG_NUM
```

`PG_NUM` ne peut pas être calculé automatiquement. Voici quelques valeurs couramment utilisées, en fonction du nombre d'OSD dans la grappe :

Moins de 5 OSD :

définissez `PG_NUM` sur 128.

Entre 5 et 10 OSD :

définissez `PG_NUM` sur 512.

Entre 10 et 50 OSD :

définissez `PG_NUM` sur 1024.

Plus le nombre d'OSD est élevé, plus il est important de bien choisir la valeur de `PG_NUM`. `PG_NUM` influence fortement le comportement de la grappe ainsi que la durabilité des données en cas de défaillance d'OSD.

17.4.2.1 Calcul du nombre de groupes de placement pour plus de 50 OSD

Si vous avez moins de 50 OSD, utilisez la présélection décrite à la [Section 17.4.2, « Détermination de la valeur de PG_NUM »](#). Si vous avez plus de 50 OSD, nous recommandons environ 50 à 100 groupes de placement par OSD pour équilibrer l'utilisation des ressources, la durabilité des données et la distribution. Pour une seule réserve d'objets, vous pouvez utiliser la formule suivante afin d'obtenir une base de référence :

```
total PGs = (OSDs * 100) / POOL_SIZE
```

`POOL_SIZE` représente soit le nombre de répliques pour les réserves répliquées, soit la somme « k » + « m » pour les réserves codées à effacement, en fonction du retour de la commande **ceph osd erasure-code-profile get**. Vous devez arrondir le résultat à la puissance de 2 la plus proche. L'arrondissement est recommandé pour l'algorithme CRUSH afin d'équilibrer uniformément le nombre d'objets entre les groupes de placement.

Par exemple, pour une grappe avec 200 OSD et une taille de réserve de 3 répliques, vous estimez le nombre de groupes de placement comme suit :

```
(200 * 100) / 3 = 6667
```

La puissance de 2 la plus proche est **8192**.

Lorsque vous utilisez plusieurs réserves de données pour stocker des objets, vous devez veiller à équilibrer le nombre de groupes de placement par réserve avec le nombre de groupes de placement par OSD. Vous devez parvenir à un nombre total raisonnable de groupes de placement qui varie suffisamment peu par OSD, sans surcharger les ressources système ni rendre le processus d'homologation trop lent.

Par exemple, une grappe de 10 réserves, dont chacune comporte 512 groupes de placement sur 10 OSD, représente un total de 5 120 groupes de placement répartis sur 10 OSD, soit 512 groupes de placement par OSD. Une telle configuration n'utilise pas trop de ressources. Toutefois, si 1 000 réserves étaient créées avec 512 groupes de placement chacune, les OSD gèreraient environ 50 000 groupes de placement chacun, ce qui nécessiterait beaucoup plus de ressources et de temps pour l'homologation.

17.4.3 Définition du nombre de groupes de placement



Note

Depuis Ceph Nautilus (v14.x), vous pouvez utiliser le module `pg_autoscaler` de Ceph Manager pour mettre à l'échelle automatiquement les groupes de placement si nécessaire. Si vous souhaitez activer cette fonction, reportez-vous au *Manuel « Deploying and Administering SUSE Enterprise Storage with Rook »*, Chapitre 8 « Configuration », Section 8.1.1.1 « Default PG and PGP counts ».

Si vous devez encore spécifier manuellement le nombre de groupes de placement dans une réserve, vous devez le faire au moment de la création de la réserve (reportez-vous à la [Section 18.1, « Création d'une réserve »](#)). Une fois que vous avez établi des groupes de placement pour une réserve, vous pouvez augmenter leur nombre à l'aide de la commande suivante :

```
# ceph osd pool set POOL_NAME pg_num PG_NUM
```

Après avoir augmenté le nombre de groupes de placement, vous devez également accroître le nombre de groupes de placement pour le placement (`PGP_NUM`) avant que votre grappe ne se rééquilibre. `PGP_NUM` correspond au nombre de groupes de placement qui seront pris en compte pour le placement par l'algorithme CRUSH. L'augmentation de `PG_NUM` divise les groupes de placement, mais les données ne sont migrées vers les groupes de placement plus récents qu'une fois `PG_NUM` augmenté. `PGP_NUM` doit être égal à `PG_NUM`. Pour augmenter le nombre de groupes de placement pour le placement, exécutez la commande suivante :

```
# ceph osd pool set POOL_NAME pgp_num PGP_NUM
```

17.4.4 Détermination du nombre de groupes de placement

Pour connaître le nombre de groupes de placement dans une réserve, exécutez la commande `get` suivante :

```
# ceph osd pool get POOL_NAME pg_num
```

17.4.5 Détermination des statistiques relatives aux groupes de placement d'une grappe

Pour connaître les statistiques relatives aux groupes de placement de votre grappe, exécutez la commande suivante :

```
# ceph pg dump [--format FORMAT]
```

Les formats valides sont « plain » (brut, valeur par défaut) et « json ».

17.4.6 Détermination des statistiques relatives aux groupes de placement bloqués

Pour déterminer les statistiques relatives à tous les groupes de placement bloqués dans un état donné, exécutez la commande suivante :

```
# ceph pg dump_stuck STATE \  
  [--format FORMAT] [--threshold THRESHOLD]
```

STATE correspond à l'une des valeurs suivantes : « inactive » (inactif - les groupes de placement ne peuvent pas traiter les lectures ou les écritures, car ils attendent qu'un OSD disposant des données les plus à jour soit opérationnel), « unclean » (impropre - les groupes de placement contiennent des objets qui ne sont pas répliqués le nombre de fois souhaité), « stale » (obsolète - les groupes de placement sont dans un état inconnu ; les OSD qui les hébergent n'ont pas rendu de compte à la grappe depuis un certain temps spécifié par l'option `mon_osd_report_timeout`), « undersized » (de taille insuffisante) ou « degraded » (altéré).

Les formats valides sont « plain » (brut, valeur par défaut) et « json ».

Le seuil définit le nombre minimum de secondes pendant lesquelles le groupe de placement doit être bloqué avant qu'il soit inclus dans les statistiques renvoyées (300 secondes par défaut).

17.4.7 Recherche d'une assignation de groupe de placement

Pour rechercher l'assignation d'un groupe de placement particulier, exécutez la commande suivante :

```
# ceph pg map PG_ID
```

Ceph renvoie alors l'assignation du groupe de placement, le groupe de placement et le statut OSD :

```
# ceph pg map 1.6c
osdmap e13 pg 1.6c (1.6c) -> up [1,0] acting [1,0]
```

17.4.8 Récupération des statistiques d'un groupe de placement

Pour récupérer les statistiques d'un groupe de placement particulier, exécutez la commande suivante :

```
# ceph pg PG_ID query
```

17.4.9 Nettoyage d'un groupe de placement

Pour nettoyer (*Section 17.6, « Nettoyage des groupes de placement »*) un groupe de placement, exécutez la commande suivante :

```
# ceph pg scrub PG_ID
```

Ceph vérifie les noeuds primaires et des répliques, génère un catalogue de tous les objets du groupe de placement et les compare pour s'assurer qu'aucun objet n'est manquant ou discordant et que son contenu est cohérent. Si toutes les répliques correspondent, un balayage sémantique final garantit que toutes les métadonnées d'objets associées à l'instantané sont cohérentes. Les erreurs sont signalées via les journaux.

17.4.10 Définition de priorités pour le renvoi et la récupération des groupes de placement

Vous pouvez vous retrouver dans une situation où plusieurs groupes de placement nécessitent une récupération et/ou un renvoi, alors que certains groupes hébergent des données plus importantes que celles d'autres groupes. Par exemple, vous pouvez avoir des groupes de placement qui contiennent des données pour des images utilisées par les machines en cours d'exécution, tandis que d'autres groupes de placement peuvent être utilisés par des machines inactives ou héberger des données moins essentielles. Dans ce cas, vous pouvez donner la priorité à la récupération des

groupes plus critiques afin que les performances et la disponibilité des données stockées sur ces groupes soient restaurées plus rapidement. Pour marquer des groupes de placement particuliers comme prioritaires lors du renvoi ou de la récupération, exécutez la commande suivante :

```
# ceph pg force-recovery PG_ID1 [PG_ID2 ... ]  
# ceph pg force-backfill PG_ID1 [PG_ID2 ... ]
```

De cette façon, Ceph effectuera la récupération ou le renvoi sur les groupes de placement spécifiés d'abord, avant de poursuivre avec d'autres groupes de placement. Cela n'interrompt pas les renvois ou les récupérations en cours, mais permet que des groupes de placement spécifiés soient traités dès que possible. Si vous changez d'avis ou si vous avez mal défini les groupes prioritaires, annulez la définition des priorités :

```
# ceph pg cancel-force-recovery PG_ID1 [PG_ID2 ... ]  
# ceph pg cancel-force-backfill PG_ID1 [PG_ID2 ... ]
```

Les commandes **cancel-*** suppriment le drapeau « force » des groupes de placement afin qu'ils soient traités selon l'ordre par défaut. Dans ce cas également, cela n'affecte pas les groupes de placement en cours de traitement, seulement ceux qui sont encore en file d'attente. Le drapeau « force » est automatiquement effacé une fois la récupération ou le renvoi du groupe terminé.

17.4.11 Rétablissement des objets perdus

Si la grappe a perdu un ou plusieurs objets et que vous avez décidé d'abandonner la recherche des données perdues, vous devez marquer les objets introuvables comme « perdu ».

Si les objets sont toujours perdus après avoir interrogé tous les emplacements possibles, vous devrez peut-être renoncer à ces objets. Cela est possible moyennant des combinaisons inhabituelles d'échecs qui permettent à la grappe d'apprendre à partir des opérations d'écriture qui ont été effectuées avant que les écritures elles-mêmes soient récupérées.

Actuellement, la seule option prise en charge est « revert » (rétablir), qui permet soit de revenir à une version précédente de l'objet, soit de l'oublier entièrement dans le cas d'un nouvel objet. Pour marquer les objets « unfound » (introuvable) comme « perdu », exécutez la commande suivante :

```
cephuser@adm > ceph pg PG_ID mark_unfound_lost revert|delete
```

17.4.12 Activation de la mise à l'échelle automatique des groupes de placement

Les groupes de placement sont un détail d'implémentation interne de la façon dont Ceph distribue les données. En activant la mise à l'échelle automatique des groupes de placement, vous pouvez autoriser la grappe à créer ou à ajuster automatiquement des groupes de placement en fonction de l'utilisation de la grappe.

Chaque réserve du système possède une propriété `pg_autoscale_mode` qui peut être définie sur `off`, `on` ou `warn` :

La mise à l'échelle automatique est configurée pour chaque réserve et peut s'exécuter dans trois modes :

off

Désactive la mise à l'échelle automatique pour cette réserve. L'administrateur doit choisir un numéro de groupe de placement approprié pour chaque réserve.

on

Active les ajustements automatisés du nombre de groupes de placement pour la réserve donnée.

avertissement

Génère des alertes d'état de santé lorsque le nombre de groupes de placement doit être ajusté.

Pour définir le mode de mise à l'échelle automatique pour les réserves existantes :

```
cephuser@adm > ceph osd pool set POOL_NAME pg_autoscale_mode mode
```

Vous pouvez également configurer le paramètre `pg_autoscale_mode` par défaut appliqué à toutes les réserves créées par la suite comme suit :

```
cephuser@adm > ceph config set global osd_pool_default_pg_autoscale_mode MODE
```

Vous pouvez afficher chaque réserve, son utilisation relative et toute modification suggérée du nombre de groupes de placement à l'aide de cette commande :

```
cephuser@adm > ceph osd pool autoscale-status
```

17.5 Manipulation de la carte CRUSH

Cette section décrit des méthodes simples de manipulation de carte CRUSH, telles que la modification d'une carte CRUSH, la modification de paramètres de carte CRUSH et l'ajout/le déplacement/la suppression d'un OSD.

17.5.1 Modification d'une carte CRUSH

Pour modifier une carte CRUSH existante, procédez comme suit :

1. Obtenez une carte CRUSH. Pour obtenir la carte CRUSH pour votre grappe, exécutez la commande suivante :

```
cephuser@adm > ceph osd getcrushmap -o compiled-crushmap-filename
```

Ceph associe (-o) une carte CRUSH compilée au nom de fichier que vous avez indiqué. Comme la carte CRUSH est compilée, vous devez la décompiler pour pouvoir la modifier.

2. Décompilez une carte CRUSH. Pour décompiler une carte CRUSH, exécutez la commande suivante :

```
cephuser@adm > crushtool -d compiled-crushmap-filename \  
-o decompiled-crushmap-filename
```

Ceph décompile (-d) la carte CRUSH compilée et l'associe (-o) au nom de fichier que vous avez indiqué.

3. Modifiez au moins l'un des paramètres des périphériques, des compartiments et des règles.
4. Compilez une carte CRUSH. Pour compiler une carte CRUSH, exécutez la commande suivante :

```
cephuser@adm > crushtool -c decompiled-crush-map-filename \  
-o compiled-crush-map-filename
```

Ceph stocke alors une carte CRUSH compilée et l'associe au nom de fichier que vous avez indiqué.

5. Définissez une carte CRUSH. Pour définir la carte CRUSH pour votre grappe, exécutez la commande suivante :

```
cephuser@adm > ceph osd setcrushmap -i compiled-crushmap-filename
```


Ceph considérera la carte CRUSH compilée du nom de fichier que vous avez spécifié comme la carte CRUSH de la grappe.



Astuce : utilisation du système de contrôle de version

Utilisez un système de contrôle de version, comme git ou svn, pour les fichiers de carte CRUSH exportés et modifiés. Cela permet un éventuel retour à l'état initial.



Astuce : test de la nouvelle carte CRUSH

Testez la nouvelle carte CRUSH ajustée à l'aide de la commande **crushtool --test** et comparez avec l'état avant l'application de la nouvelle carte CRUSH. Les paramètres de commande suivants pourraient vous être utiles : --show-statistics, --show-mappings, --show-bad-mappings, --show-utilization, --show-utilization-all, --show-choose-tries

17.5.2 Ajout ou déplacement d'un OSD

Pour ajouter ou déplacer un OSD dans la carte CRUSH d'une grappe en cours d'exécution, exécutez la commande suivante :

```
cephuser@adm > ceph osd crush set id_or_name weight root=pool-name  
bucket-type=bucket-name ...
```

id

Nombre entier. ID numérique de l'OSD. Cette option est obligatoire.

name

Chaîne. Nom complet de l'OSD. Cette option est obligatoire.

weight

Nombre de type double. Pondération CRUSH de l'OSD. Cette option est obligatoire.

root

Paire clé/valeur. Par défaut, la racine de la hiérarchie CRUSH correspond à la valeur par défaut de la réserve. Cette option est obligatoire.

bucket-type

Paires clé/valeur. Vous pouvez indiquer l'emplacement de l'OSD dans la hiérarchie CRUSH.

L'exemple suivant ajoute `osd.0` à la hiérarchie ou déplace l'OSD à partir d'un emplacement précédent.

```
cephuser@adm > ceph osd crush set osd.0 1.0 root=data datacenter=dc1 room=room1 \
row=foo rack=bar host=foo-bar-1
```

17.5.3 Différence entre **ceph osd reweight** et **ceph osd crush reweight**

Il existe deux commandes similaires qui modifient la pondération (« weight ») d'un Ceph OSD. Le contexte de leur utilisation est différent et peut causer de la confusion.

17.5.3.1 **ceph osd reweight**

Syntaxe :

```
cephuser@adm > ceph osd reweight OSD_NAME NEW_WEIGHT
```

ceph osd reweight définit une pondération de remplacement pour le Ceph OSD. Cette valeur est comprise entre 0 et 1, et oblige CRUSH à repositionner les données qui, autrement, seraient sur cette unité. Elle ne modifie **pas** les pondérations assignées aux compartiments au-dessus de l'OSD ; c'est une mesure corrective pour le cas où la distribution CRUSH normale ne fonctionne pas correctement. Par exemple, si l'un de vos OSD est à 90 % et les autres à 40 %, vous pourriez réduire cette pondération pour essayer de compenser.



Note : la pondération OSD est temporaire

Notez **ceph osd reweight** n'est pas un paramètre persistant. Lorsqu'un OSD est marqué comme sorti, sa pondération est définie sur 0 et lorsqu'il est à nouveau marqué comme rentré, sa pondération passe à 1.

17.5.3.2 **ceph osd crush reweight**

Syntaxe :

```
cephuser@adm > ceph osd crush reweight OSD_NAME NEW_WEIGHT
```

ceph osd crush reweight définit la pondération **CRUSH** de l'OSD. Cette pondération est une valeur arbitraire (généralement la taille du disque en To) et contrôle la quantité de données que le système tente d'allouer à l'OSD.

17.5.4 Suppression d'un OSD

Pour supprimer un OSD de la carte CRUSH d'une grappe en cours d'exécution, exécutez la commande suivante :

```
cephuser@adm > ceph osd crush remove OSD_NAME
```

17.5.5 Ajout d'un compartiment

Pour ajouter un compartiment à la carte CRUSH d'une grappe en cours d'exécution, exécutez la commande **ceph osd crush add-bucket** :

```
cephuser@adm > ceph osd crush add-bucket BUCKET_NAME BUCKET_TYPE
```

17.5.6 Déplacement d'un compartiment

Pour déplacer un compartiment vers un autre emplacement ou une autre position dans la hiérarchie de la carte CRUSH, exécutez la commande suivante :

```
cephuser@adm > ceph osd crush move BUCKET_NAME BUCKET_TYPE=BUCKET_NAME [...]
```

Par exemple :

```
cephuser@adm > ceph osd crush move bucket1 datacenter=dc1 room=room1 row=foo rack=bar  
host=foo-bar-1
```

17.5.7 Suppression d'un compartiment

Pour supprimer un compartiment de la hiérarchie de la carte CRUSH, exécutez la commande suivante :

```
cephuser@adm > ceph osd crush remove BUCKET_NAME
```



Note : compartiment vide uniquement

Un compartiment doit être vide pour pouvoir le retirer de la hiérarchie CRUSH.

17.6 Nettoyage des groupes de placement

En plus de réaliser plusieurs copies d'objets, Ceph assure l'intégrité des données en *nettoyant* les groupes de placement (pour en savoir plus sur les groupes de placement, voir *Manuel « Guide de déploiement », Chapitre 1 « SES et Ceph », Section 1.3.2 « Groupes de placement »*). Le nettoyage que réalise Ceph est analogue à l'exécution de **fsck** sur la couche de stockage d'objets. Pour chaque groupe de placement, Ceph génère un catalogue de tous les objets et compare chaque objet principal et ses répliques pour s'assurer qu'aucun objet n'est manquant ou discordant. Le nettoyage léger réalisé quotidiennement vérifie la taille et les attributs de l'objet, tandis que le nettoyage approfondi hebdomadaire lit les données et utilise les sommes de contrôle pour garantir l'intégrité de celles-ci.

Le nettoyage est essentiel au maintien de l'intégrité des données, mais il peut réduire les performances. Vous pouvez ajuster les paramètres suivants pour augmenter ou réduire la fréquence des opérations de nettoyage :

osd max scrubs

Nombre maximum d'opérations de nettoyage simultanées pour un Ceph OSD. La valeur par défaut est 1.

osd scrub begin hour,osd scrub end hour

Heures du jour (0 à 24) qui définissent une fenêtre temporelle pendant laquelle le nettoyage peut avoir lieu. Par défaut, elle commence à 0 et se termine à 24.



Important

Si l'intervalle de nettoyage du groupe de placement dépasse la valeur du paramètre osd scrub max interval, le nettoyage se produit quelle que soit la fenêtre temporelle que vous avez définie.

osd scrub during recovery

Autorise les nettoyages durant la récupération. Si vous définissez cette option sur « false », la planification de nouveaux nettoyages ne sera pas possible tant qu'une récupération est active. L'exécution des nettoyages déjà en cours se poursuivra. Cette option est utile pour réduire la charge sur les grappes occupées. La valeur par défaut est « true ».

osd scrub thread timeout

Durée maximale en secondes avant le timeout d'un thread de nettoyage. La valeur par défaut est 60.

osd scrub finalize thread timeout

Durée maximale en secondes avant le timeout d'un thread de finalisation de nettoyage. La valeur par défaut est 60*10.

osd scrub load threshold

Charge maximale normalisée. Ceph n'effectue pas d'opération de nettoyage lorsque la charge du système (définie par le rapport de `getloadavg()`/nombre d'`online cpus`) est supérieure à ce nombre. La valeur par défaut est 0,5.

osd scrub min interval

Intervalle minimal en secondes pour le nettoyage de Ceph OSD lorsque la charge de la grappe Ceph est faible. La valeur par défaut est 60*60*24 (une fois par jour).

osd scrub max interval

Intervalle maximal en secondes pour le nettoyage de Ceph OSD indépendamment de la charge de la grappe. La valeur par défaut est 7*60*60*24 (une fois par semaine).

osd scrub chunk min

Nombre minimal de tranches de magasin d'objets à nettoyer en une seule opération. L'écriture des blocs Ceph porte sur une seule tranche pendant un nettoyage. La valeur par défaut est 5.

osd scrub chunk max

Nombre maximal de tranches de magasin d'objets à nettoyer en une seule opération. La valeur par défaut est 25.

osd scrub sleep

Temps de veille avant le nettoyage du prochain groupe de tranches. L'augmentation de cette valeur ralentit toute l'opération de nettoyage alors que les opérations client sont moins impactées. La valeur par défaut est 0.

osd deep scrub interval

Intervalle de nettoyage approfondi (lecture complète de toutes les données). L'option osd scrub load threshold n'a pas d'effet sur ce paramètre. La valeur par défaut est $60*60*24*7$ (une fois par semaine).

osd scrub interval randomize ratio

Ajoute un délai aléatoire à la valeur osd scrub interval randomize ratio lors de la planification du prochain travail de nettoyage d'un groupe de placement. Le délai est une valeur aléatoire inférieure au résultat du produit osd scrub min interval * osd scrub interval randomized ratio. Par conséquent, le paramètre par défaut répartit de manière aléatoire les nettoyages dans la fenêtre temporelle autorisée de $[1, 1,5] * \text{osd scrub min interval}$. La valeur par défaut est 0,5.

osd deep scrub stride

Taille des données à lire lors d'un nettoyage en profondeur. La valeur par défaut est 524288 (512 Ko).

18 Gestion des réserves de stockage

Ceph stocke les données dans des réserves. Les réserves sont des groupes logiques pour le stockage des objets. Lorsque vous déployez une grappe pour la première fois sans créer de réserve, Ceph utilise les réserves par défaut pour stocker les données. Les points importants suivants concernent les réserves Ceph :

- *Résilience* : les réserves Ceph garantissent une résilience en répliquant ou en codant les données qu'elles contiennent. Chaque réserve peut être configurée pour être répliquée (« replicated ») ou codée à effacement (« erasure coded »). Pour les réserves répliquées, vous devez également définir le nombre de répliques, ou copies, dont disposera chaque objet de données de la réserve. Le nombre de copies (OSD, compartiments/feuilles CRUSH) pouvant être perdues est inférieur d'une unité au nombre de répliques. Avec le codage à effacement, vous devez définir les valeurs de k et m, k correspondant au nombre de tranches de données et m au nombre de tranches de codage. Pour les réserves utilisant le codage à effacement, c'est le nombre de tranches de codage qui détermine le nombre d'OSD (compartiments/feuilles CRUSH) pouvant être perdus sans perte de données.
- *Groupes de placement* : vous pouvez définir le nombre de groupes de placement pour la réserve. Une configuration type utilise environ 100 groupes de placement par OSD pour fournir un équilibrage optimal sans nécessiter trop de ressources informatiques. Lors de la configuration de plusieurs grappes, veillez à définir un nombre raisonnable de groupes de placement pour la réserve et la grappe dans leur ensemble.
- *Règles CRUSH* : lorsque vous stockez des données dans une réserve, les objets et leurs répliques (ou tranches en cas de réserves codées à effacement) sont placés selon l'ensemble de règles CRUSH assignées à la réserve. Vous pouvez créer une règle CRUSH personnalisée pour votre réserve.
- *Instantanés* : lorsque vous créez des instantanés avec `ceph osd pool mksnap`, vous prenez effectivement un instantané d'une réserve en particulier.

Pour organiser les données en réserves, vous pouvez répertorier, créer et supprimer des réserves. Vous pouvez également afficher les statistiques d'utilisation pour chaque réserve.

18.1 Création d'une réserve

Une réserve peut être de type replicated (répliquée) pour récupérer des OSD perdus en conservant plusieurs copies des objets, ou de type erasure (à effacement) pour obtenir une sorte de fonctionnalité RAID5/6 généralisée. Les réserves répliquées nécessitent plus de stockage brut, tandis que les réserves codées à effacement en exigent moins. Le paramètre par défaut est replicated. Pour plus d'informations sur les réserves codées à effacement, reportez-vous au [Chapitre 19, Réserves codées à effacement](#).

Pour créer une réserve répliquée, exécutez :

```
cephuser@adm > ceph osd pool create POOL_NAME
```



Note

La mise à l'échelle automatique se chargera des arguments facultatifs restants. Pour plus d'informations, reportez-vous à la [Section 17.4.12, « Activation de la mise à l'échelle automatique des groupes de placement »](#).

Pour créer une réserve codée à effacement, exécutez :

```
cephuser@adm > ceph osd pool create POOL_NAME erasure CRUSH_RULESET_NAME \
EXPECTED_NUM_OBJECTS
```

La commande **ceph osd pool create** peut échouer si vous dépassez la limite de groupes de placement par OSD. La limite est définie avec l'option mon_max_pg_per_osd.

POOL_NAME

Nom de la réserve. Il doit être unique. Cette option est obligatoire.

POOL_TYPE

Le type de réserve qui peut être soit répliqué pour récupérer des OSD perdus en conservant plusieurs copies des objets ou à effacement pour obtenir une sorte de fonctionnalité RAID5 généralisée. Les réserves répliquées nécessitent plus de stockage brut mais implémentent toutes les opérations Ceph. Les réserves à effacement nécessitent moins de stockage brut, mais implémentent uniquement un sous-ensemble des opérations disponibles. La valeur par défaut de POOL_TYPE est replicated.

CRUSH_RULESET_NAME

Nom de l'ensemble de règles CRUSH de cette réserve. Si l'ensemble de règles indiqué n'existe pas, la création de réserves répliquées échoue avec -ENOENT. Pour les réserves répliquées, il s'agit de l'ensemble de règles spécifié par la variable de configuration `osd pool default CRUSH replicated ruleset`. Cet ensemble de règles doit exister. Pour les réserves à effacement, il s'agit de « erasure-code » si le profil de code à effacement par défaut est utilisé, sinon de `NOM_RÉSERVE`. Cet ensemble de règles sera créé implicitement s'il n'existe pas déjà.

erasure_code_profile=profile

Pour les réserves codées à effacement uniquement. Utilisez le profil de code d'effacement. Il doit s'agir d'un profil existant tel que défini par `osd erasure-code-profile set`.



Note

Si, pour une raison quelconque, la mise à l'échelle automatique a été désactivée (`pg_autoscale_mode` désactivé) sur une réserve, vous pouvez calculer et définir manuellement le nombre de groupes de placement. Reportez-vous à la [Section 17.4, « Groupes de placement »](#) pour plus d'informations sur le calcul du nombre de groupes de placement approprié pour votre réserve.

EXPECTED_NUM_OBJECTS

Nombre d'objets attendu pour cette réserve. En définissant cette valeur (avec un seuil `filestore merge threshold` négatif), le fractionnement du dossier de groupes de placement se produit au moment de la création de la réserve. Cela évite l'impact de latence lié au fractionnement du dossier à l'exécution.

18.2 Liste des réserves

Pour afficher la liste des réserves de la grappe, exécutez :

```
cephuser@adm > ceph osd pool ls
```

18.3 Modification du nom d'une réserve

Pour renommer une réserve, exécutez :

```
cephuser@adm > ceph osd pool rename CURRENT_POOL_NAME NEW_POOL_NAME
```

Si vous renommez une réserve et que vous disposez de fonctions de réserve pour un utilisateur authentifié, vous devez mettre à jour les fonctions de l'utilisateur avec le nouveau nom de réserve.

18.4 Suppression d'une réserve



Avertissement : la suppression d'une réserve est irréversible

Les réserves peuvent contenir des données importantes. La suppression d'une réserve entraîne la disparition de toutes les données qu'elle contient et l'impossibilité de la récupérer.

Comme la suppression accidentelle d'une réserve constitue un réel danger, Ceph implémente deux mécanismes qui empêchent cette suppression. Ces deux mécanismes doivent être désactivés avant la suppression d'une réserve.

Le premier mécanisme consiste à utiliser l'indicateur `NODELETE`. Chaque réserve possède cet indicateur dont la valeur par défaut est « false ». Pour connaître la valeur de cet indicateur sur une réserve, exécutez la commande suivante :

```
cephuser@adm > ceph osd pool get pool_name nodelete
```

Si elle génère `nodelete: true`, il n'est pas possible de supprimer la réserve tant que vous ne modifiez pas l'indicateur à l'aide de la commande suivante :

```
cephuser@adm > ceph osd pool set pool_name nodelete false
```

Le second mécanisme est le paramètre de configuration de la grappe `mon allow pool delete`, qui est « false » par défaut. Cela signifie que, par défaut, il n'est pas possible de supprimer une réserve. Le message d'erreur affiché est le suivant :

```
Error EPERM: pool deletion is disabled; you must first set the
```

```
mon_allow_pool_delete config option to true before you can destroy a pool
```

Pour supprimer la réserve malgré ce paramètre de sécurité, vous pouvez définir temporairement `mon allow pool delete` sur « true », supprimer la réserve, puis renvoyer le paramètre avec « false » :

```
cephuser@adm > ceph tell mon.* injectargs --mon-allow-pool-delete=true
cephuser@adm > ceph osd pool delete pool_name pool_name --yes-i-really-really-mean-it
cephuser@adm > ceph tell mon.* injectargs --mon-allow-pool-delete=false
```

La commande **`injectargs`** affiche le message suivant :

```
injectargs:mon_allow_pool_delete = 'true' (not observed, change may require restart)
```

Cela confirme simplement que la commande a été exécutée avec succès. Il ne s'agit pas d'une erreur.

Si vous avez défini vos propres ensembles de règles et règles pour une réserve que vous avez créée, vous devez envisager de les supprimer lorsque vous n'avez plus besoin de la réserve.

18.5 Autres opérations

18.5.1 Association de réserves à une application

Pour pouvoir utiliser les réserves, vous devez les associer à une application. Les réserves qui seront utilisées avec CephFS ou les réserves créées automatiquement par Object Gateway sont automatiquement associées.

Pour les autres cas, vous pouvez associer manuellement un nom de l'application de format libre à une réserve :

```
cephuser@adm > ceph osd pool application enable POOL_NAME APPLICATION_NAME
```



Astuce : noms d'application par défaut

CephFS utilise le nom de l'application `cephfs`, le périphérique de bloc RADOS emploie `rd` et la passerelle Object Gateway fait appel à `rgw`.

Une réserve peut être associée à plusieurs applications et chaque application peut avoir ses propres métadonnées. Pour lister l'application (ou les applications) associée(s) à une réserve, exécutez la commande suivante :

```
cephuser@adm > ceph osd pool application get pool_name
```

18.5.2 Définition de quotas de réserve

Vous pouvez définir des quotas de réserve pour le nombre maximal d'octets et/ou le nombre maximal d'objets par réserve.

```
cephuser@adm > ceph osd pool set-quota POOL_NAME MAX_OBJECTS OBJ_COUNT MAX_BYTES BYTES
```

Par exemple :

```
cephuser@adm > ceph osd pool set-quota data max_objects 10000
```

Pour supprimer un quota, définissez sa valeur sur 0.

18.5.3 Affichage des statistiques d'une réserve

Pour afficher les statistiques d'utilisation d'une réserve, exécutez :

```
cephuser@adm > rados df
```

| POOL_NAME | DEGRADED | RD_OPS | RD | WR_OPS | WR | USED | CLONES | COPIES | MISSING_ON_PRIMARY | UNFOUND |
|---------------------------|----------|---------|---------|---------|-----|------|--------|--------|--------------------|---------|
| .rgw.root | | | | 768 KiB | 4 | 0 | 12 | | 0 | 0 |
| 0 | 44 | 44 KiB | 4 | 4 KiB | 0 B | | 0 B | | | |
| cephfs_data | | | | 960 KiB | 5 | 0 | 15 | | 0 | 0 |
| 0 | 5502 | 2.1 MiB | 14 | 11 KiB | 0 B | | 0 B | | | |
| cephfs_metadata | | | | 1.5 MiB | 22 | 0 | 66 | | 0 | 0 |
| 0 | 26 | 78 KiB | 176 | 147 KiB | 0 B | | 0 B | | | |
| default.rgw.buckets.index | | | | 0 B | 1 | 0 | 3 | | 0 | 0 |
| 0 | 4 | 4 KiB | 1 | 0 B | 0 B | | 0 B | | | |
| default.rgw.control | | | | 0 B | 8 | 0 | 24 | | 0 | 0 |
| 0 | 0 | 0 B | 0 | 0 B | 0 B | | 0 B | | | |
| default.rgw.log | | | | 0 B | 207 | 0 | 621 | | 0 | 0 |
| 0 | 5372132 | 5.1 GiB | 3579618 | 0 B | 0 B | | 0 B | | | |
| default.rgw.meta | | | | 961 KiB | 6 | 0 | 18 | | 0 | 0 |
| 0 | 155 | 140 KiB | 14 | 7 KiB | 0 B | | 0 B | | | |
| example_rbd_pool | | | | 2.1 MiB | 18 | 0 | 54 | | 0 | 0 |
| 0 | 3350841 | 2.7 GiB | 118 | 98 KiB | 0 B | | 0 B | | | |
| iscsi-images | | | | 769 KiB | 8 | 0 | 24 | | 0 | 0 |
| 0 | 1559261 | 1.3 GiB | 61 | 42 KiB | 0 B | | 0 B | | | |

| | | | | | | |
|-------------------|---------------|-----|---|-----|---|---|
| mirrored-pool | 1.1 MiB | 10 | 0 | 30 | 0 | 0 |
| 0 475724 395 MiB | 54 48 KiB | 0 B | | 0 B | | |
| pool2 | 0 B | 0 | 0 | 0 | 0 | 0 |
| 0 0 0 B | 0 0 B | 0 B | | 0 B | | |
| pool3 | 333 MiB | 37 | 0 | 111 | 0 | 0 |
| 0 3169308 2.5 GiB | 14847 118 MiB | 0 B | | 0 B | | |
| pool4 | 1.1 MiB | 13 | 0 | 39 | 0 | 0 |
| 0 1379568 1.1 GiB | 16840 16 MiB | 0 B | | 0 B | | |

Une description de chaque colonne suit :

USED

Nombre d'octets utilisés par la réserve.

OBJECTS

Nombre d'objets stockés dans la réserve.

CLONES

Nombre de clones stockés dans la réserve. Lorsqu'un instantané est créé et que l'on écrit dans un objet, au lieu de modifier l'objet d'origine, son clone est créé de sorte que le contenu de l'objet instantané d'origine n'est pas modifié.

COPIES

Nombre de répliques d'objets. Par exemple, si une réserve répliquée avec le facteur de réplification 3 a « x » objets, elle aura normalement $3 * x$ copies.

MISSING_ON_PRIMARY

Nombre d'objets dans l'état altéré (toutes les copies n'existent pas) alors que la copie est manquante sur l'OSD primaire.

UNFOUND

Nombre d'objets introuvables.

DEGRADED

Nombre d'objets altérés.

RD_OPS

Nombre total d'opérations de lecture demandées pour cette réserve.

RD

Nombre total d'octets lus à partir de cette réserve.

WR_OPS

Nombre total d'opérations d'écriture demandées pour cette réserve.

WR

Nombre total d'octets écrits dans la réserve. Notez que cela n'est pas la même chose que l'utilisation de la réserve, car vous pouvez écrire plusieurs fois sur le même objet. Au final, l'utilisation de la réserve restera la même, mais le nombre d'octets qui y sont écrits augmentera.

USED COMPR

Nombre d'octets alloués aux données compressées.

UNDER COMPR

Nombre d'octets occupés par les données compressées lorsqu'elles ne sont pas compressées.

18.5.4 Obtention de valeurs d'une réserve

Pour obtenir une valeur à partir d'une réserve, exécutez la commande **get** suivante :

```
cephuser@adm > ceph osd pool get POOL_NAME KEY
```

Vous pouvez obtenir des valeurs pour les clés répertoriées à la [Section 18.5.5, « Définition des valeurs d'une réserve »](#) ainsi que les clés suivantes :

PG_NUM

Nombre de groupes de placement pour la réserve.

PGP_NUM

Nombre effectif de groupes de placement à utiliser lors du calcul du placement des données. La plage valide est inférieure ou égale à PG_NUM.



Astuce : toutes les valeurs d'une réserve

Pour répertorier toutes les valeurs associées à une réserve spécifique, exécutez :

```
cephuser@adm > ceph osd pool get POOL_NAME all
```

18.5.5 Définition des valeurs d'une réserve

Pour définir une valeur d'une réserve, exécutez :

```
cephuser@adm > ceph osd pool set POOL_NAME KEY VALUE
```

La liste ci-dessous répertorie les valeurs de réserve triées par type de réserve :

VALEURS DE RÉSERVE COMMUNE

crash_replay_interval

Nombre de secondes pendant lesquelles les clients peuvent relire les requêtes acquittées mais non validées.

pg_num

Nombre de groupes de placement pour la réserve. Si vous ajoutez des OSD à la grappe, vérifiez la valeur des groupes de placement sur toutes les réserves ciblées pour les nouveaux OSD.

pgp_num

Nombre effectif de groupes de placement à utiliser lors du calcul du placement des données.

crush_ruleset

Ensemble de règles à utiliser pour l'assignation de placement d'objets dans la grappe.

hashpspool

Définissez (1) ou désélectionnez (0) l'indicateur HASHPSPOOL sur une réserve donnée. L'activation de cet indicateur modifie l'algorithme pour mieux répartir les PG sur les OSD. Après avoir activé ce drapeau sur une réserve dont le drapeau HASHPSPOOL avait été défini par défaut sur 0, la grappe commence à effectuer des renvois afin de rétablir le placement correct de tous les groupes de placement. Cette activation pouvant créer une charge d'E/S assez importante sur une grappe, ne passez pas le drapeau de 0 à 1 sur les grappes de production très chargées.

nodelete

Empêche la suppression de la réserve.

nopgchange

Empêche la modification des options pg_num et pgp_num de la réserve.

noscrub, nodeep-scrub

Désactive le nettoyage (en profondeur) des données pour la réserve en particulier afin de résoudre une charge d'E/S élevée temporaire.

write_fadvise_dontneed

Sélectionnez ou désélectionnez l'indicateur WRITE_FADVISE_DONTNEED sur les requêtes de lecture/d'écriture d'une réserve donnée afin de contourner la mise en cache des données. La valeur par défaut est false. S'applique aux réserves répliquées et EC.

scrub_min_interval

Intervalle minimal en secondes pour le nettoyage des réserves lorsque la charge de la grappe est faible. La valeur par défaut 0 signifie que la valeur de `osd_scrub_min_interval` du fichier de configuration Ceph est utilisée.

scrub_max_interval

Intervalle maximal en secondes pour le nettoyage des réserves, quelle que soit la charge de la grappe. La valeur par défaut 0 signifie que la valeur de `osd_scrub_max_interval` du fichier de configuration Ceph est utilisée.

deep_scrub_interval

Intervalle en secondes pour le nettoyage *en profondeur* de la grappe. La valeur par défaut 0 signifie que la valeur de `osd_deep_scrub` du fichier de configuration Ceph est utilisée.

VALEURS DE RÉSERVE RÉPLIQUÉE

size

Définit le nombre de répliques d'objets dans la réserve. Pour plus d'informations, reportez-vous à la [Section 18.5.6, « Définition du nombre de répliques d'objets »](#). Réserves répliquées uniquement.

min_size

Définit le nombre minimum de répliques requises pour les E/S. Reportez-vous à la [Section 18.5.6, « Définition du nombre de répliques d'objets »](#) pour plus de détails. Réserves répliquées uniquement.

nosizechange

Empêche la modification de la taille de la réserve. Lorsqu'une réserve est créée, la valeur par défaut est tirée de la valeur du paramètre `osd_pool_default_flag_nosizechange`, lequel est défini par défaut sur `false`. S'applique uniquement aux réserves répliquées, car la taille des réserves EC ne peut pas être modifiée.

hit_set_type

Active le suivi des jeux d'accès pour les réserves de cache. Reportez-vous à l'article [Filtre de Bloom](http://en.wikipedia.org/wiki/Bloom_filter) (http://en.wikipedia.org/wiki/Bloom_filter) [↗](#) pour obtenir des informations complémentaires. Cette option accepte l'une des valeurs suivantes : `bloom`, `explicit_hash`, `explicit_object`. La valeur par défaut est `bloom`, les autres valeurs sont utilisées à des fins de test uniquement.


hit_set_count

Nombre de jeux d'accès à stocker dans les réserves de cache. Plus le nombre est élevé, plus le daemon `ceph-osd` consomme une quantité importante de mémoire vive. La valeur par défaut est 0.

hit_set_period

Durée d'une période de jeu d'accès définie en secondes pour les réserves de cache. Plus le nombre est élevé, plus le daemon `ceph-osd` consomme une quantité importante de mémoire vive. Lorsqu'une réserve est créée, la valeur par défaut est tirée de la valeur du paramètre `osd_tier_default_cache_hit_set_period`, lequel est défini par défaut sur 1200. S'applique uniquement aux réserves répliquées, car les réserves EC ne peuvent pas être utilisées en tant que niveau de cache.

hit_set_fpp

Probabilité de faux positifs pour le type de jeu d'accès de filtre de Bloom. Reportez-vous à l'article [Filtre de Bloom](http://en.wikipedia.org/wiki/Bloom_filter) (http://en.wikipedia.org/wiki/Bloom_filter)  pour obtenir des informations complémentaires. La plage valide est comprise entre 0.0 et 1.0. La valeur par défaut est 0.05.

use_gmt_hitset

Forcez les OSD à utiliser les horodatages GMT (Greenwich Mean Time) lors de la création d'un jeu d'accès pour la hiérarchisation du cache. Cela garantit que les noeuds situés dans des fuseaux horaires différents retournent le même résultat. La valeur par défaut est 1. Cette valeur ne doit pas être modifiée.

cache_target_dirty_ratio

Pourcentage de la réserve de cache contenant des objets modifiés (altérés) avant que l'agent de hiérarchisation du cache les transfère à la réserve de stockage de sauvegarde. La valeur par défaut est 0.4.

cache_target_dirty_high_ratio

Vous pouvez indiquer l'âge minimal d'un objet récemment modifié (altéré) avant que l'agent de hiérarchisation du cache le transfère à la réserve de stockage de sauvegarde à une vitesse supérieure. La valeur par défaut est 0.6.

cache_target_full_ratio

Pourcentage de la réserve de cache contenant des objets non modifiés (propres) avant que l'agent de hiérarchisation du cache les élimine de la réserve de cache. La valeur par défaut est 0.8.

target_max_bytes

Ceph commence à vider ou à éliminer des objets lorsque le seuil max_bytes est déclenché.

target_max_objects

Ceph commence à vider ou à éliminer des objets lorsque le seuil max_objects est déclenché.

hit_set_grade_decay_rate

Taux de baisse de la température entre deux hit_set successifs. Valeur par défaut : 20.

hit_set_search_last_n

Comptez au plus N apparitions dans les valeurs de hit_set pour le calcul de la température. La valeur par défaut est 1.

cache_min_flush_age

Durée (en secondes) avant que l'agent de hiérarchisation du cache vide un objet de la réserve de cache vers la réserve de stockage.

cache_min_evict_age

Durée (en secondes) avant que l'agent de hiérarchisation du cache élimine un objet de la réserve de cache.

VALEURS DE RÉSERVE CODÉE À EFFACEMENT

fast_read

Si cet indicateur est activé sur les réserves de codage à effacement, la demande de lecture émet des sous-lectures sur toutes les partitions et attend de recevoir un nombre suffisant de fragments à décoder afin de servir le client. Dans le cas des plug-ins *jerasure* et *isa*, lorsque les premières réponses K sont retournées, la requête du client est servie immédiatement avec les données décodées issues de ces réponses. Cette approche augmente la charge de processeur et diminue la charge de disque/réseau. Pour le moment, cet indicateur est pris en charge uniquement pour les réserves de codage à effacement. La valeur par défaut est 0.

18.5.6 Définition du nombre de répliques d'objets

Pour définir le nombre de répliques d'objets sur un réserve répliquée, exécutez la commande suivante :

```
cephuser@adm > ceph osd pool set poolname size num-replicas
```

`num-replicas` inclut l'objet lui-même. Si vous souhaitez, par exemple, l'objet et deux copies de l'objet pour obtenir au total trois instances de l'objet, indiquez 3.



Avertissement : ne définissez pas moins de 3 répliques

Si vous définissez `num-replicas` sur 2, *une seule* copie de vos données est disponible. Si vous perdez une instance d'objet, vous devez être sûr que l'autre copie n'a pas été endommagée, par exemple depuis le dernier nettoyage pendant la récupération (pour plus d'informations, reportez-vous à la [Section 17.6, « Nettoyage des groupes de placement »](#)).

La définition d'une réserve à réplique unique implique qu'il existe exactement *une* instance de l'objet de données dans la réserve. Si l'OSD échoue, vous perdez les données. Une utilisation possible d'une réserve avec une réplique consiste à stocker des données temporaires pendant une courte période.



Astuce : définition de plus de 3 répliques

La définition de 4 répliques pour une réserve augmente la fiabilité de 25 %.

Dans le cas de deux centres de données, vous devez définir au moins 4 répliques pour une réserve de sorte à avoir deux copies dans chaque centre de données. De cette façon, en cas de perte d'un centre de données, il existe toujours deux copies et vous pouvez perdre un disque sans perdre de données.



Note

Un objet peut accepter des E/S en mode dégradé avec moins de `pool size` répliques. Pour définir un nombre minimum de répliques requis pour les E/S, vous devez utiliser le paramètre `min_size`. Par exemple :

```
cephuser@adm > ceph osd pool set data min_size 2
```

Cela garantit qu'aucun objet de la réserve de données ne recevra d'E/S avec moins de `min_size` répliques.



Astuce : obtention du nombre de répliques d'objets

Pour obtenir le nombre de répliques d'objet, exécutez la commande suivante :

```
cephuser@adm > ceph osd dump | grep 'replicated size'
```

Ceph dresse la liste des réserves en mettant en surbrillance l'attribut `replicated size`. Par défaut, Ceph crée deux répliques d'un objet (soit un total de trois copies ou une taille de 3).

18.6 Migration d'une réserve

Lors de la création d'une réserve (voir [Section 18.1, « Création d'une réserve »](#)), vous devez indiquer ses paramètres initiaux, tels que le type de réserve ou le nombre de groupes de placement. Si vous décidez ultérieurement de modifier l'un de ces paramètres, par exemple lors de la conversion d'une réserve répliquée en réserve codée à effacement ou de la diminution du nombre de groupes de placement, vous devez migrer les données de réserve vers une autre réserve dont les paramètres conviennent à votre déploiement.

Cette section décrit deux méthodes de migration : une méthode de *niveau de cache* pour la migration générale des données d'une réserve, et une méthode utilisant des sous-commandes **`rbd migrate`** pour migrer des images RBD vers une nouvelle réserve. Chaque méthode a ses spécificités et ses limites.

18.6.1 Limites

- Vous pouvez utiliser la méthode de *niveau de cache* pour migrer une réserve répliquée vers une réserve EC ou vers une autre réserve répliquée. La migration à partir d'une réserve EC n'est pas prise en charge.
- Vous ne pouvez pas migrer des images RBD et des exportations CephFS depuis une réserve répliquée vers une réserve codée à effacement (EC), car les réserves EC ne prennent pas en charge `omap`, alors que RBD et CephFS utilisent `omap` pour stocker leurs métadonnées. Par exemple, l'objet d'en-tête de RBD ne sera pas vidé. En revanche, vous pouvez migrer des données vers une réserve EC, en laissant les métadonnées dans la réserve répliquée.
- La méthode **`rbd migration`** permet de migrer des images avec un temps hors service minimal du client. Vous ne devez arrêter le client qu'avant l'étape de *préparation* et pouvez le redémarrer après. Notez que seul un client `librbd` qui prend en charge cette fonction (Ceph Nautilus ou plus récent) sera en mesure d'ouvrir l'image juste après l'étape de *préparation*. Les clients `librbd` plus anciens ou les clients `krbd` ne pourront pas ouvrir l'image avant l'exécution de l'étape de *validation*.

18.6.2 Migration à l'aide du niveau de cache

Le principe est simple : incluez la réserve dont vous avez besoin pour migrer dans un niveau de cache dans l'ordre inverse. L'exemple suivant illustre la migration d'une réserve répliquée appelée « testpool » vers une réserve codée à effacement :

PROCÉDURE 18.1 : MIGRATION D'UNE RÉSERVE RÉPLIQUÉE VERS UNE RÉSERVE CODÉE À EFFACEMENT

1. Créez une réserve codée à effacement nommée « newpool ». Pour plus d'informations sur les paramètres de création d'une réserve, reportez-vous à la [Section 18.1, « Création d'une réserve »](#).

```
cephuser@adm > ceph osd pool create newpool erasure default
```

Vérifiez que le trousseau de clés client utilisé fournit au moins les mêmes fonctionnalités pour « newpool » que pour « testpool ».

Vous avez maintenant deux réserves : la réserve répliquée initiale « testpool » contenant des données et la nouvelle réserve codée à effacement « newpool » :

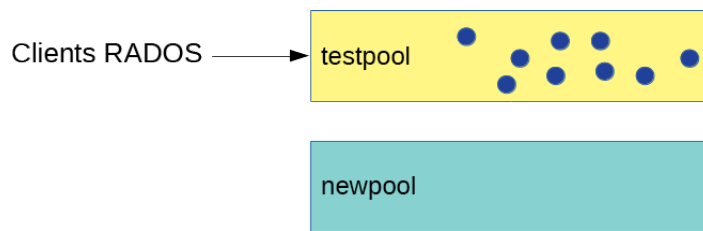


FIGURE 18.1 : RÉSERVES AVANT MIGRATION

2. Configurez le niveau de cache et la réserve répliquée « testpool » en tant que réserve de cache. L'option `--force-nonempty` permet d'ajouter un niveau de cache même si la réserve a déjà des données :

```
cephuser@adm > ceph tell mon.* injectargs \
'--mon_debug_unsafe_allow_tier_with_nonempty_snaps=1'
cephuser@adm > ceph osd tier add newpool testpool --force-nonempty
cephuser@adm > ceph osd tier cache-mode testpool proxy
```

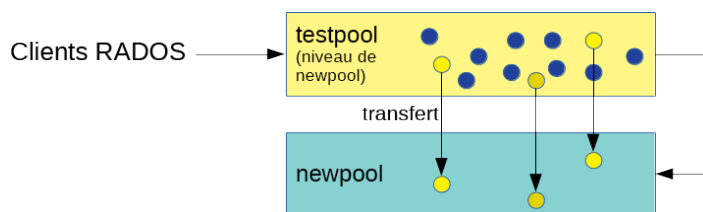


FIGURE 18.2 : CONFIGURATION DU NIVEAU DE CACHE

3. Forcez la réserve de cache à déplacer tous les objets vers la nouvelle réserve :

```
cephuser@adm > rados -p testpool cache-flush-evict-all
```

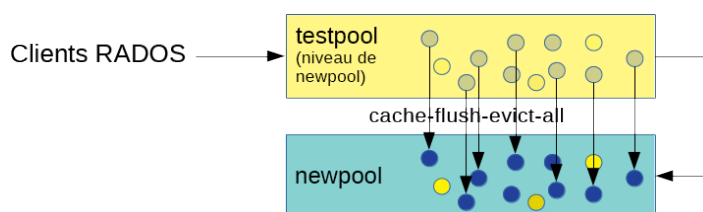


FIGURE 18.3 : VIDAGE DES DONNÉES

4. Tant que toutes les données n'ont pas été vidées vers la nouvelle réserve codée à effacement, vous devez indiquer une superposition afin que les recherches d'objets s'effectuent dans l'ancienne réserve :

```
cephuser@adm > ceph osd tier set-overlay newpool testpool
```

Avec la superposition, toutes les opérations sont réacheminées vers l'ancienne réserve « testpool » répliquée :

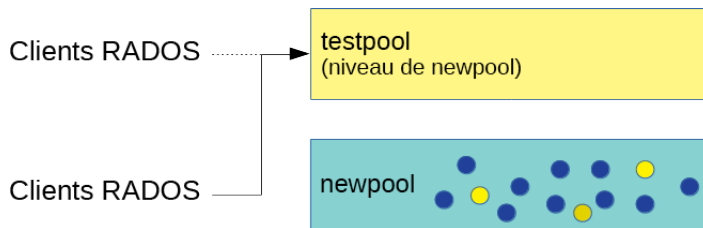


FIGURE 18.4 : DÉFINITION DE LA SUPERPOSITION

Vous pouvez maintenant basculer tous les clients pour accéder aux objets de la nouvelle réserve.

5. Une fois toutes les données migrées vers la réserve codée à effacement « newpool », supprimez la superposition et l'ancienne réserve de cache « testpool » :

```
cephuser@adm > ceph osd tier remove-overlay newpool  
cephuser@adm > ceph osd tier remove newpool testpool
```

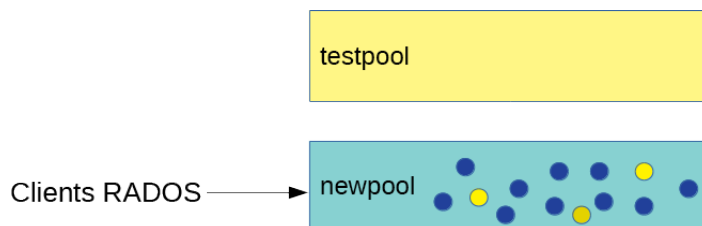


FIGURE 18.5 : MIGRATION EFFECTUÉE

6. Exécutez :

```
cephuser@adm > ceph tell mon.* injectargs \  
'--mon_debug_unsafe_allow_tier_with_nonempty_snaps=0'
```

18.6.3 Migration d'images RBD

Voici la méthode recommandée pour migrer des images RBD d'une réserve répliquée vers une autre réserve répliquée.

1. Empêchez les clients (une machine virtuelle, par exemple) d'accéder à l'image RBD.
2. Créez une image dans la réserve cible, avec le parent défini sur l'image source :

```
cephuser@adm > rbd migration prepare SRC_POOL/IMAGE TARGET_POOL/IMAGE
```



Astuce : migration de données uniquement vers une réserve codée à effacement

Si vous devez migrer uniquement les données d'image vers une nouvelle réserve EC et laisser les métadonnées dans la réserve répliquée d'origine, exécutez la commande suivante à la place :

```
cephuser@adm > rbd migration prepare SRC_POOL/IMAGE \  
--data-pool TARGET_POOL/IMAGE
```

3. Laissez les clients accéder à l'image dans la réserve cible.
4. Migrez les données vers la réserve cible :

```
cephuser@adm > rbd migration execute SRC_POOL/IMAGE
```

5. Supprimez l'ancienne image :

```
cephuser@adm > rbd migration commit SRC_POOL/IMAGE
```

18.7 Instantanés de réserve

Les instantanés de réserve sont des instantanés de l'état de l'ensemble de la réserve Ceph. Avec les instantanés de réserve, vous pouvez conserver l'historique de l'état de la réserve. La création d'instantanés de réserve consomme un espace de stockage proportionnel à la taille de la réserve. Vérifiez toujours que le stockage associé possède un espace disque suffisant avant de créer un instantané d'une réserve.

18.7.1 Création d'un instantané d'une réserve

Pour créer un instantané d'une réserve, exécutez :

```
cephuser@adm > ceph osd pool mksnap POOL-NAME SNAP-NAME
```

Par exemple :

```
cephuser@adm > ceph osd pool mksnap pool1 snap1  
created pool pool1 snap snap1
```

18.7.2 Liste des instantanés d'une réserve

Pour lister les instantanés existants d'une réserve, exécutez :

```
cephuser@adm > rados lssnap -p POOL_NAME
```

Par exemple :

```
cephuser@adm > rados lssnap -p pool1  
1 snap1 2018.12.13 09:36:20  
2 snap2 2018.12.13 09:46:03  
2 snaps
```

18.7.3 Suppression d'un instantané d'une réserve

Pour supprimer un instantané d'une réserve, exécutez :

```
cephuser@adm > ceph osd pool rmsnap POOL-NAME SNAP-NAME
```

18.8 Compression des données

BlueStore (voir *Manuel « Guide de déploiement », Chapitre 1 « SES et Ceph », Section 1.4 « BlueStore »* pour plus de détails) fournit la compression des données à la volée pour économiser de l'espace disque. Le rapport de compression dépend des données stockées sur le système. Notez que la compression/décompression nécessite davantage de ressources processeur.

Vous pouvez configurer la compression des données globalement (voir [Section 18.8.3, « Options de compression globales »](#)), puis remplacer les paramètres de compression spécifiques pour chaque réserve.

Vous pouvez activer ou désactiver la compression des données de réserve, ou modifier l'algorithme et le mode de compression à tout moment, que la réserve contienne des données ou non. Aucune compression ne sera appliquée aux données existantes après avoir activé la compression de la réserve.

Après avoir désactivé la compression d'une réserve, toutes ses données seront décompressées.

18.8.1 Activation de la compression

Pour activer la compression des données pour une réserve nommée *PPOOL_NAME*, exécutez la commande suivante :

```
cephuser@adm > ceph osd pool set PPOOL_NAME compression_algorithm COMPRESSION_ALGORITHM  
cephuser@adm > ceph osd pool set PPOOL_NAME compression_mode COMPRESSION_MODE
```



Astuce : désactivation de la compression d'une réserve

Pour désactiver la compression des données pour une réserve, utilisez « none » comme algorithme de compression :

```
cephuser@adm > ceph osd pool set PPOOL_NAME compression_algorithm none
```

18.8.2 Options de compression de réserve

Liste complète de paramètres de compression :

compression_algorithm

Les valeurs possibles sont none, zstd, snappy. La valeur par défaut est snappy.

L'algorithme de compression à utiliser dépend du cas d'utilisation particulier. Voici quelques recommandations :

- Utilisez la valeur par défaut snappy tant que vous n'avez pas une raison valable d'en changer.
- zstd offre un bon rapport de compression, mais provoque un overhead important du processeur lors de la compression de petites quantités de données.
- Effectuez un banc d'essai de ces algorithmes sur un échantillon de vos données réelles tout en gardant un oeil sur l'utilisation du processeur et de la mémoire de votre grappe.

compression_mode

Les valeurs possibles sont none, aggressive, passive et force. La valeur par défaut est none.

- none : jamais de compression
- passive : compresser si COMPRESSIBLE est suggéré
- aggressive : compresser sauf si INCOMPRESSIBLE est suggéré
- force : compresser toujours

compression_required_ratio

Valeur : Double, $\text{Ratio} = \text{SIZE_COMPRESSED} / \text{SIZE_ORIGINAL}$. La valeur par défaut est 0,875, ce qui signifie que si la compression ne réduit pas l'espace occupé d'au moins 12,5 %, l'objet ne sera pas compressé.

Les objets au-dessus de ce ratio ne seront pas stockés dans un format compressé en raison du faible gain net.

compression_max_blob_size

Valeur : entier non signé, taille en octets. Valeur par défaut : 0
Taille maximale des objets compressés.

compression_min_blob_size

Valeur : entier non signé, taille en octets. Valeur par défaut : 0
Taille minimale des objets compressés.

18.8.3 Options de compression globales

Les options de configuration suivantes peuvent être définies dans la configuration Ceph et s'appliquent à tous les OSD et non pas seulement à une réserve. La configuration spécifique de la réserve répertoriée à la [Section 18.8.2, « Options de compression de réserve »](#) prévaut.

`bluestore_compression_algorithm`

Reportez-vous à la [compression_algorithm](#)

`bluestore_compression_mode`

Reportez-vous à la [compression_mode](#)

`bluestore_compression_required_ratio`

Reportez-vous à la [compression_required_ratio](#)

`bluestore_compression_min_blob_size`

Valeur : entier non signé, taille en octets. Valeur par défaut : `0`

Taille minimale des objets compressés. Le paramètre est ignoré par défaut en faveur de `bluestore_compression_min_blob_size_hdd` et `bluestore_compression_min_blob_size_ssd`. Il est prioritaire lorsqu'il est défini sur une valeur différente de zéro.

`bluestore_compression_max_blob_size`

Valeur : entier non signé, taille en octets. Valeur par défaut : `0`

Taille maximale des objets qui sont compressés avant d'être divisés en petites tranches. Le paramètre est ignoré par défaut en faveur de `bluestore_compression_max_blob_size_hdd` et `bluestore_compression_max_blob_size_ssd`. Il est prioritaire lorsqu'il est défini sur une valeur différente de zéro.

`bluestore_compression_min_blob_size_ssd`

Valeur : entier non signé, taille en octets. Valeur par défaut : `8 000`

Taille minimale des objets compressés et stockés sur une unité SSD.

`bluestore_compression_max_blob_size_ssd`

Valeur : entier non signé, taille en octets. Valeur par défaut : `64 000`

Taille maximale des objets qui sont compressés et stockés sur disque SSD (Solid-State Drive) avant qu'ils ne soient divisés en plus petites tranches.

`bluestore_compression_min_blob_size_hdd`

Valeur : entier non signé, taille en octets. Valeur par défaut : `128 000`

Taille minimale des objets compressés et stockés sur disques durs.

bluestore_compression_max_blob_size_hdd

Valeur : entier non signé, taille en octets. Valeur par défaut : 512 000

Taille maximale des objets qui sont compressés et stockés sur des disques durs avant qu'ils ne soient divisés en plus petites tranches.

19 Réserves codées à effacement

Ceph fournit une alternative à la réplication normale des données dans les réserves : elle est appelée *réserve à effacement* ou *réserve codée à effacement*. Les réserves à effacement ne fournissent pas toutes les fonctionnalités des réserves *répliquées* (par exemple, elles ne peuvent pas stocker les métadonnées pour les réserves RBD), mais nécessitent moins de stockage brut. Une réserve à effacement par défaut capable de stocker 1 To de données requiert 1,5 To de stockage brut, ce qui permet une défaillance de disque. C'est un constat favorable par rapport à une réserve répliquée qui nécessite 2 To de stockage brut pour la même finalité.

Pour plus d'informations sur le code à effacement, reportez-vous à la page https://en.wikipedia.org/wiki/Erasure_code ↗.

Pour obtenir la liste des valeurs de réserve associées aux réserves codées à effacement, reportez-vous à la section *Valeurs de réserve codée à effacement*.

19.1 Conditions préalables pour les réserves codées à effacement

Pour utiliser le codage à effacement, vous devez :

- définir une règle d'effacement dans la carte CRUSH ;
- définir un profil de code à effacement qui spécifie l'algorithme de codage à utiliser ;
- créer une réserve utilisant la règle et le profil susmentionnés.

Gardez à l'esprit que la modification du profil et de ses détails ne sera pas possible une fois que la réserve aura été créée et contiendra des données.

Assurez-vous que les règles CRUSH des *réserves à effacement* utilisent indep pour step. Pour plus d'informations, reportez-vous à la *Section 17.3.2, « firstn et indep »*.

19.2 Création d'un exemple de réserve codée à effacement

La réserve codée à effacement la plus simple équivaut à une configuration RAID5 et nécessite au moins trois hôtes. Cette procédure décrit la création d'une réserve à des fins de test.

1. La commande **ceph osd pool create** permet de créer une réserve de type effacement (*erasure*). 12 représente le nombre de groupes de placement. Avec les paramètres par défaut, la réserve est en mesure de gérer l'échec d'un OSD.

```
cephuser@adm > ceph osd pool create ecpool 12 12 erasure  
pool 'ecpool' created
```

2. La chaîne ABCDEFGHI est écrite dans un objet appelé NYAN.

```
cephuser@adm > echo ABCDEFGHI | rados --pool ecpool put NYAN -
```

3. À des fins de test, les OSD peuvent alors être désactivés, par exemple en les déconnectant du réseau.
4. Pour tester si la réserve peut gérer l'échec des périphériques, il est possible d'accéder au contenu du fichier à l'aide de la commande **rados**.

```
cephuser@adm > rados --pool ecpool get NYAN -  
ABCDEFGHI
```

19.3 Profils de code à effacement

Lorsque la commande **ceph osd pool create** est appelée pour créer une *réserve à effacement*, le profil par défaut est utilisé, sauf si un autre profil est indiqué. Les profils définissent la redondance des données, à l'aide de deux paramètres, arbitrairement nommés k et m. k et m définissent en combien de tranches une donnée est divisée et combien de tranches de codage sont créées. Les tranches redondantes sont ensuite stockées sur des OSD différents.

Définitions requises pour les profils de réserves à effacement :

tranche

Lorsqu'elle est appelée, la fonction de codage renvoie des tranches (« chunks ») de même taille : des tranches de données pouvant être concaténées pour reconstruire l'objet d'origine et des tranches de codage pouvant être utilisées pour la reconstruction d'une tranche perdue.

k

Nombre de tranches de données, c'est-à-dire le nombre de tranches divisant l'objet original. Par exemple si $k = 2$, un objet de 10 Ko sera divisé en k objets de 5 Ko chacun. La valeur par défaut de `min_size` sur les réserves codées à effacement est $k + 1$. Cependant, nous recommandons une valeur `min_size` $k + 2$ ou plus pour éviter la perte d'écritures et de données.

m

Nombre de tranches de codage, c'est-à-dire le nombre de tranches supplémentaires calculées par les fonctions de codage. S'il existe 2 tranches de codage, cela autorise l'échec de 2 OSD sans perte de données.

crush-failure-domain

Définit les périphériques auxquels les tranches sont distribuées. Un type de compartiment doit être défini en tant que valeur. Pour tous les types de compartiment, reportez-vous à la [Section 17.2, « Compartiments »](#). Si le domaine en échec est de type `rack`, les tranches seront stockées sur des racks différents afin d'augmenter la résistance en cas de défaillances de racks. Gardez à l'esprit que cela nécessite $k + m$ racks.

Avec le profil de code à effacement par défaut utilisé à la [Section 19.2, « Création d'un exemple de réserve codée à effacement »](#), vous ne perdrez pas de données de grappe si un seul OSD ou hôte échoue. Par conséquent, pour stocker 1 To de données, il faut encore 0,5 To de stockage brut. Cela signifie que 1,5 To de stockage brut est nécessaire pour 1 To de données (étant donné que $k=2$, $m=1$). Cette configuration équivaut à une configuration RAID 5 courante. À titre de comparaison, une réserve répliquée nécessite 2 To de stockage brut pour stocker 1 To de données.

Les paramètres du profil par défaut peuvent être affichés avec les commandes suivantes :

```
cephuser@adm > ceph osd erasure-code-profile get default
directory=.libs
k=2
m=1
plugin=jerasure
crush-failure-domain=host
technique=reed_sol_van
```

Le choix du bon profil est important, car il ne peut pas être modifié après la création de la réserve. Une réserve doit être créée avec un profil différent, et tous les objets de la réserve précédente doivent être déplacés vers la nouvelle (voir [Section 18.6, « Migration d'une réserve »](#)).

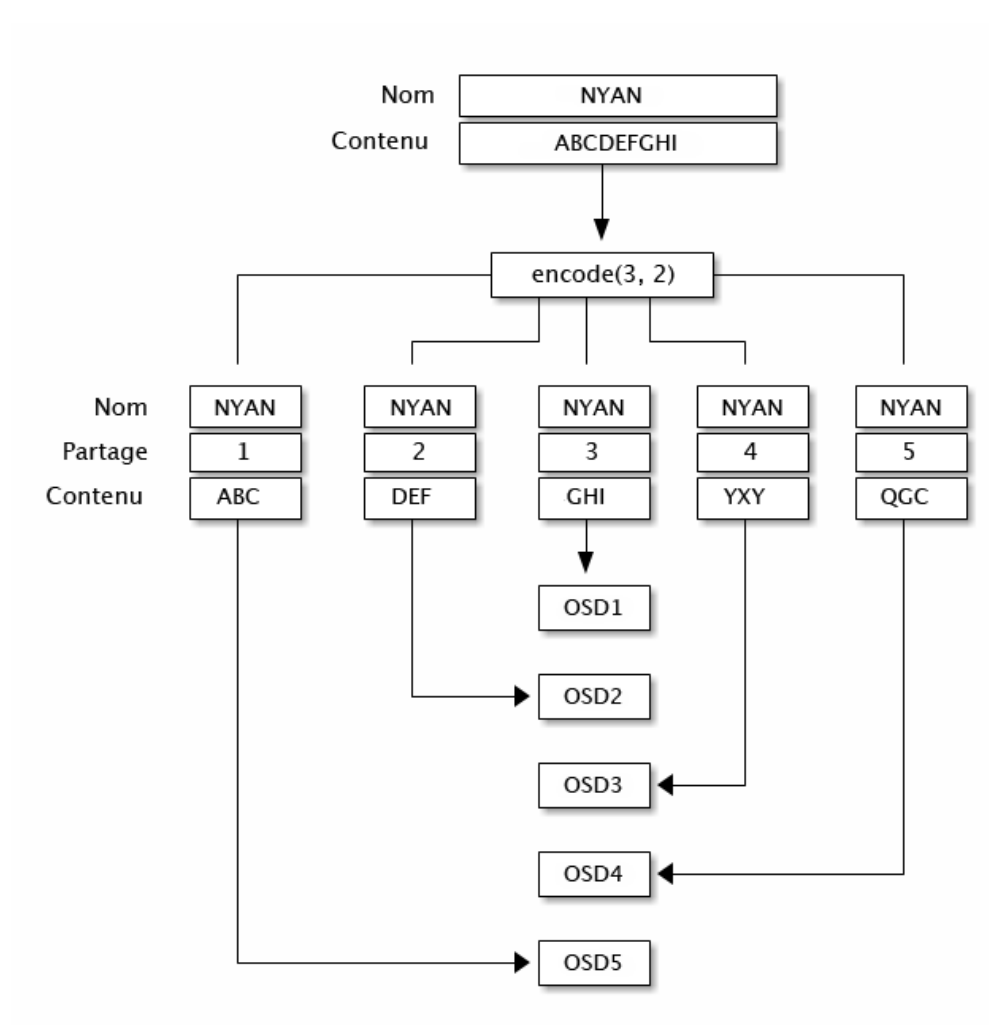
Les paramètres les plus importants du profil sont k, m et crush-failure-domain, car ils définissent l'overhead de stockage et la durabilité des données. Par exemple, si l'architecture souhaitée doit supporter la perte de deux racks avec un overhead de stockage de 66 %, le profil suivant peut être défini. Notez que cela n'est valable qu'avec une carte CRUSH comportant des compartiments de type « rack » :

```
cephuser@adm > ceph osd erasure-code-profile set myprofile \  
k=3 \  
m=2 \  
crush-failure-domain=rack
```

L'exemple de la [Section 19.2](#), « *Création d'un exemple de réserve codée à effacement* » peut être répété avec ce nouveau profil :

```
cephuser@adm > ceph osd pool create ecpool 12 12 erasure myprofile  
cephuser@adm > echo ABCDEFGHI | rados --pool ecpool put NYAN -  
cephuser@adm > rados --pool ecpool get NYAN -  
ABCDEFGHI
```

L'objet NYAN est divisé en trois (k=3) et deux tranches supplémentaires sont créées (m=2). La valeur de m définit le nombre d'OSD pouvant être perdus simultanément sans perte de données. crush-failure-domain=rack crée un ensemble de règles CRUSH qui garantit que deux tranches ne sont pas stockées dans le même rack.



19.3.1 Création d'un profil de code à effacement

La commande suivante crée un profil de code à effacement :

```
# ceph osd erasure-code-profile set NAME \
  directory=DIRECTORY \
  plugin=PLUGIN \
  stripe_unit=STRIPE_UNIT \
  KEY=VALUE ... \
  --force
```

DIRECTORY

Facultatif. Définissez le nom du répertoire à partir duquel le plug-in de code à effacement est chargé. Sa valeur par défaut est /usr/lib/ceph/erasure-code.

PLUGIN

Facultatif. Utilisez le plug-in de code à effacement pour calculer les tranches de codage et récupérer les tranches manquantes. Les plug-ins disponibles sont « jerasure », « jisa », « jlrc » et « jshes ». Le plug-in par défaut est « jerasure ».

STRIPE_UNIT

Facultatif. Quantité de données dans une tranche de données, par segment. Par exemple, un profil avec 2 tranches de données et `stripe_unit=4K` placerait la plage 0-4K dans la tranche 0, 4K-8K dans la tranche 1, puis 8K-12K de nouveau dans la tranche 0. Il doit s'agir d'un multiple de 4K pour de meilleures performances. La valeur par défaut est prise à partir de l'option de configuration du moniteur `osd_pool_erasure_code_stripe_unit` lors de la création d'une réserve. La valeur « `stripe_width` » d'une réserve utilisant ce profil sera le nombre de tranches de données multiplié par ce « `stripe_unit` ».

KEY=VALUE

Paires clé/valeur des options spécifiques au plug-in de code à effacement sélectionné.

--force

Facultatif. Permet de remplacer un profil existant portant le même nom et de définir une valeur `stripe_unit` non alignée sur 4K.

19.3.2 Suppression d'un profil de code à effacement

La commande suivante supprime un profil de code à effacement tel qu'identifié par son nom (*NAME*) :

```
# ceph osd erasure-code-profile rm NAME
```



Important

Si le profil est référencé par une réserve, la suppression échoue.

19.3.3 Affichage des détails d'un profil de code à effacement

La commande suivante affiche les détails d'un profil de code à effacement tel qu'identifié par son nom (*NAME*) :

```
# ceph osd erasure-code-profile get NAME
```

19.3.4 Liste des profils de code à effacement

La commande suivante répertorie les noms de tous les profils de code à effacement :

```
# ceph osd erasure-code-profile ls
```

19.4 Marquage des réserves codées à effacement avec périphérique de bloc RADOS (RBD)

Pour marquer une réserve EC (« Erasure Coded », codée à effacement) en tant que réserve RBD, étiquetez-la en conséquence :

```
cephuser@adm > ceph osd pool application enable rbd ec_pool_name
```

RBD peut stocker des *données* d'image dans des réserves EC. Cependant, l'en-tête et les métadonnées d'image doivent encore être stockées dans une réserve répliquée. En supposant que vous ayez la réserve nommée « rbd » à cet effet :

```
cephuser@adm > rbd create rbd/image_name --size 1T --data-pool ec_pool_name
```

Vous pouvez utiliser l'image normalement comme toute autre image, hormis que toutes les données seront stockées dans la réserve *ec_pool_name* et non pas dans la réserve « rbd ».

20 Périphérique de bloc RADOS

Un bloc est une séquence d'octets, par exemple un bloc de 4 Mo de données. Les interfaces de stockage basées sur des blocs constituent le moyen le plus courant de stocker des données sur des supports rotatifs, tels que des disques durs, des CD ou des disquettes. L'omniprésence des interfaces de périphériques de bloc fait d'un périphérique de bloc virtuel un candidat idéal pour interagir avec un système de stockage de données de masse, tel que Ceph.

Les périphériques de bloc Ceph permettent le partage de ressources physiques et sont redimensionnables. Ils stockent les données réparties sur plusieurs OSD dans une grappe Ceph. Les périphériques de bloc Ceph exploitent les fonctionnalités RADOS, telles que les instantanés, la réplique et la cohérence. Les périphériques de bloc RADOS (RADOS Block Devices, RBD) Ceph interagissent avec les OSD utilisant des modules de kernel ou la bibliothèque `librbd`.

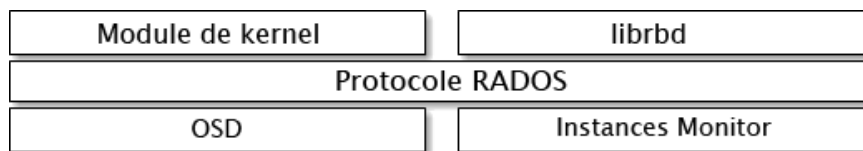


FIGURE 20.1 : PROTOCOLE RADOS

Les périphériques de bloc Ceph offrent des performances exceptionnelles ainsi qu'une évolutivité infinie des modules de kernel. Ils prennent en charge des solutions de virtualisation, telles que QEMU, ou des systèmes basés sur le cloud, tels qu'OpenStack, qui reposent sur `libvirt`. Vous pouvez utiliser la même grappe pour faire fonctionner Object Gateway, CephFS et les périphériques de bloc RADOS simultanément.

20.1 Commandes de périphériques de bloc

La commande `rbd` permet de créer, de répertorier, d'explorer et de supprimer des images de périphérique de bloc. Vous pouvez également l'utiliser, par exemple, pour cloner des images, créer des instantanés, restaurer une image dans un instantané ou afficher un instantané.

20.1.1 Création d'une image de périphérique de bloc dans une réserve répliquée

Avant de pouvoir ajouter un périphérique de bloc à un client, vous devez créer une image associée dans une réserve existante (voir [Chapitre 18, Gestion des réserves de stockage](#)) :

```
cephuser@adm > rbd create --size MEGABYTES POOL-NAME/IMAGE-NAME
```

Par exemple, pour créer une image de 1 Go nommée « myimage » qui stocke des informations dans une réserve nommée « mypool », exécutez la commande suivante :

```
cephuser@adm > rbd create --size 1024 mypool/myimage
```



Astuce : unités de taille d'image

Si vous n'indiquez pas de raccourci d'unité de taille (« G » ou « T »), la taille de l'image est en mégaoctets. Indiquez « G » ou « T » après le chiffre de la taille pour spécifier des gigaoctets ou des téraoctets.

20.1.2 Création d'une image de périphérique de bloc dans une réserve codée à effacement

Il est possible de stocker les données d'une image de périphérique de bloc directement dans des réserves codées à effacement (Erasure Coded, EC). Une image de périphérique de bloc RADOS se compose de *données* et de *métadonnées*. Seules les données d'une image de périphérique de bloc peuvent être stockées dans une réserve EC. Le drapeau `overwrite` (écraser) de la réserve doit être défini sur `true` (vrai), ce qui est possible uniquement si tous les OSD sur lesquels la réserve est stockée utilisent BlueStore.

La partie « métadonnées » de l'image ne peut pas être stockée dans une réserve EC. Vous pouvez spécifier la réserve répliquée pour stocker les métadonnées de l'image à l'aide de l'option `--pool=` de la commande `rbd create` ou spécifier `pool/` comme préfixe du nom de l'image.

Créez une réserve EC :

```
cephuser@adm > ceph osd pool create EC_POOL 12 12 erasure
cephuser@adm > ceph osd pool set EC_POOL allow_ec_overwrites true
```

Spécifiez la réserve répliquée dans laquelle stocker les métadonnées :

```
cephuser@adm > rbd create IMAGE_NAME --size=1G --data-pool EC_POOL --pool=POOL
```

Ou :

```
cephuser@adm > rbd create POOL/IMAGE_NAME --size=1G --data-pool EC_POOL
```

20.1.3 Liste des images de périphériques de bloc

Pour lister les périphériques de bloc dans une réserve nommée « mypool », exécutez la commande suivante :

```
cephuser@adm > rbd ls mypool
```

20.1.4 Récupération d'informations sur l'image

Pour récupérer des informations à partir d'une image « myimage » dans une réserve nommée « mypool », exécutez la commande suivante :

```
cephuser@adm > rbd info mypool/myimage
```

20.1.5 Redimensionnement d'une image de périphérique de bloc

Les images de périphérique de bloc RADOS sont provisionnées dynamiquement : en effet, elles n'utilisent aucun stockage physique tant que vous n'y avez pas enregistré des données. Cependant, elles possèdent une capacité maximale que vous définissez à l'aide de l'option `--size`. Si vous souhaitez augmenter (ou diminuer) la taille maximale de l'image, exécutez la commande suivante :

```
cephuser@adm > rbd resize --size 2048 POOL_NAME/IMAGE_NAME # to increase  
cephuser@adm > rbd resize --size 2048 POOL_NAME/IMAGE_NAME --allow-shrink # to decrease
```

20.1.6 Suppression d'une image de périphérique de bloc

Pour supprimer un périphérique de bloc qui correspond à une image « myimage » dans une réserve nommée « mypool », exécutez la commande suivante :

```
cephuser@adm > rbd rm mypool/myimage
```

20.2 Montage et démontage

Après avoir créé un périphérique de bloc RADOS, vous pouvez l'utiliser comme n'importe quel autre périphérique de disque et le formater, le monter pour pouvoir échanger des fichiers et le démonter une fois que vous avez terminé.

La commande **rbd** accède par défaut à la grappe à l'aide du compte utilisateur Ceph **admin**. Ce compte dispose d'un accès administratif complet à la grappe. Il existe un risque de causer accidentellement des dommages, comme lorsque vous vous connectez à un poste de travail Linux en tant que **root**. Par conséquent, il est préférable de créer des comptes utilisateur avec moins de privilèges et d'utiliser ces comptes pour un accès normal en lecture/écriture aux périphériques de bloc RADOS.

20.2.1 Création d'un compte utilisateur Ceph

Pour créer un nouveau compte utilisateur avec les fonctionnalités Ceph Manager, Ceph Monitor et Ceph OSD, utilisez la commande **ceph** avec la sous-commande **auth get-or-create** :

```
cephuser@adm > ceph auth get-or-create client.ID mon 'profile rbd' osd 'profile profile
name \
[pool=pool-name] [, profile ...]' mgr 'profile rbd [pool=pool-name]'
```

Par exemple, pour créer un utilisateur appelé **qemu** avec un accès en lecture-écriture aux **vms** de la réserve et un accès en lecture seule aux **images** de la réserve, exécutez la commande suivante :

```
ceph auth get-or-create client.qemu mon 'profile rbd' osd 'profile rbd pool=vms, profile
rbd-read-only pool=images' \
mgr 'profile rbd pool=images'
```

La sortie de la commande **ceph auth get-or-create** sera le trousseau de clés de l'utilisateur spécifié, qui peut être écrit dans **/etc/ceph/ceph.client.ID.keyring**.



Note

Lorsque vous utilisez la commande **rbd**, vous pouvez spécifier l'ID utilisateur en fournissant l'argument facultatif **--id ID**.

Pour plus d'informations sur la gestion des comptes utilisateur Ceph, reportez-vous au [Chapitre 30, Authentification avec cephx](#).

20.2.2 Authentification des utilisateurs

Pour indiquer un nom d'utilisateur, utilisez `--id nom-utilisateur`. Si vous utilisez l'authentification `cephx`, vous devez également indiquer un secret. Il peut provenir d'un trousseau de clés ou d'un fichier contenant le secret :

```
cephuser@adm > rbd device map --pool rbd myimage --id admin --keyring /path/to/keyring
```

ou

```
cephuser@adm > rbd device map --pool rbd myimage --id admin --keyfile /path/to/file
```

20.2.3 Préparation du périphérique de bloc RADOS à utiliser

1. Assurez-vous que votre grappe Ceph inclut une réserve avec l'image disque que vous souhaitez assigner. Supposons que la réserve soit appelée `mypool` et l'image `myimage`.

```
cephuser@adm > rbd list mypool
```

2. Assignez l'image à un nouveau périphérique de bloc:

```
cephuser@adm > rbd device map --pool mypool myimage
```

3. Dressez la liste de tous les périphériques assignés :

```
cephuser@adm > rbd device list
id pool  image  snap device
0  mypool myimage -   /dev/rbd0
```

Le périphérique sur lequel nous voulons travailler est `/dev/rbd0`.



Astuce : chemin du périphérique RBD

Au lieu de `/dev/rbdNUMÉRO_PÉRIPHÉRIQUE`, vous pouvez utiliser `/dev/rbd/NOM_RÉSERVE/NOM_IMAGE` comme chemin de périphérique persistant. Par exemple :

```
/dev/rbd/mypool/myimage
```

4. Créez un système de fichiers XFS sur le périphérique `/dev/rbd0`:

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

```
log stripe unit (4194304 bytes) is too large (maximum is 256KiB)
log stripe unit adjusted to 32KiB
meta-data=/dev/rbd0          isize=256    agcount=9, agsize=261120 blks
=                               sectsz=512   attr=2, projid32bit=1
=                               crc=0        finobt=0
data      =                               bsize=4096   blocks=2097152, imaxpct=25
=                               sunit=1024   swidth=1024 blks
naming    =version 2               bsize=4096   ascii-ci=0 ftype=0
log       =internal log           bsize=4096   blocks=2560, version=2
=                               sectsz=512   sunit=8 blks, lazy-count=1
realtime  =none                   extsz=4096   blocks=0, rtextents=0
```

5. En remplaçant `/mnt` par votre point de montage, montez le périphérique et vérifiez qu'il est correctement monté :

```
# mount /dev/rbd0 /mnt
# mount | grep rbd0
/dev/rbd0 on /mnt type xfs (rw,relatime,attr2,inode64,sunit=8192,...
```

Vous pouvez maintenant déplacer des données vers et depuis le périphérique comme s'il s'agissait d'un répertoire local.



Astuce : augmentation de la taille du périphérique RBD

Si vous trouvez que la taille du périphérique RBD n'est plus suffisante, vous pouvez facilement l'augmenter.

1. Augmentez la taille de l'image RBD, par exemple jusqu'à 10 Go.

```
cephuser@adm > rbd resize --size 10000 mypool/myimage
Resizing image: 100% complete...done.
```

2. Développez le système de fichiers à la nouvelle taille du périphérique:

```
# xfs_growfs /mnt
[...]
data blocks changed from 2097152 to 2560000
```

6. Après avoir accédé au périphérique, vous pouvez annuler son assignation et le démonter.

```
cephuser@adm > rbd device unmap /dev/rbd0
# umount /mnt
```



Astuce : montage et démontage manuels

Un script **rbdmmap** et une unité **systemd** sont fournis pour faciliter le processus d'assignation et de montage des RBD après le démarrage et de leur démontage avant l'arrêt. Reportez-vous à la [Section 20.2.4, « rbdmap : assignation de périphériques RBD au moment du démarrage »](#).

20.2.4 **rbdmmap** : assignation de périphériques RBD au moment du démarrage

rbdmmap est un script shell qui automatise les opérations **rbd map** et **rbd unmap** sur une ou plusieurs images RBD. Bien que vous puissiez exécuter le script manuellement à tout moment, les principaux avantages sont l'assignation et le montage automatiques des images RBD au démarrage (ainsi que le démontage et la désassignation à l'arrêt) déclenchés par le système Init. À cet effet, un fichier unité **systemd**, **rbdmmap.service**, est fourni avec le paquetage **ceph-common**.

Le script accepte un argument unique, qui peut être **map** ou **unmap**. Dans les deux cas, le script analyse un fichier de configuration. Il pointe vers **/etc/ceph/rbdmap** par défaut, mais peut être remplacé par le biais d'une variable d'environnement **RBDMAPIFILE**. Chaque ligne du fichier de configuration correspond à une image RBD qui doit être assignée ou dont l'assignation doit être annulée.

Le fichier de configuration possède le format suivant :

```
image_specification rbd_options
```

image_specification

Chemin d'accès à une image dans une réserve. Indiquez-le en tant que nom_réserve/nom_image.

rbd_options

Liste facultative de paramètres à transmettre à la commande **rbd device map** sous-jacente. Ces paramètres et leurs valeurs doivent être indiqués en tant que chaîne séparée par des virgules, par exemple :

```
PARAM1=VAL1,PARAM2=VAL2,...
```

Dans cet exemple suivant, le script **rbdmmap** exécute la commande suivante :

```
cephuser@adm > rbd device map POOL_NAME/IMAGE_NAME --PARAM1 VAL1 --PARAM2 VAL2
```

L'exemple suivant illustre comment spécifier un nom d'utilisateur et un trousseau de clés avec un secret correspondant :

```
cephuser@adm > rbdmap device map mypool/myimage id=rbd_user,keyring=/etc/ceph/ceph.client.rbd.keyring
```

Lorsqu'il est exécuté en tant que **rbdmap map**, le script analyse le fichier de configuration et, pour chaque image RBD indiquée, tente d'abord d'assigner l'image (à l'aide de la commande **rbd device map**), puis de la monter.

Lorsqu'elles sont exécutées en tant que **rbdmap unmap**, les images répertoriées dans le fichier de configuration seront démontées et désassignées.

rbdmap unmap-all tente de démonter puis de désassigner toutes les images RBD actuellement assignées, qu'elles soient ou non répertoriées dans le fichier de configuration.

En cas de réussite, l'opération **rbd device map** assigne l'image à un périphérique `/dev/rbdX` ; une règle udev est alors déclenchée afin de créer un lien symbolique de nom de périphérique convivial `/dev/rbd/nom_réserve/nom_image` pointant vers le périphérique réellement assigné.

Pour que le montage et le démontage réussissent, le nom de périphérique « convivial » doit être répertorié dans le fichier `/etc/fstab`. Lors de l'écriture d'entrées `/etc/fstab` pour les images RBD, indiquez l'option de montage « noauto » (ou « nofail »). Cela empêche le système Init d'essayer de monter le périphérique trop tôt, avant même que le périphérique en question existe, car `rbdmap.service` est généralement déclenché assez tard dans la séquence de démarrage.

Pour obtenir la liste complète des options de **rbd**, reportez-vous à la page de manuel **rbd** (**man 8 rbd**).

Pour obtenir des exemples de l'utilisation de **rbd**, reportez-vous à la page de manuel de **rbd** (**man 8 rbd**).

20.2.5 Augmentation de la taille des périphériques RBD

Si vous trouvez que la taille du périphérique RBD n'est plus suffisante, vous pouvez facilement l'augmenter.

1. Augmentez la taille de l'image RBD, par exemple jusqu'à 10 Go.

```
cephuser@adm > rbd resize --size 10000 mypool/myimage  
Resizing image: 100% complete...done.
```

2. Développez le système de fichiers à la nouvelle taille du périphérique.

```
# xfs_growfs /mnt
[...]  
data blocks changed from 2097152 to 2560000
```

20.3 Images instantanées

Un instantané RBD est un instantané d'une image de périphérique de bloc RADOS. Avec les instantanés, vous conservez l'historique de l'état de l'image. Ceph prend également en charge la superposition d'instantanés, ce qui vous permet de cloner des images de machine virtuelle rapidement et facilement. Ceph prend en charge les instantanés de périphériques de bloc en utilisant la commande **rbd** et de nombreuses interfaces de niveau supérieur, notamment QEMU, libvirt, OpenStack et CloudStack.



Note

Arrêtez les opérations d'entrée et de sortie et videz toutes les écritures en attente avant de créer l'instantané d'une image. Si l'image contient un système de fichiers, celui-ci doit être cohérent lors de la création de l'instantané.

20.3.1 Activation et configuration de cephx

Quand cephx est activé, vous devez spécifier un nom ou un ID d'utilisateur et un chemin d'accès au trousseau de clés contenant la clé correspondante pour l'utilisateur. Pour plus d'informations, reportez-vous au [Chapitre 30, Authentification avec cephx](#). Vous pouvez également ajouter la variable d'environnement CEPH_ARGS pour ne pas avoir à saisir à nouveau les paramètres suivants.

```
cephuser@adm > rbd --id user-ID --keyring=/path/to/secret commands  
cephuser@adm > rbd --name username --keyring=/path/to/secret commands
```

Par exemple :

```
cephuser@adm > rbd --id admin --keyring=/etc/ceph/ceph.keyring commands  
cephuser@adm > rbd --name client.admin --keyring=/etc/ceph/ceph.keyring commands
```



Astuce

Ajoutez l'utilisateur et le secret à la variable d'environnement `CEPH_ARGS` afin de ne pas avoir à les saisir à chaque fois.

20.3.2 Notions de base sur les instantanés

Les procédures suivantes montrent comment créer, répertorier et supprimer des instantanés à l'aide de la commande `rbd` sur la ligne de commande.

20.3.2.1 Création d'instantanés

Pour créer un instantané avec `rbd`, indiquez l'option `snap create`, le nom de la réserve et le nom de l'image.

```
cephuser@adm > rbd --pool pool-name snap create --snap snap-name image-name
cephuser@adm > rbd snap create pool-name/image-name@snap-name
```

Par exemple :

```
cephuser@adm > rbd --pool rbd snap create --snap snapshot1 image1
cephuser@adm > rbd snap create rbd/image1@snapshot1
```

20.3.2.2 Liste des instantanés

Pour répertorier les instantanés d'une image, spécifiez le nom de la réserve et le nom de l'image.

```
cephuser@adm > rbd --pool pool-name snap ls image-name
cephuser@adm > rbd snap ls pool-name/image-name
```

Par exemple :

```
cephuser@adm > rbd --pool rbd snap ls image1
cephuser@adm > rbd snap ls rbd/image1
```

20.3.2.3 Restauration de l'état initial des instantanés

Pour rétablir l'état initial d'un instantané avec `rbd`, indiquez l'option `snap rollback`, le nom de la réserve, le nom de l'image et le nom de l'instantané.

```
cephuser@adm > rbd --pool pool-name snap rollback --snap snap-name image-name
cephuser@adm > rbd snap rollback pool-name/image-name@snap-name
```

Par exemple :

```
cephuser@adm > rbd --pool pool1 snap rollback --snap snapshot1 image1
cephuser@adm > rbd snap rollback pool1/image1@snapshot1
```



Note

Le retour à l'état initial d'une image dans un instantané revient à écraser la version actuelle de l'image avec les données d'un instantané. Le temps nécessaire à l'exécution d'un retour à l'état initial augmente avec la taille de l'image. Il est *plus rapide de cloner* à partir d'un instantané *que de rétablir* une image vers un instantané, cette méthode étant recommandée pour revenir à un état préexistant.

20.3.2.4 Suppression d'un instantané

Pour supprimer un instantané avec **rbd**, indiquez l'option `snap rm`, le nom de la réserve, le nom de l'image et le nom de l'utilisateur.

```
cephuser@adm > rbd --pool pool-name snap rm --snap snap-name image-name
cephuser@adm > rbd snap rm pool-name/image-name@snap-name
```

Par exemple :

```
cephuser@adm > rbd --pool pool1 snap rm --snap snapshot1 image1
cephuser@adm > rbd snap rm pool1/image1@snapshot1
```



Note

Les instances Ceph OSD suppriment les données de manière asynchrone de sorte que la suppression d'un instantané ne libère pas immédiatement l'espace disque.

20.3.2.5 Purge des instantanés

Pour supprimer tous les instantanés d'une image avec **rbd**, indiquez l'option `snap purge` et le nom de l'image.

```
cephuser@adm > rbd --pool pool-name snap purge image-name
cephuser@adm > rbd snap purge pool-name/image-name
```

Par exemple :

```
cephuser@adm > rbd --pool pool1 snap purge image1
cephuser@adm > rbd snap purge pool1/image1
```

20.3.3 Superposition d'instantanés

Ceph prend en charge la possibilité de créer plusieurs clones de copie à l'écriture (COW) d'un instantané de périphérique de bloc. La superposition d'instantanés donne aux clients de périphériques de bloc Ceph les moyens de créer des images très rapidement. Par exemple, vous pouvez créer une image de périphérique de bloc avec une machine virtuelle Linux écrite, puis capturer l'image, protéger l'instantané et créer autant de clones COW que vous le souhaitez. Un instantané étant en lecture seule, le clonage d'un instantané simplifie la sémantique et permet de créer rapidement des clones.



Note

Dans les exemples de ligne de commande ci-dessous, les termes « parent » et « child » (enfant) désignent un instantané de périphérique de bloc Ceph (parent) et l'image correspondante clonée à partir de l'instantané (enfant).

Chaque image clonée (enfant) stocke une référence à son image parent, ce qui permet à l'image clonée d'ouvrir l'instantané parent et de le lire.

Un clone COW d'un instantané se comporte exactement comme n'importe quelle autre image de périphérique Ceph. Vous pouvez lire, écrire, cloner et redimensionner des images clonées. Il n'y a pas de restrictions spéciales avec les images clonées. Cependant, le clone copy-on-write d'un instantané fait référence à l'instantané, donc vous *devez* protéger l'instantané avant de le cloner.



Note : --image-format 1 non pris en charge

Vous ne pouvez pas créer d'instantanés d'images créés avec l'option **rbd create --image-format 1** obsolète. Ceph ne prend en charge que le clonage des images *format 2* par défaut.

20.3.3.1 Démarrage de la superposition

La superposition de périphériques de bloc Ceph est un processus simple. Vous devez disposer d'une image. Vous devez créer un instantané de l'image. Vous devez protéger l'instantané. Après avoir effectué ces étapes, vous pouvez commencer le clonage de l'instantané.

L'image clonée contient une référence à l'instantané parent et inclut l'ID de la réserve, l'ID de l'image et l'ID de l'instantané. L'inclusion de l'ID de réserve signifie que vous pouvez cloner des instantanés d'une réserve vers des images d'une autre réserve.

- *Modèle d'image* : un cas d'utilisation courant de la superposition de périphériques de bloc consiste à créer une image principale et un instantané servant de modèle aux clones. Par exemple, un utilisateur peut créer une image pour une distribution Linux (par exemple, SUSE Linux Enterprise Server (SLES)) et créer un instantané correspondant. Périodiquement, l'utilisateur peut mettre à jour l'image et créer un instantané (par exemple, `zypper ref && zypper patch` suivi de `rbid snap create`). Au fur et à mesure que l'image mûrit, l'utilisateur peut cloner l'un des instantanés.
- *Modèle étendu* : un cas d'utilisation plus avancée inclut l'extension d'une image modèle fournissant plus d'informations qu'une image de base. Par exemple, un utilisateur peut cloner une image (un modèle de machine virtuelle) et installer d'autres logiciels (par exemple, une base de données, un système de gestion de contenu ou un système d'analyse), puis prendre un instantané de l'image agrandie, qui peut elle-même être mise à jour de la même manière que l'image de base.
- *Réserve de modèles* : une façon d'utiliser la superposition de périphériques de bloc consiste à créer une réserve contenant des images principales agissant comme des modèles et des instantanés de ces modèles. Vous pouvez ensuite étendre les privilèges de lecture seule aux utilisateurs afin qu'ils puissent cloner les instantanés sans possibilité d'écriture ou d'exécution dans la réserve.
- *Migration/récupération d'image* : une façon d'utiliser la superposition de périphériques de bloc consiste à migrer ou récupérer des données d'une réserve vers une autre.

20.3.3.2 Protection d'un instantané

Les clones accèdent aux instantanés parents. Tous les clones seraient endommagés si un utilisateur supprimait par inadvertance l'instantané parent. Pour éviter toute perte de données, vous devez protéger l'instantané avant de pouvoir le cloner.

```
cephuser@adm > rbd --pool pool-name snap protect \  
--image image-name --snap snapshot-name  
cephuser@adm > rbd snap protect pool-name/image-name@snapshot-name
```

Par exemple :

```
cephuser@adm > rbd --pool pool1 snap protect --image image1 --snap snapshot1  
cephuser@adm > rbd snap protect pool1/image1@snapshot1
```



Note

Vous ne pouvez pas supprimer un instantané protégé.

20.3.3.3 Clonage d'un instantané

Pour cloner un instantané, vous devez spécifier la réserve parent, l'image, l'instantané, la réserve enfant et le nom de l'image. Vous devez protéger l'instantané avant de pouvoir le cloner.

```
cephuser@adm > rbd clone --pool pool-name --image parent-image \  
--snap snap-name --dest-pool pool-name \  
--dest child-image  
cephuser@adm > rbd clone pool-name/parent-image@snap-name \  
pool-name/child-image-name
```

Par exemple :

```
cephuser@adm > rbd clone pool1/image1@snapshot1 pool1/image2
```



Note

Vous pouvez cloner un instantané d'une réserve vers une image d'une autre réserve. Par exemple, vous pouvez gérer des images en lecture seule et des instantanés en tant que modèles dans une réserve, d'une part, et des clones inscriptibles dans une autre réserve, d'autre part.

20.3.3.4 Suppression de la protection d'un instantané

Avant de pouvoir supprimer un instantané, vous devez d'abord le déprotéger. En outre, vous pouvez *ne pas* supprimer des instantanés contenant des références de clones. Vous devez fusionner chaque clone d'un instantané avant de pouvoir supprimer celui-ci.

```
cephuser@adm > rbd --pool pool-name snap unprotect --image image-name \
--snap snapshot-name
cephuser@adm > rbd snap unprotect pool-name/image-name@snapshot-name
```

Par exemple :

```
cephuser@adm > rbd --pool pool1 snap unprotect --image image1 --snap snapshot1
cephuser@adm > rbd snap unprotect pool1/image1@snapshot1
```

20.3.3.5 Liste des enfants d'un instantané

Pour dresser la liste des enfants d'un instantané, exécutez :

```
cephuser@adm > rbd --pool pool-name children --image image-name --snap snap-name
cephuser@adm > rbd children pool-name/image-name@snapshot-name
```

Par exemple :

```
cephuser@adm > rbd --pool pool1 children --image image1 --snap snapshot1
cephuser@adm > rbd children pool1/image1@snapshot1
```

20.3.3.6 Mise à plat d'une image clonée

Les images clonées conservent une référence à l'instantané parent. Lorsque vous supprimez la référence du clone enfant dans l'instantané parent, vous « aplatissez » (fusionnez) l'image en copiant les informations de l'instantané sur le clone. Le temps nécessaire à la fusion d'un clone augmente avec la taille de l'instantané. Pour supprimer un instantané, vous devez d'abord fusionner les images enfant.

```
cephuser@adm > rbd --pool pool-name flatten --image image-name
cephuser@adm > rbd flatten pool-name/image-name
```

Par exemple :

```
cephuser@adm > rbd --pool pool1 flatten --image image1
cephuser@adm > rbd flatten pool1/image1
```



Note

Comme une image fusionnée contient toutes les informations de l'instantané, elle occupe plus d'espace de stockage qu'un clone en couches.

20.4 Miroirs d'image RBD

Les images RBD peuvent être mises en miroir de manière asynchrone entre deux grappes Ceph. Cette fonctionnalité est disponible en deux modes :

Mode basé sur un journal

Ce mode utilise la fonctionnalité de journalisation de l'image RBD afin de garantir une réplication ponctuelle, cohérente entre les grappes en cas de panne. Chaque écriture dans l'image RBD est d'abord enregistrée dans le journal associé avant de réellement modifier l'image. La grappe remote lira le journal et relira les mises à jour de sa copie locale de l'image. Étant donné que chaque écriture dans l'image RBD entraîne deux écritures dans la grappe Ceph, attendez-vous à ce que les temps de latence en écriture soient pratiquement multipliés par deux lorsque vous utilisez la fonctionnalité de journalisation de l'image RBD.

Mode basé sur des instantanés

Ce mode utilise des instantanés en miroir d'image RBD planifiés ou créés manuellement pour répliquer des images RBD cohérentes entre les grappes en cas de panne. La grappe remote détermine toutes les mises à jour de données ou de métadonnées entre deux instantanés-miroir et copie les différences dans sa copie locale de l'image. La fonctionnalité d'image RBD fast-diff permet de calculer rapidement les blocs de données mis à jour sans devoir analyser toute l'image RBD. Étant donné que ce mode n'est pas cohérent à un moment donné, la différence de l'instantané complet devra être synchronisée avant d'être utilisée pendant un scénario de basculement. Toutes les différences d'instantanés partiellement appliquées sont restaurées vers l'état du dernier instantané entièrement synchronisé avant utilisation.

La mise en miroir est configurée réserve par réserve au sein des grappes homologues. Elle peut être configurée sur un sous-ensemble spécifique d'images dans la réserve ou configurée pour mettre en miroir automatiquement toutes les images d'une réserve lorsque vous utilisez la mise en miroir basée sur le journal uniquement. La mise en miroir est configurée à l'aide de la commande **rbd**. Le daemon `rbd-mirror` est chargé d'extraire les mises à jour de l'image de la grappe homologue remote (distante) et de les appliquer à l'image dans la grappe local (locale).

Selon les besoins de réplication souhaités, la mise en miroir RBD peut être configurée pour une réplication unidirectionnelle ou bidirectionnelle :

Réplication unidirectionnelle

Lorsque les données sont mises en miroir uniquement à partir d'une grappe primaire vers une grappe secondaire, le daemon `rbd-mirror` s'exécute uniquement sur la grappe secondaire.

Réplication bidirectionnelle

Lorsque les données sont mises en miroir à partir des images primaires sur une grappe vers des images non primaires sur une autre grappe (et inversement), le daemon `rbd-mirror` s'exécute sur les deux grappes.



Important

Chaque instance du daemon `rbd-mirror` doit pouvoir se connecter simultanément aux grappes Ceph locales (`local`) et distantes (`remote`). Par exemple, tous les hôtes du moniteur et OSD. En outre, le réseau doit disposer de suffisamment de bande passante entre les deux centres de données pour gérer le workload en miroir.

20.4.1 Configuration de la réserve

Les procédures suivantes montrent comment effectuer les tâches d'administration de base pour configurer la mise en miroir à l'aide de la commande `rbd`. La mise en miroir est configurée réserve par réserve au sein des grappes Ceph.

Vous devez effectuer les étapes de configuration de la réserve sur les deux grappes homologues. Ces procédures supposent que deux grappes, nommées `local` et `remote`, sont accessibles depuis un seul hôte pour plus de clarté.

Reportez-vous à la page de manuel `rbd` ([man 8 rbd](#)) pour plus de détails sur la procédure de connexion à des grappes Ceph différentes.



Astuce : grappes multiples

Le nom de la grappe dans les exemples suivants correspond à un fichier de configuration Ceph du même nom `/etc/ceph/remote.conf` et à un fichier de trousseau de clés Ceph du même nom `/etc/ceph/remote.client.admin.keyring`.

20.4.1.1 Activation de la mise en miroir sur une réserve

Pour activer la mise en miroir sur une grappe, indiquez la sous-commande **mirror pool enable**, le nom de la réserve et le mode de mise en miroir. Le mode de mise en miroir peut être **pool** ou **image** :

pool

Toutes les images de la réserve dont la fonctionnalité de journalisation est activée sont mises en miroir.

image

La mise en miroir doit être explicitement activée sur chaque image. Pour plus d'informations, reportez-vous à la [Section 20.4.2.1, « Activation de la mise en miroir d'images »](#).

Par exemple :

```
cephuser@adm > rbd --cluster local mirror pool enable POOL_NAME pool
cephuser@adm > rbd --cluster remote mirror pool enable POOL_NAME pool
```

20.4.1.2 Désactivation de la mise en miroir

Pour désactiver la mise en miroir sur une grappe, indiquez la sous-commande **mirror pool disable** et le nom de la réserve. Lorsque la mise en miroir est désactivée sur une réserve de cette manière, la mise en miroir est également désactivée sur toutes les images (dans la réserve) pour lesquelles la mise en miroir a été explicitement activée.

```
cephuser@adm > rbd --cluster local mirror pool disable POOL_NAME
cephuser@adm > rbd --cluster remote mirror pool disable POOL_NAME
```

20.4.1.3 Démarrage des homologues

Pour que le daemon **rbd-mirror** découvre sa grappe homologue, l'homologue doit être enregistré dans la réserve et un compte utilisateur doit être créé. Ce processus peut être automatisé avec **rbd** et les commandes **mirror pool peer bootstrap create** ainsi que **mirror pool peer bootstrap import**.

Pour créer manuellement un nouveau jeton de démarrage avec **rbd**, spécifiez la commande **mirror pool peer bootstrap create**, un nom de réserve, ainsi qu'un nom de site convivial facultatif pour décrire la grappe **local** :

```
cephuser@local > rbd mirror pool peer bootstrap create \
```

```
[--site-name LOCAL_SITE_NAME] POOL_NAME
```

La sortie de la commande **mirror pool peer bootstrap create** sera un jeton qui doit être fourni à la commande **mirror pool peer bootstrap import**. Par exemple, sur la grappe local :

```
cephuser@local > rbd --cluster local mirror pool peer bootstrap create --site-name local
image-pool
eyJmc2lkIjoioWY1MjgyZGItYjg5OS00NTk2LTgwOTgtMzIwYzFmYzM5NmYzIiwiaY2xpZW50X2lkIjoicmJkLWlpcnJvcilwZWVyIiw
\
joiQVFBUnczOWQwdkhvQmhBQVlMM1I4RmR5dHNJQU50bkFTZ0l0TVE9PSIsIm1vbl9ob3N0Ijoiw3Yy0jE5Mi4xNjguMS4z0jY4MjAs
```

Pour importer manuellement le jeton de démarrage créé par une autre grappe avec la commande **rbd**, utilisez la syntaxe suivante :

```
rbd mirror pool peer bootstrap import \
[--site-name LOCAL_SITE_NAME] \
[--direction DIRECTION] \
POOL_NAME TOKEN_PATH
```

Où :

LOCAL_SITE_NAME

Nom facultatif convivial du site pour décrire la grappe local.

DIRECTION

Direction de la mise en miroir. La valeur par défaut est rx-tx pour la mise en miroir bidirectionnelle, mais peut également être définie sur rx-only pour la mise en miroir unidirectionnelle.

POOL_NAME

Nom de la réserve.

TOKEN_PATH

Chemin du fichier pour accéder au jeton créé (ou - pour le lire à partir de l'entrée standard).

Par exemple, sur la grappe remote :

```
cephuser@remote > cat <<EOF > token
eyJmc2lkIjoioWY1MjgyZGItYjg5OS00NTk2LTgwOTgtMzIwYzFmYzM5NmYzIiwiaY2xpZW50X2lkIjoicmJkLWlpcnJvcilwZWVyIiw
EOF
```

```
cephuser@adm > rbd --cluster remote mirror pool peer bootstrap import \
```

```
--site-name remote image-pool token
```

20.4.1.4 Ajout manuel d'un homologue de grappe

Au lieu de démarrer des homologues comme décrit à la [Section 20.4.1.3, « Démarrage des homologues »](#), vous pouvez spécifier des homologues manuellement. Le daemon `rbd-mirror` distant doit accéder à la grappe locale pour effectuer la mise en miroir. Créez un nouvel utilisateur Ceph local que le daemon `rbd-mirror` distant utilisera, par exemple, `rbd-mirror-peer` :

```
cephuser@adm > ceph auth get-or-create client.rbd-mirror-peer \
mon 'profile rbd' osd 'profile rbd'
```

Utilisez la syntaxe suivante pour ajouter une grappe Ceph homologue en miroir avec la commande `rbd` :

```
rbd mirror pool peer add POOL_NAME CLIENT_NAME@CLUSTER_NAME
```

Par exemple :

```
cephuser@adm > rbd --cluster site-a mirror pool peer add image-pool client.rbd-mirror-
peer@site-b
cephuser@adm > rbd --cluster site-b mirror pool peer add image-pool client.rbd-mirror-
peer@site-a
```

Par défaut, le daemon `rbd-mirror` doit avoir accès au fichier de configuration Ceph situé à l'emplacement `/etc/ceph/.NOM_GRAPPE.conf`. Il fournit les adresses IP des instances MON de la grappe homologue et un trousseau de clés pour un client nommé `NOM_CLIENT` situé dans les chemins de recherche par défaut ou personnalisés du trousseau de clés, par exemple `/etc/ceph/NOM_GRAPPE.NOM_CLIENT.keyring`.

Sinon, l'instance MON et/ou la clé du client de la grappe homologue peut être stockée en toute sécurité dans la zone de stockage locale `config-key` de Ceph. Pour spécifier les attributs de connexion de la grappe homologue lors de l'ajout d'un homologue en miroir, utilisez les options `--remote-mon-host` et `--remote-key-file`. Par exemple :

```
cephuser@adm > rbd --cluster site-a mirror pool peer add image-pool \
client.rbd-mirror-peer@site-b --remote-mon-host 192.168.1.1,192.168.1.2 \
--remote-key-file /PATH/TO/KEY_FILE
cephuser@adm > rbd --cluster site-a mirror pool info image-pool --all
Mode: pool
Peers:
  UUID      NAME      CLIENT      MON_HOST      KEY
```



```
587b08db... site-b client.rbd-mirror-peer 192.168.1.1,192.168.1.2 AQAeuZdb...
```

20.4.1.5 Suppression d'un homologue de grappe

Pour supprimer une grappe homologue de mise en miroir, indiquez la sous-commande **mirror pool peer remove**, le nom de la réserve et l'UUID de l'homologue (disponible dans le résultat de la commande **rbd mirror pool info**) :

```
cephuser@adm > rbd --cluster local mirror pool peer remove POOL_NAME \
55672766-c02b-4729-8567-f13a66893445
cephuser@adm > rbd --cluster remote mirror pool peer remove POOL_NAME \
60c0e299-b38f-4234-91f6-eed0a367be08
```

20.4.1.6 Réserves de données

Lors de la création d'images dans la grappe cible, **rbd-mirror** sélectionne une réserve de données comme suit :

- Si une réserve de données par défaut est configurée pour la grappe cible (avec l'option de configuration **rbd_default_data_pool**), cette réserve sera utilisée.
- Dans le cas contraire, si l'image source utilise une réserve de données distincte et qu'une réserve portant le même nom existe sur la grappe cible, cette réserve est utilisée.
- Si aucune des conditions ci-dessus n'est vraie, aucune réserve de données n'est configurée.

20.4.2 Configuration de l'image RBD

Contrairement à la configuration de réserve, la configuration d'image ne doit être effectuée que par rapport à une seule grappe Ceph homologue de mise en miroir.

Les images RBD en miroir sont désignées comme étant soit *primaires*, soit *non primaires*. Il s'agit d'une propriété de l'image et non pas de la réserve. Les images qui sont désignées comme non primaires ne sont pas modifiables.

Les images sont automatiquement promues au rang d'images primaires lorsque la mise en miroir est activée pour la première fois sur une image (soit implicitement si le mode de mise en miroir de la réserve était « pool » et que la fonctionnalité de journalisation de l'image a été activée, soit explicitement – reportez-vous à la [Section 20.4.2.1, « Activation de la mise en miroir d'images »](#) – à l'aide de la commande **rbd**).

20.4.2.1 Activation de la mise en miroir d'images

Si la mise en miroir est configurée en mode image, il est nécessaire d'activer explicitement la mise en miroir pour chaque image de la réserve. Pour activer la mise en miroir d'une image en particulier avec la commande **rbd**, indiquez la sous-commande **mirror image enable** ainsi que le nom de la réserve et le nom de l'image :

```
cephuser@adm > rbd --cluster local mirror image enable \  
POOL_NAME/IMAGE_NAME
```

Le mode d'image en miroir peut être journal ou snapshot :

journal (valeur par défaut)

Lorsqu'elle est configurée en mode journal, la mise en miroir utilise la fonctionnalité de journalisation de l'image RBD pour répliquer le contenu de l'image. Si la fonction de journalisation de l'image RBD n'est pas encore activée sur l'image, elle sera activée automatiquement.

snapshot

Lorsqu'elle est configurée en mode snapshot (instantané), la mise en miroir utilise des instantanés-miroir de l'image RBD pour répliquer le contenu de l'image. Une fois activé, un instantané-miroir initial est automatiquement créé. Des instantanés-miroir de l'image RBD supplémentaires peuvent être créés à l'aide de la commande **rbd**.

Par exemple :

```
cephuser@adm > rbd --cluster local mirror image enable image-pool/image-1 snapshot  
cephuser@adm > rbd --cluster local mirror image enable image-pool/image-2 journal
```

20.4.2.2 Activation de la fonctionnalité de journalisation des images

La mise en miroir RBD utilise la fonctionnalité de journalisation RBD pour garantir que l'image répliquée est préservée en cas de panne. Lorsque vous utilisez le mode de mise en miroir image, la fonctionnalité de journalisation est automatiquement activée si la mise en miroir est activée sur l'image. Lorsque vous utilisez le mode de mise en miroir pool, avant qu'une image puisse être mise en miroir sur une grappe homologue, la fonction de journalisation d'image RBD doit être activée. La fonctionnalité peut être activée au moment de la création de l'image en indiquant l'option --image-feature exclusive-lock, journaling dans la commande **rbd**.

Le cas échéant, la fonction de journalisation peut être dynamiquement activée sur des images RBD préexistantes. Pour activer la journalisation, indiquez la sous-commande **feature enable**, le nom de la réserve et de l'image, et le nom de l'entité :

```
cephuser@adm > rbd --cluster local feature enable POOL_NAME/IMAGE_NAME exclusive-lock
cephuser@adm > rbd --cluster local feature enable POOL_NAME/IMAGE_NAME journaling
```



Note : dépendance des options

La fonctionnalité journaling dépend de la fonctionnalité exclusive-lock. Si la fonctionnalité exclusive-lock n'est pas encore activée, vous devez l'activer avant la fonctionnalité journaling.



Astuce

Vous pouvez activer la journalisation sur toutes les nouvelles images par défaut en ajoutant rd default features = layering,exclusive-lock,object-map,deep-flat-ten,journaling à votre fichier de configuration Ceph.

20.4.2.3 Création d'instantanés-miroir d'images

Lors de l'utilisation de la mise en miroir basée sur des instantanés, des instantanés-miroir doivent être créés chaque fois que vous souhaitez mettre en miroir le contenu modifié de l'image RBD. Pour créer manuellement un instantané-miroir à l'aide de la commande **rd**, spécifiez la commande **mirror image snapshot** ainsi que le nom de la réserve et de l'image :

```
cephuser@adm > rbd mirror image snapshot POOL_NAME/IMAGE_NAME
```

Par exemple :

```
cephuser@adm > rbd --cluster local mirror image snapshot image-pool/image-1
```

Par défaut, seuls trois instantanés-miroir sont créés par image. L'instantané-miroir le plus récent est automatiquement nettoyé si la limite est atteinte. La limite peut être remplacée par l'option de configuration rd mirroring_max_mirroring_snapshots si nécessaire. En outre, les instantanés-miroir sont automatiquement supprimés lors du retrait de l'image ou de la désactivation de la mise en miroir.

Des instantanés-miroir peuvent également être créés automatiquement à intervalles réguliers si des planifications d'instantanés-miroir sont définies. L'instantané-miroir peut être planifié de manière globale, par réserve ou par image. Plusieurs planifications d'instantanés-miroir peuvent être définies à n'importe quel niveau, mais seules les planifications d'instantanés les plus spécifiques qui correspondent à une image en miroir individuelle seront exécutées.

Pour créer une planification d'instantanés-miroir avec **rbd**, spécifiez la commande **mirror snapshot schedule add** ainsi qu'un nom de réserve ou d'image, un intervalle et une heure de début facultative.

L'intervalle peut être spécifié en jours, heures ou minutes respectivement à l'aide des suffixes **d**, **h** ou **m**. L'heure de début facultative peut être spécifiée à l'aide du format d'heure ISO 8601. Par exemple :

```
cephuser@adm > rbd --cluster local mirror snapshot schedule add --pool image-pool 24h
14:00:00-05:00
cephuser@adm > rbd --cluster local mirror snapshot schedule add --pool image-pool --image
image1 6h
```

Pour supprimer une planification d'instantané-miroir avec **rbd**, spécifiez la commande **mirror snapshot schedule remove** avec des options qui correspondent à la commande d'ajout de planification correspondante.

Pour répertorier toutes les planifications d'instantanés d'un niveau spécifique (global, réserve ou image) avec la commande **rbd**, spécifiez la commande **mirror snapshot schedule ls** avec un nom facultatif de réserve ou d'image. En outre, l'option **--recursive** peut être spécifiée pour répertorier toutes les planifications du niveau spécifié et des niveaux inférieurs. Par exemple :

```
cephuser@adm > rbd --cluster local mirror schedule ls --pool image-pool --recursive
POOL      NAMESPACE IMAGE  SCHEDULE
image-pool -          -    every 1d starting at 14:00:00-05:00
image-pool          image1 every 6h
```

Pour savoir quand les prochains instantanés seront créés pour les images RBD en miroir basées sur des instantanés avec **rbd**, spécifiez la commande **mirror snapshot schedule status** ainsi qu'un nom de réserve ou d'image facultatif. Par exemple :

```
cephuser@adm > rbd --cluster local mirror schedule status
SCHEDULE TIME      IMAGE
2020-02-26 18:00:00 image-pool/image1
```

20.4.2.4 Désactivation de la mise en miroir d'images

Pour désactiver la mise en miroir d'une image en particulier, indiquez la sous-commande **mirror image disable** avec le nom de la réserve et le nom de l'image :

```
cephuser@adm > rbd --cluster local mirror image disable POOL_NAME/IMAGE_NAME
```

20.4.2.5 Promotion et rétrogradation d'images

Dans un scénario de basculement où la désignation principale doit être déplacée sur l'image dans la grappe homologue, vous devez arrêter l'accès à l'image primaire, rétrograder l'image primaire actuelle, promouvoir la nouvelle image primaire et reprendre l'accès à l'image sur la grappe alternative.



Note : promotion forcée

La promotion peut être forcée à l'aide de l'option **--force**. La promotion forcée est nécessaire lorsque la rétrogradation ne peut pas être propagée à la grappe homologue (par exemple, en cas d'échec de la grappe ou de panne de communication). Cela se traduira par un scénario de divergence entre les deux homologues, et l'image ne sera plus synchronisée jusqu'à l'émission de la sous-commande **resync**.

Pour rétrograder une image non primaire spécifique, indiquez la sous-commande **mirror image demote** ainsi que le nom de la réserve et le nom de l'image :

```
cephuser@adm > rbd --cluster local mirror image demote POOL_NAME/IMAGE_NAME
```

Pour rétrograder toutes les images primaires, indiquez la sous-commande **mirror image demote** ainsi que le nom de la réserve :

```
cephuser@adm > rbd --cluster local mirror pool demote POOL_NAME
```

Pour promouvoir une image spécifique au rang d'image primaire, indiquez la sous-commande **mirror image promote** ainsi que le nom de la réserve et le nom de l'image :

```
cephuser@adm > rbd --cluster remote mirror image promote POOL_NAME/IMAGE_NAME
```

Pour promouvoir toutes les images non primaires d'une réserve au rang d'images primaires, indiquez la sous-commande **mirror image promote** ainsi que le nom de la réserve :

```
cephuser@adm > rbd --cluster local mirror pool promote POOL_NAME
```



Astuce : division de la charge d'E/S

Comme l'état primaire ou non primaire s'applique au niveau de l'image, il est possible que deux grappes divisent le chargement des E/S et le basculement ou la restauration par phases.

20.4.2.6 Resynchronisation forcée de l'image

Si le daemon `rbd-mirror` détecte un événement de divergence, il n'y aura pas de tentative de mettre en miroir l'image concernée jusqu'à ce que celle-ci soit corrigée. Pour reprendre la mise en miroir d'une image, commencez par rétrograder l'image jugée obsolète, puis demandez une resynchronisation avec l'image principale. Pour demander une resynchronisation de l'image, indiquez la sous-commande `mirror image resync` avec le nom de la réserve et le nom de l'image :

```
cephuser@adm > rbd mirror image resync POOL_NAME/IMAGE_NAME
```

20.4.3 Vérification de l'état du miroir

L'état de réplication de la grappe homologue est stocké pour chaque image en miroir principale. Cet état peut être récupéré à l'aide des sous-commandes `mirror image status` et `mirror pool status` :

Pour demander l'état de l'image miroir, indiquez la sous-commande `mirror image status` avec le nom de la réserve et le nom de l'image :

```
cephuser@adm > rbd mirror image status POOL_NAME/IMAGE_NAME
```

Pour demander l'état du résumé de la réserve miroir, indiquez la sous-commande `mirror pool status` avec le nom de la réserve :

```
cephuser@adm > rbd mirror pool status POOL_NAME
```



Astuce :

L'option `--verbose` de la sous-commande `mirror pool status` permet d'afficher des informations détaillées sur l'état de chaque image de mise en miroir présente dans la réserve.

20.5 Paramètres de cache

L'implémentation de l'espace utilisateur du périphérique de bloc Ceph (`librbd`) ne peut pas profiter du cache de page Linux. Il comprend donc son propre caching en mémoire. Le caching RBD se comporte comme le caching de disque dur. Lorsque le système d'exploitation envoie une demande de barrière ou de vidage, toutes les données altérées (« dirty ») sont écrites sur l'OSD. Cela signifie que l'utilisation du caching à écriture différée est tout aussi sûre que celle d'un disque dur physique correct avec une machine virtuelle qui envoie correctement des demandes de vidage. Le cache utilise un algorithme *Moins récemment utilisée* (LRU) et peut, en mode d'écriture différée, fusionner les demandes adjacentes pour un meilleur débit.

Ceph prend en charge le caching à écriture différée pour RBD. Pour l'activer, exécutez

```
cephuser@adm > ceph config set client rbd_cache true
```

Par défaut, `librbd` n'effectue aucun caching. Les écritures et les lectures sont envoyées directement à la grappe de stockage, et les écritures ne reviennent que lorsque les données sont sur disque sur toutes les répliques. Lorsque le caching est activé, les écritures reviennent immédiatement sauf si le volume d'octets non vidés est supérieur à celui défini par l'option `rbd_cache_max_dirty`. Dans un tel cas, l'écriture déclenche l'écriture différée et les blocs jusqu'à ce que suffisamment d'octets soient vidés.

Ceph prend en charge le caching à écriture immédiate pour RBD. Vous pouvez définir la taille du cache ainsi que des objectifs et des limites pour passer du caching à écriture différée au caching à écriture immédiate. Pour activer le mode d'écriture immédiate, exécutez

```
cephuser@adm > ceph config set client rbd_cache_max_dirty 0
```

Cela signifie que les écritures ne reviennent que lorsque les données sont sur disque sur toutes les répliques, mais que les lectures peuvent provenir du cache. Le cache est en mémoire sur le client, et chaque image RBD a son propre cache. Étant donné que le cache est en local sur le client, il n'y a pas de cohérence s'il y a d'autres accès à l'image. L'exécution de GFS ou d'OCFS sur RBD ne fonctionnera pas avec le caching activé.

Les paramètres suivants affectent le comportement des périphériques de bloc RADOS. Pour les définir, utilisez la catégorie `client` :

```
cephuser@adm > ceph config set client PARAMETER VALUE
```

rbd cache

Permet d'activer le caching pour le périphérique de bloc RADOS (RBD). La valeur par défaut est « true ».

rbd cache size

Taille du cache RBD en octets. La valeur par défaut est 32 Mo.

rbd cache max dirty

Limite « dirty » en octets à laquelle le cache déclenche l'écriture différée. rbd cache max dirty doit être inférieur à rbd cache size. Si la valeur est définie sur 0, le caching à écriture immédiate est utilisé. La valeur par défaut est 24 Mo.

rbd cache target dirty

Valeur « dirty target » avant que le cache commence à écrire des données sur le stockage de données. Ne bloque pas les écritures dans le cache. La valeur par défaut est 16 Mo.

rbd cache max dirty age

Temps en secondes pendant lequel les données altérées sont dans le cache avant le début de l'écriture différée. La valeur par défaut est 1.

rbd cache writethrough until flush

Indique de commencer en mode d'écriture immédiate et de passer à l'écriture différée après la réception de la première demande de vidage. Cette configuration classique est judicieuse lorsque les machines virtuelles qui s'exécutent sur rbd sont trop anciennes pour envoyer des vidages (par exemple, le pilote virtio dans Linux avant le kernel 2.6.32). La valeur par défaut est « true ».

20.6 Paramètres QoS

En règle générale, la qualité de service (QoS) fait référence aux méthodes de priorisation du trafic et de réservation des ressources. Elle est particulièrement importante pour le transport du trafic avec des exigences spéciales.



Important : non pris en charge par iSCSI

Les paramètres QoS suivants sont utilisés uniquement par l'implémentation RBD de l'espace utilisateur librbd et *non* par l'implémentation kRBD. Étant donné qu'iSCSI utilise kRBD, il n'emploie pas les paramètres QoS. Toutefois, pour iSCSI, vous pouvez configurer la qualité de service sur la couche des périphériques de bloc du kernel à l'aide des fonctionnalités standard du kernel.

rbd qos iops limit

Limite souhaitée des opérations d'E/S par seconde. La valeur par défaut est 0 (pas de limite).

rbd qos bps limit

Limite souhaitée d'octets en E/S par seconde. La valeur par défaut est 0 (pas de limite).

rbd qos read iops limit

Limite souhaitée des opérations de lecture par seconde. La valeur par défaut est 0 (pas de limite).

rbd qos write iops limit

Limite souhaitée des opérations d'écriture par seconde. La valeur par défaut est 0 (pas de limite).

rbd qos read bps limit

Limite souhaitée des octets en lecture par seconde. La valeur par défaut est 0 (pas de limite).

rbd qos write bps limit

Limite souhaitée des octets en écriture par seconde. La valeur par défaut est 0 (pas de limite).

rbd qos iops burst

Limite de rafales souhaitée des opérations d'E/S. La valeur par défaut est 0 (pas de limite).

rbd qos bps burst

Limite de rafales souhaitée des octets en E/S. La valeur par défaut est 0 (pas de limite).

rbd qos read iops burst

Limite de rafales souhaitée des opérations de lecture. La valeur par défaut est 0 (pas de limite).

rbd qos write iops burst

Limite de rafales souhaitée des opérations d'écriture. La valeur par défaut est 0 (pas de limite).

rbd qos read bps burst

Limite de rafales souhaitée des octets en lecture. La valeur par défaut est 0 (pas de limite).

rbd qos write bps burst

Limite de rafales souhaitée des octets en écriture. La valeur par défaut est 0 (pas de limite).

rbd qos schedule tick min

Cycle d'horloge de planification minimal (en millisecondes) pour la qualité de service. La valeur par défaut est 50.

20.7 Paramètres de la lecture anticipée

Le périphérique de bloc RADOS prend en charge la lecture anticipée/la prérécupération pour optimiser les petites lectures séquentielles. Ces opérations devraient normalement être gérées par le système d'exploitation invité dans le cas d'une machine virtuelle, mais les chargeurs de démarrage peuvent ne pas émettre des lectures efficaces. La lecture anticipée est automatiquement désactivée si le caching est désactivé.



Important : non pris en charge par iSCSI

Les paramètres de lecture anticipée suivants sont utilisés uniquement par l'implémentation RBD de l'espace utilisateur librbd et *non* par l'implémentation krbd. Étant donné qu'iSCSI utilise krbd, il n'emploie pas les paramètres de lecture anticipée. Toutefois, pour iSCSI, vous pouvez configurer la lecture anticipée sur la couche des périphériques de bloc du kernel à l'aide des fonctionnalités standard du kernel.

rbd readahead trigger requests

Nombre de demandes de lecture séquentielle nécessaires pour déclencher la lecture anticipée. La valeur par défaut est 10.

rbd readahead max bytes

Taille maximale d'une demande de lecture anticipée. Lorsque la valeur est 0, la lecture anticipée est désactivée. La valeur par défaut est 512 Ko.

rbd readahead disable after bytes

Après la lecture de tous ces octets à partir d'une image RBD, la lecture anticipée est désactivée pour cette image jusqu'à ce qu'elle soit fermée. Cela permet à l'OS invité de prendre en charge la lecture anticipée quand il est démarré. Lorsque la valeur est 0, la lecture anticipée reste activée. La valeur par défaut est 50 Mo.

20.8 Fonctions avancées

Le périphérique de bloc RADOS prend en charge les fonctions avancées qui améliorent la fonctionnalité des images RBD. Vous pouvez spécifier les fonctions sur la ligne de commande lors de la création d'une image RBD ou dans le fichier de configuration Ceph à l'aide de l'option `rbd_default_features`.

Vous pouvez spécifier les valeurs de l'option `rbd_default_features` de deux façons :

- Comme une somme de valeurs internes des fonctions. Chaque fonction a sa propre valeur interne, par exemple 1 pour « layering » et 16 pour « fast-diff ». Par conséquent, pour activer ces deux fonctions par défaut, incluez la ligne suivante :

```
rbd_default_features = 17
```

- Comme une liste de fonctions séparées par des virgules. L'exemple précédent se présentera comme suit :

```
rbd_default_features = layering,fast-diff
```



Note : fonctions non prises en charge par iSCSI

Les images RBD avec les fonctions suivantes ne seront pas prises en charge par iSCSI : `deep-flatten`, `object-map`, `journaling`, `fast-diff` et `striping`.

Voici une liste de fonctions RBD avancées :

layering

La création de couches, ou superposition (layering), permet d'utiliser le clonage.

La valeur interne est 1, la valeur par défaut est « yes ».

striping

La segmentation (striping) propage les données sur plusieurs objets et contribue au parallélisme pour les workloads séquentiels de lecture/écriture. Elle empêche les goulots d'étranglement de noeud unique pour les périphériques de bloc RADOS volumineux ou fort occupés.

La valeur interne est 2, la valeur par défaut est « yes ».

exclusive-lock

Lorsque cette fonction est activée, il faut qu'un client obtienne un verrouillage sur un objet avant d'effectuer une écriture. Activez le verrouillage exclusif uniquement lorsqu'un seul client accède à une image en même temps. La valeur interne est 4. La valeur par défaut est « yes ».

object-map

La prise en charge de l'assignation d'objet dépend de la prise en charge du verrouillage exclusif. Les périphériques de bloc sont provisionnés dynamiquement, ce qui signifie qu'ils ne stockent que les données qui existent réellement. La prise en charge de l'assignation d'objet permet de suivre quels objets existent réellement (ont des données stockées sur un disque). L'activation de la prise en charge de l'assignation d'objet permet d'accélérer les opérations d'E/S pour le clonage, l'importation et l'exportation d'une image peu peuplée, et pour la suppression.

La valeur interne est 8, la valeur par défaut est « yes ».

fast-diff

La prise en charge de la fonction fast-diff dépend de la prise en charge de l'assignation d'objet et du verrouillage exclusif. Elle ajoute une propriété à l'assignation d'objet, ce qui la rend beaucoup plus rapide pour générer des différentiels entre les instantanés d'une image et l'utilisation réelle des données d'un instantané.

La valeur interne est 16, la valeur par défaut est « yes ».

deep-flatten

La fonction deep-flatten rend **rbd flatten** (voir la [Section 20.3.3.6, « Mise à plat d'une image clonée »](#)) opérationnel sur tous les instantanés d'une image, en plus de l'image elle-même. Sans elle, les instantanés d'une image s'appuieront toujours sur le parent, et vous ne pourrez pas supprimer l'image parent avant que les instantanés soient supprimés. La fonction deep-flatten rend un parent indépendant de ses clones, même s'ils ont des instantanés.

La valeur interne est 32, la valeur par défaut est « yes ».

journaling

La prise en charge de la fonction de journalisation (journaling) dépend de la prise en charge du verrouillage exclusif. La journalisation enregistre toutes les modifications d'une image dans l'ordre où elles se produisent. La mise en miroir RBD (voir la [Section 20.4, « Miroirs d'image RBD »](#)) utilise le journal pour répliquer une image cohérente sur une grappe distante en cas de panne.

La valeur interne est 64, la valeur par défaut est « no ».

20.9 Assignation RBD à l'aide d'anciens clients de kernel

Les anciens clients (par exemple, SLE 11 SP4) peuvent ne pas être en mesure d'assigner les images RBD parce qu'une grappe déployée avec SUSE Enterprise Storage 7.1 force certaines fonctions (à la fois les fonctions de niveau image RBD et celles de niveau RADOS) que ces anciens clients ne prennent pas en charge. Dans ce cas, les journaux OSD afficheront des messages semblables à ce qui suit :

```
2019-05-17 16:11:33.739133 7fcb83a2e700 0 -- 192.168.122.221:0/1006830 >> \
192.168.122.152:6789/0 pipe(0x65d4e0 sd=3 :57323 s=1 pgs=0 cs=0 l=1 c=0x65d770).connect \
protocol feature mismatch, my 2fffffffffff < peer 4010ff8ffacffff missing 401000000000000
```



Avertissement : la modification des types de compartiment de carte CRUSH provoque un rééquilibrage massif

Si vous avez l'intention de commuter les types de compartiment de carte CRUSH « straw » et « straw2 », procédez de manière méthodique. Attendez-vous à un impact significatif sur la charge de la grappe, car un tel changement provoque un rééquilibrage massif des grappes.

1. Désactivez toutes les fonctions d'image RBD qui ne sont pas prises en charge. Par exemple :

```
cephuser@adm > rbd feature disable pool1/image1 object-map
cephuser@adm > rbd feature disable pool1/image1 exclusive-lock
```

2. Remplacez les types de compartiment de carte CRUSH « straw2 » par « straw » :

- a. Enregistrez la carte CRUSH :

```
cephuser@adm > ceph osd getcrushmap -o crushmap.original
```

- b. Décompilez la carte CRUSH :

```
cephuser@adm > crushtool -d crushmap.original -o crushmap.txt
```

- c. Modifiez la carte CRUSH et remplacez « straw2 » par « straw ».

d. Recompilez la carte CRUSH :

```
cephuser@adm > crushtool -c crushmap.txt -o crushmap.new
```

e. Définissez la nouvelle carte CRUSH :

```
cephuser@adm > ceph osd setcrushmap -i crushmap.new
```

20.10 Activation des périphériques de bloc et de Kubernetes

Vous pouvez utiliser Ceph RBD avec Kubernetes v1.13 et versions ultérieures via le pilote `ceph-csi`. Ce pilote provisionne dynamiquement des images RBD pour soutenir les volumes Kubernetes et assigne ces images RBD en tant que périphériques de bloc (éventuellement en montant un système de fichiers contenu dans l'image) sur des noeuds de travail exécutant des pods faisant référence à un volume soutenu par RBD.

Pour utiliser des périphériques de bloc Ceph avec Kubernetes, vous devez installer et configurer `ceph-csi` dans votre environnement Kubernetes.



Important

`ceph-csi` utilise les modules de kernel RBD par défaut qui peuvent ne pas prendre en charge tous les paramètres Ceph CRUSH ou les fonctions d'image RBD.

1. Par défaut, les périphériques de bloc Ceph utilisent la réserve RBD. Créez une réserve pour stocker les volumes Kubernetes. Assurez-vous que votre grappe Ceph est en cours d'exécution, puis créez la réserve :

```
cephuser@adm > ceph osd pool create kubernetes
```

2. Utilisez l'outil RBD pour initialiser la réserve :

```
cephuser@adm > rbd pool init kubernetes
```

3. Créez un nouvel utilisateur pour Kubernetes et ceph-csi. Exécutez la commande suivante et enregistrez la clé générée :

```
cephuser@adm > ceph auth get-or-create client.kubernetes mon 'profile rbd' osd
'profile rbd pool=kubernetes' mgr 'profile rbd pool=kubernetes'
[client.kubernetes]
key = AQD9o0Fd6hQRChAA7fMaSZXduT3NWEqylNpmg==
```

4. ceph-csi nécessite un objet ConfigMap stocké dans Kubernetes pour définir les adresses de moniteur Ceph pour la grappe Ceph. Collectez le fsid unique de la grappe Ceph et les adresses de moniteur :

```
cephuser@adm > ceph mon dump
<...>
fsid b9127830-b0cc-4e34-aa47-9d1a2e9949a8
<...>
0: [v2:192.168.1.1:3300/0,v1:192.168.1.1:6789/0] mon.a
1: [v2:192.168.1.2:3300/0,v1:192.168.1.2:6789/0] mon.b
2: [v2:192.168.1.3:3300/0,v1:192.168.1.3:6789/0] mon.c
```

5. Générez un fichier csi-config-map.yaml similaire à l'exemple ci-dessous, en remplaçant le FSID par clusterID et les adresses de moniteur pour monitors :

```
kubect1@adm > cat <<EOF > csi-config-map.yaml
---
apiVersion: v1
kind: ConfigMap
data:
  config.json: |-
    [
      {
        "clusterID": "b9127830-b0cc-4e34-aa47-9d1a2e9949a8",
        "monitors": [
          "192.168.1.1:6789",
          "192.168.1.2:6789",
          "192.168.1.3:6789"
        ]
      }
    ]
metadata:
  name: ceph-csi-config
EOF
```

6. Une fois généré, stockez le nouvel objet ConfigMap dans Kubernetes :

```
kubect@adm > kubectl apply -f csi-config-map.yaml
```

7. ceph-csi a besoin des informations d'identification cephx pour communiquer avec la grappe Ceph. Générez un fichier csi-rbd-secret.yaml similaire à l'exemple ci-dessous, en utilisant l'ID utilisateur Kubernetes et la clé cephx que vous venez de créer :

```
kubect@adm > cat <<EOF > csi-rbd-secret.yaml
---
apiVersion: v1
kind: Secret
metadata:
  name: csi-rbd-secret
  namespace: default
stringData:
  userID: kubernetes
  userKey: AQD9o0Fd6hQRChAAt7fMaSZXduT3NWEqylNpmg==
EOF
```

8. Une fois généré, stockez le nouvel objet Secret dans Kubernetes :

```
kubect@adm > kubectl apply -f csi-rbd-secret.yaml
```

9. Créez les objets ServiceAccount et RBAC ClusterRole/ClusterRoleBinding Kubernetes requis. Ces objets ne doivent pas nécessairement être personnalisés pour votre environnement Kubernetes et peuvent donc être utilisés directement à partir des fichiers YAML de déploiement ceph-csi :

```
kubect@adm > kubectl apply -f https://raw.githubusercontent.com/ceph/ceph-csi/master/deploy/rbd/kubernetes/csi-provisioner-rbac.yaml
kubect@adm > kubectl apply -f https://raw.githubusercontent.com/ceph/ceph-csi/master/deploy/rbd/kubernetes/csi-nodeplugin-rbac.yaml
```

10. Créez l'outil de déploiement ceph-csi et les plug-ins de noeud :

```
kubect@adm > wget https://raw.githubusercontent.com/ceph/ceph-csi/master/deploy/rbd/kubernetes/csi-rbdplugin-provisioner.yaml
kubect@adm > kubectl apply -f csi-rbdplugin-provisioner.yaml
kubect@adm > wget https://raw.githubusercontent.com/ceph/ceph-csi/master/deploy/rbd/kubernetes/csi-rbdplugin.yaml
kubect@adm > kubectl apply -f csi-rbdplugin.yaml
```




Important

Par défaut, les fichiers YAML de l'outil de déploiement et du plug-in de noeud récupèrent la version de développement du conteneur `ceph-csi`. Les fichiers YAML doivent être mis à jour pour utiliser une version commerciale.

20.10.1 Utilisation de périphériques de bloc Ceph dans Kubernetes

Kubernetes StorageClass définit une classe de stockage. Plusieurs objets StorageClass peuvent être créés pour être assignés à différents niveaux et fonctionnalités de qualité de service. Par exemple, NVMe par rapport aux réserves sur disque dur.

Pour créer une classe de stockage `ceph-csi` assignée à la réserve Kubernetes créée ci-dessus, le fichier YAML suivant peut être utilisé, après avoir vérifié que la propriété `clusterID` correspond au FSID de votre grappe Ceph :

```
kubect@adm > cat <<EOF > csi-rbd-sc.yaml
---
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: csi-rbd-sc
provisioner: rbd.csi.ceph.com
parameters:
  clusterID: b9127830-b0cc-4e34-aa47-9d1a2e9949a8
  pool: kubernetes
  csi.storage.k8s.io/provisioner-secret-name: csi-rbd-secret
  csi.storage.k8s.io/provisioner-secret-namespace: default
  csi.storage.k8s.io/node-stage-secret-name: csi-rbd-secret
  csi.storage.k8s.io/node-stage-secret-namespace: default
reclaimPolicy: Delete
mountOptions:
  - discard
EOF
kubect@adm > kubectl apply -f csi-rbd-sc.yaml
```

`PersistentVolumeClaim` est une requête de ressources de stockage abstrait émise par un utilisateur. Le paramètre `PersistentVolumeClaim` serait alors associé à une ressource de pod pour provisionner un volume `PersistentVolume`, qui serait soutenu par une image de bloc Ceph. Un mode de volume `volumeMode` facultatif peut être inclus pour choisir entre un système de fichiers monté (par défaut) ou un volume basé sur un périphérique de bloc brut.

À l'aide de `ceph-csi`, la spécification de `Filesystem` pour `volumeMode` peut prendre en charge les réclamations `ReadWriteOnce` et `ReadOnlyMode` `accessMode` et la spécification de `Block` pour `volumeMode` peut prendre en charge les réclamations `ReadWriteOnce`, `ReadWriteMany` et `ReadOnlyMany` `accessMode`.

Par exemple, pour créer une réclamation `PersistentVolumeClaim` basée sur des blocs qui utilise la classe `ceph-csi-based` `StorageClass` créée ci-dessus, le fichier YAML suivant peut être utilisé pour demander un stockage de bloc brut à partir de la classe de stockage `csi-rbd-sc` `StorageClass` :

```
kubectl@adm > cat <<EOF > raw-block-pvc.yaml
---
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: raw-block-pvc
spec:
  accessModes:
    - ReadWriteOnce
  volumeMode: Block
  resources:
    requests:
      storage: 1Gi
  storageClassName: csi-rbd-sc
EOF
kubectl@adm > kubectl apply -f raw-block-pvc.yaml
```

L'exemple suivant illustre la liaison d'une réclamation `PersistentVolumeClaim` à une ressource de pod en tant que périphérique de bloc brut :

```
kubectl@adm > cat <<EOF > raw-block-pod.yaml
---
apiVersion: v1
kind: Pod
metadata:
  name: pod-with-raw-block-volume
spec:
  containers:
    - name: fc-container
      image: fedora:26
      command: ["/bin/sh", "-c"]
      args: ["tail -f /dev/null"]
      volumeDevices:
        - name: data
          devicePath: /dev/xvda
```

```

volumes:
  - name: data
    persistentVolumeClaim:
      claimName: raw-block-pvc
EOF
kubectl@adm > kubectl apply -f raw-block-pod.yaml

```

Pour créer une réclamation PersistentVolumeClaim basée sur le système de fichiers qui utilise la classe ceph-csi-based StorageClass créée ci-dessus, le fichier YAML suivant peut être utilisé pour demander un système de fichiers monté (soutenu par une image RBD) à partir de la classe de stockage csi-rbd-sc StorageClass :

```

kubectl@adm > cat <<EOF > pvc.yaml
---
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: rbd-pvc
spec:
  accessModes:
    - ReadWriteOnce
  volumeMode: Filesystem
  resources:
    requests:
      storage: 1Gi
  storageClassName: csi-rbd-sc
EOF
kubectl@adm > kubectl apply -f pvc.yaml

```

L'exemple suivant illustre la liaison d'une réclamation PersistentVolumeClaim à une ressource de pod en tant que système de fichiers monté :

```

kubectl@adm > cat <<EOF > pod.yaml
---
apiVersion: v1
kind: Pod
metadata:
  name: csi-rbd-demo-pod
spec:
  containers:
    - name: web-server
      image: nginx
      volumeMounts:
        - name: mypvc
          mountPath: /var/lib/www/html
  volumes:

```

```
- name: mypvc
  persistentVolumeClaim:
    claimName: rbd-pvc
    readOnly: false
EOF
kubectl@adm > kubectl apply -f pod.yaml
```

IV Accès aux données de la grappe

- 21 Ceph Object Gateway **267**
- 22 Passerelle Ceph iSCSI **323**
- 23 Système de fichiers en grappe **341**
- 24 Exportation des données Ceph via Samba **352**
- 25 NFS Ganesha **371**

21 Ceph Object Gateway

Ce chapitre présente des détails sur les tâches d'administration liées à Object Gateway, telles que la vérification de l'état du service, la gestion des comptes, les passerelles multisites ou l'authentification LDAP.

21.1 Restrictions d'Object Gateway et règles de dénomination

Voici la liste des limites importantes d'Object Gateway :

21.1.1 Limitations des compartiments

Lors de l'approche d'Object Gateway via l'API S3, les noms de compartiment doivent être conformes au DNS et peuvent contenir le caractère tiret (« - »). Lors de l'approche d'Object Gateway via l'API Swift, vous pouvez utiliser n'importe quelle combinaison de caractères UTF-8 à l'exception du caractère « / ». Le nom d'un compartiment peut comporter jusqu'à 255 caractères. Chaque nom de compartiment doit être unique.



Astuce : utilisation de noms de compartiment conformes au DNS

Bien que vous puissiez utiliser n'importe quel nom de compartiment basé sur UTF-8 dans l'API Swift, il est recommandé de nommer les compartiments conformément aux règles de dénomination S3 afin d'éviter les problèmes d'accès au même compartiment via l'API S3.

21.1.2 Limitations des objets stockés

Nombre maximal d'objets par utilisateur

Aucune restriction par défaut (limite de $\sim 2^{63}$).

Nombre maximal d'objets par compartiment

Aucune restriction par défaut (limite de $\sim 2^{63}$).

Taille maximale d'un objet à charger/stocker

La limite est de 5 Go par chargement. Utilisez le chargement en plusieurs parties pour les objets plus volumineux. Le nombre maximal s'élève à 10 000 pour les tranches en plusieurs parties.

21.1.3 Limitations d'en-tête HTTP

La limitation de requête et d'en-tête HTTP dépend de l'interface Web utilisée. L'entité par défaut limite la taille de l'en-tête HTTP à 16 ko.

21.2 Déploiement de la passerelle Object Gateway

Le déploiement de la passerelle Ceph Object Gateway suit la même procédure que le déploiement des autres services Ceph, à l'aide de `cephdm`. Pour plus de détails, reportez-vous au *Manuel « Guide de déploiement », Chapitre 8 « Déploiement des services essentiels restants à l'aide de cephadm », Section 8.2 « Spécification de service et de placement », en particulier au Manuel « Guide de déploiement », Chapitre 8 « Déploiement des services essentiels restants à l'aide de cephadm », Section 8.3.4 « Déploiement d'instances Object Gateway ».*

21.3 Exploitation du service Object Gateway

Vous pouvez utiliser les passerelles Object Gateway de la même manière que les autres services Ceph en identifiant d'abord le nom du service avec la commande `ceph orch ps`, puis en exécutant la commande suivante pour l'exploitation des services, par exemple :

```
ceph orch daemon restart OGW_SERVICE_NAME
```

Reportez-vous au *Chapitre 14, Exécution des services Ceph* pour obtenir des informations complètes sur le fonctionnement des services Ceph.

21.4 Options de configuration

Reportez-vous à la *Section 28.5, « Ceph Object Gateway »* pour obtenir une liste des options de configuration d'Object Gateway.

21.5 Gestion des accès à la passerelle Object Gateway

Vous pouvez communiquer avec Object Gateway en utilisant une interface compatible avec S3 ou Swift. L'interface S3 est compatible avec un vaste sous-ensemble de l'API RESTful d'Amazon S3. L'interface Swift est compatible avec un vaste sous-ensemble de l'API OpenStack Swift.

Les deux interfaces nécessitent la création d'un utilisateur spécifique et l'installation du logiciel client approprié pour communiquer avec la passerelle à l'aide de la clé secrète de l'utilisateur.

21.5.1 Accès à Object Gateway

21.5.1.1 Accès à l'interface S3

Pour accéder à l'interface S3, un client REST est nécessaire. **S3cmd** est un client S3 de ligne de commande. Il est disponible dans [OpenSUSE Build Service \(https://build.opensuse.org/package/show/Cloud:Tools/s3cmd\)](https://build.opensuse.org/package/show/Cloud:Tools/s3cmd). Le dépôt contient des versions de distributions SUSE Linux Enterprise et de distributions openSUSE.

Si vous voulez tester votre accès à l'interface S3, vous pouvez aussi écrire un petit script Python. Le script se connecte à Object Gateway, crée un compartiment et dresse la liste de tous les compartiments. Les valeurs de `aws_access_key_id` (ID_clé_accès_AWS) et de `aws_secret_access_key` (Clé_accès_secret_AWS) sont issues des valeurs de `access_key` (clé_accès) et de `secret_key` (clé_secret) renvoyées par la commande `radosgw_admin` de la [Section 21.5.2.1, « Ajout d'utilisateurs S3 et Swift »](#).

1. Installez le paquetage `python-boto` :

```
# zypper in python-boto
```

2. Créez un script Python appelé `s3test.py` avec le contenu suivant :

```
import boto
import boto.s3.connection
access_key = '11BS02LGFB6AL6H1ADMW'
secret_key = 'vzCEkuryfn060dfec4fgQPqFrncKEIkh3Zcd0ANY'
conn = boto.connect_s3(
    aws_access_key_id = access_key,
    aws_secret_access_key = secret_key,
    host = 'HOSTNAME',
```



```
is_secure=False,
calling_format = boto.s3.connection.OrdinaryCallingFormat(),
)
bucket = conn.create_bucket('my-new-bucket')
for bucket in conn.get_all_buckets():
    print "NAME\tCREATED".format(
        name = bucket.name,
        created = bucket.creation_date,
    )
```

Remplacez *HOSTNAME* par le nom de l'hôte sur lequel vous avez configuré le service Object Gateway, par exemple `gateway_host`.

3. Exécutez le script :

```
python s3test.py
```

Le script produit un résultat similaire à ceci :

```
my-new-bucket 2015-07-22T15:37:42.000Z
```

21.5.1.2 Accès à l'interface Swift

Pour accéder à Object Gateway via l'interface Swift, vous devez disposer du client de ligne de commande **swift**. La page de manuel **man 1 swift** fournit des informations complémentaires sur les options de ligne de commande du client.

Le paquetage est inclus dans le module « Public Cloud » (Cloud public) pour SUSE Linux Enterprise 12 à partir de SP3 et SUSE Linux Enterprise 15. Avant d'installer le paquetage, vous devez activer le module et rafraîchir le dépôt des logiciels :

```
# SUSEConnect -p sle-module-public-cloud/12/SYSTEM-ARCH
sudo zypper refresh
```

Ou

```
# SUSEConnect -p sle-module-public-cloud/15/SYSTEM-ARCH
# zypper refresh
```

Pour installer la commande **swift**, exécutez ce qui suit :

```
# zypper in python-swiftclient
```

L'accès swift utilise la syntaxe suivante :

```
> swift -A http://IP_ADDRESS/auth/1.0 \
```

```
-U example_user:swift -K 'SWIFT_SECRET_KEY' list
```

Remplacez `IP_ADDRESS` par l'adresse IP du serveur de passerelle, et `_SECRET_KEY` par la valeur de la clé secrète SWIFT dans la sortie de la commande **radosgw-admin key create** exécutée pour l'utilisateur `swiftswift` à la [Section 21.5.2.1, « Ajout d'utilisateurs S3 et Swift »](#).

Par exemple :

```
> swift -A http://gateway.example.com/auth/1.0 -U example_user:swift \
-K 'r5wWIXj0CeE07DixD1FjTLmNYIViaC6JVhi3013h' list
```

La sortie est la suivante :

```
my-new-bucket
```

21.5.2 Gestion des comptes S3 et Swift

21.5.2.1 Ajout d'utilisateurs S3 et Swift

Vous devez créer un utilisateur, une clé d'accès et un secret pour permettre aux utilisateurs finaux d'interagir avec la passerelle. Il existe deux types d'utilisateur : *user* et *subuser*. Alors que des *users* sont employés pour les interactions avec l'interface S3, les *subusers* sont les utilisateurs de l'interface Swift. Chaque subuser est associé à un user.

Pour créer un utilisateur Swift, procédez de la façon suivante :

1. Pour créer un utilisateur Swift (qui est un *subuser* suivant notre terminologie), vous devez d'abord créer le *user* qui lui est associé.

```
cephuser@adm > radosgw-admin user create --uid=USERNAME \
--display-name="DISPLAY-NAME" --email=EMAIL
```

Par exemple :

```
cephuser@adm > radosgw-admin user create \
--uid=example_user \
--display-name="Example User" \
--email=penguin@example.com
```

2. Pour créer un subuser (interface Swift) pour l'utilisateur, vous devez indiquer l'ID utilisateur (`--uid = USERNAME`), un ID subuser et le niveau d'accès du subuser.

```
cephuser@adm > radosgw-admin subuser create --uid=UID \  
--subuser=UID \  
--access=[ read | write | readwrite | full ]
```

Par exemple :

```
cephuser@adm > radosgw-admin subuser create --uid=example_user \  
--subuser=example_user:swift --access=full
```

3. Générez une clé secrète pour l'utilisateur.

```
cephuser@adm > radosgw-admin key create \  
--gen-secret \  
--subuser=example_user:swift \  
--key-type=swift
```

4. Les deux commandes afficheront des données au format JSON indiquant l'état de l'utilisateur. Notez les lignes suivantes et souvenez-vous de la valeur de secret_key (clé_secrète) :

```
"swift_keys": [  
  { "user": "example_user:swift",  
    "secret_key": "r5wWIXj0CeE07DixD1FjTlMNYIViaC6JVhi3013h"}],
```

Lorsque vous accédez à Object Gateway via l'interface S3, vous devez créer un utilisateur S3 en exécutant :

```
cephuser@adm > radosgw-admin user create --uid=USERNAME \  
--display-name="DISPLAY-NAME" --email=EMAIL
```

Par exemple :

```
cephuser@adm > radosgw-admin user create \  
--uid=example_user \  
--display-name="Example User" \  
--email=penguin@example.com
```

La commande crée également l'accès et la clé secrète de l'utilisateur. Vérifiez le résultat des mots clés access_key (clé_accès) et secret_key (clé_secret), et leurs valeurs :

```
[...]  
"keys": [  
  { "user": "example_user",  
    "access_key": "11BS02LGFB6AL6H1ADMW",  
    "secret_key": "vzCEkuryfn060dfec4fgQPqFrncKEIkh3Zcd0ANY"}],  
[...]
```

21.5.2.2 Suppression d'utilisateurs S3 et Swift

La procédure de suppression des utilisateurs est similaire pour les utilisateurs S3 et Swift. Dans le cas d'utilisateurs Swift cependant, vous devrez peut-être supprimer l'utilisateur ainsi que ses subusers.

Pour supprimer un utilisateur S3 ou Swift (y compris tous ses subusers), spécifiez `user rm` et l'ID utilisateur dans la commande suivante :

```
cephuser@adm > radosgw-admin user rm --uid=example_user
```

Pour supprimer un subuser, spécifiez `subuser rm` et l'ID de celui-ci.

```
cephuser@adm > radosgw-admin subuser rm --uid=example_user:swift
```

Vous pouvez utiliser les options suivantes :

`--purge-data`

Purge toutes les données associées à l'ID utilisateur.

`--purge-keys`

Purge toutes les clés associées à l'ID utilisateur.



Astuce : suppression d'un subuser

Lorsque vous supprimez un subuser, vous supprimez également l'accès à l'interface Swift. L'utilisateur est conservé dans le système.

21.5.2.3 Modification de l'accès utilisateur S3 et Swift et des clés secrètes

Les paramètres `access_key` (clé_accès) et `secret_key` (clé_secret) identifient l'utilisateur Object Gateway lors de l'accès à la passerelle. La modification des clés utilisateur existantes revient à créer de nouvelles clés, car les anciennes clés sont écrasées.

Pour les utilisateurs S3, exécutez la commande suivante :

```
cephuser@adm > radosgw-admin key create --uid=EXAMPLE_USER --key-type=s3 --gen-access-key --gen-secret
```

Pour les utilisateurs Swift, exécutez la commande suivante :

```
cephuser@adm > radosgw-admin key create --subuser=EXAMPLE_USER:swift --key-type=swift --gen-secret
```

--key-type=TYPE

Indique le type de clé. Soit swift, soit s3.

--gen-access-key

Génère une clé d'accès aléatoire (pour l'utilisateur S3 par défaut).

--gen-secret

Génère une clé secrète aléatoire.

--secret=KEY

Indique une clé secrète, par exemple générée manuellement.

21.5.2.4 Activation de la gestion des quotas utilisateur

La passerelle Ceph Object Gateway vous permet de définir des quotas sur les utilisateurs et les compartiments appartenant aux utilisateurs. Les quotas incluent le nombre maximal d'objets dans un compartiment et la taille de stockage maximale en mégaoctets.

Avant d'activer un quota utilisateur, vous devez tout d'abord définir ses paramètres :

```
cephuser@adm > radosgw-admin quota set --quota-scope=user --uid=EXAMPLE_USER \
--max-objects=1024 --max-size=1024
```

--max-objects

Indique le nombre maximal d'objets. Une valeur négative désactive la vérification.

--max-size

Indique le nombre maximal d'octets. Une valeur négative désactive la vérification.

--quota-scope

Définit l'étendue du quota. Les options sont bucket (compartiment) et user (utilisateur).
Les quotas de compartiment s'appliquent aux compartiments appartenant à un utilisateur.
Les quotas utilisateur s'appliquent à un utilisateur.

Une fois que vous avez défini un quota utilisateur, vous pouvez l'activer :

```
cephuser@adm > radosgw-admin quota enable --quota-scope=user --uid=EXAMPLE_USER
```

Pour désactiver un quota :

```
cephuser@adm > radosgw-admin quota disable --quota-scope=user --uid=EXAMPLE_USER
```

Pour dresser la liste des paramètres de quota :

```
cephuser@adm > radosgw-admin user info --uid=EXAMPLE_USER
```

Pour mettre à jour les statistiques de quota :

```
cephuser@adm > radosgw-admin user stats --uid=EXAMPLE_USER --sync-stats
```

21.6 Interfaces clients HTTP

La passerelle Ceph Object Gateway prend en charge deux interfaces clients HTTP : *Beast* et *Civetweb*.

L'interface client Beast utilise la bibliothèque Boost.Beast pour l'analyse HTTP et la bibliothèque Boost.Asio pour les entrées/sorties (I/O) réseau asynchrones.

L'interface client Civetweb utilise la bibliothèque HTTP Civetweb, qui est un dérivé de Mongoose.

Vous pouvez les configurer avec l'option `rgw_frontends`. Reportez-vous à la [Section 28.5, « Ceph Object Gateway »](#) pour obtenir une liste des options de configuration.

21.7 Activation de HTTPS/SSL pour les passerelles Object Gateway

Pour activer la passerelle Object Gateway qui permet de communiquer en toute sécurité par le biais du protocole SSL, vous devez disposer d'un certificat émis par une autorité de certification ou créer un certificat auto-signé.

21.7.1 Création d'un certificat auto-signé



Astuce

Ignorez cette section si vous avez déjà un certificat valide signé par une autorité de certification.

La procédure suivante décrit comment générer un certificat SSL auto-signé sur Salt Master.

1. Si Object Gateway doit être connu par d'autres identités de sujet, ajoutez-les à l'option `subjectAltName` dans la section `[v3_req]` du fichier `/etc/ssl/openssl.cnf` :

```
[...]
[ v3_req ]
subjectAltName = DNS:server1.example.com DNS:server2.example.com
[...]
```



Astuce : adresses IP dans `subjectAltName`

Pour utiliser des adresses IP à la place de noms de domaine dans l'option `subjectAltName`, remplacez la ligne d'exemple par la suivante :

```
subjectAltName = IP:10.0.0.10 IP:10.0.0.11
```

2. Créez la clé et le certificat à l'aide d'**openssl**. Entrez toutes les données que vous devez inclure dans votre certificat. Il est recommandé d'entrer le nom de domaine complet (FQDN) comme nom commun. Avant de signer le certificat, vérifiez que « X509v3 Subject Alternative Name: » est inclus dans les extensions requises et que « X509v3 Subject Alternative Name: » est défini dans le certificat généré.

```
root@master # openssl req -x509 -nodes -days 1095 \
-newkey rsa:4096 -keyout rgw.key
-out rgw.pem
```

3. Ajoutez la clé à la fin du fichier de certificat :

```
root@master # cat rgw.key >> rgw.pem
```

21.7.2 Configuration d'Object Gateway avec SSL

Pour configurer Object Gateway afin d'utiliser des certificats SSL, utilisez l'option `rgw_frontends`. Par exemple :

```
cephuser@adm > ceph config set WHO rgw_frontends \
beast ssl_port=443 ssl_certificate=config://CERT ssl_key=config://KEY
```

Si vous ne spécifiez pas les clés de configuration `CERT` et `KEY`, le service Object Gateway recherche le certificat et la clé SSL sous les clés de configuration suivantes :

```
rgw/cert/RGW_REALM/RGW_ZONE.key
```

```
rgw/cert/RGW_REALM/RGW_ZONE.crt
```

Si vous souhaitez remplacer l'emplacement par défaut de la clé et du certificat SSL, importez-les dans la base de données de configuration à l'aide de la commande suivante :

```
ceph config-key set CUSTOM_CONFIG_KEY -i PATH_TO_CERT_FILE
```

Utilisez ensuite vos clés de configuration personnalisées à l'aide de la directive `config://`.

21.8 Modules de synchronisation

Object Gateway est déployé en tant que service multisite tandis que vous pouvez mettre en miroir des données et des métadonnées entre les zones. Les *modules de synchronisation* sont construits au sommet de la structure multisite qui permet de transmettre des données et des métadonnées à un niveau externe différent. Un module de synchronisation permet d'effectuer un ensemble d'opérations chaque fois qu'un changement se produit dans les données (par exemple, opérations de métadonnées telles que la création d'un compartiment ou d'un utilisateur). Comme les modifications multisite Object Gateway sont finalement cohérentes sur les sites distants, elles sont propagées de manière asynchrone. Cela couvre des cas d'utilisation tels que la sauvegarde du stockage d'objets sur une grappe cloud externe, une solution de sauvegarde personnalisée à l'aide de lecteurs de bande ou encore l'indexation de métadonnées dans Elasticsearch.

21.8.1 Configuration des modules de synchronisation

Tous les modules de synchronisation sont configurés d'une manière similaire. Vous devez créer une zone (voir [Section 21.13, « Passerelles Object Gateway multisites »](#) pour plus de détails) et définir son option `--tier_type`, par exemple `--tier-type=cloud`, pour le module de synchronisation cloud :

```
cephuser@adm > radosgw-admin zone create --rgw-zonegroup=ZONE-GROUP-NAME \  
--rgw-zone=ZONE-NAME \  
--endpoints=http://endpoint1.example.com,http://endpoint2.example.com, [...] \  
--tier-type=cloud
```

Vous pouvez configurer le niveau spécifique à l'aide de la commande suivante :

```
cephuser@adm > radosgw-admin zone modify --rgw-zonegroup=ZONE-GROUP-NAME \  
--rgw-zone=ZONE-NAME \  
--tier-config=KEY1=VALUE1,KEY2=VALUE2
```


KEY dans la configuration spécifie la variable de configuration que vous souhaitez mettre à jour et VALUE indique sa nouvelle valeur. Les valeurs imbriquées peuvent être accédées à l'aide d'un point. Par exemple :

```
cephuser@adm > radosgw-admin zone modify --rgw-zonegroup=ZONE-GROUP-NAME \  
--rgw-zone=ZONE-NAME \  
--tier-config=connection.access_key=KEY,connection.secret=SECRET
```

Vous pouvez accéder aux entrées de tableau en ajoutant des crochets « [] » avec l'entrée référencée et vous pouvez ajouter une nouvelle entrée de tableau à l'aide de crochets « [] ». La valeur d'index -1 fait référence à la dernière entrée du tableau. Il n'est pas possible de créer une entrée et de la référencer à nouveau dans la même commande. Par exemple, une commande pour créer un profil pour les compartiments commençant par PREFIX se présenterait comme suit :

```
cephuser@adm > radosgw-admin zone modify --rgw-zonegroup=ZONE-GROUP-NAME \  
--rgw-zone=ZONE-NAME \  
--tier-config=profiles[].source_bucket=PREFIX'*'  
cephuser@adm > radosgw-admin zone modify --rgw-zonegroup=ZONE-GROUP-NAME \  
--rgw-zone=ZONE-NAME \  
--tier-config=profiles[-1].connection_id=CONNECTION_ID,profiles[-1].acls_id=ACLS_ID
```



Astuce : ajout et suppression d'entrées de configuration

Vous pouvez ajouter une nouvelle entrée de configuration de niveau à l'aide du paramètre --tier-config-add=KEY=VALUE.

Vous pouvez supprimer une entrée existante à l'aide de --tier-config-rm=KEY.

21.8.2 Synchronisation des zones

Une configuration de module de synchronisation est locale pour une zone en particulier. Le module de synchronisation détermine si la zone exporte des données ou ne peut consommer que des données modifiées dans une autre zone. Depuis la version « Luminous », les plug-ins de synchronisation pris en charge sont ElasticSearch, rgw (qui est le plug-in de synchronisation par défaut des données entre les zones) et log (qui est un plug-in générique de synchronisation consignnant les opérations de métadonnées entre zones distantes). Les sections suivantes s'appuient sur l'exemple d'une zone qui utilise le module de synchronisation ElasticSearch. Le processus serait similaire pour la configuration de tout autre plug-in de synchronisation.



Note : plug-in de synchronisation par défaut

`rgw` est le plug-in de synchronisation par défaut et il n'est pas nécessaire de le configurer explicitement.

21.8.2.1 Exigences et hypothèses

Supposons que vous ayez la configuration multisite simple décrite à la [Section 21.13, « Passerelles Object Gateway multisites »](#) et composée de deux zones : `us-east` et `us-west`. Maintenant, nous ajoutons une troisième zone `us-east-es`, qui traite uniquement les métadonnées des autres sites. Cette zone peut résider dans la même grappe Ceph ou dans une autre que `us-east`. Cette zone ne consommera que les métadonnées d'autres zones et les passerelles Object Gateway de cette zone ne serviront pas directement les requêtes des utilisateurs finaux.

21.8.2.2 Configuration des zones

1. Créez la troisième zone similaire à celles décrites dans la [Section 21.13, « Passerelles Object Gateway multisites »](#), par exemple :

```
cephuser@adm > radosgw-admin zone create --rgw-zonegroup=us --rgw-zone=us-east-es \
--access-key=SYSTEM-KEY --secret=SECRET --endpoints=http://rgw-es:80
```

2. Il est possible de configurer un module de synchronisation pour cette zone avec la commande suivante :

```
cephuser@adm > radosgw-admin zone modify --rgw-zone=ZONE-NAME --tier-type=TIER-TYPE \
--tier-config={set of key=value pairs}
```

3. Par exemple, dans le module de synchronisation `ElasticSearch` :

```
cephuser@adm > radosgw-admin zone modify --rgw-zone=ZONE-NAME --tier-
type=elasticsearch \
--tier-config=endpoint=http://localhost:9200,num_shards=10,num_replicas=1
```

Pour les différentes options `tier-config` prises en charge, reportez-vous à la [Section 21.8.3, « Module de synchronisation ElasticSearch »](#).

4. Enfin, mettez à jour la période :

```
cephuser@adm > radosgw-admin period update --commit
```

5. Démarrez ensuite la passerelle Object Gateway dans la zone :

```
cephuser@adm > ceph orch start rgw.REALM-NAME.ZONE-NAME
```

21.8.3 Module de synchronisation ElasticSearch

Ce module de synchronisation écrit les métadonnées d'autres zones dans ElasticSearch. À partir de la version « Luminous », voici les champs de données JSON qui sont stockés dans ElasticSearch.

```
{
  "_index" : "rgw-gold-ee5863d6",
  "_type" : "object",
  "_id" : "34137443-8592-48d9-8ca7-160255d52ade.34137.1:object1:null",
  "_score" : 1.0,
  "_source" : {
    "bucket" : "testbucket123",
    "name" : "object1",
    "instance" : "null",
    "versioned_epoch" : 0,
    "owner" : {
      "id" : "user1",
      "display_name" : "user1"
    },
    "permissions" : [
      "user1"
    ],
    "meta" : {
      "size" : 712354,
      "mtime" : "2017-05-04T12:54:16.462Z",
      "etag" : "7ac66c0f148de9519b8bd264312c4d64"
    }
  }
}
```

21.8.3.1 Paramètres de configuration du type de niveau ElasticSearch

endpoint

Indique le noeud d'extrémité du serveur ElasticSearch auquel accéder.

num_shards

(entier) Nombre de partitions avec lesquelles ElasticSearch sera configuré lors de l'initialisation de la synchronisation des données. Notez que cela ne peut pas être modifié après l'initialisation. Toute modification nécessite la reconstruction de l'index ElasticSearch et la réinitialisation du processus de synchronisation des données.

num_replicas

(entier) Nombre de répliques avec lesquelles ElasticSearch sera configuré lors de l'initialisation de la synchronisation des données.

explicit_custom_meta

(true | false) Indique si toutes les métadonnées personnalisées de l'utilisateur seront indexées ou si l'utilisateur devra configurer (au niveau compartiment) les entrées de métadonnées client à indexer. La valeur par défaut est « false ».

index_buckets_list

(liste de chaînes séparées par des virgules) Si elle est vide, tous les compartiments seront indexés. Dans le cas contraire, l'indexation portera uniquement sur les compartiments spécifiés ici. Il est possible de fournir des préfixes de compartiment (« foo * », par exemple) ou des suffixes de compartiment (« *bar », par exemple).

approved_owners_list

(liste de chaînes séparées par des virgules) Si elle est vide, les compartiments de tous les propriétaires seront indexés (sous réserve d'autres restrictions) ; dans le cas contraire, l'indexation portera uniquement sur les compartiments appartenant aux propriétaires spécifiés. Il est également possible de définir des suffixes et des préfixes.

override_index_path

(chaîne) Si elle n'est pas vide, cette chaîne sera utilisée comme chemin d'index ElasticSearch. Dans le cas contraire, le chemin d'index sera déterminé et généré lors de l'initialisation de la synchronisation.

username

Spécifie un nom d'utilisateur pour ElasticSearch si l'authentification est nécessaire.

password

Spécifie un mot de passe pour ElasticSearch si l'authentification est nécessaire.

21.8.3.2 Requêtes de métadonnées

Étant donné que la grappe ElasticSearch stocke désormais les métadonnées d'objet, il est important que le noeud d'extrémité ElasticSearch ne soit pas exposé à tous les utilisateurs, mais accessible uniquement aux administrateurs de la grappe. L'exposition de requêtes de métadonnées à l'utilisateur final lui-même engendre un problème, car celui-ci doit interroger uniquement ses métadonnées et non pas celles des autres utilisateurs. Cette opération requiert que la grappe ElasticSearch authentifie les utilisateurs d'une manière similaire à RGW, ce qui pose problème.

À partir de la version « Luminous » de RGW, la zone maître des métadonnées peut traiter les demandes des utilisateurs finaux. Cette approche présente l'avantage de masquer le noeud d'extrémité ElasticSearch pour le public et de résoudre le problème d'authentification et d'autorisation, puisque RGW peut authentifier lui-même les requêtes de l'utilisateur final. Dans cette optique, RGW introduit une nouvelle requête dans les API de compartiment pouvant desservir les requêtes ElasticSearch. Toutes ces requêtes doivent être envoyées à la zone maître de métadonnées.

Obtention d'une requête ElasticSearch

```
GET /BUCKET?query=QUERY-EXPR
```

Paramètres de requête :

- max-keys : nombre maximal d'entrées à renvoyer
- marker : marqueur de pagination

```
expression := [((<arg> <op> <value> [ ])[<and|or> ...]
```

op est l'un des opérateurs suivants : <, <=, ==, >=, >

Par exemple :

```
GET /?query=name==foo
```

Renvoie toutes les clés indexées pour lesquelles l'utilisateur dispose d'une autorisation de lecture et qui s'appellent « foo ». Dans ce cas, la sortie se compose d'une liste XML de clés similaire à la sortie des compartiments de liste S3.

Configuration des champs de métadonnées personnalisés

Définissez quelles entrées de métadonnées personnalisées doivent être indexées (sous le compartiment indiqué) et précisez le type de ces clés. Si l'indexation explicite de métadonnées personnalisées est configurée, cette opération est nécessaire pour que rgw indexe les valeurs de ces métadonnées personnalisées. Dans le cas contraire, cette opération est nécessaire lorsque les clés de métadonnées indexées ne sont pas du type chaîne.

```
POST /BUCKET?mdsearch
x-amz-meta-search: <key [; type]> [, ...]
```

Les champs de métadonnées doivent être séparés les uns des autres par des virgules ; pour forcer le type d'un champ, indiquez « ; ». Les types actuellement autorisés sont « string » (chaîne - valeur par défaut), « integer » (entier) et « date ». Par exemple, si vous souhaitez indexer des métadonnées d'objet personnalisées x-amz-meta-year comme entier, x-amz-meta-date comme date et x-amz-meta-title comme chaîne, vous exécuteriez la commande suivante :

```
POST /mybooks?mdsearch
x-amz-meta-search: x-amz-meta-year;int, x-amz-meta-release-date;date, x-amz-meta-title;string
```

Suppression de la configuration des métadonnées personnalisée

Supprimez la configuration de compartiment de métadonnées personnalisée.

```
DELETE /BUCKET?mdsearch
```

Obtention de la configuration des métadonnées personnalisée

Récupérez la configuration de compartiment de métadonnées personnalisée.

```
GET /BUCKET?mdsearch
```

21.8.4 Module de synchronisation cloud

Cette section présente un module qui synchronise les données de zone avec un service cloud distant. La synchronisation est unidirectionnelle : la date n'est pas resynchronisée à partir de la zone distante. L'objectif principal de ce module est de permettre la synchronisation des données auprès de plusieurs fournisseurs de services cloud. Actuellement, il prend en charge les fournisseurs de services cloud qui sont compatibles avec AWS (S3).

Pour synchroniser les données auprès d'un service cloud distant, vous devez configurer les informations d'identification de l'utilisateur. Étant donné que de nombreux services cloud introduisent des limites sur le nombre de compartiments que chaque utilisateur peut créer, vous

pouvez configurer l'assignation des objets et compartiments sources, des cibles différentes pour différents compartiments et des préfixes de compartiment. Notez que les listes d'accès (ACL) sources ne seront pas conservées. Il est possible d'assigner les autorisations d'utilisateurs sources spécifiques à des utilisateurs cibles spécifiques.

En raison de limitations d'API, il n'existe aucun moyen de conserver l'heure de modification de l'objet d'origine ni la balise d'entité HTTP (Etag). Le module de synchronisation cloud enregistre ces informations sous forme d'attributs de métadonnées sur les objets cibles.

21.8.4.1 Configuration du module de synchronisation cloud

Voici des exemples d'une configuration générique et non générique pour le module de synchronisation cloud. Notez que la configuration générique peut entrer en conflit avec la configuration non générique.

EXEMPLE 21.1 : CONFIGURATION GÉNÉRIQUE

```
{
  "connection": {
    "access_key": ACCESS,
    "secret": SECRET,
    "endpoint": ENDPOINT,
    "host_style": path | virtual,
  },
  "acls": [ { "type": id | email | uri,
    "source_id": SOURCE_ID,
    "dest_id": DEST_ID } ... ],
  "target_path": TARGET_PATH,
}
```

EXEMPLE 21.2 : CONFIGURATION NON GÉNÉRIQUE

```
{
  "default": {
    "connection": {
      "access_key": ACCESS,
      "secret": SECRET,
      "endpoint": ENDPOINT,
      "host_style" path | virtual,
    },
    "acls": [
      {
        "type": id | email | uri,  # optional, default is id
        "source_id": ID,
      }
    ]
  }
}
```

```

    "dest_id": ID
  } ... ]
  "target_path": PATH # optional
},
"connections": [
{
  "connection_id": ID,
  "access_key": ACCESS,
  "secret": SECRET,
  "endpoint": ENDPOINT,
  "host_style": path | virtual, # optional
} ... ],
"acl_profiles": [
{
  "acls_id": ID, # acl mappings
  "acls": [ {
    "type": id | email | uri,
    "source_id": ID,
    "dest_id": ID
  } ... ]
}
],
"profiles": [
{
  "source_bucket": SOURCE,
  "connection_id": CONNECTION_ID,
  "acls_id": MAPPINGS_ID,
  "target_path": DEST,          # optional
} ... ],
}

```

Voici l'explication des termes de configuration utilisés :

connection

Représente une connexion au service cloud distant. Contient les valeurs « connection_id », « access_key », « secret », « endpoint » et « host_style ».

access_key

Clé d'accès cloud distant qui sera utilisée pour la connexion spécifique.

secret

Clé secrète du service cloud distant.

endpoint

URL du noeud d'extrémité du service cloud distant.

host_style

Type de style hôte (« path » [chemin] ou « virtual » [virtuel]) à utiliser lors de l'accès au noeud d'extrémité cloud distant. La valeur par défaut est « path ».

acls

Tableau des assignations de liste d'accès.

acl_mapping

Chaque structure « acl_mapping » contient les valeurs « type », « source_id » et « dest_id ». Celles-ci définissent la mutation ACL pour chaque objet. Une mutation ACL permet de convertir l'ID utilisateur source en ID cible.

type

Type d'ACL : « id » définit l'identifiant utilisateur, « email » définit l'utilisateur à l'aide de son adresse électronique et « uri » définit l'utilisateur selon son URI (groupe).

source_id

ID de l'utilisateur dans la zone source.

dest_id

ID de l'utilisateur dans la destination.

target_path

Chaîne qui définit la façon dont le chemin cible est créé. Le chemin cible spécifie un préfixe auquel le nom de l'objet source est ajouté. Les valeurs du chemin cible pouvant être configurées sont les variables suivantes :

SID

Chaîne unique qui représente l'ID d'instance de synchronisation.

ZONEGROUP

Nom du groupe de zones.

ZONEGROUP_ID

ID du groupe de zones.

ZONE

Nom de la zone.

ZONE_ID

ID de la zone.

BUCKET

Nom du compartiment source.

OWNER

ID du propriétaire du compartiment source.

Par exemple : `target_path = rgwx-ZONE-SID/OWNER/BUCKET`

acl_profiles

Tableau des profils de listes d'accès.

acl_profile

Chaque profil contient une valeur « `acls_id` » qui représentent le profil et un tableau « `acls` » qui reprend une liste des « `acl_mappings` ».

profiles

Liste des profils. Chaque profil contient les éléments suivants :

source_bucket

Nom ou préfixe (si se termine avec `*`) de compartiment qui définit le ou les compartiments sources de ce profil.

target_path

Voir ci-dessus pour l'explication.

connection_id

ID de la connexion qui sera utilisée pour ce profil.

acls_id

ID du profil d'ACL qui sera utilisé pour ce profil.

21.8.4.2 Valeurs configurables spécifiques à S3

Le module de synchronisation cloud fonctionne uniquement avec les interfaces dorsales compatibles avec AWS S3. Quelques valeurs configurables peuvent être utilisées pour modifier son comportement lors de l'accès aux services cloud S3 :

```
{
  "multipart_sync_threshold": OBJECT_SIZE,
  "multipart_min_part_size": PART_SIZE
}
```

`multipart_sync_threshold`

Les objets dont la taille est égale ou supérieure à cette valeur seront synchronisés avec le service cloud à l'aide du téléchargement en plusieurs parties.

`multipart_min_part_size`

Taille minimale des parties à utiliser lors de la synchronisation d'objets à l'aide du téléchargement en plusieurs parties.

21.8.5 Module de synchronisation de l'archivage

Le *module de synchronisation de l'archivage* utilise la fonction de contrôle de version des objets S3 dans Object Gateway. Vous pouvez configurer une *zone d'archivage* qui capture les différentes versions des objets S3 au fur et à mesure qu'elles apparaissent dans d'autres zones. L'historique des versions que la zone d'archivage conserve ne peut être éliminé que par le biais des passerelles associées à cette dernière.

Avec une telle architecture, plusieurs zones sans contrôle de version peuvent mettre en miroir leurs données et métadonnées via leurs passerelles de zone de manière à offrir une haute disponibilité aux utilisateurs finaux, tandis que la zone d'archivage capture toutes les mises à jour de données pour les consolider en tant que versions d'objets S3.

En incluant la zone d'archivage dans une configuration multizone, vous bénéficiez de la flexibilité d'un historique des objets S3 au sein d'une zone unique, tout en économisant l'espace que les répliques des objets S3 avec contrôle de version consommeraient dans les autres zones.

21.8.5.1 Configuration du module de synchronisation de l'archivage



Astuce : pour en savoir plus

Reportez-vous à la [Section 21.13, « Passerelles Object Gateway multisites »](#) pour plus de détails sur la configuration des passerelles multisites.

Reportez-vous à la [Section 21.8, « Modules de synchronisation »](#) pour plus de détails sur la configuration des modules de synchronisation.

Pour utiliser le module de synchronisation de l'archivage, vous devez créer une zone dont le type de niveau est défini sur `archive` :

```
cephuser@adm > radosgw-admin zone create --rgw-zonegroup=ZONE_GROUP_NAME \
```

```
--rgw-zone=OGW_ZONE_NAME \  
--endpoints=http://OGW_ENDPOINT1_URL[,http://OGW_ENDPOINT2_URL,...]  
--tier-type=archive
```

21.9 authentication LDAP

Outre l'authentification par défaut des utilisateurs locaux, Object Gateway peut également utiliser les services du serveur LDAP pour authentifier les utilisateurs.

21.9.1 Mécanisme d'authentification

Object Gateway extrait les informations d'identification LDAP de l'utilisateur à partir d'un jeton. Un filtre de recherche est défini à partir du nom d'utilisateur. La passerelle Object Gateway utilise le compte de service configuré pour rechercher une entrée correspondante dans l'annuaire. Si une entrée est trouvée, la passerelle Object Gateway tente d'établir une liaison avec le nom distinctif trouvé et le mot de passe à partir du jeton. Si les informations d'identification sont valides, la liaison réussit et Object Gateway accorde l'accès.

Vous pouvez limiter les utilisateurs autorisés en définissant la base de la recherche sur une unité organisationnelle spécifique ou en spécifiant un filtre de recherche personnalisé, qui, par exemple, exige l'appartenance à un groupe spécifique, des classes d'objets personnalisées ou des attributs.

21.9.2 Configuration requise

- *LDAP ou Active Directory* : instance LDAP en cours d'exécution accessible par Object Gateway.
- *Compte de service* : informations d'identification LDAP que la passerelle Object Gateway doit utiliser avec les autorisations de recherche.
- *Compte utilisateur* : au moins un compte utilisateur dans l'annuaire LDAP.

Important : différenciation des utilisateurs LDAP et des utilisateurs locaux

Vous ne devez pas utiliser les mêmes noms d'utilisateur pour les utilisateurs locaux et les utilisateurs authentifiés à l'aide de LDAP. La passerelle Object Gateway ne peut pas les distinguer et les traite comme un même utilisateur.

Astuce : contrôle d'intégrité

Vérifiez le compte de service ou la connexion LDAP à l'aide de l'utilitaire **ldapsearch**.
Par exemple :

```
> ldapsearch -x -D "uid=ceph,ou=system,dc=example,dc=com" -W \  
-H ldaps://example.com -b "ou=users,dc=example,dc=com" 'uid=*' dn
```

Veillez à utiliser les mêmes paramètres LDAP que dans le fichier de configuration Ceph pour éliminer les problèmes éventuels.

21.9.3 Configuration de la passerelle Object Gateway en vue de l'utilisation de l'authentification LDAP

Les paramètres suivants sont liés à l'authentification LDAP :

rgw_s3_auth_use_ldap

Définissez cette option sur true pour activer l'authentification S3 avec LDAP.

rgw_ldap_uri

Indique le serveur LDAP à utiliser. Veillez à utiliser le paramètre ldaps://FQDN:PORT afin d'éviter de transmettre les informations d'identification en texte brut sur une liaison non sécurisée.

rgw_ldap_binddn

Nom distinctif (DN) du compte de service utilisé par Object Gateway.

rgw_ldap_secret

Mot de passe du compte de service.

rgw_ldap_searchdn

Indique la base dans l'arborescence des informations d'annuaire pour la recherche d'utilisateurs. Il peut s'agir de l'unité organisationnelle de vos utilisateurs ou d'une unité organisationnelle plus spécifique.

rgw_ldap_dnattr

Attribut utilisé dans le filtre de recherche en vue de correspondre à un nom d'utilisateur. Il s'agit le plus souvent de uid ou cn selon votre arborescence d'annuaire.

rgw_search_filter

Si cette information n'est pas indiquée, la passerelle Object Gateway s'appuie sur le paramètre rgw_ldap_dnattr pour définir automatiquement le filtre de recherche. Utilisez ce paramètre pour limiter librement la liste des utilisateurs autorisés. Reportez-vous à la [Section 21.9.4, « Utilisation d'un filtre de recherche personnalisé pour limiter l'accès des utilisateurs »](#) pour plus d'informations.

21.9.4 Utilisation d'un filtre de recherche personnalisé pour limiter l'accès des utilisateurs

Il existe deux moyens d'utiliser le paramètre rgw_search_filter.

21.9.4.1 Filtre partiel de restriction supplémentaire du filtre de recherche construit

Voici un exemple de filtre partiel :

```
"objectclass=inetorgperson"
```

La passerelle Object Gateway génère le filtre de recherche comme d'habitude avec le nom d'utilisateur issu du jeton et la valeur de rgw_ldap_dnattr. Le filtre construit est ensuite combiné au filtre partiel à partir de l'attribut rgw_search_filter. En fonction du nom d'utilisateur et des paramètres, le filtre de recherche final peut devenir ce qui suit :

```
"(&(uid=hari)(objectclass=inetorgperson))"
```

Dans ce cas, l'utilisateur « hari » ne recevra un accès que s'il se trouve dans l'annuaire LDAP, possède la classe d'objets « inetorgperson » et a fourni un mot de passe valide.

21.9.4.2 Filtre complet

Un filtre complet doit contenir un jeton `USERNAME` qui sera remplacé par le nom d'utilisateur lors de la tentative d'authentification. Le paramètre `rgw_ldap_dnattr` n'est plus utilisé dans ce cas. Par exemple, pour limiter les utilisateurs valides à un groupe spécifique, utilisez le filtre suivant :

```
" (&(uid=USERNAME) (memberOf=cn=ceph-users,ou=groups,dc=mycompany,dc=com)) "
```



Note : attribut `memberOf`

L'utilisation de l'attribut `memberOf` dans les recherches LDAP requiert la prise en charge côté serveur dans votre implémentation de serveur LDAP spécifique.

21.9.5 Génération d'un jeton d'accès pour l'authentification LDAP

L'utilitaire **`radosgw-token`** génère le jeton d'accès basé sur le nom d'utilisateur et le mot de passe LDAP. Il génère une chaîne codée en base 64 correspondant au jeton d'accès physique. Utilisez votre client S3 favori (voir [Section 21.5.1, « Accès à Object Gateway »](#)) et spécifiez le jeton en tant que clé d'accès ainsi qu'une clé secrète vide.

```
> export RGW_ACCESS_KEY_ID="USERNAME"
> export RGW_SECRET_ACCESS_KEY="PASSWORD"
cephuser@adm > radosgw-token --encode --ttype=ldap
```



Important : informations d'identification en texte clair

Le jeton d'accès est une structure JSON codée en base 64 qui contient les informations d'identification LDAP en texte clair.



Note : Active Directory

Pour Active Directory, utilisez le paramètre `--ttype=ad`.

21.10 Partitionnement d'index de compartiment

Object Gateway stocke les données d'index de compartiment dans une réserve d'index, par défaut `.rgw.buckets.index`. Si vous placez trop d'objets (plusieurs centaines de milliers) dans un même compartiment et que le quota du nombre maximal d'objets par compartiment (`rgw bucket default quota max objects`) n'est pas défini, les performances de la réserve d'index risquent de se dégrader. Le *partitionnement d'index de compartiment* maintient le niveau de performances et autorise un nombre élevé d'objets par compartiment.

21.10.1 Repartitionnement d'index de compartiment

Si un compartiment est devenu volumineux et que sa configuration initiale n'est plus suffisante, la réserve d'index du compartiment doit être redéfinie. Vous pouvez soit utiliser le repartitionnement d'index de compartiment en ligne automatique (voir [Section 21.10.1.1, « Repartitionnement dynamique »](#)), soit repartitionner l'index de compartiment hors ligne manuellement (voir [Section 21.10.1.2, « Repartitionnement manuel »](#)).

21.10.1.1 Repartitionnement dynamique

Depuis la version 5 de SUSE Enterprise Storage, le repartitionnement de compartiments en ligne est pris en charge. Il vérifie si le nombre d'objets par compartiment atteint un certain seuil et augmente automatiquement le nombre de partitions utilisées par l'index de compartiment. Ce processus réduit le nombre d'entrées dans chaque partition d'index de compartiment.

Le processus de détection s'exécute :

- Lorsque les nouveaux objets sont ajoutés au compartiment.
- Dans un processus d'arrière-plan qui analyse périodiquement tous les compartiments. Cela est nécessaire pour traiter les compartiments existants qui ne sont pas mis à jour.

Un compartiment devant être partitionné est ajouté à la file d'attente `reshard_log` en vue de son traitement ultérieur. Les threads de repartitionnement s'exécutent en arrière-plan et effectuent un repartitionnement planifié à la fois.

CONFIGURATION DU REPARTITIONNEMENT DYNAMIQUE

`rgw_dynamic_resharding`

Active ou désactive le repartitionnement dynamique d'index de compartiment. Les valeurs admises sont « true » ou « false ». La valeur par défaut est « true ».

rgw_reshard_num_logs

Nombre de partitions pour le journal de repartitionnement. La valeur par défaut est 16.

rgw_reshard_bucket_lock_duration

Durée de verrouillage sur l'objet Compartiment pendant le repartitionnement. La valeur par défaut est de 120 secondes.

rgw_max_objs_per_shard

Nombre maximal d'objets par partition d'index de compartiment. La valeur par défaut est de 100 000 objets.

rgw_reshard_thread_interval

Durée maximale entre deux exécutions du repartitionnement de threads. La valeur par défaut est de 600 secondes.

COMMANDES D'ADMINISTRATION DU PROCESSUS DE REPARTITIONNEMENT

Ajouter un compartiment à la file d'attente du repartitionnement :

```
cephuser@adm > radosgw-admin reshard add \  
--bucket BUCKET_NAME \  
--num-shards NEW_NUMBER_OF_SHARDS
```

Dresser la liste de la file d'attente de repartitionnement :

```
cephuser@adm > radosgw-admin reshard list
```

Traiter/planifier un repartitionnement de compartiment :

```
cephuser@adm > radosgw-admin reshard process
```

Afficher l'état du repartitionnement de compartiment :

```
cephuser@adm > radosgw-admin reshard status --bucket BUCKET_NAME
```

Annuler la mise en attente du repartitionnement de compartiment :

```
cephuser@adm > radosgw-admin reshard cancel --bucket BUCKET_NAME
```

21.10.1.2 Repartitionnement manuel

Le repartitionnement dynamique mentionné à la [Section 21.10.1.1, « Repartitionnement dynamique »](#) est pris en charge uniquement pour les configurations Object Gateway simples. Pour les configurations multisites, utilisez le repartitionnement manuel décrit dans cette section.

Pour repartitionner l'index de compartiment manuellement hors ligne, utilisez la commande suivante :

```
cephuser@adm > radosgw-admin bucket reshard
```

La commande **bucket reshard** effectue les opérations suivantes :

- Elle crée un nouvel ensemble d'objets d'index de compartiment pour l'objet spécifié.
- Elle propage toutes les entrées de ces objets d'index.
- Elle crée une nouvelle instance de compartiment.
- Elle lie la nouvelle occurrence de compartiment au compartiment afin que toutes les nouvelles opérations d'index passent par les nouveaux index de compartiment.
- Elle copie l'ancien ID et le nouvel ID de compartiment vers la sortie standard.



Astuce

Lorsque vous choisissez un certain nombre de partitions, tenez compte des points suivants : visez un maximum de 100 000 entrées par partition. Les nombres premiers de partitions d'index de compartiment ont tendance à mieux répartir les entrées d'index de compartiment entre les partitions. Par exemple, 503 partitions d'index de compartiment fonctionnent mieux que 500 puisque 503 est un nombre premier.

PROCÉDURE 21.1 : REPARTITIONNEMENT DE L'INDEX DE COMPARTIMENT

1. Assurez-vous que toutes les opérations à effectuer dans le compartiment sont bien arrêtées.
2. Sauvegardez l'index de compartiment original :

```
cephuser@adm > radosgw-admin bi list \  
--bucket=BUCKET_NAME \  
> BUCKET_NAME.list.backup
```

3. Repartitionnez l'index de compartiment :

```
cephuser@adm > radosgw-admin bucket reshard \  
--bucket=BUCKET_NAME \  
--num-shards=NEW_SHARDS_NUMBER
```



Astuce : ancien ID de compartiment

Dans le cadre de sa sortie, cette commande affiche également le nouvel ID et l'ancien ID du compartiment.

21.10.2 Partitionnement d'index des nouveaux compartiments

Deux options affectent le partitionnement d'index de compartiment :

- Utilisez l'option `rgw_override_bucket_index_max_shards` pour les configurations simples.
- Utilisez l'option `bucket_index_max_shards` pour les configurations multisites.

Définir les options sur `0` désactive le partitionnement d'index de compartiment. Une valeur supérieure à `0` active le partitionnement d'index de compartiment et définit le nombre maximal de partitions.

La formule suivante permet de calculer le nombre recommandé de partitions :

```
number_of_objects_expected_in_a_bucket / 100000
```

Sachez que le nombre maximal de partitions est 7877.

21.10.2.1 Configurations multisites

Les configurations multisites peuvent disposer d'une réserve d'index différente pour gérer le basculement. Pour configurer un nombre de partitions cohérent pour les zones d'un groupe de zones, définissez l'option `bucket_index_max_shards` dans la configuration du groupe de zones :

1. Exportez la configuration du groupe de zones dans le fichier `zonegroup.json` :

```
cephuser@adm > radosgw-admin zonegroup get > zonegroup.json
```

2. Modifiez le fichier `zonegroup.json` et définissez l'option `bucket_index_max_shards` pour chaque zone nommée.
3. Réinitialisez le groupe de zones :

```
cephuser@adm > radosgw-admin zonegroup set < zonegroup.json
```

4. Mettez à jour la période. Reportez-vous à la [Section 21.13.2.6, « Mise à jour de la période »](#).

21.11 Intégration à OpenStack Keystone

OpenStack Keystone est un service d'identité pour le produit OpenStack. Vous pouvez intégrer la passerelle Object Gateway à Keystone pour configurer une passerelle qui accepte un jeton d'authentification Keystone. Un utilisateur autorisé par Keystone à accéder à la passerelle est vérifié du côté de Ceph Object Gateway et créé automatiquement, si nécessaire. Object Gateway interroge Keystone périodiquement pour obtenir la liste des jetons révoqués.

21.11.1 Configuration d'OpenStack

Avant de configurer la passerelle Ceph Object Gateway, vous devez configurer OpenStack Keystone afin d'activer le service Swift et de le faire pointer vers la passerelle Ceph Object Gateway :

1. *Définissez le service Swift.* Pour utiliser OpenStack en vue de la validation des utilisateurs Swift, créez d'abord le service Swift :

```
> openstack service create \
  --name=swift \
  --description="Swift Service" \
  object-store
```

2. *Définissez les noeuds d'extrémité.* Après avoir créé le service Swift, pointez vers la passerelle Ceph Object Gateway. Remplacez REGION_NAME par le nom du groupe de zones ou de la région de la passerelle.

```
> openstack endpoint create --region REGION_NAME \
  --publicurl "http://radosgw.example.com:8080/swift/v1" \
  --adminurl "http://radosgw.example.com:8080/swift/v1" \
  --internalurl "http://radosgw.example.com:8080/swift/v1" \
  swift
```

3. *Vérifiez les paramètres.* Après avoir créé le service Swift et défini les noeuds d'extrémité, affichez ceux-ci pour vérifier que tous les paramètres sont corrects.

```
> openstack endpoint show object-store
```

21.11.2 Configuration de la passerelle Ceph Object Gateway

21.11.2.1 Configuration des certificats SSL

Ceph Object Gateway interroge Keystone périodiquement pour obtenir la liste des jetons révoqués. Ces requêtes sont codées et signées. Il est également possible de configurer Keystone pour fournir des jetons signés automatiquement, qui sont aussi codés et signés. Vous devez configurer la passerelle afin qu'elle puisse décoder et vérifier ces messages signés. Par conséquent, les certificats OpenSSL que Keystone utilise pour créer les requêtes doivent être convertis au format « nss db » :

```
# mkdir /var/ceph/nss
# openssl x509 -in /etc/keystone/ssl/certs/ca.pem \
  -pubkey | certutil -d /var/ceph/nss -A -n ca -t "TCu,Cu,Tuw"
rootopenssl x509 -in /etc/keystone/ssl/certs/signing_cert.pem \
  -pubkey | certutil -A -d /var/ceph/nss -n signing_cert -t "P,P,P"
```

Pour permettre à Ceph Object Gateway d'interagir avec OpenStack Keystone, ce dernier peut utiliser un certificat SSL auto-signé. Installez le certificat SSL de Keystone sur le noeud exécutant Ceph Object Gateway ou définissez l'option `rgw keystone verify ssl` sur « false ». Définir `rgw keystone verify ssl` sur « false » signifie que la passerelle ne tentera pas de vérifier le certificat.

21.11.2.2 Configuration des options de la passerelle Object Gateway

Vous pouvez configurer l'intégration de Keystone à l'aide des options suivantes :

`rgw keystone api version`

Version de l'API Keystone. Les options valides sont 2 ou 3. La valeur par défaut est 2.

`rgw keystone url`

URL et numéro de port de l'API RESTful d'administration sur le serveur Keystone. Suit le modèle `URL_SERVEUR:NUMÉRO_PORT`.

`rgw keystone admin token`

Jeton ou secret partagé qui est configuré en interne dans Keystone pour les requêtes d'administration.

`rgw keystone accepted roles`

Rôles nécessaires pour répondre aux requêtes. Par défaut « Member, admin ».

rgw keystone accepted admin roles

Liste des rôles autorisant un utilisateur à obtenir des privilèges d'administration.

rgw keystone token cache size

Nombre maximal d'entrées dans le cache de jetons Keystone.

rgw keystone revocation interval

Nombre de secondes avant le contrôle de jetons révoqués. Par défaut, 15 * 60.

rgw keystone implicit tenants

Créez des utilisateurs dans leurs locataires du même nom. La valeur par défaut est « false ».

rgw s3 auth use keystone

Si ce paramètre est défini sur « false », Ceph Object Gateway authentifie les utilisateurs à l'aide de Keystone. La valeur par défaut est « false ».

nss db path

Chemin d'accès à la base de données NSS.

Il est également possible de configurer le locataire du service Keystone, le nom d'utilisateur et le mot de passe Keystone (pour la version 2.0 de l'API Identity OpenStack) d'une façon similaire aux services OpenStack. De cette façon, vous pouvez éviter de définir le secret partagé rgw keystone admin token dans le fichier de configuration, qui doit être désactivé dans les environnements de production. Les informations d'identification du locataire du service doivent disposer de privilèges d'administration. Pour plus de détails, reportez-vous à la [documentation officielle OpenStack Keystone \(https://docs.openstack.org/keystone/latest/#setting-up-projects-users-and-roles\)](https://docs.openstack.org/keystone/latest/#setting-up-projects-users-and-roles)⁷. Les options de configuration associées sont les suivantes :

rgw keystone admin user

Nom d'utilisateur de l'administrateur Keystone.

rgw keystone admin password

Mot de passe de l'administrateur Keystone.

rgw keystone admin tenant

Locataire de l'utilisateur administrateur Keystone version 2.0.

Un utilisateur Ceph Object Gateway est assigné à un locataire Keystone. Un utilisateur Keystone possède plusieurs rôles qui lui sont assignés, sur plusieurs locataires, le cas échéant. Lorsque la passerelle Ceph Object Gateway reçoit le ticket, elle examine le locataire et les rôles utilisateur qui lui sont attribués, et elle accepte ou rejette la demande en fonction de la valeur de l'option rgw keystone accepted roles.



Astuce : assignation aux locataires OpenStack

Les locataires Swift sont assignés à l'utilisateur Object Gateway par défaut, mais peuvent l'être également aux locataires OpenStack grâce à l'option `rgw keystone implicit tenants`. Les conteneurs utiliseront alors l'espace de noms du locataire à la place de l'espace de noms global S3 associé par défaut à Object Gateway. Il est recommandé de choisir la méthode d'assignation au stade de la planification afin d'éviter toute confusion. En effet, l'affectation ultérieure de l'option concerne uniquement les requêtes plus récentes qui sont alors assignées sous un locataire, alors que les compartiments créés précédemment continuent à résider dans un espace de noms global.

Pour la version 3 de l'API Identity OpenStack, vous devez remplacer l'option `rgw keystone admin tenant` par :

`rgw keystone admin domain`

Domaine de l'utilisateur administrateur Keystone.

`rgw keystone admin project`

Projet de l'utilisateur administrateur Keystone.

21.12 Placement de réserve et classes de stockage

21.12.1 Affichage des cibles de placement

Les cibles de placement contrôlent les réserves associées à un compartiment particulier. La cible de placement d'un compartiment est sélectionnée lors de la création et ne peut pas être modifiée. Vous pouvez afficher son paramètre `placement_rule` en exécutant la commande suivante :

```
cephuser@adm > radosgw-admin bucket stats
```

La configuration du groupe de zones contient une liste de cibles de placement avec une cible initiale nommée « default-placement ». La configuration de la zone assigne ensuite le nom de la cible de placement de chaque groupe de zones à son stockage local. Ces informations de placement de zone incluent le nom « `index_pool` » pour l'index de compartiment, le nom « `data_extra_pool` » pour les métadonnées sur les téléchargements multiparties incomplets et un nom « `data_pool` » pour chaque classe de stockage.

21.12.2 Classes de stockage

Les classes de stockage aident à personnaliser le placement des données d'objets. Les règles de cycle de vie de compartiment S3 peuvent automatiser la transition des objets entre les classes de stockage.

Les classes de stockage sont définies en fonction des cibles de placement. Chaque cible de placement de groupe de zones répertorie ses classes de stockage disponibles avec une classe initiale nommée « STANDARD ». La configuration de zone fournit un nom de réserve « data_pool » pour chacune des classes de stockage du groupe de zones.

21.12.3 Configuration des groupes de zones et des zones

Utilisez la commande **radosgw-admin** sur les groupes de zones et les zones pour configurer leur placement. Vous pouvez interroger la configuration du placement de groupe de zones à l'aide de la commande suivante :

```
cephuser@adm > radosgw-admin zonegroup get
{
  "id": "ab01123f-e0df-4f29-9d71-b44888d67cd5",
  "name": "default",
  "api_name": "default",
  ...
  "placement_targets": [
    {
      "name": "default-placement",
      "tags": [],
      "storage_classes": [
        "STANDARD"
      ]
    }
  ],
  "default_placement": "default-placement",
  ...
}
```

Pour interroger la configuration du placement de zone, exécutez la commande suivante :

```
cephuser@adm > radosgw-admin zone get
{
  "id": "557cdcee-3aae-4e9e-85c7-2f86f5eddb1f",
  "name": "default",
  "domain_root": "default.rgw.meta:root",
  ...
}
```



```

    "placement_pools": [
      {
        "key": "default-placement",
        "val": {
          "index_pool": "default.rgw.buckets.index",
          "storage_classes": {
            "STANDARD": {
              "data_pool": "default.rgw.buckets.data"
            }
          },
          "data_extra_pool": "default.rgw.buckets.non-ec",
          "index_type": 0
        }
      }
    ],
    ...
  }

```



Note : aucune configuration multisite précédente

Si vous n'avez effectué aucune configuration multisite auparavant, le système crée pour vous une zone et un groupe de zones « default » (par défaut). Les modifications apportées à cette zone/ce groupe de zones ne prennent effet qu'après avoir redémarré les passerelles Ceph Object Gateway. Si vous avez créé un domaine Kerberos pour plusieurs sites, les modifications de zone/groupe de zones entrent en vigueur une fois que vous les avez validées avec la commande **`radosgw-admin period update --commit`**.

21.12.3.1 Ajout d'une cible de placement

Pour créer une cible de placement nommée « temporary », commencez par l'ajouter au groupe de zones :

```

cephuser@adm > radosgw-admin zonegroup placement add \
  --rgw-zonegroup default \
  --placement-id temporary

```

Ensuite, fournissez les informations de placement de zone pour cette cible :

```

cephuser@adm > radosgw-admin zone placement add \
  --rgw-zone default \
  --placement-id temporary \
  --data-pool default.rgw.temporary.data \
  --index-pool default.rgw.temporary.index \

```

```
--data-extra-pool default.rgw temporary.non-ec
```

21.12.3.2 Ajout d'une classe de stockage

Pour ajouter une nouvelle classe de stockage nommée « COLD » à la cible « default-placement », commencez par l'ajouter au groupe de zones :

```
cephuser@adm > radosgw-admin zonegroup placement add \  
  --rgw-zonegroup default \  
  --placement-id default-placement \  
  --storage-class COLD
```

Ensuite, fournissez les informations de placement de zone pour cette classe de stockage :

```
cephuser@adm > radosgw-admin zone placement add \  
  --rgw-zone default \  
  --placement-id default-placement \  
  --storage-class COLD \  
  --data-pool default.rgw.cold.data \  
  --compression lz4
```

21.12.4 Personnalisation du placement

21.12.4.1 Modification du placement des groupes de zones par défaut

Par défaut, les nouveaux compartiments utilisent la cible `default_placement` du groupe de zones. Vous pouvez modifier ce paramètre de groupe de zones avec la commande suivante :

```
cephuser@adm > radosgw-admin zonegroup placement default \  
  --rgw-zonegroup default \  
  --placement-id new-placement
```

21.12.4.2 Modification du placement par défaut de l'utilisateur

Un utilisateur de Ceph Object Gateway peut remplacer la cible de placement par défaut du groupe de zones en définissant un champ `default_placement` non vide dans les informations utilisateur. De même, la valeur `default_storage_class` peut remplacer la classe de stockage `STANDARD` appliquée aux objets par défaut.

```
cephuser@adm > radosgw-admin user info --uid testid
```

```
{
  ...
  "default_placement": "",
  "default_storage_class": "",
  "placement_tags": [],
  ...
}
```

Si la cible de placement d'un groupe de zones contient des balises, les utilisateurs ne seront pas en mesure de créer des compartiments avec cette cible de placement, excepté si leurs informations utilisateur contiennent au moins une balise correspondante dans leur champ « placement_tags ». Cela peut être utile pour restreindre l'accès à certains types de stockage.

La commande **radosgw-admin** ne peut pas modifier ces champs directement, raison pour laquelle vous devez modifier manuellement le format JSON :

```
cephuser@adm > radosgw-admin metadata get user:USER-ID > user.json
> vi user.json      # edit the file as required
cephuser@adm > radosgw-admin metadata put user:USER-ID < user.json
```

21.12.4.3 Modification du placement du compartiment par défaut S3

Lors de la création d'un compartiment avec le protocole S3, une cible de placement peut être fournie dans le cadre de l'option LocationConstraint pour remplacer les cibles de placement par défaut de l'utilisateur et du groupe de zones.

Normalement, l'option LocationConstraint doit correspondre à la valeur api_name du groupe de zones :

```
<LocationConstraint>default</LocationConstraint>
```

Vous pouvez ajouter une cible de placement personnalisée à la valeur api_name en insérant à sa suite un caractère « : » suivi de la cible :

```
<LocationConstraint>default:new-placement</LocationConstraint>
```

21.12.4.4 Modification du placement du compartiment Swift

Lorsque vous créez un compartiment avec le protocole Swift, vous pouvez fournir une cible de placement dans l'option X-Storage-Policy de l'en-tête HTTP :

```
X-Storage-Policy: NEW-PLACEMENT
```

21.12.5 Utilisation des classes de stockage

Toutes les cibles de placement ont une classe de stockage `STANDARD` qui est appliquée par défaut aux nouveaux objets. Vous pouvez remplacer cette valeur par défaut avec son option `default_storage_class`.

Pour créer un objet dans une classe de stockage autre que celle par défaut, spécifiez le nom de cette classe de stockage dans un en-tête HTTP avec la requête. Le protocole S3 utilise l'en-tête `X-Amz-Storage-Class`, tandis que le protocole Swift utilise l'en-tête `X-Object-Storage-Class`.

Vous pouvez utiliser la fonction de gestion du cycle de vie des objets S3 (*S3 Object Lifecycle Management*) pour déplacer les données d'objets entre les classes de stockage à l'aide d'opérations `Transition`.

21.13 Passerelles Object Gateway multisites

Ceph prend en charge plusieurs options de configuration multisite pour la passerelle Ceph Object Gateway :

Multizone

Configuration composée d'un groupe de zones et de plusieurs zones, chacune d'elles présentant une ou plusieurs instances `ceph-radosgw`. Chaque zone est soutenue par sa propre grappe de stockage Ceph. Plusieurs zones au sein d'un groupe fournissent une reprise après sinistre pour le groupe de zones en cas de défaillance importante de l'une d'elles. Chaque zone est active et peut recevoir des opérations d'écriture. Outre la reprise après sinistre, plusieurs zones actives peuvent également servir de base aux réseaux de distribution de contenu.

Groupe multizone

Ceph Object Gateway prend en charge plusieurs groupes de zones, chacun d'entre eux comportant une ou plusieurs zones. Les objets stockés dans les zones d'un groupe de zones inclus dans le même domaine qu'un autre groupe de zones partagent un espace de noms d'objet global, ce qui permet de garantir l'unicité des ID d'objet dans les groupes de zones et les zones.



Note

Il est important de noter que les groupes de zones synchronisent *uniquement* les métadonnées entre eux. Les données et les métadonnées sont répliquées entre les zones du groupe de zones. Aucune donnée ou métadonnée n'est partagée dans un domaine.

Plusieurs domaines

Ceph Object Gateway prend en charge la notion de domaines ; un espace de noms globalement unique. Plusieurs domaines sont pris en charge et peuvent englober un ou plusieurs groupes de zones.

Vous pouvez configurer chaque passerelle Object Gateway pour qu'elle fonctionne dans une configuration de zone active-active, ce qui permet d'écrire dans des zones non maîtres. La configuration multisite est stockée dans un conteneur appelé domaine. Le domaine stocke des groupes de zones, des zones et une période avec plusieurs époques pour le suivi des modifications apportées à la configuration. Les daemons `rgw` gèrent la synchronisation, ce qui élimine le besoin d'un agent de synchronisation distinct. Cette approche de la synchronisation permet à la passerelle Ceph Object Gateway de fonctionner avec une configuration active-active plutôt qu'active-passive.

21.13.1 Exigences et hypothèses

Une configuration multisite nécessite au moins deux grappes de stockage Ceph et au moins deux instances Ceph Object Gateway, une pour chaque grappe de stockage Ceph. Dans la configuration suivante, au moins deux grappes de stockage Ceph doivent être situées dans des emplacements géographiquement distincts. Cependant, la configuration peut fonctionner sur le même site. Par exemple, sous les noms `rgw1` et `rgw2`.

Une configuration multisite nécessite un groupe de zones maître et une zone maître. Une zone maître est une source fiable pour toutes les opérations de métadonnées dans une grappe multisite. En outre, chaque groupe de zones nécessite une zone maître. Les groupes de zones peuvent avoir une ou plusieurs zones secondaires ou non maîtres. Dans ce guide, l'hôte `rgw1` fait office de zone maître du groupe de zones maître et l'hôte `rgw2` sert de zone secondaire du groupe de zones maître.

21.13.2 Configuration d'une zone maître

Toutes les passerelles d'une configuration multisite récupèrent leur configuration à partir d'un daemon `ceph - radosgw` sur un hôte au sein du groupe de zones maître et de la zone maître. Pour configurer vos passerelles dans une configuration multisite, sélectionnez une instance `ceph - radosgw` pour configurer le groupe de zones maître et la zone maître.

21.13.2.1 Création d'un domaine

Un domaine représente un espace de noms globalement unique constitué d'un ou de plusieurs groupes de zones contenant une ou plusieurs zones. Les zones comportent des compartiments qui, à leur tour, contiennent des objets. Un domaine permet à Ceph Object Gateway de prendre en charge plusieurs espaces de noms et leur configuration sur le même matériel. Un domaine contient la notion de périodes. Chaque période représente l'état de la configuration du groupe de zones et de la zone dans le temps. Chaque fois que vous apportez une modification à un groupe de zones ou à une zone, mettez à jour la période et validez-la. Par défaut, la passerelle Ceph Object Gateway ne crée pas de domaine pour des raisons de compatibilité avec les versions précédentes. Il est recommandé de créer des domaines pour les nouvelles grappes.

Créez un nouveau domaine appelé `gold` pour la configuration multisite en ouvrant une interface de ligne de commande sur un hôte identifié pour desservir le groupe de zones et la zone maîtres. Ensuite, exécutez la commande suivante :

```
cephuser@adm > radosgw-admin realm create --rgw-realm=gold --default
```

Si la grappe a un seul domaine, spécifiez l'indicateur `--default`. Si l'indicateur `--default` est spécifié, `radosgw-admin` utilise ce domaine par défaut. Si l'indicateur `--default` n'est pas spécifié, l'ajout de groupes de zones et de zones nécessite la spécification de l'indicateur `--rgw-realm` ou `--realm-id` pour identifier le domaine lors de l'ajout de groupes de zones et de zones.

Une fois le domaine créé, `radosgw-admin` renvoie la configuration du domaine :

```
{
  "id": "4a367026-bd8f-40ee-b486-8212482ddcd7",
  "name": "gold",
  "current_period": "09559832-67a4-4101-8b3f-10dfcd6b2707",
  "epoch": 1
}
```



Note

Ceph génère un ID unique pour le domaine, ce qui permet de renommer un domaine, le cas échéant.

21.13.2.2 Création d'un groupe de zones maître

Un domaine doit compter au moins un groupe de zones maître. Créez un groupe de zones maître pour la configuration multisite en ouvrant une interface de ligne de commande sur un hôte identifié pour desservir le groupe de zones et la zone maîtres. Créez un groupe de zones maître appelé us en exécutant la commande suivante :

```
cephuser@adm > radosgw-admin zonegroup create --rgw-zonegroup=us \
--endpoints=http://rgw1:80 --master --default
```

Si le domaine ne comporte qu'un seul groupe de zones, spécifiez l'indicateur --default. Si l'indicateur --default est spécifié, **radosgw-admin** utilise ce groupe de zones par défaut lors de l'ajout de nouvelles zones. Si l'indicateur --default n'est pas spécifié, l'ajout de zones nécessite l'indicateur --rgw-zonegroup ou --zonegroup-id pour identifier le groupe de zones lors de l'ajout ou de la modification de zones.

Une fois le groupe de zones maître créé, **radosgw-admin** renvoie la configuration du groupe de zones. Par exemple :

```
{
  "id": "d4018b8d-8c0d-4072-8919-608726fa369e",
  "name": "us",
  "api_name": "us",
  "is_master": "true",
  "endpoints": [
    "http://rgw1:80"
  ],
  "hostnames": [],
  "hostnames_s3website": [],
  "master_zone": "",
  "zones": [],
  "placement_targets": [],
  "default_placement": "",
  "realm_id": "4a367026-bd8f-40ee-b486-8212482ddcd7"
}
```

21.13.2.3 Création d'une zone maître

Important

Les zones doivent être créées sur un noeud Ceph Object Gateway qui se situe dans la zone.

Créez une zone maître pour la configuration multisite en ouvrant une interface de ligne de commande sur un hôte identifié pour desservir le groupe de zones et la zone maîtres. Exécutez la commande suivante :

```
cephuser@adm > radosgw-admin zone create --rgw-zonegroup=us --rgw-zone=us-east-1 \
--endpoints=http://rgw1:80 --access-key=SYSTEM_ACCESS_KEY --secret=SYSTEM_SECRET_KEY
```

Note

Les options `--access-key` et `--secret` ne sont pas spécifiées dans l'exemple ci-dessus. Ces paramètres sont ajoutés à la zone lorsque l'utilisateur est créé dans la section suivante.

Une fois la zone maître créée, **`radosgw-admin`** renvoie la configuration de la zone. Par exemple :

```
{
  "id": "56dfabbb-2f4e-4223-925e-de3c72de3866",
  "name": "us-east-1",
  "domain_root": "us-east-1.rgw.meta:root",
  "control_pool": "us-east-1.rgw.control",
  "gc_pool": "us-east-1.rgw.log:gc",
  "lc_pool": "us-east-1.rgw.log:lc",
  "log_pool": "us-east-1.rgw.log",
  "intent_log_pool": "us-east-1.rgw.log:intent",
  "usage_log_pool": "us-east-1.rgw.log:usage",
  "reshard_pool": "us-east-1.rgw.log:reshard",
  "user_keys_pool": "us-east-1.rgw.meta:users.keys",
  "user_email_pool": "us-east-1.rgw.meta:users.email",
  "user_swift_pool": "us-east-1.rgw.meta:users.swift",
  "user_uid_pool": "us-east-1.rgw.meta:users.uid",
  "otp_pool": "us-east-1.rgw.otp",
  "system_key": {
    "access_key": "1555b35654ad1656d804",
    "secret_key": "h7GhxuBLTrlhVUyxSPUKUV8r/2EI4ngqJxD7iBdBYLhwluN30JaT3Q=="
  },
  "placement_pools": [
    {
```



```

        "key": "us-east-1-placement",
        "val": {
            "index_pool": "us-east-1.rgw.buckets.index",
            "storage_classes": {
                "STANDARD": {
                    "data_pool": "us-east-1.rgw.buckets.data"
                }
            },
            "data_extra_pool": "us-east-1.rgw.buckets.non-ec",
            "index_type": 0
        }
    },
    "metadata_heap": "",
    "realm_id": ""
}

```

21.13.2.4 Suppression de la zone et du groupe par défaut

! Important

Les étapes suivantes supposent une configuration multisite utilisant des systèmes nouvellement installés qui ne stockent pas encore de données. **Ne supprimez pas** la zone par défaut et ses réserves si vous les utilisez déjà pour stocker des données, sinon les données seront supprimées et irrécupérables.

L'installation par défaut d'Object Gateway crée le groupe de zones par défaut appelé default. Supprimez la zone par défaut si elle existe. Veillez à la supprimer d'abord du groupe de zones par défaut.

```
cephuser@adm > radosgw-admin zonegroup delete --rgw-zonegroup=default
```

Supprimez les réserves par défaut de votre grappe de stockage Ceph si elles existent :

! Important

L'étape suivante suppose une configuration multisite utilisant des systèmes nouvellement installés qui ne contiennent actuellement pas de données. **Ne supprimez pas** le groupe de zones par défaut si vous l'utilisez déjà pour stocker des données.

```
cephuser@adm > ceph osd pool rm default.rgw.control default.rgw.control --yes-i-really-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.data.root default.rgw.data.root --yes-i-really-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.gc default.rgw.gc --yes-i-really-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.log default.rgw.log --yes-i-really-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.meta default.rgw.meta --yes-i-really-really-mean-it
```



Avertissement

La suppression du groupe de zones par défaut entraîne celle de l'utilisateur système. Si vos clés d'administrateur ne sont pas propagées, la fonctionnalité de gestion d'Object Gateway de Ceph Dashboard ne fonctionnera pas. Passez à la section suivante pour recréer votre utilisateur système si vous poursuivez cette étape.

21.13.2.5 Création d'utilisateurs système

Les daemons `ceph-radosgw` doivent s'authentifier avant de récupérer les informations de domaine et de période. Dans la zone maître, créez un utilisateur système pour simplifier l'authentification entre les daemons :

```
cephuser@adm > radosgw-admin user create --uid=zone.user \
--display-name="Zone User" --access-key=SYSTEM_ACCESS_KEY \
--secret=SYSTEM_SECRET_KEY --system
```

Notez les paramètres `access_key` et `secret_key` car les zones secondaires exigent l'authentification auprès de la zone maître.

Ajoutez l'utilisateur système à la zone maître :

```
cephuser@adm > radosgw-admin zone modify --rgw-zone=us-east-1 \
--access-key=ACCESS-KEY --secret=SECRET
```

Mettez à jour la période pour que les modifications prennent effet :

```
cephuser@adm > radosgw-admin period update --commit
```

21.13.2.6 Mise à jour de la période

Après avoir mis à jour la configuration de la zone maître, mettez à jour la période :

```
cephuser@adm > radosgw-admin period update --commit
```

Une fois la période mise à jour, **radosgw-admin** renvoie la configuration de la période. Par exemple :

```
{
  "id": "09559832-67a4-4101-8b3f-10dfcd6b2707", "epoch": 1, "predecessor_uuid": "",
  "sync_status": [], "period_map":
  {
    "id": "09559832-67a4-4101-8b3f-10dfcd6b2707", "zonegroups": [], "short_zone_ids": []
  }, "master_zonegroup": "", "master_zone": "", "period_config":
  {
    "bucket_quota": {
      "enabled": false, "max_size_kb": -1, "max_objects": -1
    }, "user_quota": {
      "enabled": false, "max_size_kb": -1, "max_objects": -1
    }
  }, "realm_id": "4a367026-bd8f-40ee-b486-8212482ddcd7", "realm_name": "gold",
  "realm_epoch": 1
}
```



Note

La mise à jour de la période modifie l'époque et garantit que les autres zones reçoivent la configuration mise à jour.

21.13.2.7 Démarrage de la passerelle Gateway

Sur l'hôte Object Gateway, démarrez et activez le service Ceph Object Gateway. Pour identifier le FSID unique de la grappe, exécutez **ceph fsid**. Pour identifier le nom du daemon Object Gateway, exécutez **ceph orch ps --hostname HOSTNAME**.

```
cephuser@ogw > systemctl start ceph-FSID@DAEMON_NAME
cephuser@ogw > systemctl enable ceph-FSID@DAEMON_NAME
```

21.13.3 Configuration des zones secondaires

Les zones d'un groupe de zones répliquent toutes les données pour garantir que chaque zone possède les mêmes données. Lors de la création de la zone secondaire, exécutez toutes les opérations suivantes sur un hôte identifié pour desservir cette dernière.



Note

Pour ajouter une troisième zone, suivez les mêmes procédures que pour ajouter la zone secondaire. Utilisez un nom de zone différent.



Important

Vous devez exécuter les opérations de métadonnées, telles que la création d'utilisateurs, sur un hôte de la zone maître. La zone maître et la zone secondaire peuvent recevoir des opérations de compartiment, mais la zone secondaire redirige les opérations de compartiment vers la zone maître. Si la zone maître est arrêtée, les opérations de compartiment échouent.

21.13.3.1 Extraction du domaine

À l'aide du chemin URL, de la clé d'accès et du secret de la zone maître dans le groupe de zones maître, importez la configuration du domaine vers l'hôte. Pour extraire un domaine autre que celui par défaut, spécifiez le domaine à l'aide des options de configuration `--rgw-realm` ou `--realm-id`.

```
cephuser@adm > radosgw-admin realm pull --url=url-to-master-zone-gateway --access-key=access-key --secret=secret
```



Note

L'extraction du domaine récupère également la configuration de la période en cours de l'hôte distant et en fait également la période en cours sur cet hôte.

Si ce domaine est le domaine par défaut ou le seul domaine, définissez-le comme par défaut.

```
cephuser@adm > radosgw-admin realm default --rgw-realm=REALM-NAME
```

21.13.3.2 Création d'une zone secondaire

Créez une zone secondaire pour la configuration multisite en ouvrant une interface de ligne de commande sur un hôte identifié pour desservir la zone secondaire. Spécifiez l'ID du groupe de zones, le nom de la nouvelle zone et un noeud d'extrémité pour cette dernière. *N'utilisez pas* l'indicateur `--master`. Toutes les zones s'exécutent dans une configuration active-active par défaut. Si la zone secondaire ne doit pas accepter les opérations d'écriture, spécifiez l'indicateur `--read-only` pour créer une configuration active-passive entre la zone maître et la zone secondaire. En outre, fournissez la clé d'accès (`access_key`) et la clé secrète (`secret_key`) de l'utilisateur système généré stockées dans la zone maître du groupe de zones maître. Exécutez la commande suivante :

```
cephuser@adm > radosgw-admin zone create --rgw-zonegroup=ZONE-GROUP-NAME\
--rgw-zone=ZONE-NAME --endpoints=URL \
--access-key=SYSTEM-KEY --secret=SECRET\
--endpoints=http://FQDN:80 \
[ --read-only]
```

Par exemple :

```
cephuser@adm > radosgw-admin zone create --rgw-zonegroup=us --endpoints=http://rgw2:80 \
--rgw-zone=us-east-2 --access-key=SYSTEM_ACCESS_KEY --secret=SYSTEM_SECRET_KEY
{
  "id": "950c1a43-6836-41a2-a161-64777e07e8b8",
  "name": "us-east-2",
  "domain_root": "us-east-2.rgw.data.root",
  "control_pool": "us-east-2.rgw.control",
  "gc_pool": "us-east-2.rgw.gc",
  "log_pool": "us-east-2.rgw.log",
  "intent_log_pool": "us-east-2.rgw.intent-log",
  "usage_log_pool": "us-east-2.rgw.usage",
  "user_keys_pool": "us-east-2.rgw.users.keys",
  "user_email_pool": "us-east-2.rgw.users.email",
  "user_swift_pool": "us-east-2.rgw.users.swift",
  "user_uid_pool": "us-east-2.rgw.users.uid",
  "system_key": {
    "access_key": "1555b35654ad1656d804",
    "secret_key": "h7GhxuBLTrlhVUyxSPUKUV8r\2EI4ngqJxD7iBdBYLhwluN30JaT3Q=="
  },
  "placement_pools": [
    {
      "key": "default-placement",
      "val": {
        "index_pool": "us-east-2.rgw.buckets.index",
        "data_pool": "us-east-2.rgw.buckets.data",

```

```

        "data_extra_pool": "us-east-2.rgw.buckets.non-ec",
        "index_type": 0
    }
}
],
"metadata_heap": "us-east-2.rgw.meta",
"realm_id": "815d74c2-80d6-4e63-8cfc-232037f7ff5c"
}

```

Important

Les étapes suivantes supposent une configuration multisite utilisant des systèmes nouvellement installés qui ne stockent pas encore de données. **Ne supprimez pas** la zone par défaut et ses réserves si vous les utilisez déjà pour stocker des données, sinon les données seront perdues et irrécupérables.

Supprimez la zone par défaut si nécessaire :

```
cephuser@adm > radosgw-admin zone delete --rgw-zone=default
```

Supprimez les réserves par défaut de votre grappe de stockage Ceph si nécessaire :

```

cephuser@adm > ceph osd pool rm default.rgw.control default.rgw.control --yes-i-really-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.data.root default.rgw.data.root --yes-i-really-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.gc default.rgw.gc --yes-i-really-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.log default.rgw.log --yes-i-really-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.users.uid default.rgw.users.uid --yes-i-really-really-mean-it

```

21.13.3.3 Mise à jour du fichier de configuration Ceph

Mettez à jour le fichier de configuration Ceph sur les hôtes de la zone secondaire en ajoutant l'option de configuration rgw_zone et le nom de la zone secondaire à l'entrée de l'instance.

Pour ce faire, exécutez la commande suivante :

```
cephuser@adm > ceph config set SERVICE_NAME rgw_zone us-west
```

21.13.3.4 Mise à jour de la période

Après avoir mis à jour la configuration de la zone maître, mettez à jour la période :

```
cephuser@adm > radosgw-admin period update --commit
{
  "id": "b5e4d3ec-2a62-4746-b479-4b2bc14b27d1",
  "epoch": 2,
  "predecessor_uuid": "09559832-67a4-4101-8b3f-10dfcd6b2707",
  "sync_status": [ "[...]"
],
  "period_map": {
    "id": "b5e4d3ec-2a62-4746-b479-4b2bc14b27d1",
    "zonegroups": [
      {
        "id": "d4018b8d-8c0d-4072-8919-608726fa369e",
        "name": "us",
        "api_name": "us",
        "is_master": "true",
        "endpoints": [
          "http://\rgw1:80"
        ],
        "hostnames": [],
        "hostnames_s3website": [],
        "master_zone": "83859a9a-9901-4f00-aa6d-285c777e10f0",
        "zones": [
          {
            "id": "83859a9a-9901-4f00-aa6d-285c777e10f0",
            "name": "us-east-1",
            "endpoints": [
              "http://\rgw1:80"
            ],
            "log_meta": "true",
            "log_data": "false",
            "bucket_index_max_shards": 0,
            "read_only": "false"
          },
          {
            "id": "950c1a43-6836-41a2-a161-64777e07e8b8",
            "name": "us-east-2",
            "endpoints": [
              "http://\rgw2:80"
            ],
            "log_meta": "false",
            "log_data": "true",
            "bucket_index_max_shards": 0,
            "read_only": "false"
          }
        ]
      }
    ]
  }
}
```

```

        }

        ],
        "placement_targets": [
            {
                "name": "default-placement",
                "tags": []
            }
        ],
        "default_placement": "default-placement",
        "realm_id": "4a367026-bd8f-40ee-b486-8212482ddcd7"
    }
],
"short_zone_ids": [
    {
        "key": "83859a9a-9901-4f00-aa6d-285c777e10f0",
        "val": 630926044
    },
    {
        "key": "950c1a43-6836-41a2-a161-64777e07e8b8",
        "val": 4276257543
    }
]
},
"master_zonegroup": "d4018b8d-8c0d-4072-8919-608726fa369e",
"master_zone": "83859a9a-9901-4f00-aa6d-285c777e10f0",
"period_config": {
    "bucket_quota": {
        "enabled": false,
        "max_size_kb": -1,
        "max_objects": -1
    },
    "user_quota": {
        "enabled": false,
        "max_size_kb": -1,
        "max_objects": -1
    }
},
"realm_id": "4a367026-bd8f-40ee-b486-8212482ddcd7",
"realm_name": "gold",
"realm_epoch": 2
}

```




Note

La mise à jour de la période modifie l'époque et garantit que les autres zones reçoivent la configuration mise à jour.

21.13.3.5 Démarrage de la passerelle Object Gateway

Sur l'hôte Object Gateway, démarrez et activez le service Ceph Object Gateway :

```
cephuser@adm > ceph orch start rgw.us-east-2
```

21.13.3.6 Vérification de l'état de la synchronisation

Lorsque la zone secondaire est opérationnelle, vérifiez l'état de la synchronisation. La synchronisation copie les utilisateurs et les compartiments créés dans la zone maître vers la zone secondaire.

```
cephuser@adm > radosgw-admin sync status
```

La sortie indique l'état des opérations de synchronisation. Par exemple :

```
realm f3239bc5-e1a8-4206-a81d-e1576480804d (gold)
  zonegroup c50dbb7e-d9ce-47cc-a8bb-97d9b399d388 (us)
    zone 4c453b70-4a16-4ce8-8185-1893b05d346e (us-west)
metadata sync syncing
  full sync: 0/64 shards
  metadata is caught up with master
  incremental sync: 64/64 shards
data sync source: lee9da3e-114d-4ae3-a8a4-056e8a17f532 (us-east)
  syncing
  full sync: 0/128 shards
  incremental sync: 128/128 shards
  data is caught up with source
```



Note

Les zones secondaires acceptent les opérations de compartiment ; toutefois, elles les redirigent vers la zone maître, puis se synchronisent avec la zone maître pour recevoir le résultat des opérations de compartiment. Si la zone maître est arrêtée, les opérations de compartiment exécutées sur la zone secondaire échouent, mais les opérations sur les objets réussissent.

21.13.3.7 Vérification d'un objet

Par défaut, les objets ne sont pas revérifiés si la synchronisation d'un objet réussit. Pour activer la vérification, définissez l'option `rgw_sync_obj_etag_verify` sur `true`. Après l'activation, les objets facultatifs seront synchronisés. Une somme de contrôle MD5 supplémentaire vérifie qu'elle est calculée sur la source et la destination. Cela permet de garantir l'intégrité des objets récupérés à partir d'un serveur distant via HTTP, y compris la synchronisation multisite. Cette option peut affecter les performances des RGW, car davantage de calculs sont nécessaires.

21.13.4 Maintenance générale d'Object Gateway

21.13.4.1 Vérification de l'état de la synchronisation

Pour obtenir des informations sur l'état de réplication d'une zone, exécutez la commande suivante :

```
cephuser@adm > radosgw-admin sync status
  realm b3bc1c37-9c44-4b89-a03b-04c269bea5da (gold)
  zonegroup f54f9b22-b4b6-4a0e-9211-fa6ac1693f49 (us)
    zone adcellc9-b8ed-4a90-8bc5-3fc029ff0816 (us-west)
      metadata sync syncing
        full sync: 0/64 shards
        incremental sync: 64/64 shards
        metadata is behind on 1 shards
        oldest incremental change not applied: 2017-03-22 10:20:00.0.881361s
      data sync source: 341c2d81-4574-4d08-ab0f-5a2a7b168028 (us-east)
        syncing
        full sync: 0/128 shards
        incremental sync: 128/128 shards
        data is caught up with source
```

```
source: 3b5d1a3f-3f27-4e4a-8f34-6072d4bb1275 (us-3)
syncing
full sync: 0/128 shards
incremental sync: 128/128 shards
data is caught up with source
```

La sortie peut différer selon l'état de la synchronisation. Les partitions sont décrites comme deux types différents lors de la synchronisation :

Partitions obsolètes

Les partitions obsolètes sont des partitions qui nécessitent une synchronisation complète des données et des partitions nécessitant une synchronisation incrémentielle des données, car elles ne sont pas à jour.

Partitions de récupération

Les partitions de récupération sont des partitions qui ont rencontré une erreur lors de la synchronisation et signalées comme devant effectuer une nouvelle tentative. L'erreur provient principalement de problèmes mineurs tels que l'échec du verrouillage d'un compartiment. Le problème se résout généralement spontanément.

21.13.4.2 Vérification des journaux

Uniquement dans le cadre d'une configuration multisite, vous pouvez vérifier le journal de métadonnées (`mdlog`), le journal d'index du compartiment (`biolog`) ainsi que le journal de données (`datalog`). Vous pouvez les lister et les découper. Toutefois, ce n'est généralement pas nécessaire, car, par défaut, l'option `rgw_sync_log_trim_interval` est définie sur 20 minutes. Si l'option n'est pas définie manuellement sur 0, vous ne devriez le découper à aucun moment, au risque d'entraîner des effets indésirables.

21.13.4.3 Modification de la zone maître des métadonnées



Important

Faites preuve de précaution lorsque vous modifiez la zone maître des métadonnées. Si une zone n'a pas terminé la synchronisation des métadonnées depuis la zone maître actuelle, elle ne peut pas desservir les entrées restantes une fois promue au rang de maître et ces modifications sont perdues. De ce fait, nous vous recommandons d'attendre que l'état de synchronisation `radosgw-admin` d'une zone indique que la synchronisation des métadon-

nées est terminée avant de la promouvoir au rang de maître. De même, si les modifications apportées aux métadonnées sont traitées par la zone maître actuelle alors qu'une autre zone est promue au rang de maître, ces modifications risquent d'être perdues. Pour éviter cela, nous vous recommandons de fermer toutes les instances d'Object Gateway sur l'ancienne zone maître. Une fois qu'une autre zone a été promue, sa nouvelle période peut être récupérée à l'aide de l'extraction de période `radosgw-admin` et la ou les passerelles peuvent être redémarrées.

Pour promouvoir une zone (par exemple, la zone `us-west` dans le groupe de zones `us`) en tant que maître de métadonnées, exécutez les commandes suivantes sur cette zone :

```
cephuser@ogw > radosgw-admin zone modify --rgw-zone=us-west --master
cephuser@ogw > radosgw-admin zonegroup modify --rgw-zonegroup=us --master
cephuser@ogw > radosgw-admin period update --commit
```

Cela génère une nouvelle période et les instances d'Object Gateway de la zone `us-west` envoient cette période à d'autres zones.

21.13.5 Basculement et reprise après sinistre

Si la zone maître échoue, basculez vers la zone secondaire pour la reprise après sinistre.

1. Faites de la zone secondaire la zone maître et la zone par défaut. Par exemple :

```
cephuser@adm > radosgw-admin zone modify --rgw-zone=ZONE-NAME --master --default
```

Par défaut, Ceph Object Gateway s'exécute dans une configuration active-active. Si la grappe a été configurée pour s'exécuter dans une configuration active-passive, la zone secondaire est une zone en lecture seule. Retirez l'état `--read-only` pour autoriser la zone à recevoir les opérations d'écriture. Par exemple :

```
cephuser@adm > radosgw-admin zone modify --rgw-zone=ZONE-NAME --master --default \
--read-only=false
```

2. Mettez à jour la période pour que les modifications prennent effet :

```
cephuser@adm > radosgw-admin period update --commit
```

3. Redémarrez la passerelle Ceph Object Gateway :

```
cephuser@adm > ceph orch restart rgw
```

En cas de reprise de la zone maître précédente, annulez l'opération.

1. Depuis la zone récupérée, extrayez la dernière configuration de domaine de la zone maître actuelle.

```
cephuser@adm > radosgw-admin realm pull --url=URL-TO-MASTER-ZONE-GATEWAY \
--access-key=ACCESS-KEY --secret=SECRET
```

2. Faites de la zone récupérée la zone maître et la zone par défaut:

```
cephuser@adm > radosgw-admin zone modify --rgw-zone=ZONE-NAME --master --default
```

3. Mettez à jour la période pour que les modifications prennent effet :

```
cephuser@adm > radosgw-admin period update --commit
```

4. Redémarrez la passerelle Ceph Object Gateway dans la zone récupérée :

```
cephuser@adm > ceph orch restart rgw@rgw
```

5. Si la zone secondaire doit être une configuration en lecture seule, mettez à jour la zone secondaire :

```
cephuser@adm > radosgw-admin zone modify --rgw-zone=ZONE-NAME --read-only
```

6. Mettez à jour la période pour que les modifications prennent effet :

```
cephuser@adm > radosgw-admin period update --commit
```

7. Enfin, redémarrez la passerelle Ceph Object Gateway dans la zone secondaire :

```
cephuser@adm > ceph orch restart@rgw
```

22 Passerelle Ceph iSCSI

Ce chapitre est consacré aux tâches d'administration liées à la passerelle iSCSI. Pour suivre une procédure de déploiement, reportez-vous au *Manuel « Guide de déploiement », Chapitre 8 « Déploiement des services essentiels restants à l'aide de cephadm », Section 8.3.5 « Déploiement de passerelles iSCSI »*.

22.1 Cibles gérées par ceph-iscsi


Ce chapitre décrit comment se connecter aux cibles gérées par `ceph-iscsi` à partir de clients exécutant Linux, Microsoft Windows ou VMware.

22.1.1 Connexion à open-iscsi

La connexion aux cibles iSCSI soutenues par `ceph-iscsi` avec `open-iscsi` s'effectue en deux étapes. Tout d'abord, l'initiateur doit découvrir les cibles iSCSI disponibles sur l'hôte de passerelle, puis il doit se connecter et assigner les unités logiques (LU) disponibles.

Les deux étapes exigent que le daemon `open-iscsi` soit en cours d'exécution. La façon dont vous démarrez le daemon `open-iscsi` dépend de votre distribution Linux :

- Sur les hôtes SUSE Linux Enterprise Server (SLES) et Red Hat Enterprise Linux (RHEL), exécutez `systemctl start iscsid` (ou `service iscsid start` si `systemctl` n'est pas disponible).
- Sur les hôtes Debian et Ubuntu, lancez `systemctl start open-iscsi` (ou `service open-iscsi start`).

Si votre hôte initiateur exécute SUSE Linux Enterprise Server, reportez-vous à la page <https://documentation.suse.com/sles/15-SP1/single-html/SLES-storage/#sec-iscsi-initiator>  pour plus de détails sur la façon de se connecter à une cible iSCSI.

Pour toute autre distribution Linux prenant en charge `open-iscsi`, poursuivez la découverte des cibles sur votre passerelle `ceph-iscsi` (cet exemple utilise `iscsi1.example.com` comme adresse de portail ; pour l'accès multipath, répétez ces étapes avec `iscsi2.example.com`) :

```
# iscsiadm -m discovery -t sendtargets -p iscsi1.example.com
```

```
192.168.124.104:3260,1 iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol
```

Ensuite, connectez-vous au portail. Si la connexion s'effectue correctement, les unités logiques soutenues par RBD sur le portail sont immédiatement disponibles sur le bus SCSI du système :

```
# iscsiadm -m node -p iscsil.example.com --login
Logging in to [iface: default, target: iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol, portal: 192.168.124.104,3260] (multiple)
Login to [iface: default, target: iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol, portal: 192.168.124.104,3260] successful.
```

Répétez ce processus pour les autres adresses IP ou hôtes du portail.

Si l'utilitaire `lsscsi` est installé sur votre système, vous pouvez l'utiliser pour énumérer les périphériques SCSI disponibles sur votre système :

```
lsscsi
[8:0:0:0]    disk      SUSE      RBD              4.0    /dev/sde
[9:0:0:0]    disk      SUSE      RBD              4.0    /dev/sdf
```

Dans une configuration multipath (où deux périphériques iSCSI connectés représentent une seule et même LU), vous pouvez également examiner l'état du périphérique multipath avec l'utilitaire `multipath` :

```
# multipath -ll
360014050cf9dcfcb2603933ac3298dca dm-9 SUSE,RBD
size=49G features='0' hwhandler='0' wp=rw
|+- policy='service-time 0' prio=1 status=active
|  `-- 8:0:0:0 sde 8:64 active ready running
`+- policy='service-time 0' prio=1 status=enabled
  `-- 9:0:0:0 sdf 8:80 active ready running
```

Vous pouvez désormais utiliser ce périphérique multipath comme vous le feriez pour n'importe quel périphérique de bloc. Par exemple, vous pouvez utiliser le périphérique en tant que volume physique pour la gestion de volumes logiques (LVM) ou simplement créer un système de fichiers dessus. L'exemple ci-dessous montre comment créer un système de fichiers XFS sur le volume iSCSI multipath nouvellement connecté :

```
# mkfs -t xfs /dev/mapper/360014050cf9dcfcb2603933ac3298dca
log stripe unit (4194304 bytes) is too large (maximum is 256KiB)
log stripe unit adjusted to 32KiB
meta-data=/dev/mapper/360014050cf9dcfcb2603933ac3298dca isize=256    agcount=17,
    agsize=799744 blks
    =                               sectsz=512   attr=2, projid32bit=1
```

| | | | |
|----------|---------------|------------|-----------------------------|
| | = | crc=0 | finobt=0 |
| data | = | bsize=4096 | blocks=12800000, imaxpct=25 |
| | = | sunit=1024 | swidth=1024 blks |
| naming | =version 2 | bsize=4096 | ascii-ci=0 ftype=0 |
| log | =internal log | bsize=4096 | blocks=6256, version=2 |
| | = | sectsz=512 | sunit=8 blks, lazy-count=1 |
| realtime | =none | extsz=4096 | blocks=0, rtextents=0 |

XFS étant un système de fichiers hors grappe, vous pouvez uniquement le monter sur un seul noeud initiateur iSCSI à un moment donné.

Si vous souhaitez à tout moment interrompre l'utilisation des LU iSCSI associées à une cible particulière, exécutez la commande suivante :

```
# iscsiadm -m node -p iscsil.example.com --logout
Logging out of session [sid: 18, iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol,
portal: 192.168.124.104,3260]
Logout of [sid: 18, target: iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol,
portal: 192.168.124.104,3260] successful.
```

Comme pour la découverte et la connexion, vous devez répéter les étapes de déconnexion pour toutes les adresses IP ou tous les noms d'hôte du portail.

22.1.1.1 Configuration multipath

La configuration multipath est gérée sur les clients ou les initiateurs, et elle est indépendante de toute configuration `ceph-iscsi`. Sélectionnez une stratégie avant d'utiliser le stockage de bloc. Après avoir modifié le fichier `/etc/multipath.conf`, redémarrez `multipathd` avec :

```
# systemctl restart multipathd
```

Pour une configuration active-passive avec des noms conviviaux, ajoutez

```
defaults {
    user_friendly_names yes
}
```

à votre fichier `/etc/multipath.conf`. Après vous être connecté à vos cibles, exécutez

```
# multipath -ll
mpathd (36001405dbb561b2b5e439f0aed2f8e1e) dm-0 SUSE,RBD
size=2.0G features='0' hwhandler='0' wp=rw
|+-+ policy='service-time 0' prio=1 status=active
```



```
| ` - 2:0:0:3 sdl 8:176 active ready running
| -+- policy='service-time 0' prio=1 status=enabled
| ` - 3:0:0:3 sdj 8:144 active ready running
` -+- policy='service-time 0' prio=1 status=enabled
  ` - 4:0:0:3 sdk 8:160 active ready running
```

Notez l'état de chaque liaison. Pour une configuration active-active, ajoutez

```
defaults {
    user_friendly_names yes
}

devices {
    device {
        vendor "(LIO-ORG|SUSE)"
        product "RBD"
        path_grouping_policy "multibus"
        path_checker "tur"
        features "0"
        hardware_handler "1 alua"
        prio "alua"
        failback "immediate"
        rr_weight "uniform"
        no_path_retry 12
        rr_min_io 100
    }
}
```

à votre fichier /etc/multipath.conf. Redémarrez multipathd et exécutez

```
# multipath -ll
mpathd (36001405dbb561b2b5e439f0aed2f8e1e) dm-3 SUSE,RBD
size=2.0G features='1 queue_if_no_path' hwhandler='1 alua' wp=rw
`-+- policy='service-time 0' prio=50 status=active
  |- 4:0:0:3 sdj 8:144 active ready running
  |- 3:0:0:3 sdk 8:160 active ready running
  ` - 2:0:0:3 sdl 8:176 active ready running
```

22.1.2 Connexion Microsoft Windows (initiateur iSCSI de Microsoft)

Pour vous connecter à une cible iSCSI de SUSE Enterprise Storage à partir d'un serveur Windows 2012, procédez comme suit :

1. Ouvrez Windows Server Manager. Dans le tableau de bord, sélectionnez *Outils > Initiateur iSCSI*. La boîte de dialogue des *propriétés de l'initiateur iSCSI* apparaît. Sélectionnez l'onglet *Découverte* :

The screenshot shows the 'Properties of iSCSI Initiator' dialog box with the 'Discovery' tab selected. The dialog has several tabs at the top: 'Cibles', 'Découverte', 'Cibles favorites', 'Volumes et périphériques', 'RADIUS', and 'Configuration'. The 'Découverte' tab is active, showing two main sections: 'Portails cibles' and 'Serveurs iSNS'. The 'Portails cibles' section includes a 'Rafrâchir' button, a list of portals with columns 'Adresse', 'Port', 'Adaptateur', and 'Adresse IP', and buttons 'Découvrir un portail...' and 'Supprimer'. The 'Serveurs iSNS' section includes a 'Rafrâchir' button, a list of servers with a 'Nom' column, and buttons 'Ajouter un serveur...' and 'Supprimer'. At the bottom of the dialog are 'OK', 'Annuler', and 'Appliquer' buttons.

FIGURE 22.1 : PROPRIÉTÉS DE L'INITIATEUR iSCSI

2. Dans la boîte de dialogue *Détecter un portail cible*, entrez le nom d'hôte ou l'adresse IP dans le champ *Cible* et cliquez sur *OK* :

Entrez l'adresse IP ou le nom DNS et le numéro de port du portail à ajouter.

Pour modifier les paramètres par défaut de la découverte du portail cible, cliquez sur le bouton Avancé.

Adresse IP ou nom DNS : Port : (La valeur par défaut est 3260.)

192.168.124.104 3260

Avancé... OK Annuler

FIGURE 22.2 : DÉCOUVERTE DU PORTAIL CIBLE

3. Répétez ce processus pour tous les autres noms d'hôte ou adresses IP de passerelle. Une fois terminé, passez en revue la liste *Portails cibles* :

Portails cibles

Le système recherchera des cibles sur les portails suivants :

| Adresse | Port | Adaptateur | Adresse IP |
|-----------------|------|-------------------|-------------------|
| 192.168.124.104 | 3260 | Valeur par défaut | Valeur par défaut |
| 192.168.124.105 | 3260 | Valeur par défaut | Valeur par défaut |

Pour ajouter un portail cible, cliquez sur Découvrir un portail.

Pour supprimer un portail cible, sélectionnez l'adresse ci-dessus, puis cliquez sur Supprimer.

Serveurs iSNS

Le système est enregistré sur les serveurs iSNS suivants :

| Nom |
|-----|
|-----|

Pour ajouter un serveur iSNS, cliquez sur Ajouter un serveur.

Pour supprimer un serveur iSNS, sélectionnez le serveur ci-dessus, puis cliquez sur Supprimer.

OK Annuler Appliquer

FIGURE 22.3 : PORTAILS CIBLES

4. Ensuite, basculez vers l'onglet *Cibles* et passez en revue votre ou vos cibles découvertes.

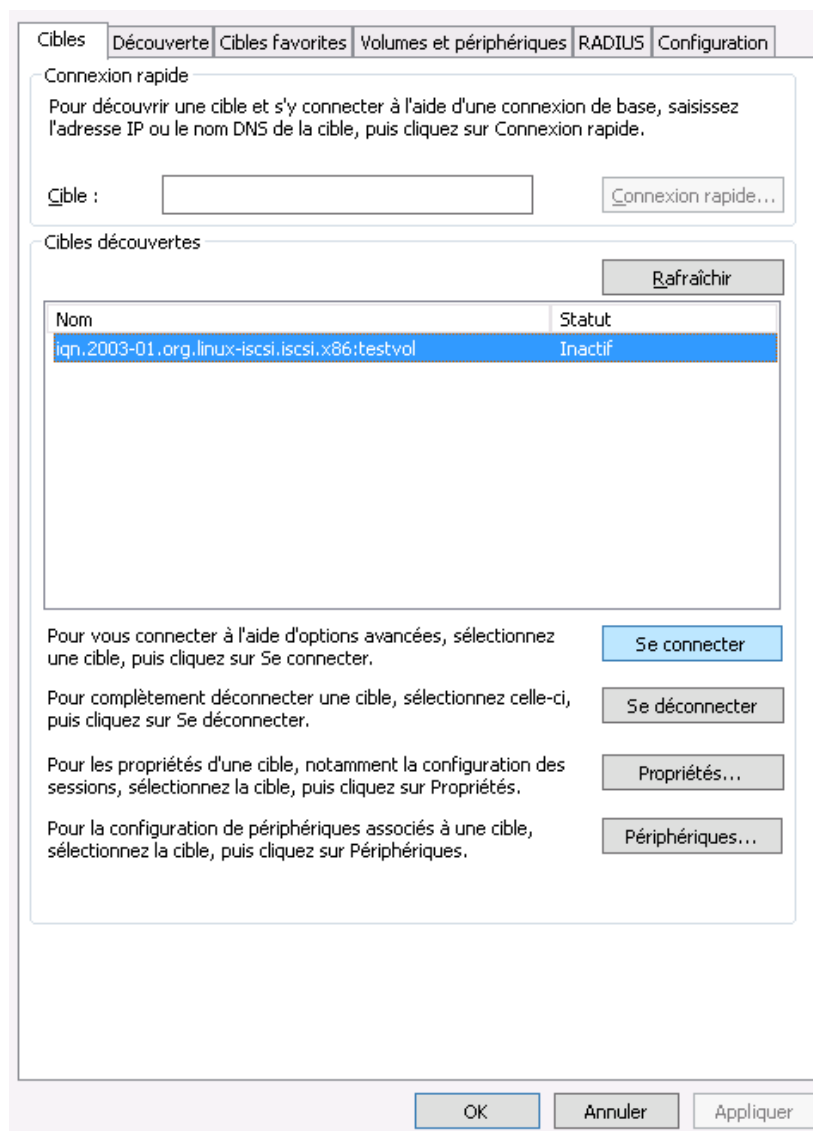


FIGURE 22.4 : CIBLES

5. Cliquez sur *Se connecter* sous l'onglet *Cibles*. La boîte de dialogue *Se connecter à la cible* apparaît. Sélectionnez la case à cocher *Activer la prise en charge de plusieurs chemins d'accès* pour activer MPIO (Multi-path I/O), puis cliquez sur *OK* :

6. Lorsque la boîte de dialogue *Se connecter à la cible* se ferme, sélectionnez *Propriétés* pour examiner les propriétés de la cible :

Sessions Groupes de portails

Rafraîchir

Identificateur

- ☒ fffff00103669020-400001370000000f
- ☒ fffff00103669020-4000013700000010

Pour ajouter une session, cliquez sur Ajouter une session.

Pour déconnecter une ou plusieurs sessions, sélectionnez chaque session, puis cliquez sur Se déconnecter.

Pour afficher les périphériques associés à une session, sélectionnez une session, puis cliquez sur Périphériques.

Ajouter une session

Se déconnecter

Périphériques...

Informations sur la session

| | |
|---|------------------|
| Balise du groupe de portails cible : | 1 |
| Statut : | Connecté |
| Nombre de connexions : | 1 |
| Nombre maximal de connexions autorisées : | 1 |
| Authentification : | Aucune spécifiée |
| Résumé d'en-tête : | Aucune spécifiée |
| Résumé des données : | Aucune spécifiée |

Configurer une session à connexions multiples (MCS)

Pour ajouter des connexions supplémentaires à une session ou pour configurer la stratégie MCS pour une session sélectionnée, cliquez sur MCS.

MCS...

OK Annuler

FIGURE 22.5 : PROPRIÉTÉS DE LA CIBLE ISCSI

7. Sélectionnez *Périphériques*, puis cliquez sur *MPIO* pour réviser la configuration entrées/sorties réparties sur plusieurs chemins :

The screenshot shows the 'MPIO' configuration window. At the top, there's a tab labeled 'MPIO'. Below it, a section titled 'Stratégie d'équilibrage de charge :' contains a dropdown menu set to 'Tourniquet avec sous-ensemble'. A 'Description' box explains that this strategy uses round-robin only on active paths, and in case of failure, it tries paths in a round-robin fashion. Below this, a section 'Ce périphérique comporte les chemins suivants :' contains a table with two rows of path information. At the bottom of the table is a scrollbar. Below the table are 'Détails' and 'Modifier...' buttons. At the very bottom are 'OK', 'Annuler', and 'Appliquer' buttons.

| ID de chemin | Statut | Type | Pondération | ID de session |
|--------------|---------|-------|-------------|-----------------------|
| 0x7703... | Conn... | Actif | s/o | ffffe00103669020-4000 |
| 0x7703... | Conn... | Actif | s/o | ffffe00103669020-4000 |

FIGURE 22.6 : DÉTAILS DU PÉRIPHÉRIQUE

La *Stratégie d'équilibrage de charge* par défaut s'appuie sur la méthode *Tourniquet avec sous-ensemble*. Si vous préférez une configuration de reprise après incident pure, choisissez *Fail Over Only* (Basculement seul).

Cela conclut la configuration de l'initiateur iSCSI. Les volumes iSCSI sont maintenant disponibles comme tous les autres périphériques SCSI et peuvent être initialisés en vue de leur utilisation en tant que volumes et lecteurs. Cliquez sur *OK* pour fermer la boîte de dialogue des *propriétés de l'initiateur iSCSI* et poursuivez avec le rôle *Services de fichiers et de stockage* à partir du tableau de bord *Gestionnaire de serveur*.

Observez le volume nouvellement connecté. Il s'identifie comme *SUSE RBD SCSI Multi-Path Drive* (Unité multipath SCSI RBD de SUSE) sur le bus iSCSI et il est marqué initialement avec l'état *Hors ligne* et une table de partitions de type *Inconnu*. Si le nouveau volume n'apparaît pas immédiatement, sélectionnez *Relancer l'analyse du stockage* dans la zone de liste déroulante *Tâches* pour relancer l'analyse du bus iSCSI.

1. Cliquez avec le bouton droit sur le volume iSCSI et sélectionnez *Nouveau volume* dans le menu contextuel. L'*Assistant Nouveau volume* apparaît. Cliquez sur *Suivant*, mettez en surbrillance le volume iSCSI nouvellement connecté et cliquez sur *Suivant* pour commencer.

Sélectionner le serveur et le disque

FIGURE 22.7 : ASSISTANT NOUVEAU VOLUME

2. Initialement, le périphérique est vide et ne contient pas de table de partitions. Lorsque vous y êtes invité, vérifiez la boîte de dialogue indiquant que le volume est initialisé avec une table de partitions GPT :

FIGURE 22.8 : INVITE DE DISQUE HORS LIGNE

3. Sélectionnez la taille du volume. En règle générale, vous pouvez utiliser la capacité totale du périphérique. Assignez ensuite une lettre d'unité ou un nom de répertoire dans lequel le volume nouvellement créé sera disponible. Sélectionnez un système de fichiers à créer sur le nouveau volume et, enfin, confirmez vos sélections en cliquant sur *Créer* pour achever la création du volume :

Confirmer les sélections

Avant de commencer
Serveur et disque
Taille
Lettre d'unité ou dossier
Paramètres du système de fichiers
Confirmation
Résultats

Vérifiez que les paramètres suivants sont corrects, puis cliquez sur Créer.

| | |
|--------------------------------------|-------------------|
| EMPLACEMENT DU VOLUME | |
| Serveur : | WIN-U3AILLIMUEE |
| Disque : | Disque 3 |
| Espace libre : | 48,8 Go |
| PROPRIÉTÉS DU VOLUME | |
| Taille du volume : | 48,8 Go |
| Lettre d'unité ou dossier : | D:\ |
| Étiquette du volume : | Nouveau volume |
| PARAMÈTRES DU SYSTÈME DE FICHIERS | |
| Système de fichiers : | NTFS |
| Création d'un nom de fichier court : | Désactivé |
| Taille d'unité d'allocation : | Valeur par défaut |

< Précédent Suivant > Créer Annuler

FIGURE 22.9 : CONFIRMATION DES SÉLECTIONS DE VOLUME

Une fois la procédure terminée, vérifiez les résultats, puis cliquez sur *Fermer* afin de conclure l'initialisation de l'unité. Une fois l'initialisation terminée, le volume (ainsi que son système de fichiers NTFS) devient disponible en tant qu'unité locale nouvellement initialisée.

22.1.3 Connexion de VMware

1. Pour vous connecter aux volumes iSCSI gérés par `ceph-iscsi`, vous devez disposer d'un adaptateur logiciel iSCSI configuré. Si aucun adaptateur n'est disponible dans votre configuration vSphere, créez-en un en sélectionnant *Configuration > Adaptateurs de stockage > Ajouter > Initiateur de logiciel iSCSI*.

-
2. Lorsqu'elles sont disponibles, sélectionnez les propriétés de l'adaptateur en cliquant avec le bouton droit sur l'adaptateur et en sélectionnant *Propriétés* depuis le menu contextuel :

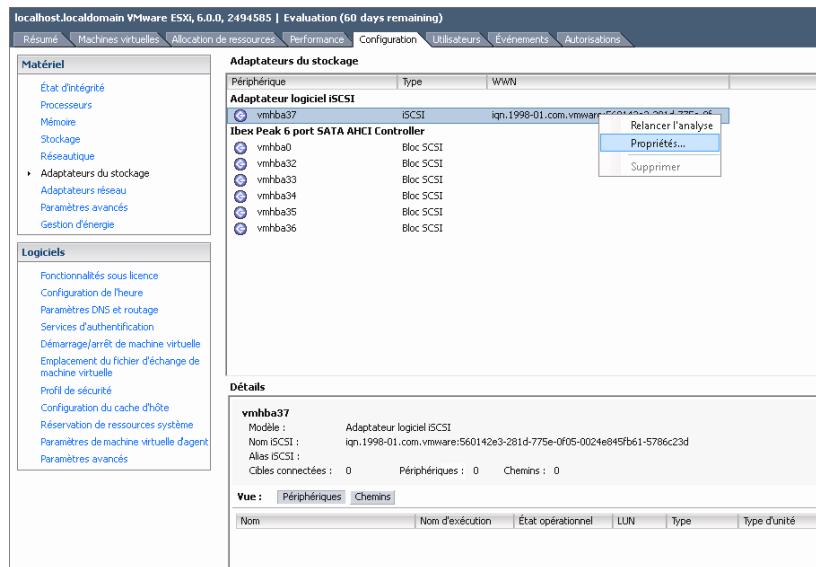


FIGURE 22.10 : PROPRIÉTÉS DE L'INITIATEUR ISCSI

-
-
3. Dans la boîte de dialogue *iSCSI Software Initiator* (Initiateur logiciel iSCSI), cliquez sur le bouton *Configurer*. Accédez ensuite à l'onglet *Découverte dynamique* et sélectionnez *Ajouter*.
4. Entrez l'adresse IP ou le nom d'hôte de votre passerelle iSCSI `ceph - i s c s i`. Si vous exécutez plusieurs passerelles iSCSI dans une configuration de basculement, répétez cette étape autant de fois que vous avez des passerelles dont vous êtes responsable.

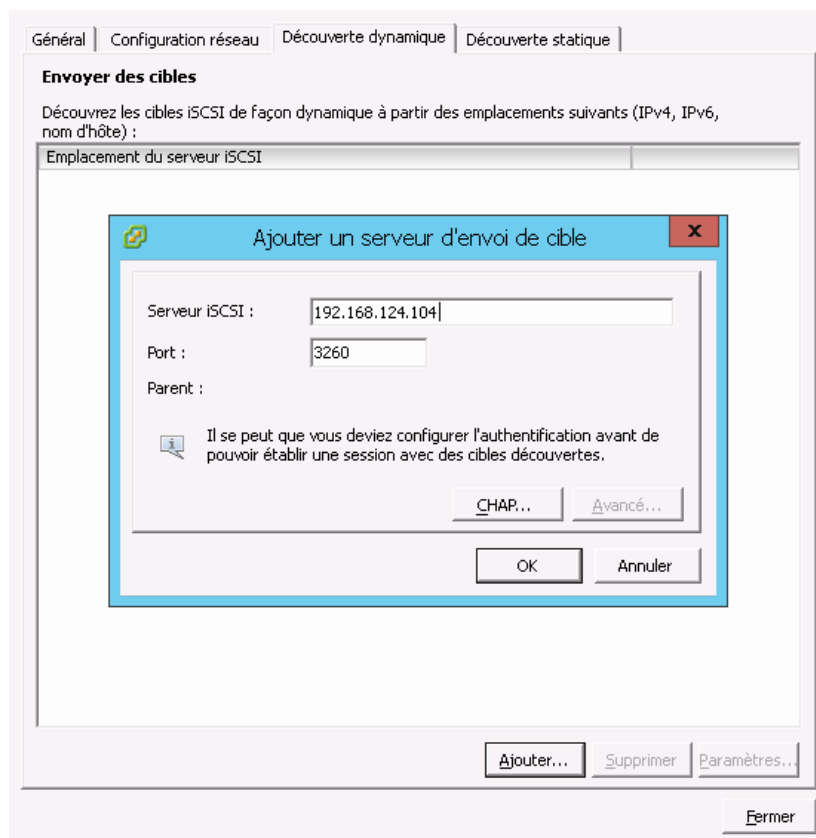


FIGURE 22.11 : AJOUT D'UN SERVEUR CIBLE

Lorsque vous avez entré toutes les passerelles iSCSI, cliquez sur *OK* dans la boîte de dialogue pour lancer une nouvelle analyse de l'adaptateur iSCSI.

5. Une fois cette analyse terminée, le nouveau périphérique iSCSI apparaît dans la liste *Adaptateurs du stockage* du volet *Détails*. Pour les périphériques multipath, vous pouvez à présent cliquer avec le bouton droit sur l'adaptateur et sélectionnez *Gérer les chemins* dans le menu contextuel :

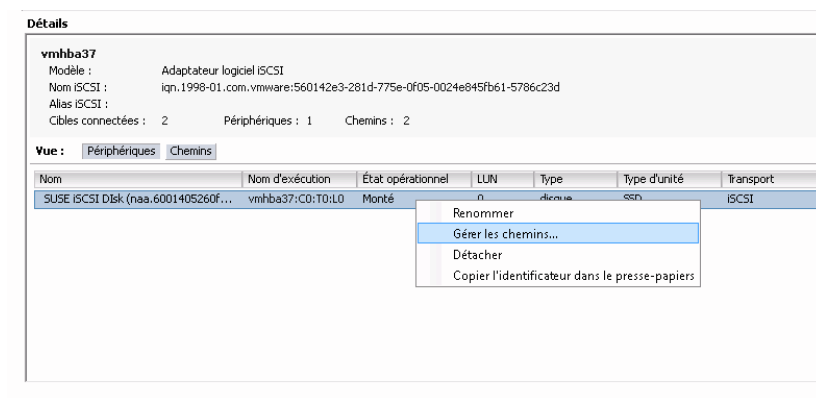


FIGURE 22.12 : GESTION DES PÉRIPHÉRIQUES MULTIPATH

Tous les chemins visibles doivent être signalés par une diode verte sous l'entête *État*. L'un de vos chemins d'accès doit être signalé par *Actif (E/S)* et tous les autres simplement par *Actif* :

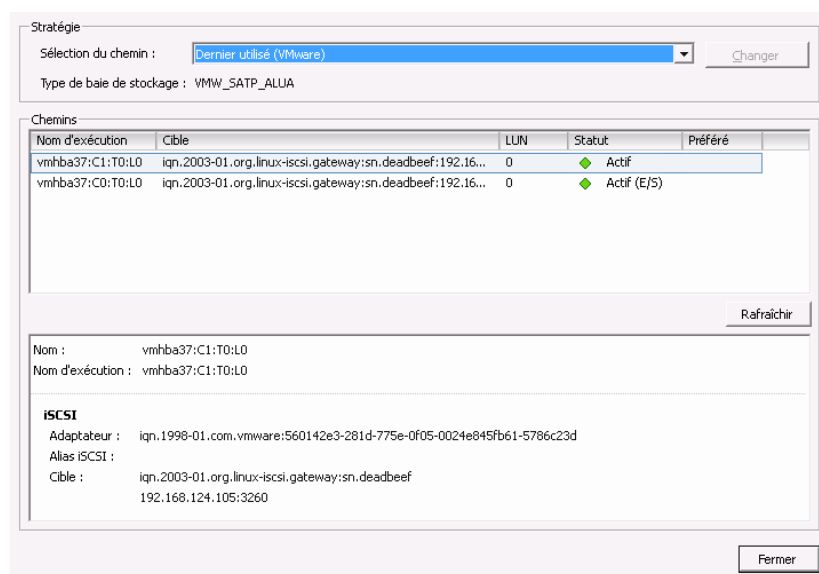


FIGURE 22.13 : LISTE DES CHEMINS POUR MULTIPATH

6. Vous pouvez maintenant passer d'*Adaptateurs du stockage* à l'élément étiqueté *Stockage*. Sélectionnez *Ajouter un stockage...* dans le coin supérieur droit du volet pour afficher la boîte de dialogue *Ajouter un stockage*. Sélectionnez ensuite *Disque/LUN*, puis cliquez sur *Suivant*. Le périphérique iSCSI nouvellement ajouté apparaît dans la liste *Sélectionner un disque/LUN*. Sélectionnez-le, puis cliquez sur *Suivant* pour continuer :

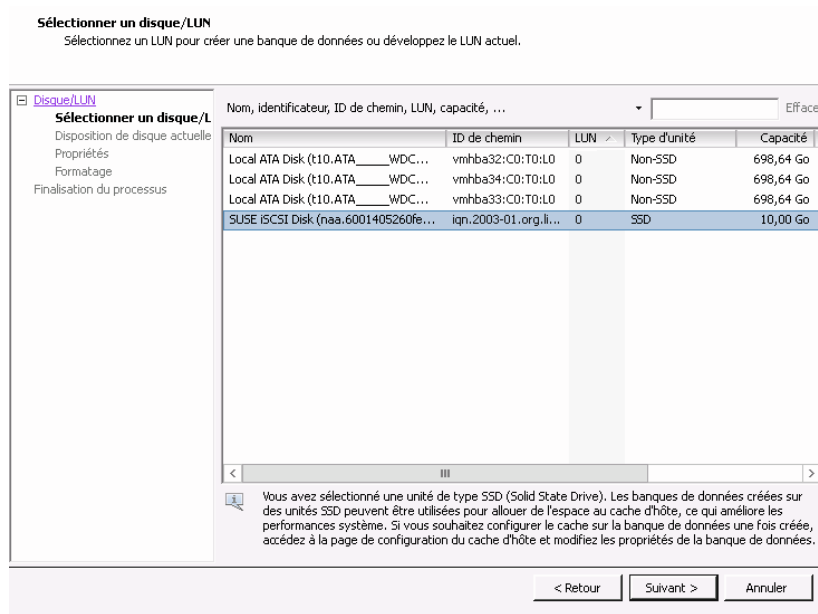


FIGURE 22.14 : BOÎTE DE DIALOGUE AJOUTER UN STOCKAGE

Cliquez sur *Suivant* pour accepter la disposition de disque par défaut.

7. Dans le volet *Propriétés*, assignez un nom à la nouvelle banque de données, puis cliquez sur *Suivant*. Acceptez le paramétrage par défaut pour utiliser l'intégralité de l'espace de volume pour la banque de données ou sélectionnez *Configuration de l'espace personnalisé* pour une banque de données plus petite :

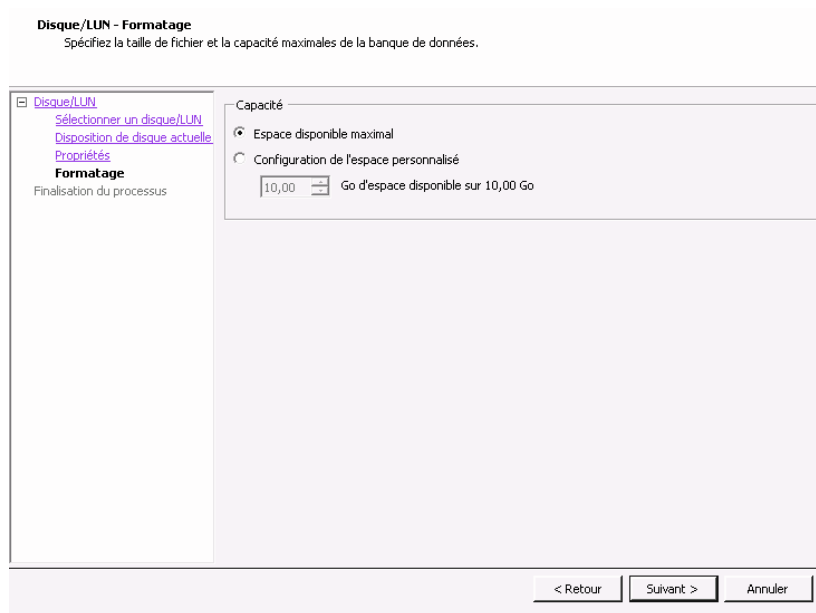


FIGURE 22.15 : CONFIGURATION DE L'ESPACE PERSONNALISÉ

Cliquez sur *Terminer* pour achever la création de la banque de données.

La nouvelle banque de données apparaît maintenant dans la liste de banques de données, et vous pouvez la sélectionner afin d'afficher les informations détaillées qui s'y rapportent. Vous pouvez maintenant utiliser le volume iSCSI soutenu par ceph-iscsi comme toute autre banque de données vSphere.

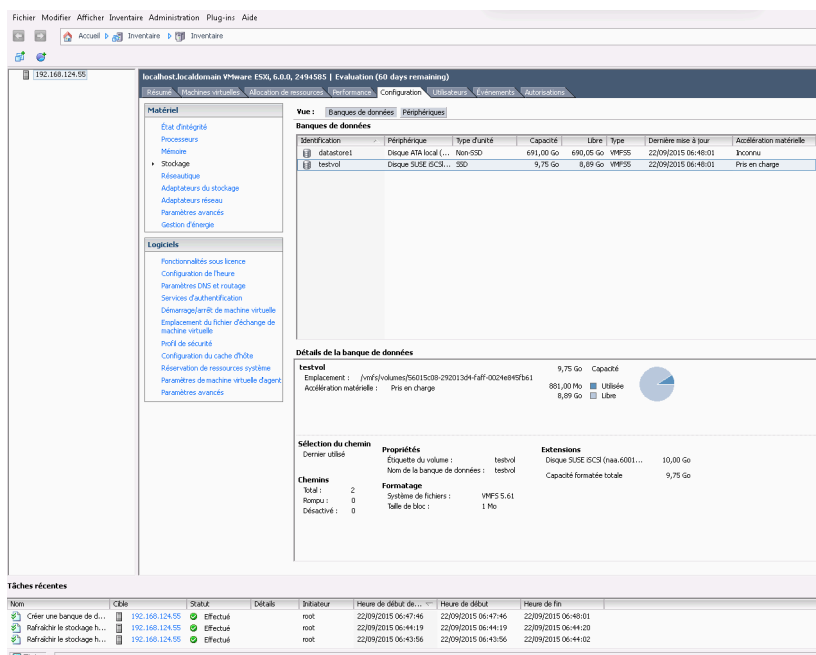


FIGURE 22.16 : PRÉSENTATION DE LA BANQUE DE DONNÉES ISCSI

22.2 Conclusion

`ceph-iscsi` est un composant clé de SUSE Enterprise Storage 7.1 qui permet d'accéder à un stockage de bloc distribué hautement disponible à partir de n'importe quel serveur ou client compatible avec le protocole iSCSI. En utilisant `ceph-iscsi` sur un ou plusieurs hôtes de passerelle iSCSI, les images Ceph RBD deviennent disponibles sous la forme d'unités logiques (LU) associées à des cibles iSCSI, dont l'accès peut être hautement disponible et régi par l'équilibrage de charge.

Puisque la configuration de `ceph-iscsi` est entièrement stockée dans la zone de stockage des objets Ceph RADOS, les hôtes de passerelle `ceph-iscsi` ne possèdent pas intrinsèquement d'état persistant, ce qui permet de les remplacer, les augmenter ou les réduire à volonté. En conséquence, SUSE Enterprise Storage 7.1 permet aux clients SUSE de déployer une technologie de stockage d'entreprise véritablement distribuée, hautement disponible, résiliente et à autodépannage sur du matériel de base et une plate-forme entièrement Open Source.

23 Système de fichiers en grappe

Ce chapitre décrit les tâches d'administration normalement effectuées après la configuration de la grappe et l'exportation de CephFS. Pour plus d'informations sur la configuration de CephFS, reportez-vous au *Manuel « Guide de déploiement », Chapitre 8 « Déploiement des services essentiels restants à l'aide de cephadm », Section 8.3.3 « Déploiement de serveurs de métadonnées ».*

23.1 Montage de CephFS

Lorsque le système de fichiers est créé et que MDS est actif, vous êtes prêt à monter le système de fichiers à partir d'un hôte client.

23.1.1 Préparation du client

Si l'hôte client exécute SUSE Linux Enterprise 12 SP2 ou une version ultérieure, le système est prêt à monter CephFS dans sa version « prête à l'emploi ».

Si l'hôte client exécute SUSE Linux Enterprise 12 SP1, vous devez appliquer tous les correctifs les plus récents avant de monter CephFS.

Dans tous les cas, SUSE Linux Enterprise inclut tout ce qui est nécessaire au montage de CephFS. Le produit SUSE Enterprise Storage 7.1 n'est pas nécessaire.

Pour prendre en charge la syntaxe de **mount** complète, le paquetage `ceph-common` (fourni avec SUSE Linux Enterprise) doit être installé avant d'essayer de monter CephFS.



Important

En l'absence du paquetage `ceph-common` (et donc sans le programme auxiliaire **mount.ceph**), les adresses IP des moniteurs doivent être utilisées plutôt que leurs noms. Cela est dû au fait que le client de kernel ne pourra pas effectuer la résolution de nom.

La syntaxe de montage de base est la suivante :

```
# mount -t ceph MON1_IP[:PORT],MON2_IP[:PORT],...:CEPHFS_MOUNT_TARGET \
MOUNT_POINT -o name=CEPHX_USER_NAME,secret=SECRET_STRING
```


23.1.2 Création d'un fichier de secret

L'authentification est activée par défaut pour la grappe Ceph active. Vous devez créer un fichier qui stocke votre clé secrète (et non pas le trousseau de clés lui-même). Pour obtenir la clé secrète d'un utilisateur particulier et créer ensuite le fichier, procédez comme suit :

PROCÉDURE 23.1 : CRÉATION D'UNE CLÉ SECRÈTE

1. Affichez la clé d'un utilisateur particulier dans un fichier de trousseau de clés :

```
cephuser@adm > cat /etc/ceph/ceph.client.admin.keyring
```

2. Copiez la clé de l'utilisateur qui emploiera le système de fichiers Ceph FS monté. La clé ressemble généralement à ceci :

```
AQCj2YpRiAe6CxAA7/ETt7Hcl9IyxyYciVs47w==
```

3. Créez un fichier en indiquant le nom de l'utilisateur dans le nom de fichier, par exemple `/etc/ceph/admin.secret` pour l'utilisateur *admin*.
4. Collez la valeur de la clé dans le fichier créé à l'étape précédente.
5. Définissez les droits d'accès appropriés au fichier. L'utilisateur doit être le seul à pouvoir lire le fichier, les autres n'ont aucun droit d'accès.

23.1.3 Montage de CephFS

La commande **mount** permet de monter CephFS. Vous devez indiquer le nom d'hôte ou l'adresse IP du moniteur. Comme l'authentification `cephx` est activée par défaut dans SUSE Enterprise Storage, vous devez également spécifier un nom d'utilisateur et le secret qui lui est associé :

```
# mount -t ceph ceph_mon1:6789:/ /mnt/cephfs \  
-o name=admin,secret=AQATSKdNGBnwLhAAAnNDKnH65FmVKpXZJVASueQ==
```

Étant donné que la commande précédente reste dans l'historique du shell, une approche plus sécurisée consiste à lire le secret d'un fichier :

```
# mount -t ceph ceph_mon1:6789:/ /mnt/cephfs \  
-o name=admin,secretfile=/etc/ceph/admin.secret
```

Notez que le fichier de secret ne doit contenir que le secret du trousseau de clés. Dans notre exemple, le fichier contient uniquement la ligne suivante :

```
AQATSKdNGBnwLhAAAnNDKnH65FmVKpXZJVASueQ==
```



Astuce : spécification de plusieurs moniteurs

Il est judicieux d'indiquer plusieurs moniteurs séparés par des virgules sur la ligne de commande **mount** dans le cas où un moniteur est arrêté au moment du montage. Chaque adresse de moniteur figure sous la forme hôte[:port]. Si le port n'est pas indiqué, le port par défaut est le port 6789.

Créez le point de montage sur l'hôte local :

```
# mkdir /mnt/cephfs
```

Montez le système de fichiers CephFS :

```
# mount -t ceph ceph_mon1:6789:/ /mnt/cephfs \
-o name=admin,secretfile=/etc/ceph/admin.secret
```

Un sous-répertoire subdir peut être indiqué si un sous-ensemble du système de fichiers doit être monté :

```
# mount -t ceph ceph_mon1:6789:/subdir /mnt/cephfs \
-o name=admin,secretfile=/etc/ceph/admin.secret
```

Vous pouvez spécifier plusieurs hôtes de moniteur dans la commande **mount** :

```
# mount -t ceph ceph_mon1,ceph_mon2,ceph_mon3:6789:/ /mnt/cephfs \
-o name=admin,secretfile=/etc/ceph/admin.secret
```



Important : accès en lecture au répertoire racine

Si des clients avec restriction de chemin d'accès sont utilisés, les fonctionnalités MDS doivent inclure un accès en lecture au répertoire racine. Par exemple, un trousseau de clés peut ressembler à ceci :

```
client.bar
key: supersecretkey
caps: [mds] allow rw path=/barjail, allow r path=/
caps: [mon] allow r
caps: [osd] allow rwx
```

La partie allow r path=/ signifie que les clients dont le chemin est restreint peuvent voir le volume racine sans être autorisés à y écrire des données. Cela peut être un problème dans les cas où une isolation complète est requise.

23.2 Démontage de CephFS

Pour démonter le système de fichiers CephFS, utilisez la commande `umount` :

```
# umount /mnt/cephfs
```

23.3 Montage de CephFS dans `/etc/fstab`

Pour monter automatiquement CephFS au démarrage du client, insérez la ligne correspondante dans sa table des systèmes de fichiers `/etc/fstab` :

```
mon1:6790,mon2:/subdir /mnt/cephfs ceph name=admin,secretfile=/etc/ceph/  
secret.key,noatime,_netdev 0 2
```

23.4 Daemons MDS actifs multiples (MDS actif-actif)

Par défaut, CephFS est configuré pour un seul daemon MDS actif. Pour mettre à l'échelle les performances des métadonnées pour les systèmes vastes, vous pouvez activer plusieurs daemons MDS actifs, qui partageront entre eux la charge de travail des métadonnées.

23.4.1 Utilisation de MDS actif-actif

Vous pouvez envisager d'utiliser plusieurs daemons MDS actifs lorsque les performances de vos métadonnées sont bloquées sur le MDS unique par défaut.

L'ajout de plusieurs daemons n'améliore pas les performances pour tous les types de charge de travail. Par exemple, une seule application s'exécutant sur un seul client ne bénéficie pas d'un nombre accru de daemons MDS à moins d'effectuer de nombreuses opérations de métadonnées en parallèle.

Les charges de travail qui bénéficient généralement d'un nombre supérieur de daemons MDS actifs sont celles qui possèdent de nombreux clients, travaillant, le cas échéant, sur plusieurs répertoires distincts.

23.4.2 Augmentation de la taille de la grappe active MDS

Chaque système de fichiers CephFS possède un paramètre `max_mds` qui détermine le nombre de rangs à créer. Le nombre réel de rangs dans le système de fichiers n'augmentera que si un daemon supplémentaire est en mesure de prendre le nouveau rang créé. Par exemple, si un seul daemon MDS s'exécute et que la valeur `max_mds` est définie sur 2, aucun second rang n'est créé. Dans l'exemple suivant, nous définissons `max_mds` sur 2 pour créer un rang en dehors du rang par défaut. Pour voir les changements, lancez **ceph status** avant et après avoir défini `max_mds` et reportez-vous à la ligne contenant `fsmap` :

```
cephuser@adm > ceph status
[...]
services:
  [...]
  mds: cephfs-1/1/1 up {0=node2=up:active}, 1 up:standby
  [...]
cephuser@adm > ceph fs set cephfs max_mds 2
cephuser@adm > ceph status
[...]
services:
  [...]
  mds: cephfs-2/2/2 up {0=node2=up:active,1=node1=up:active}
  [...]
```

Le rang nouvellement créé (1) passe à l'état « creating » (création en cours), puis à l'état « active » (actif).



Important : daemons de secours

Même avec plusieurs daemons MDS actifs, un système hautement disponible nécessite toujours des daemons de secours qui prennent le relais si l'un des serveurs exécutant un daemon actif tombe en panne.

Par conséquent, la valeur maximale pratique de `max_mds` pour les systèmes hautement disponibles est égale au nombre total de serveurs MDS de votre système moins 1. Pour rester disponible en cas de défaillance de plusieurs serveurs, augmentez le nombre de daemons de secours du système pour qu'ils correspondent au nombre de pannes de serveur que vous devez pouvoir surmonter.

23.4.3 Diminution du nombre de rangs

Tous les rangs, y compris les rangs à supprimer, doivent d'abord être actifs. Cela signifie que vous devez disposer d'au moins `max_mds` daemons MDS disponibles.

Commencez par définir `max_mds` sur un nombre inférieur. Par exemple, définissez un seul MDS actif :

```
cephuser@adm > ceph status
[...]
services:
  [...]
  mds: cephfs-2/2/2 up {0=node2=up:active,1=node1=up:active}
  [...]
cephuser@adm > ceph fs set cephfs max_mds 1
cephuser@adm > ceph status
[...]
services:
  [...]
  mds: cephfs-1/1/1 up {0=node2=up:active}, 1 up:standby
  [...]
```

23.4.4 Épinglage manuel d'arborescences de répertoires à un rang

Dans plusieurs configurations de serveur de métadonnées actives, l'équilibreur permet de répartir la charge de métadonnées uniformément dans la grappe. Cela fonctionne généralement assez bien pour la plupart des utilisateurs, mais il est parfois souhaitable de remplacer l'équilibreur dynamique par des assignations explicites de métadonnées à des rangs particuliers. L'administrateur ou les utilisateurs peuvent ainsi répartir uniformément la charge d'applications ou limiter l'impact des demandes de métadonnées des utilisateurs sur l'ensemble de la grappe.

Le mécanisme prévu à cet effet s'appelle une « épingle d'exportation ». Il s'agit d'un attribut étendu des répertoires. Cet attribut étendu s'appelle `ceph.dir.pin`. Les utilisateurs peuvent définir cet attribut à l'aide de commandes standard :

```
# setfattr -n ceph.dir.pin -v 2 /path/to/dir
```

La valeur (`-v`) de l'attribut étendu correspond au rang à attribuer à la sous-arborescence de répertoires. La valeur par défaut (`-1`) indique que le répertoire n'est pas épinglé.

Une épingle d'exportation de répertoire est héritée de son parent le plus proche ayant une épingle d'exportation définie. Par conséquent, la définition de l'épingle d'exportation sur un répertoire affecte tous ses enfants. Cependant, il est possible d'annuler l'épingle du parent en définissant l'épingle d'exportation du répertoire enfant. Par exemple :

```
# mkdir -p a/b # "a" and "a/b" start with no export pin set.
setfattr -n ceph.dir.pin -v 1 a/ # "a" and "b" are now pinned to rank 1.
setfattr -n ceph.dir.pin -v 0 a/b # "a/b" is now pinned to rank 0
                                # and "a/" and the rest of its children
                                # are still pinned to rank 1.
```

23.5 Gestion du basculement

Si un daemon MDS cesse de communiquer avec le moniteur, le moniteur attend `mds_beacon_grace` secondes (par défaut, 15 secondes) avant de marquer le daemon comme *laggy* (lent à réagir). Vous pouvez configurer un ou plusieurs daemons de secours qui prendront le relais lors du basculement du daemon MDS.

23.5.1 Configuration de daemons de secours avec relecture

Chaque système de fichiers CephFS peut être configuré pour ajouter des daemons de secours avec relecture. Ces daemons de secours suivent le journal de métadonnées du MDS actif pour réduire le temps de basculement en cas d'indisponibilité du MDS actif. Chaque MDS actif ne peut avoir qu'un seul daemon de secours avec relecture.

Configurez le daemon de secours avec relecture sur un système de fichiers à l'aide de la commande suivante :

```
cephuser@adm > ceph fs set FS-NAME allow_standby_replay BOOL
```

Une fois les daemons de secours avec relecture définis, les moniteurs les assigneront pour suivre les MDS actifs dans ce système de fichiers.

Lorsqu'un MDS passe à l'état `standbyReplay`, il ne peut être utilisé que comme daemon en veille pour le rang auquel il est associé. En cas d'échec d'un autre rang, ce daemon de secours avec relecture ne peut pas jouer le rôle de remplaçant même si tous les autres daemons de secours sont indisponibles. Pour cette raison, il est conseillé d'utiliser un daemon de secours avec relecture pour chaque MDS actif.

23.6 Définition des quotas CephFS

Vous pouvez définir des quotas sur n'importe quel sous-répertoire du système de fichiers Ceph. Le quota limite le nombre d'**octets** ou de **fichiers** stockés sous le point spécifié dans la hiérarchie de répertoires.

23.6.1 Limites des quotas CephFS

L'utilisation de quotas avec CephFS présente les limites suivantes :

Les quotas sont coopératifs et non concurrentiels.

Les quotas Ceph s'appuient sur le client qui monte le système de fichiers pour arrêter d'écrire sur ce dernier lorsqu'une limite est atteinte. La partie serveur ne peut pas empêcher un client malveillant d'écrire autant de données qu'il en a besoin. N'utilisez pas de quotas pour empêcher la saturation du système de fichiers dans des environnements où les clients ne sont pas pleinement approuvés.

Les quotas sont imprécis.

Les processus qui écrivent sur le système de fichiers seront arrêtés peu de temps après que la limite de quota aura été atteinte. Ils seront inévitablement autorisés à écrire une certaine quantité de données au-delà de la limite configurée. Les systèmes d'écriture clients seront arrêtés quelques dixièmes de seconde après avoir franchi la limite configurée.

Les quotas sont implémentés dans le client du kernel à partir de la version 4.17.

Les quotas sont pris en charge par le client de l'espace utilisateur (libcephfs, ceph-fuse). Les clients de kernel Linux des versions 4.17 et ultérieures prennent en charge les quotas CephFS sur les grappes SUSE Enterprise Storage 7.1. Les clients de kernel (y compris les versions récentes) ne sont pas en mesure de gérer les quotas sur des grappes plus anciennes, même s'ils parviennent à définir les attributs étendus des quotas. Les kernels SLE12-SP3 (et versions ultérieures) incluent déjà les rétroports requis pour gérer les quotas.

Configurez les quotas avec prudence lorsqu'ils sont utilisés avec des restrictions de montage basées sur le chemin.

Le client doit avoir accès à l'inode de répertoire sur lequel les quotas sont configurés afin de les appliquer. Si le client dispose d'un accès restreint à un chemin spécifique (par exemple /home/user) sur la base de la fonction MDS et qu'un quota est configuré sur un répertoire ancêtre auquel il n'a pas accès (/home), le client n'appliquera pas ce quota. Lorsque

vous utilisez des restrictions d'accès basées sur les chemins, veuillez à configurer le quota sur le répertoire auquel le client peut accéder (par exemple /home/user ou /home/user/quota_dir).

23.6.2 Configuration des quotas CephFS

Vous pouvez configurer les quotas CephFS à l'aide d'attributs étendus virtuels :

ceph.quota.max_files

Configure une limite de *fichiers*.

ceph.quota.max_bytes

Configure une limite d'*octets*.

Si les attributs apparaissent sur un inode de répertoire, un quota y est configuré. S'ils ne sont pas présents, aucun quota n'est défini sur ce répertoire (mais il est encore possible d'en configurer un sur un répertoire parent).

Pour définir un quota de 100 Mo, exécutez :

```
cephuser@mds > setfattr -n ceph.quota.max_bytes -v 100000000 /SOME/DIRECTORY
```

Pour définir un quota de 10 000 fichiers, exécutez :

```
cephuser@mds > setfattr -n ceph.quota.max_files -v 10000 /SOME/DIRECTORY
```

Pour afficher la configuration de quota, exécutez :

```
cephuser@mds > getfattr -n ceph.quota.max_bytes /SOME/DIRECTORY
```

```
cephuser@mds > getfattr -n ceph.quota.max_files /SOME/DIRECTORY
```



Note : quota non défini

Si la valeur de l'attribut étendu est « 0 », le quota n'est pas défini.

Pour supprimer un quota, exécutez :

```
cephuser@mds > setfattr -n ceph.quota.max_bytes -v 0 /SOME/DIRECTORY
cephuser@mds > setfattr -n ceph.quota.max_files -v 0 /SOME/DIRECTORY
```


23.7 Gestion des instantanés CephFS

Les instantanés CephFS créent une vue en lecture seule du système de fichiers au moment où ils sont réalisés. Vous pouvez créer un instantané dans n'importe quel répertoire. L'instantané couvrira toutes les données du système de fichiers sous le répertoire spécifié. Après avoir créé un instantané, les données mises en mémoire tampon sont vidées de façon asynchrone à partir de divers clients. La création d'un instantané est dès lors très rapide.



Important : systèmes de fichiers multiples

Si vous avez plusieurs systèmes de fichiers CephFS partageant une réserve unique (via des espaces de noms), leurs instantanés entreraient en collision, et la suppression d'un instantané entraînerait des données de fichier manquantes pour d'autres instantanés partageant la même réserve.

23.7.1 Création d'instantanés

La fonction d'instantané CephFS est activée par défaut sur les nouveaux systèmes de fichiers. Pour l'activer sur les systèmes de fichiers existants, exécutez la commande suivante :

```
cephuser@adm > ceph fs set CEPHFS_NAME allow_new_snaps true
```

Une fois que vous activez des instantanés, tous les répertoires de CephFS auront un sous-répertoire spécial `.snap`.



Note

Il s'agit d'un sous-répertoire *virtuel*. Il n'apparaît pas dans la liste des répertoires du répertoire parent, mais le nom `.snap` ne peut pas être utilisé comme nom de fichier ou de répertoire. Pour accéder au répertoire `.snap`, vous devez y accéder explicitement, par exemple :

```
> ls -la /CEPHFS_MOUNT/.snap/
```



Important : limite des clients de kernel

Les clients de kernel CephFS ont une limite : ils ne peuvent pas gérer plus de 400 instantanés dans un système de fichiers. Le nombre d'instantanés doit toujours être maintenu en dessous de cette limite, quel que soit le client que vous utilisez. Si vous utilisez des clients CephFS plus anciens, tels que SLE12-SP3, gardez à l'esprit que dépasser la limite de 400 instantanés est préjudiciable pour les opérations, car le client va se bloquer.



Astuce : nom personnalisé pour le sous-répertoire d'instantanés

Vous pouvez configurer un nom différent pour le sous-répertoire d'instantanés en définissant le paramètre `client snapdir`.

Pour créer un instantané, créez un sous-répertoire sous le répertoire `.snap` avec un nom personnalisé. Par exemple, pour créer un instantané du répertoire `/CEPHFS_MOUNT/2/3/`, exécutez la commande suivante :

```
> mkdir /CEPHFS_MOUNT/2/3/.snap/CUSTOM_SNAPSHOT_NAME
```

23.7.2 Suppression d'instantanés

Pour supprimer un instantané, supprimez son sous-répertoire au sein du répertoire `.snap` :

```
> rmdir /CEPHFS_MOUNT/2/3/.snap/CUSTOM_SNAPSHOT_NAME
```

24 Exportation des données Ceph via Samba

Ce chapitre décrit comment exporter des données stockées dans une grappe Ceph via un partage Samba/CIFS afin que vous puissiez facilement y accéder à partir des machines clientes Windows*. Il comprend également des informations qui vous aideront à configurer une passerelle Ceph Samba pour joindre Active Directory dans le domaine Windows* afin d'authentifier et d'autoriser les utilisateurs.



Note : performances de la passerelle Samba

En raison du overhead accru du protocole et de la latence supplémentaire causée par des sauts de réseau additionnels entre le client et le stockage, l'accès à CephFS via la passerelle Samba peut réduire considérablement les performances de l'application par rapport aux clients Ceph natifs.

24.1 Exportation de CephFS via un partage Samba



Avertissement : accès interprotocole

Les clients CephFS et NFS natifs ne sont pas limités par les verrouillages de fichiers obtenus via Samba, et inversement. Les applications qui s'appuient sur le verrouillage de fichiers interprotocole peuvent présenter des altérations des données si les chemins de partage Samba soutenus par CephFS sont accessibles par d'autres moyens.

24.1.1 Configuration et exportation de paquetages Samba

Pour configurer et exporter un partage Samba, les paquetages suivants doivent être installés : samba-ceph et samba-winbind. Si ces paquetages ne sont pas installés, installez-les :

```
cephuser@smb > zypper install samba-ceph samba-winbind
```

24.1.2 Exemple de passerelle unique

Pour préparer l'exportation d'un partage Samba, choisissez un noeud approprié afin de faire office de passerelle Samba. Le noeud doit avoir accès au réseau client Ceph et disposer de ressources suffisantes en termes de processeur, de mémoire et de réseau.

La fonctionnalité de basculement peut être fournie par CTDB et l'extension SUSE Linux Enterprise High Availability Extension. Reportez-vous à la [Section 24.1.3, « Configuration de la haute disponibilité »](#) pour plus d'informations sur la configuration HA.

1. Assurez-vous qu'un CephFS actif existe déjà dans votre grappe.
2. Créez un trousseau de clés spécifique à la passerelle Samba sur le noeud Admin de Ceph et copiez-le sur les deux noeuds de la passerelle Samba :

```
cephuser@adm > ceph auth get-or-create client.samba.gw mon 'allow r' \
  osd 'allow *' mds 'allow *' -o ceph.client.samba.gw.keyring
cephuser@adm > scp ceph.client.samba.gw.keyring SAMBA_NODE:/etc/ceph/
```

Remplacez *SAMBA_NODE* par le nom du noeud de la passerelle Samba.

3. Les étapes suivantes sont exécutées sur le noeud de la passerelle Samba. Installez Samba avec le paquetage d'intégration de Ceph :

```
cephuser@smb > sudo zypper in samba samba-ceph
```

4. Remplacez le contenu par défaut du fichier `/etc/samba/smb.conf` par ce qui suit :

```
[global]
  netbios name = SAMBA-GW
  clustering = no
  idmap config * : backend = tdb2
  passdb backend = tdbsam
  # disable print server
  load printers = no
  smbda: backgroundqueue = no

[SHARE_NAME]
  path = CEPHFS_MOUNT
  read only = no
  oplocks = no
  kernel share modes = no
```

Le chemin `CEPHFS_MOUNT` ci-dessus doit être monté avant de démarrer Samba avec une configuration de partage CephFS du kernel. Reportez-vous à la [Section 23.3, « Montage de CephFS dans /etc/fstab »](#).

La configuration de partage ci-dessus utilise le client CephFS du kernel Linux, recommandé pour des raisons de performances. Comme alternative, le module `vfs_ceph` de Samba peut également être utilisé pour communiquer avec la grappe Ceph. Les instructions sont indiquées ci-dessous pour des installation existantes, mais ne sont pas recommandées pour les nouveaux déploiements de Samba :

```
[SHARE_NAME]
path = /
vfs objects = ceph
ceph: config_file = /etc/ceph/ceph.conf
ceph: user_id = samba.gw
read only = no
oplocks = no
kernel share modes = no
```



Astuce : modes oplocks et share

Les verrous optionnels (`oplocks` - également connus sous le terme de baux SMB2+) permettent d'améliorer les performances grâce à un caching client agressif, mais à l'heure actuelle, ils sont risqués lorsque Samba est déployé avec d'autres clients CephFS, tels que le kernel `mount.ceph`, FUSE ou NFS Ganesha.

Si tous les accès au chemin du système de fichiers CephFS sont exclusivement gérés par Samba, le paramètre `oplocks` peut être activé en toute sécurité.

Actuellement, pour que les fichiers soient desservis correctement, l'option `kernel share modes` doit être désactivée sur un partage en cours d'exécution avec le module CephFS `vfs`.



Important : autorisation d'accès

Samba assigne les utilisateurs et les groupes SMB à des comptes locaux. Les utilisateurs locaux peuvent se voir assigner un mot de passe pour l'accès au partage Samba via :

```
# smbpasswd -a USERNAME
```

Pour que les entrées/sorties se déroulent correctement, la liste de contrôle d'accès (ACL) du chemin du partage doit autoriser l'accès à l'utilisateur connecté via Samba. Vous pouvez modifier l'ACL en effectuant un montage temporaire via le client de kernel CephFS et en employant les utilitaires **chmod**, **chown** ou **setfacl** par rapport au chemin du partage. Par exemple, pour permettre l'accès à tous les utilisateurs, exécutez la commande suivante :

```
# chmod 777 MOUNTED_SHARE_PATH
```

24.1.2.1 Démarrage des services Samba

Pour démarrer ou redémarrer les services Samba autonomes, utilisez les commandes suivantes :

```
# systemctl restart smb.service
# systemctl restart nmb.service
# systemctl restart winbind.service
```

Pour vous assurer que les services Samba se lancent au démarrage, activez-les via :

```
# systemctl enable smb.service
# systemctl enable nmb.service
# systemctl enable winbind.service
```



Astuce : services nmb et winbind facultatifs

Si vous n'avez pas besoin de parcourir un partage réseau, il n'est pas nécessaire d'activer et de démarrer le service nmb.

Le service winbind est uniquement requis en cas de configuration en tant que membre du domaine Active Directory. Reportez-vous à la [Section 24.2, « Jointure de la passerelle Samba et d'Active Directory »](#).

24.1.3 Configuration de la haute disponibilité



Important : basculement transparent non pris en charge

Bien qu'un déploiement Samba + CTDB à plusieurs noeuds soit plus disponible que le noeud unique (voir [Chapitre 24, Exportation des données Ceph via Samba](#)), le basculement transparent côté client n'est pas pris en charge. Les applications connaîtront probablement une brève interruption de service en cas de défaillance d'un noeud de passerelle Samba.

Cette section fournit un exemple de configuration à haute disponibilité à deux noeuds des serveurs Samba. La configuration requiert l'extension SUSE Linux Enterprise High Availability Extension. Les deux noeuds sont appelés [earth192.168.1.1](#) et [mars \(192.168.1.2\)](#).

Pour plus d'informations sur l'extension SUSE Linux Enterprise High Availability Extension, reportez-vous à la page <https://documentation.suse.com/sle-ha/15-SP1/>.

Par ailleurs, deux adresses IP virtuelles flottantes permettent aux clients de se connecter au service quel que soit le noeud physique sur lequel il s'exécute. L'adresse IP [192.168.1.10](#) est utilisée pour l'administration des grappes avec Hawk2 et [192.168.2.1](#) exclusivement pour les exportations CIFS. Cela facilite l'application des restrictions de sécurité ultérieurement.

La procédure suivante décrit l'exemple d'installation. Pour plus d'informations, reportez-vous à la page <https://documentation.suse.com/sle-ha/15-SP1/single-html/SLE-HA-install-quick/>.

1. Créez un trousseau de clés spécifique à la passerelle Samba sur le noeud Admin et copiez-le sur les deux noeuds :

```
cephuser@adm > ceph auth get-or-create client.samba.gw mon 'allow r' \
    osd 'allow *' mds 'allow *' -o ceph.client.samba.gw.keyring
cephuser@adm > scp ceph.client.samba.gw.keyring earth:/etc/ceph/
cephuser@adm > scp ceph.client.samba.gw.keyring mars:/etc/ceph/
```

2. La configuration de SLE-HA nécessite un périphérique d'isolation pour éviter une situation de *split brain* (cerveau divisé) lorsque les noeuds de grappe actifs ne sont plus synchronisés. Pour cela, vous pouvez utiliser une image Ceph RBD avec un périphérique de traitement par blocs Stonith (SBD). Pour plus d'informations, reportez-vous à la page <https://documentation.suse.com/sle-ha/15-SP1/single-html/SLE-HA-guide/#sec-ha-storage-protect-fencing-setup>.

Si elle n'existe pas encore, créez une réserve RBD appelée rbd (voir [Section 18.1, « Création d'une réserve »](#)) et associez-la à l'application rbd (voir [Section 18.5.1, « Association de réserves à une application »](#)). Créez ensuite une image RBD associée appelée sbd01 :

```
cephuser@adm > ceph osd pool create rbd
cephuser@adm > ceph osd pool application enable rbd rbd
cephuser@adm > rbd -p rbd create sbd01 --size 64M --image-shared
```

3. Préparez earth et mars pour héberger le service Samba :

- a. Assurez-vous que les paquetages suivants sont installés avant de poursuivre : ctdb, tdb-tools et samba.

```
# zypper in ctdb tdb-tools samba samba-ceph
```

- b. Assurez-vous que les services Samba et CTDB sont arrêtés et désactivés :

```
# systemctl disable ctdb
# systemctl disable smb
# systemctl disable nmb
# systemctl disable winbind
# systemctl stop ctdb
# systemctl stop smb
# systemctl stop nmb
# systemctl stop winbind
```

- c. Ouvrez le port 4379 de votre pare-feu sur tous les noeuds. Cette ouverture est nécessaire pour que CTDB communique avec d'autres noeuds de la grappe.

4. Sur earth, créez les fichiers de configuration de Samba. Ils seront ensuite automatiquement synchronisés avec mars.

- a. Insérez une liste d'adresses IP privées des noeuds de la passerelle Samba dans le fichier /etc/ctdb/nodes. Pour plus d'informations, reportez-vous à la page de manuel ctdb (**man 7 ctdb**).

```
192.168.1.1
192.168.1.2
```


- b. Configurez Samba. Ajoutez les lignes suivantes à la section `[global]` de `/etc/samba/smb.conf`. Utilisez le nom d'hôte de votre choix à la place de `CTDB-SERVER` (tous les noeuds de la grappe apparaîtront sous la forme d'un gros noeud portant ce nom). Ajoutez également une définition de partage. Prenons `SHARE_NAME` (NOM_PARTAGE) comme exemple :

```
[global]
    netbios name = SAMBA-HA-GW
    clustering = yes
    idmap config * : backend = tdb2
    passdb backend = tdbsam
    ctdbd socket = /var/lib/ctdb/ctdb.socket
    # disable print server
    load printers = no
    smbd: backgroundqueue = no

[SHARE_NAME]
    path = /
    vfs objects = ceph
    ceph: config_file = /etc/ceph/ceph.conf
    ceph: user_id = samba.gw
    read only = no
    oplocks = no
    kernel share modes = no
```

Notez que les fichiers `/etc/ctdb/nodes` et `/etc/samba/smb.conf` doivent correspondre sur tous les noeuds de passerelle Samba.

5. Installez et amorcez la grappe SUSE Linux Enterprise High Availability.

- a. Enregistrez l'extension SUSE Linux Enterprise High Availability sur `earth` et `mars` :

```
root@earth # SUSEConnect -r ACTIVATION_CODE -e E_MAIL
```

```
root@mars # SUSEConnect -r ACTIVATION_CODE -e E_MAIL
```

- b. Installez `ha-cluster-bootstrap` sur les deux noeuds :

```
root@earth # zypper in ha-cluster-bootstrap
```

```
root@mars # zypper in ha-cluster-bootstrap
```

- c. Assignez l'image RBD `sbd01` sur les deux passerelles Samba via `rbdmap.service`.

Modifiez `/etc/ceph/rbdmap` et ajoutez une entrée pour l'image SBD :

```
rbd/sbd01 id=samba.gw,keyring=/etc/ceph/ceph.client.samba.gw.keyring
```

Activez et démarrez `rbdmap.service` :

```
root@earth # systemctl enable rbdmap.service && systemctl start rbdmap.service
root@mars # systemctl enable rbdmap.service && systemctl start rbdmap.service
```

Le périphérique `/dev/rbd/rbd/sbd01` doit être disponible sur les deux passerelles Samba.

d. Initialisez la grappe sur `earth` et laissez `mars` la rejoindre.

```
root@earth # ha-cluster-init
```

```
root@mars # ha-cluster-join -c earth
```



Important

Au cours du processus d'initialisation et de jointure de la grappe, il vous sera demandé de manière interactive si vous souhaitez utiliser SBD. Confirmez avec `y`, puis spécifiez `/dev/rbd/rbd/sbd01` comme chemin d'accès au périphérique de stockage.

6. Vérifiez l'état de la grappe. Vous devriez voir deux noeuds ajoutés à la grappe :

```
root@earth # crm status
2 nodes configured
1 resource configured

Online: [ earth mars ]

Full list of resources:

admin-ip          (ocf::heartbeat:IPaddr2):      Started earth
```

7. Exécutez les commandes suivantes sur `earth` pour configurer la ressource CTDB :

```
root@earth # crm configure
crm(live)configure# primitive ctdb ocf:heartbeat:CTDB params \
    ctdb_manages_winbind="false" \
    ctdb_manages_samba="false" \
    ctdb_recovery_lock="/usr/lib64/ctdb/ctdb_mutex_ceph_rados_helper
```

```

ceph client.samba.gw cephfs_metadata ctdb-mutex"
ctdb_socket="/var/lib/ctdb/ctdb.socket" \
  op monitor interval="10" timeout="20" \
  op start interval="0" timeout="200" \
  op stop interval="0" timeout="100"
crm(live)configure# primitive smb systemd:smb \
  op start timeout="100" interval="0" \
  op stop timeout="100" interval="0" \
  op monitor interval="60" timeout="100"
crm(live)configure# primitive nmb systemd:nmb \
  op start timeout="100" interval="0" \
  op stop timeout="100" interval="0" \
  op monitor interval="60" timeout="100"
crm(live)configure# primitive winbind systemd:winbind \
  op start timeout="100" interval="0" \
  op stop timeout="100" interval="0" \
  op monitor interval="60" timeout="100"
crm(live)configure# group g-ctdb ctdb winbind nmb smb
crm(live)configure# clone cl-ctdb g-ctdb meta interleave="true"
crm(live)configure# commit

```



Astuce : primitives nmb et winbind facultatives

Si vous n'avez pas besoin de parcourir un partage réseau, il n'est pas nécessaire d'ajouter la primitive nmb.

La primitive winbind est uniquement requise en cas de configuration en tant que membre du domaine Active Directory. Reportez-vous à la [Section 24.2, « Jointure de la passerelle Samba et d'Active Directory »](#).

Le fichier binaire /usr/lib64/ctdb/ctdb_mutex_ceph_rados_helper de l'option de configuration ctdb_recovery_lock possède les paramètres CLUSTER_NAME (NOM_GRAPPE), CEPHX_USER (UTILISATEUR_CEPHX), RADOS_POOL (RÉSERVE_RADOS) et RADOS_OBJECT (OBJET_RADOS), dans cet ordre.

Un paramètre de timeout de verrouillage supplémentaire peut être ajouté à la suite pour remplacer la valeur par défaut utilisée (10 secondes). Une valeur plus élevée augmente le délai de basculement du maître de récupération CTDB, tandis qu'une valeur plus faible peut entraîner une détection incorrecte de l'arrêt du maître de récupération, déclenchant des basculements bagottants.

8. Ajoutez une adresse IP en grappe :

```
crm(live)configure# primitive ip ocf:heartbeat:IPaddr2
    params ip=192.168.2.1 \
    unique_clone_address="true" \
    op monitor interval="60" \
    meta resource-stickiness="0"
crm(live)configure# clone cl-ip ip \
    meta interleave="true" clone-node-max="2" globally-unique="true"
crm(live)configure# colocation col-with-ctdb 0: cl-ip cl-ctdb
crm(live)configure# order o-with-ctdb 0: cl-ip cl-ctdb
crm(live)configure# commit
```

Si `unique_clone_address` est défini sur `true`, l'agent de ressource `IPaddr2` ajoute un ID de clone à l'adresse spécifiée, ce qui permet d'obtenir trois adresses IP différentes. Elles ne sont généralement pas nécessaires, mais aident à l'équilibrage de la charge. Pour plus d'informations sur ce sujet, reportez-vous à l'adresse <https://documentation.suse.com/sle-ha/15-SP1/single-html/SLE-HA-guide/#cha-ha-lb>.

9. Vérifiez le résultat :

```
root@earth # crm status
Clone Set: base-clone [dlm]
    Started: [ factory-1 ]
    Stopped: [ factory-0 ]
Clone Set: cl-ctdb [g-ctdb]
    Started: [ factory-1 ]
    Started: [ factory-0 ]
Clone Set: cl-ip [ip] (unique)
    ip:0      (ocf:heartbeat:IPaddr2):      Started factory-0
    ip:1      (ocf:heartbeat:IPaddr2):      Started factory-1
```

10. Effectuez un test à partir d'une machine client. Sur un client Linux, exécutez la commande suivante pour vérifier si vous pouvez copier des fichiers depuis et vers le système :

```
# smbclient //192.168.2.1/myshare
```

24.1.3.1 Redémarrage des ressources HA Samba

Après toute modification de la configuration de Samba ou CTDB, un redémarrage des ressources HA peut s'avérer nécessaire pour que les modifications prennent effet. Cette opération peut être effectuée via :

```
# crm resource restart cl-ctdb
```

24.2 Jointure de la passerelle Samba et d'Active Directory

Vous pouvez configurer la passerelle Ceph Samba pour qu'elle devienne membre du domaine Samba avec prise en charge d'Active Directory (AD). En tant que membre du domaine Samba, vous pouvez utiliser les utilisateurs et groupes du domaine des listes d'accès local (ACL) sur les fichiers et répertoires du CephFS exporté.

24.2.1 Préparation de l'installation de Samba

Cette section présente les étapes préparatoires à effectuer avant de configurer Samba. Commencer avec un environnement propre évite la confusion et permet de vous assurer qu'aucun fichier de l'installation Samba précédente n'est mélangé avec l'installation du nouveau membre du domaine.



Astuce : synchronisation des horloges

Les horloges de tous les noeuds de passerelle Samba doivent être synchronisées avec le contrôleur du domaine Active Directory. Les décalages d'horloge peuvent entraîner des échecs d'authentification.

Vérifiez qu'aucun processus de caching de nom ou Samba n'est en cours d'exécution :

```
cephuser@smb > ps ax | egrep "samba|smbd|nmbd|winbindd|nscd"
```

Si la sortie répertorie des processus samba, smbd, nmbd, winbindd ou nscd, arrêtez-les.

Si vous avez déjà exécuté une installation Samba sur cet hôte, supprimez le fichier `/etc/samba/smb.conf`. Supprimez également tous les fichiers de base de données Samba, tels que les fichiers `*.tdb` et `*.ldb`. Pour lister les répertoires contenant des bases de données Samba, exécutez la commande suivante :

```
cephuser@smb > smb -b | egrep "LOCKDIR|STATEDIR|CACHEDIR|PRIVATE_DIR"
```

24.2.2 Vérification de DNS

Active Directory (AD) utilise DNS pour localiser d'autres contrôleurs de domaine (DC) et services, tels que Kerberos. Par conséquent, les membres et les serveurs du domaine AD doivent être en mesure de résoudre les zones DNS AD.

Vérifiez que DNS est correctement configuré et que la recherche directe et inversée donne une résolution correcte, par exemple :

```
cephuser@adm > nslookup DC1.domain.example.com
Server:      10.99.0.1
Address:     10.99.0.1#53

Name:   DC1.domain.example.com
Address: 10.99.0.1
```

```
cephuser@adm > 10.99.0.1
Server:      10.99.0.1
Address:     10.99.0.1#53

1.0.99.10.in-addr.arpa name = DC1.domain.example.com.
```

24.2.3 Résolution des enregistrements SRV

AD utilise les enregistrements SRV pour localiser des services, tels que Kerberos et LDAP. Pour vérifier que les enregistrements SRV sont résolus correctement, utilisez le shell interactif **nslookup**, par exemple :

```
cephuser@adm > nslookup
Default Server:  10.99.0.1
Address:         10.99.0.1

> set type=SRV
```

```
> _ldap._tcp.domain.example.com.
Server: UnKnown
Address: 10.99.0.1

_ldap._tcp.domain.example.com SRV service location:
    priority      = 0
    weight        = 100
    port          = 389
    svr hostname  = dc1.domain.example.com
domain.example.com nameserver = dc1.domain.example.com
dc1.domain.example.com internet address = 10.99.0.1
```

24.2.4 Configuration de Kerberos

Samba prend en charge les interfaces dorsales Heimdal et MIT Kerberos. Pour configurer Kerberos sur le membre du domaine, définissez ce qui suit dans votre fichier /etc/krb5.conf :

```
[libdefaults]
    default_realm = DOMAIN.EXAMPLE.COM
    dns_lookup_realm = false
    dns_lookup_kdc = true
```

L'exemple précédent configure Kerberos pour le domaine DOMAIN.EXAMPLE.COM. Nous déconseillons de définir d'autres paramètres dans le fichier /etc/krb5.conf. Si /etc/krb5.conf contient une ligne include, cela ne fonctionnera pas ; vous **devez** supprimer cette ligne.

24.2.5 Résolution du nom de l'hôte local

Lorsque vous joignez un hôte au domaine, Samba essaie d'enregistrer le nom d'hôte dans la zone DNS AD. Pour cela, l'utilitaire **net** doit être en mesure de résoudre le nom de l'hôte à l'aide de DNS ou d'une entrée correcte dans le fichier /etc/hosts.

Pour vérifier que votre nom d'hôte se résout correctement, utilisez la commande **getent hosts** :

```
cephuser@adm > getent hosts example-host
10.99.0.5      example-host.domain.example.com    example-host
```

La résolution du nom d'hôte et du nom de domaine complet (FQDN) ne doit pas renvoyer l'adresse IP 127.0.0.1 ni quelconque adresse IP autre que celle utilisée sur l'interface LAN du membre du domaine. Si aucune sortie n'est retournée ou si la résolution de l'hôte donne une adresse IP incorrecte et que vous n'utilisez pas DHCP, configurez l'entrée correcte dans le fichier `/etc/hosts` :

```
127.0.0.1    localhost
10.99.0.5    example-host.samdom.example.com  example-host
```



Astuce : DHCP et `/etc/hosts`

Si vous utilisez DHCP, vérifiez que le fichier `/etc/hosts` contient uniquement la ligne « 127.0.0.1 ». Si les problèmes persistent, contactez l'administrateur de votre serveur DHCP.

Si vous avez besoin d'ajouter des alias au nom d'hôte de la machine, ajoutez-les à la fin de la ligne qui commence avec l'adresse IP de la machine, et non à la ligne « 127.0.0.1 ».

24.2.6 Configuration de Samba

Cette section présente des informations sur les options de configuration spécifiques que vous devez inclure dans la configuration de Samba.

L'adhésion au domaine Active Directory est principalement configurée en définissant `security = ADS` avec les paramètres d'assignation de domaine et d'ID Kerberos appropriés dans la section `[global]` du fichier `/etc/samba/smb.conf`.

```
[global]
security = ADS
workgroup = DOMAIN
realm = DOMAIN.EXAMPLE.COM
...
```

24.2.6.1 Choix de l'interface dorsale pour l'assignation d'ID dans `winbindd`

Si vos utilisateurs doivent avoir des shells de connexion et/ou des chemins de répertoire privé Unix différents, ou si vous souhaitez qu'ils aient le même ID partout, vous devez utiliser l'interface dorsale « ad » winbind et ajouter des attributs RFC2307 à AD.



Important : attributs RFC2307 et numéros d'ID

Les attributs RFC2307 ne sont pas ajoutés automatiquement lors de la création d'utilisateurs ou de groupes.

Les numéros d'ID trouvés sur un contrôleur de domaine (nombres compris dans la plage de 3000000) ne sont *pas* des attributs RFC2307 et ne seront pas utilisés sur les membres du domaine Unix. Si vous avez besoin d'avoir les mêmes numéros d'ID partout, ajoutez les attributs `uidNumber` et `gidNumber` à AD et utilisez l'interface dorsale « ad » winbind sur les membres du domaine Unix. Si vous décidez d'ajouter des attributs `uidNumber` et `gidNumber` à AD, n'utilisez pas de numéros de la plage de 3000000.

Si vos utilisateurs emploient le contrôleur de domaine AD Samba uniquement pour l'authentification, sans y stocker de données ni s'y connecter, vous pouvez utiliser l'interface dorsale « rid » winbind. Celle-ci calcule les ID utilisateur et de groupe à partir du RID Windows*. Si vous utilisez la même section `[globale]` du fichier `smb.conf` sur chaque membre du domaine Unix, vous obtiendrez les mêmes ID. Si vous utilisez l'interface dorsale « rid », vous n'avez pas besoin d'ajouter quoi que ce soit à AD et les attributs RFC2307 sont ignorés. En cas d'utilisation de l'interface dorsale « rid », définissez les paramètres `template shell` et `template homedir` dans `smb.conf`. Ces paramètres sont globaux de sorte que tout le monde obtient le même shell de connexion et le même chemin de répertoire privé Unix (contrairement aux attributs RFC2307 qui permettent de définir des shells et des chemins de répertoire privé Unix individuels).

Il existe une autre façon de configurer Samba, lorsque vos utilisateurs et groupes doivent avoir le même ID partout, mais que vous avez besoin que vos utilisateurs aient le même shell de connexion et emploient le même chemin de répertoire privé Unix. Pour ce faire, utilisez l'interface dorsale « ad » winbind et les lignes de modèle dans `smb.conf`. De cette façon, vous n'avez qu'à ajouter les attributs `uidNumber` et `gidNumber` à AD.



Astuce : supplément d'informations sur les interfaces dorsales pour l'assignation d'ID

Pour des informations plus détaillées sur les interfaces dorsales disponibles pour l'assignation d'ID, reportez-vous aux pages de manuel correspondantes : `man 8 idmap_ad`, `man 8 idmap_rid` et `man 8 idmap_autorid`.

24.2.6.2 Définition des plages d'ID utilisateur et de groupe

Après avoir décidé de l'interface dorsale winbind à employer, vous devez spécifier les plages à utiliser avec l'option `idmap config` dans `smb.conf`. Par défaut, un membre du domaine Unix comporte plusieurs blocs d'utilisateurs et de groupes :

TABEAU 24.1 : BLOCS D'ID UTILISATEUR ET DE GROUPE PAR DÉFAUT

| ID | Plage |
|-------------------|---|
| 0-999 | Utilisateurs et groupes du système local. |
| À partir de 1000 | Utilisateurs et groupes Unix locaux. |
| À partir de 10000 | Utilisateurs et groupes DOMAIN |

Comme vous pouvez le constater d'après les plages ci-dessus, vous ne devez pas définir les plages « * » ou « DOMAIN » pour commencer à 999 ou moins, car elles interfèreraient avec les utilisateurs et les groupes du système local. Vous devez également laisser un espace pour tous les utilisateurs et groupes Unix locaux, de sorte que commencer les plages `idmap config` à 3000 semble être un bon compromis.

Vous devez décider de la taille que votre domaine « DOMAIN » est susceptible d'atteindre et si vous prévoyez d'avoir des domaines approuvés. Ensuite, vous pouvez définir les plages `idmap config` comme suit :

TABEAU 24.2 : PLAGES D'ID

| Domaine | Plage |
|----------|-----------------|
| * | 3000-7999 |
| DOMAINE | 10000-999999 |
| APPROUVÉ | 1000000-9999999 |

24.2.6.3 Assignment du compte d'administrateur de domaine à l'utilisateur `root` local

Samba vous permet d'assigner des comptes de domaine à un compte local. Utilisez cette fonctionnalité pour exécuter des opérations de fichier sur le système de fichiers du membre du domaine en tant qu'utilisateur différent du compte qui a demandé l'opération sur le client.



Astuce : assignation de l'administrateur de domaine (facultative)

L'assignation de l'administrateur de domaine au compte `root` local est facultative. Configurez l'assignation uniquement si l'administrateur de domaine doit être en mesure d'exécuter des opérations de fichier sur le membre du domaine à l'aide d'autorisations `root`. Sachez que l'assignation de l'administrateur au compte `root` ne vous permet pas de vous connecter aux membres du domaine Unix en tant qu'administrateur.

Pour assigner l'administrateur de domaine au compte `root` local, procédez comme suit :

1. Ajoutez le paramètre suivant à la section `[global]` de votre fichier `smb.conf` :

```
username map = /etc/samba/user.map
```

2. Créez le fichier `/etc/samba/user.map` avec le contenu suivant :

```
!root = DOMAIN\Administrator
```



Important

Si vous utilisez l'interface dorsale d'assignation d'ID « ad », ne configurez pas l'attribut `uidNumber` pour le compte d'administrateur de domaine. Si l'attribut est défini pour le compte, la valeur remplace l'UID local « 0 » de l'utilisateur `root`, de sorte que l'assignation échoue.

Pour plus de détails, reportez-vous au paramètre `username map` sur la page de manuel de `smb.conf` (**man 5 smb.conf**).

24.2.7 Jointure du domaine Active Directory

Pour joindre l'hôte à un annuaire Active Directory, exécutez la commande suivante :

```
cephuser@smb > net ads join -U administrator
Enter administrator's password: PASSWORD
Using short domain name -- DOMAIN
Joined EXAMPLE-HOST to dns domain 'DOMAIN.example.com'
```

24.2.8 Configuration de NSS

Afin que les utilisateurs et groupes du domaine soient disponibles pour le système local, vous devez activer la bibliothèque NSS (Name Service Switch). Ajoutez l'entrée `winbind` à la fin des bases de données suivantes dans le fichier `/etc/nsswitch.conf` :

```
passwd: files winbind
group:  files winbind
```



Important : considérations

- Gardez l'entrée `file` comme première source pour les deux bases de données. Cela permet à NSS de rechercher les utilisateurs et groupes du domaine à partir des fichiers `/etc/passwd` et `/etc/group` avant d'interroger le service `winbind`.
- N'ajoutez pas l'entrée `winbind` à la base de données `shadow` NSS. Cela peut entraîner l'échec de l'utilitaire `wbinfo`.
- N'utilisez pas les mêmes noms d'utilisateur dans le fichier `/etc/passwd` local et dans le domaine.

24.2.9 Démarrage des services

Après avoir modifié la configuration, redémarrez les services Samba conformément à la [Section 24.1.2.1, « Démarrage des services Samba »](#) ou à la [Section 24.1.3.1, « Redémarrage des ressources HA Samba »](#).

24.2.10 Test de la connectivité winbindd

24.2.10.1 Envoi d'une commande ping winbindd

Pour vérifier si le service `winbindd` est en mesure de se connecter aux contrôleurs de domaine (DC) AD ou à un contrôleur de domaine primaire (PDC), entrez :

```
cephuser@smb > wbinfo --ping-dc
checking the NETLOGON for domain[DOMAIN] dc connection to "DC.DOMAIN.EXAMPLE.COM"
succeeded
```

Si la commande précédente échoue, vérifiez que le service `winbindd` est en cours d'exécution et que le fichier `smb.conf` est configuré correctement.

24.2.10.2 Recherche d'utilisateurs et de groupes du domaine

La bibliothèque `libnss_winbind` vous permet de rechercher des utilisateurs et groupes du domaine. Par exemple, pour rechercher l'utilisateur du domaine « `DOMAIN\demo01` » :

```
cephuser@smb > getent passwd DOMAIN\\demo01
DOMAIN\demo01:*:10000:10000:demo01:/home/demo01:/bin/bash
```

Pour rechercher le groupe du domaine « `Domain Users` » :

```
cephuser@smb > getent group "DOMAIN\\Domain Users"
DOMAIN\domain users:x:10000:
```

24.2.10.3 Assignation d'autorisations de fichier aux utilisateurs et groupes du domaine

La bibliothèque `NSS` vous permet d'utiliser des comptes d'utilisateurs du domaine et des groupes dans les commandes. Par exemple, pour définir le propriétaire d'un fichier sur l'utilisateur du domaine « `demo01` » et le groupe sur le groupe de domaine « `Domain Users` », entrez :

```
cephuser@smb > chown "DOMAIN\\demo01:DOMAIN\\domain users" file.txt
```

25 NFS Ganesha

NFS Ganesha est un serveur NFS qui s'exécute dans un espace d'adressage utilisateur et non pas dans le kernel du système d'exploitation. Avec NFS Ganesha, vous pouvez brancher votre propre mécanisme de stockage, tel que Ceph, et y accéder depuis n'importe quel client NFS. Pour connaître la procédure d'installation, reportez-vous au *Manuel « Guide de déploiement », Chapitre 8 « Déploiement des services essentiels restants à l'aide de cephadm », Section 8.3.6 « Déploiement de NFS Ganesha »*.



Note : performances de NFS Ganesha

En raison du overhead accru du protocole et de la latence supplémentaire causée par des sauts de réseau additionnels entre le client et le stockage, l'accès à Ceph via NFS Gateway peut réduire considérablement les performances de l'application par rapport aux clients CephFS natifs.

Chaque service NFS Ganesha se compose d'une hiérarchie de configuration contenant :

- Un fichier Bootstrap `ganesha.conf`
- Un objet de configuration commun RADOS par service
- Un objet de configuration RADOS par exportation

La configuration Bootstrap est la configuration minimale pour démarrer le daemon `nfs-ganesha` dans un conteneur. Chaque configuration Bootstrap contient une directive `%url` qui inclut toute configuration supplémentaire de l'objet de configuration commun RADOS. L'objet de configuration commun peut inclure des directives `%url` supplémentaires pour chacune des exportations NFS définies dans les objets de configuration RADOS d'exportation.

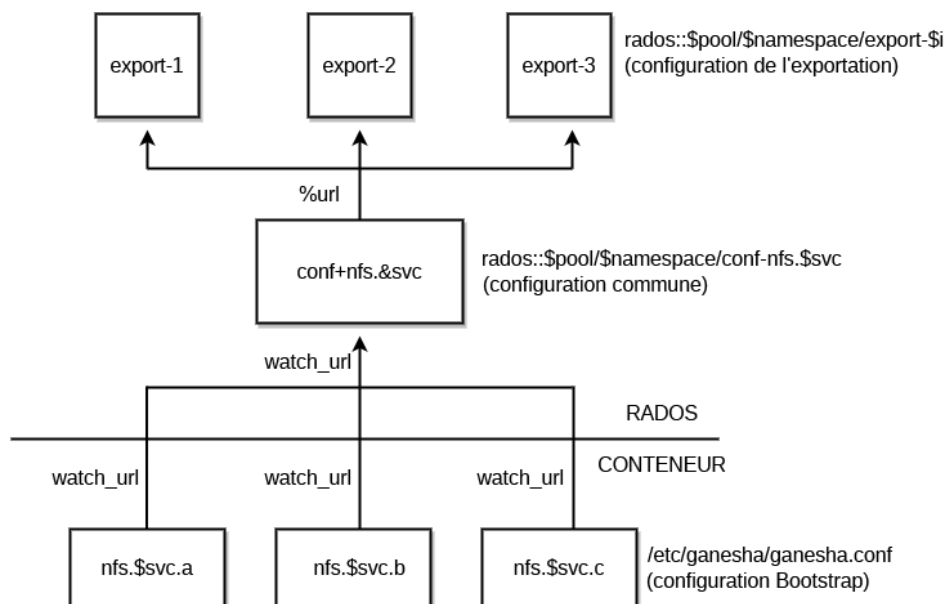


FIGURE 25.1 : STRUCTURE DE NFS GANESHA

25.1 Création d'un service NFS

La méthode recommandée pour spécifier le déploiement des services Ceph consiste à créer un fichier au format YAML spécifiant les services que vous avez l'intention de déployer. Vous pouvez créer un fichier de spécification distinct pour chaque type de service ou spécifier plusieurs (ou tous les) types de services dans un seul fichier.

Selon ce que vous avez choisi de faire, vous devrez mettre à jour ou créer un fichier au format YAML approprié pour créer un service NFS Ganesha. Pour en savoir plus sur la création du fichier, reportez-vous au *Manuel « Guide de déploiement », Chapitre 8 « Déploiement des services essentiels restants à l'aide de cephadm », Section 8.2 « Spécification de service et de placement ».*

Une fois le fichier mis à jour ou créé, exécutez la commande suivante pour créer un service `nfs-ganesha` :

```
cephuser@adm > ceph orch apply -i FILE_NAME
```

25.2 Démarrage ou redémarrage de NFS Ganesha



Important

Le démarrage du service NFS Ganesha n'exporte pas automatiquement un système de fichiers CephFS. Pour exporter un système de fichiers CephFS, créez un fichier de configuration d'exportation. Pour plus d'informations, reportez-vous à la [Section 25.4, « Création d'une exportation NFS »](#).

Pour démarrer le service NFS Ganesha, exécutez :

```
cephuser@adm > ceph orch start nfs.SERVICE_ID
```

Pour redémarrer le service NFS Ganesha, exécutez :

```
cephuser@adm > ceph orch restart nfs.SERVICE_ID
```

Si vous ne souhaitez redémarrer qu'un seul daemon NFS Ganesha, exécutez :

```
cephuser@adm > ceph orch daemon restart nfs.SERVICE_ID
```

Lorsque NFS Ganesha est démarré ou redémarré, le délai de grâce est de 90 secondes pour NFS v4. Au cours de cette période bonus, les nouvelles requêtes provenant des clients sont activement rejetées. Par conséquent, les clients peuvent être confrontés au ralentissement des requêtes lorsque NFS se trouve dans la période bonus.

25.3 Liste des objets dans la réserve de récupération NFS

Exécutez la commande suivante pour lister les objets figurant dans la réserve de récupération NFS :

```
cephuser@adm > rados --pool POOL_NAME --namespace NAMESPACE_NAME ls
```

25.4 Création d'une exportation NFS

Vous pouvez créer une exportation NFS dans le tableau de bord Ceph ou manuellement sur la ligne de commande. Pour créer l'exportation à l'aide du tableau de bord Ceph, reportez-vous au [Chapitre 7, Gestion de NFS Ganesha](#) et en particulier à la [Section 7.1, « Création d'exportations NFS »](#).

Pour créer manuellement une exportation NFS, créez un fichier de configuration à exporter. Par exemple, un fichier `/tmp/export-1` avec le contenu suivant :

```
EXPORT {
    export_id = 1;
    path = "/";
    pseudo = "/";
    access_type = "RW";
    squash = "no_root_squash";
    protocols = 3, 4;
    transports = "TCP", "UDP";
    FSAL {
        name = "CEPH";
        user_id = "admin";
        filesystem = "a";
        secret_access_key = "SECRET_ACCESS_KEY";
    }
}
```

Après avoir créé et enregistré le fichier de configuration pour la nouvelle exportation, exécutez la commande suivante pour créer l'exportation :

```
rados --pool POOL_NAME --namespace NAMESPACE_NAME put EXPORT_NAME EXPORT_CONFIG_FILE
```

Par exemple :

```
cephuser@adm > rados --pool example_pool --namespace example_namespace put export-1 /tmp/export-1
```



Note

Le bloc FSAL doit être modifié pour inclure l'ID utilisateur `cephx` et la clé d'accès secrète souhaités.

25.5 Vérification de l'exportation NFS

NFS v4 crée une liste d'exportations à la racine d'un pseudo-système de fichiers. Vous pouvez vérifier que les partages NFS sont exportés en montant `/` du noeud du serveur NFS Ganesha :

```
# mount -t nfs nfs_ganesha_server_hostname:/ /path/to/local/mountpoint
# ls /path/to/local/mountpoint cephfs
```



Note : NFS Ganesha est uniquement v4

Par défaut, `cephadm` configurera un serveur NFS v4. NFS v4 n'interagit pas avec `rpcbind` ni avec le daemon `mountd`. Les outils de client NFS, tels que `showmount` n'afficheront aucune exportation configurée.

25.6 Montage de l'exportation NFS

Pour monter le partage NFS exporté sur un hôte client, exécutez :

```
# mount -t nfs nfs_ganesha_server_hostname:/ /path/to/local/mountpoint
```

25.7 Plusieurs grappes NFS Ganesha

Plusieurs grappes NFS Ganesha peuvent être définies. Cela permet :

- À des grappes NFS Ganesha distinctes d'accéder à CephFS.

V Intégration des outils de virtualisation

26 libvirt et Ceph **377**

27 Ceph comme support de l'instance QEMU/KVM **383**

26 libvirt et Ceph

La bibliothèque `libvirt` crée une couche d'abstraction entre les interfaces d'hyperviseur et les applications logicielles qui les utilisent. Avec `libvirt`, les développeurs et administrateurs système peuvent se concentrer sur un cadre de gestion commun, une API commune et une interface shell commune (`virsh`) vers de nombreux hyperviseurs différents, notamment QEMU/KVM, Xen, LXC ou VirtualBox.

Les périphériques de bloc Ceph prennent en charge QEMU/KVM. Vous pouvez utiliser des périphériques de bloc Ceph avec un logiciel qui interagit avec `libvirt`. La solution cloud utilise `libvirt` pour interagir avec QEMU/KVM, lequel interagit avec les périphériques de bloc Ceph via `librbd`.

Pour créer des machines virtuelles qui utilisent des périphériques de bloc Ceph, utilisez les procédures décrites dans les sections suivantes. Dans ces exemples, nous avons utilisé `libvirt-pool` comme nom de réserve, `client.libvirt` comme nom d'utilisateur et `new-libvirt-image` comme nom d'image. Vous pouvez utiliser n'importe quelle valeur, mais assurez-vous de remplacer ces valeurs lors de l'exécution des commandes dans les procédures suivantes.

26.1 Configuration de Ceph avec libvirt

Pour configurer Ceph en vue de son utilisation avec `libvirt`, procédez comme suit :

1. Créez une réserve. L'exemple suivant utilise le nom de réserve `libvirt-pool` avec 128 groupes de placement.

```
cephuser@adm > ceph osd pool create libvirt-pool 128 128
```

Vérifiez que la réserve existe.

```
cephuser@adm > ceph osd lspools
```

2. Créez un utilisateur Ceph. L'exemple suivant utilise le nom d'utilisateur Ceph `client.libvirt` et fait référence à `libvirt-pool`.

```
cephuser@adm > ceph auth get-or-create client.libvirt mon 'profile rbd' osd \
'profile rbd pool=libvirt-pool'
```

Vérifiez que le nom existe.

```
cephuser@adm > ceph auth list
```



Note : nom d'utilisateur ou ID

libvirt accédera à Ceph en utilisant l'identifiant libvirt et non pas le nom Ceph client.libvirt. La [Section 30.2.1.1, « Utilisateur »](#) fournit une explication détaillée de la différence entre l'ID et le nom.

3. Utilisez QEMU pour créer une image dans votre réserve RBD. L'exemple suivant utilise le nom d'image new-libvirt-image et fait référence à libvirt-pool.



Astuce : emplacement du fichier de trousseau de clés

La clé utilisateur libvirt est stockée dans un fichier de trousseau de clés placé dans le répertoire /etc/ceph. Ce fichier doit avoir un nom approprié qui inclut le nom de la grappe Ceph à laquelle il appartient. Pour le nom de grappe par défaut « ceph », le nom de fichier de trousseau de clés est /etc/ceph/ceph.client.libvirt.keyring.

Si le trousseau de clés n'existe pas, créez-le avec :

```
cephuser@adm > ceph auth get client.libvirt > /etc/ceph/  
ceph.client.libvirt.keyring
```

```
# qemu-img create -f raw rbd:libvirt-pool/new-libvirt-image:id=libvirt 2G
```

Vérifiez que l'image existe.

```
cephuser@adm > rbd -p libvirt-pool ls
```

26.2 Préparation du gestionnaire de machines virtuelles

Vous pouvez utiliser libvirt sans gestionnaire de machines virtuelles, mais il est généralement plus simple de créer le premier domaine à l'aide de virt-manager.

1. Installez un gestionnaire de machines virtuelles.

```
# zypper in virt-manager
```

2. Préparez/téléchargez une image du système d'exploitation que vous souhaitez virtualiser.
3. Lancez le gestionnaire de machines virtuelles.

```
virt-manager
```

26.3 Création d'une machine virtuelle

Pour créer une machine virtuelle avec **virt-manager**, procédez comme suit :

1. Choisissez la connexion dans la liste, cliquez avec le bouton droit dessus, puis sélectionnez *New* (Nouvelle).
2. *Importez l'image disque existante* en fournissant le chemin du stockage existant. Indiquez le type de système d'exploitation, les paramètres de mémoire et assignez le *nom* de la machine virtuelle, par exemple libvirt-virtual-machine.
3. Terminez la configuration et démarrez la machine virtuelle.
4. Vérifiez que le domaine nouvellement créé existe avec **sudo virsh list**. Si nécessaire, indiquez la chaîne de connexion, par exemple :

```
virsh -c qemu+ssh://root@vm_host_hostname/system list
Id      Name                                State
-----
[...]
9       libvirt-virtual-machine             running
```

5. Connectez-vous à la machine virtuelle et arrêtez-la avant de la configurer en vue de son utilisation avec Ceph.

26.4 Configuration de la machine virtuelle

Dans ce chapitre, nous nous concentrons sur la configuration des machines virtuelles en vue de leur intégration avec Ceph à l'aide de **virsh**. Les commandes **virsh** nécessitent souvent des privilèges root (**sudo**) ; si vous ne possédez pas ces privilèges, elles ne vous renverront pas les

résultats appropriés et ne vous informeront pas que vous devez posséder ces privilèges root. Pour consulter une référence sur les commandes **virsh**, reportez-vous à la page [man 1 virsh](#) (nécessite l'installation du paquetage `libvirt-client`).

1. Ouvrez le fichier de configuration avec **virsh edit** `vm-domain-name`.

```
# virsh edit libvirt-virtual-machine
```

2. Une entrée `<disk>` doit figurer sous `<devices>`.

```
<devices>
  <emulator>/usr/bin/qemu-system-SYSTEM-ARCH</emulator>
  <disk type='file' device='disk'>
    <driver name='qemu' type='raw' />
    <source file='/path/to/image/recent-linux.img' />
    <target dev='vda' bus='virtio' />
    <address type='drive' controller='0' bus='0' unit='0' />
  </disk>
```

Remplacez `/path/to/image/recent-linux.img` par le chemin d'accès à l'image du système d'exploitation.

! Important

Utilisez **sudo virsh edit** au lieu d'un éditeur de texte. Si vous modifiez le fichier de configuration sous `/etc/qemu` avec un éditeur de texte, `libvirt` risque de ne pas reconnaître la modification. En cas de divergence entre le contenu du fichier XML sous `/etc/libvirt/qemu` et le résultat de **sudo virsh dumpxml** `vm-domain-name`, votre machine virtuelle risque de ne pas fonctionner correctement.

3. Ajoutez l'image Ceph RBD que vous avez précédemment créée en tant qu'entrée `<disk>`.

```
<disk type='network' device='disk'>
  <source protocol='rbd' name='libvirt-pool/new-libvirt-image'>
    <host name='monitor-host' port='6789' />
  </source>
  <target dev='vda' bus='virtio' />
</disk>
```

Remplacez `monitor-host` par le nom de votre hôte et remplacez le nom de réserve et/ou d'image, si nécessaire. Vous pouvez ajouter plusieurs entrées `<host>` pour vos moniteurs Ceph. L'attribut `dev` est le nom du périphérique logique qui apparaît sous le répertoire /

dev de votre machine virtuelle. L'attribut de bus facultatif indique le type de périphérique de disque à émuler. Les paramètres valides sont spécifiques au pilote (par exemple, ide, scsi, virtio, xen, usb ou sata).

4. Enregistrez le fichier.
5. Si l'authentification de votre grappe Ceph est activée (ce qui est le cas par défaut), vous devez générer un secret. Dans l'éditeur de votre choix, créez un fichier appelé secret.xml avec le contenu suivant :

```
<secret ephemeral='no' private='no'>
  <usage type='ceph'>
    <name>client.libvirt secret</name>
  </usage>
</secret>
```

6. Définissez le secret.

```
# virsh secret-define --file secret.xml
<uuid of secret is output here>
```

7. Récupérez la clé client.libvirt et enregistrez la chaîne de clé dans un fichier.

```
cephuser@adm > ceph auth get-key client.libvirt | sudo tee client.libvirt.key
```

8. Définissez l'UUID du secret.

```
# virsh secret-set-value --secret uuid of secret \
--base64 $(cat client.libvirt.key) && rm client.libvirt.key secret.xml
```

Vous devez également définir le secret manuellement en ajoutant l'entrée <auth> à l'élément <disk> que vous avez saisi précédemment (en remplaçant la valeur uuid par le résultat de l'exemple de ligne de commande ci-dessus).

```
# virsh edit libvirt-virtual-machine
```

Ajoutez ensuite l'élément <auth></auth> au fichier de configuration du domaine :

```
...
</source>
<auth username='libvirt'>
  <secret type='ceph' uuid='9ec59067-fdbc-a6c0-03ff-df165c0587b8' />
</auth>
<target ...
```




Note

L'ID à utiliser est `libvirt`, et non pas le nom Ceph `client.libvirt` généré à l'étape 2 de la [Section 26.1, « Configuration de Ceph avec libvirt »](#). Assurez-vous d'utiliser le composant ID du nom Ceph que vous avez généré. Si, pour une raison quelconque, vous devez régénérer le secret, il vous faudra exécuter `sudo virsh secret-undefine uuid` avant de relancer `sudo virsh secret-set-value`.

26.5 Résumé

Une fois que vous avez configuré la VM pour l'utiliser avec Ceph, vous pouvez la démarrer. Pour vérifier que la machine virtuelle et Ceph communiquent, vous pouvez effectuer les procédures suivantes.

1. Vérifiez si Ceph est en cours d'exécution :

```
cephuser@adm > ceph health
```

2. Vérifiez si la machine virtuelle est en cours d'exécution :

```
# virsh list
```

3. Vérifiez si la machine virtuelle communique avec Ceph. Remplacez `vm-domain-name` par le nom du domaine de votre machine virtuelle :

```
# virsh qemu-monitor-command --hmp vm-domain-name 'info block'
```

4. Vérifiez si le périphérique de `&target dev='hdb' bus='ide' />` figure sous `/dev` ou sous `/proc/partitions` :

```
> ls /dev  
> cat /proc/partitions
```

27 Ceph comme support de l'instance QEMU/KVM

Le cas d'utilisation le plus courant de Ceph consiste à fournir des images de périphériques de bloc aux machines virtuelles. Par exemple, un utilisateur peut créer une image « golden » avec un système d'exploitation et tout logiciel pertinent dans une configuration idéale. Il réalise ensuite un instantané de l'image. Enfin, l'utilisateur clone l'instantané (généralement plusieurs fois, voir [Section 20.3, « Images instantanées »](#) pour plus de détails). La possibilité de créer des clones de copie sur écritures (copy-on-write, COW) d'un instantané signifie que Ceph peut rapidement bloquer les images de périphérique de bloc sur des machines virtuelles, car le client n'a pas besoin de télécharger une image entière chaque fois qu'il fait tourner une nouvelle machine virtuelle.

Les périphériques de bloc Ceph peuvent s'intégrer aux machines virtuelles QEMU. Pour plus d'informations sur QEMU KVM, reportez-vous à la page <https://documentation.suse.com/sles/15-SP1/single-html/SLES-virtualization/#part-virt-qemu>.

27.1 Installation qemu-block-rbd

Pour pouvoir utiliser les périphériques de bloc Ceph, QEMU doit disposer du pilote approprié. Vérifiez si le paquetage `qemu-block-rbd` est installé ; installez-le si nécessaire :

```
# zypper install qemu-block-rbd
```

27.2 Utilisation de QEMU

Vous devez indiquer le nom de la réserve et le nom de l'image sur la ligne de commande QEMU. Vous pouvez également spécifier un nom d'instantané.

```
qemu-img command options \  
rbd:pool-name/image-name@snapshot-name:option1=value1:option2=value2...
```

Par exemple, les options `id` et `conf` peuvent se présenter ainsi :

```
qemu-img command options \  
rbd:pool_name/image_name:id=glance:conf=/etc/ceph/ceph.conf
```

27.3 Création d'images avec QEMU

Vous pouvez créer une image de périphérique de bloc à partir de QEMU. Vous devez spécifier rbd, le nom de la réserve et le nom de l'image que vous souhaitez créer. Vous devez également indiquer la taille de l'image.

```
qemu-img create -f raw rbd:pool-name/image-name size
```

Par exemple :

```
qemu-img create -f raw rbd:pool1/image1 10G
Formatting 'rbd:pool1/image1', fmt=raw size=10737418240 nocow=off cluster_size=0
```



Important

Le format de données raw (brut) est le seul format qu'il est raisonnable d'utiliser avec RBD. Techniquement, vous pouvez utiliser d'autres formats pris en charge par QEMU, tels que qcow2, mais cela ajoutera un overhead supplémentaire et fera que le volume ne sera plus sécurisé pour la migration en direct de la machine virtuelle si le caching est activé.

27.4 Redimensionnement d'images avec QEMU

Vous pouvez redimensionner une image de périphérique de bloc à partir de QEMU. Vous devez spécifier rbd, le nom de la réserve et le nom de l'image que vous souhaitez redimensionner. Vous devez également indiquer la taille de l'image.

```
qemu-img resize rbd:pool-name/image-name size
```

Par exemple :

```
qemu-img resize rbd:pool1/image1 9G
Image resized.
```

27.5 Récupération d'informations d'image avec QEMU

Vous pouvez récupérer des informations d'image de périphérique de bloc à partir de QEMU. Vous devez spécifier rbd, le nom de la réserve et le nom de l'image.

```
qemu-img info rbd:pool-name/image-name
```

Par exemple :

```
qemu-img info rbd:pool1/image1
image: rbd:pool1/image1
file format: raw
virtual size: 9.0G (9663676416 bytes)
disk size: unavailable
cluster_size: 4194304
```

27.6 Exécution de QEMU avec RBD

QEMU peut accéder à une image en tant que périphérique de bloc virtuel directement via [librbd](#). Cela évite un changement de contexte supplémentaire et permet de tirer parti du caching RBD.

Vous pouvez utiliser **qemu-img** pour convertir des images de machines virtuelles existantes en images de périphériques de bloc Ceph. Par exemple, si vous disposez d'une image qcow2, vous pouvez exécuter :

```
qemu-img convert -f qcow2 -O raw sles12.qcow2 rbd:pool1/sles12
```

Pour exécuter un démarrage de machine virtuelle à partir de cette image, vous pouvez exécuter :

```
# qemu -m 1024 -drive format=raw,file=rbd:pool1/sles12
```

Le caching RBD peut améliorer considérablement les performances. Les options de cache de QEMU contrôlent le caching [librbd](#) :

```
# qemu -m 1024 -drive format=rbd,file=rbd:pool1/sles12,cache=writeback
```

Pour plus d'informations sur le caching RBD, reportez-vous à la [Section 20.5, « Paramètres de cache »](#).

27.7 Activation du rejet et de TRIM

Les périphériques de bloc Ceph prennent en charge l'opération de rejet. Cela signifie qu'un invité peut envoyer des requêtes TRIM pour permettre à un périphérique de bloc Ceph de récupérer l'espace inutilisé. Cette fonction peut être activée sur l'invité en montant le système de fichiers [XFS](#) avec l'option `discard`.

Afin que cette option soit disponible pour l'invité, elle doit être activée explicitement pour le périphérique de bloc. Pour ce faire, vous devez spécifier une option `discard_granularity` associée à l'unité :

```
# qemu -m 1024 -drive format=raw,file=rbd:pool1/sles12,id=drive1,if=none \  
-device driver=ide-hd,drive=drive1,discard_granularity=512
```



Note

L'exemple ci-dessus utilise le pilote IDE. Le pilote virtio ne prend pas en charge le rejet.

Si vous utilisez `libvirt`, modifiez le fichier de configuration de votre domaine `libvirt` à l'aide de **`virsh edit`** afin d'inclure la valeur `xmlns:qemu`. Ajoutez ensuite un `qemu:commandline block` en tant qu'enfant de ce domaine. L'exemple suivant montre comment définir deux périphériques avec `qemu id =` et associer des valeurs `discard_granularity` différentes à cet identifiant.

```
<domain type='kvm' xmlns:qemu='http://libvirt.org/schemas/domain/qemu/1.0'>  
  <qemu:commandline>  
    <qemu:arg value='-set' />  
    <qemu:arg value='block.scsi0-0-0.discard_granularity=4096' />  
    <qemu:arg value='-set' />  
    <qemu:arg value='block.scsi0-0-1.discard_granularity=65536' />  
  </qemu:commandline>  
</domain>
```

27.8 Définition des options du cache QEMU

Les options du cache QEMU correspondent aux paramètres du cache Ceph RBD suivants.

Writeback (Écriture différée) :

```
rbd_cache = true
```

WriteThrough (Écriture immédiate) :

```
rbd_cache = true  
rbd_cache_max_dirty = 0
```

Aucun :

```
rbd_cache = false
```

Les paramètres du cache QEMU remplacent les paramètres par défaut de Ceph (paramètres qui ne sont pas explicitement définis dans le fichier de configuration de Ceph). Si vous définissez explicitement les paramètres de cache RBD dans votre fichier de configuration Ceph (voir [Section 20.5, « Paramètres de cache »](#)), vos paramètres Ceph remplacent les paramètres du cache QEMU. Si vous définissez les paramètres du cache sur la ligne de commande QEMU, les paramètres de ligne de commande QEMU remplacent les paramètres du fichier de configuration Ceph.

VI Configuration d'une grappe

- 28 Configuration de la grappe Ceph **389**
- 29 Modules Ceph Manager **410**
- 30 Authentification avec cephx **415**

28 Configuration de la grappe Ceph

Ce chapitre décrit comment configurer la grappe Ceph à l'aide des options de configuration.

28.1 Configuration du fichier `ceph.conf`

cephadm utilise un fichier `ceph.conf` de base qui ne contient qu'un ensemble minimal d'options pour la connexion aux instances MON, l'authentification et la récupération des informations de configuration. Dans la plupart des cas, cela se limite à l'option `mon_host` (bien que cela puisse être évité en utilisant des enregistrements DNS SRV).



Important

Le fichier `ceph.conf` ne fait plus office d'emplacement central pour stocker la configuration de la grappe, et est remplacé par la base de données de configuration (voir [Section 28.2, « Base de données de configuration »](#)).

Si vous devez encore modifier la configuration de la grappe avec le fichier `ceph.conf` (par exemple, parce que vous utilisez un client qui ne prend pas en charge les options de lecture de la base de données de configuration), vous devez exécuter la commande suivante et veiller à gérer et distribuer le fichier `ceph.conf` sur l'ensemble de la grappe :

```
cephuser@adm > ceph config set mgr mgr/cephadm/manage_etc_ceph_ceph_conf false
```

28.1.1 Accès à `ceph.conf` dans des images de conteneur

Même si les daemons Ceph s'exécutent au sein de conteneurs, vous pouvez toujours accéder à leur fichier de configuration `ceph.conf`. Il est *monté en liaison* en tant que fichier suivant sur le système hôte :

```
/var/lib/ceph/CLUSTER_FSID/DAEMON_NAME/config
```

Remplacez `CLUSTER_FSID` par le FSID unique de la grappe en cours d'exécution tel que renvoyé par la commande `ceph fsid`, et `DAEMON_NAME` par le nom du daemon spécifique répertorié par la commande `ceph orch ps`. Par exemple :

```
/var/lib/ceph/b4b30c6e-9681-11ea-ac39-525400d7702d/osd.2/config
```


Pour modifier la configuration d'un daemon, éditez son fichier config et redémarrez-le :

```
# systemctl restart ceph-CLUSTER_FSID-DAEMON_NAME
```

Par exemple :

```
# systemctl restart ceph-b4b30c6e-9681-11ea-ac39-525400d7702d-osd.2
```



Important

Tous les paramètres personnalisés seront perdus après le redéploiement du daemon par cephadm.

28.2 Base de données de configuration

Les instances Ceph Monitor gèrent une base de données centrale d'options de configuration qui affectent le comportement de l'ensemble de la grappe.

28.2.1 Configuration des sections et des masques

Les options de configuration stockées par l'instance MON peuvent résider dans une section *globale*, une section *type de daemon* ou une section *daemon spécifique*. En outre, les options peuvent également être associées à un *masque* afin de limiter davantage les daemons ou clients auxquels l'option s'applique. Les masques ont deux formes :

- TYPE:LOCATION où TYPE est une propriété CRUSH telle que rack ou host, tandis que LOCATION est une valeur pour cette propriété.
Par exemple, host:example_host limitera l'option aux daemons ou aux clients s'exécutant sur un hôte particulier.
- CLASS:DEVICE_CLASS où DEVICE_CLASS est le nom d'une classe de périphérique CRUSH telle que hdd ou ssd. Par exemple, class:ssd limitera l'option aux OSD sauvegardés par des disques SSD. Ce masque n'a aucun effet sur les daemons ou clients non-OSD.

28.2.2 Définition et lecture des options de configuration

Utilisez les commandes suivantes pour définir ou lire les options de configuration de la grappe. Le paramètre *WHO* peut être un nom de section, un masque ou une combinaison des deux, séparés par une barre oblique (/). Par exemple, *osd/rack:foo* représente tous les daemons OSD dans le rack appelé *foo*.

ceph config dump

Vide l'ensemble de la base de données de configuration pour l'ensemble d'une grappe.

ceph config get WHO

Vide la configuration d'un daemon ou d'un client spécifique (par exemple, *mds.a*), telle que stockée dans la base de données de configuration.

ceph config set WHO OPTION VALUE

Définit l'option de configuration sur la valeur spécifiée dans la base de données de configuration.

ceph config show WHO

Affiche la configuration en cours d'exécution signalée pour un daemon en cours d'exécution. Ces paramètres peuvent différer de ceux stockés par les moniteurs si des fichiers de configuration locaux sont également utilisés ou si des options ont été remplacées sur la ligne de commande ou lors de l'exécution. La source des valeurs d'option est signalée comme faisant partie de la sortie.

ceph config assimilate-conf -i INPUT_FILE -o OUTPUT_FILE

Importe un fichier de configuration spécifié comme *INPUT_FILE* et stocke toutes les options valides dans la base de données de configuration. Tous les paramètres qui ne sont pas reconnus, qui ne sont pas valides ou qui ne peuvent pas être contrôlés par le moniteur sont renvoyés dans un fichier abrégé stocké comme *OUTPUT_FILE*. Cette commande est utile pour la transition des fichiers de configuration hérités vers la configuration centralisée basée sur le moniteur.

28.2.3 Configuration des daemons lors de l'exécution

Dans la plupart des cas, Ceph vous permet de modifier la configuration d'un daemon lors de l'exécution. Cela est utile, par exemple, lorsque vous devez augmenter ou réduire la quantité de sortie de consigne ou lors de l'optimisation de la grappe d'exécution.

Vous pouvez mettre à jour les valeurs des options de configuration avec la commande suivante :

```
cephuser@adm > ceph config set DAEMON OPTION VALUE
```

Par exemple, pour ajuster le niveau de consignation de débogage sur un OSD spécifique, exécutez :

```
cephuser@adm > ceph config set osd.123 debug_ms 20
```



Note

Si la même option est également personnalisée dans un fichier de configuration local, le paramètre du moniteur est ignoré car sa priorité est inférieure à celle du fichier de configuration.

28.2.3.1 Remplacement de valeurs

Vous pouvez modifier temporairement la valeur d'une option à l'aide des sous-commandes **tell** ou **daemon**. Cette modification affecte uniquement le processus en cours d'exécution et est ignorée après le redémarrage du daemon ou du processus.

Il existe deux façons de remplacer des valeurs :

- Utilisez la sous-commande **tell** pour envoyer un message à un daemon spécifique à partir de n'importe quel noeud de grappe :

```
cephuser@adm > ceph tell DAEMON config set OPTION VALUE
```

Par exemple :

```
cephuser@adm > ceph tell osd.123 config set debug_osd 20
```



Astuce

La sous-commande **tell** accepte les caractères joker comme identificateurs de daemon. Par exemple, pour régler le niveau de débogage sur tous les daemons OSD, exécutez :

```
cephuser@adm > ceph tell osd.* config set debug_osd 20
```

- Utilisez la sous-commande **daemon** pour vous connecter à un processus daemon spécifique via un socket dans `/var/run/ceph` à partir du noeud sur lequel le processus est en cours d'exécution :

```
cephuser@adm > cephadm enter --name osd.ID -- ceph daemon DAEMON config  
set OPTION VALUE
```

Par exemple :

```
cephuser@adm > cephadm enter --name osd.4 -- ceph daemon osd.4 config set debug_osd  
20
```



Astuce

Lorsque vous affichez les paramètres d'exécution avec la commande **ceph config show** (voir [Section 28.2.3.2, « Affichage des paramètres d'exécution »](#)), les valeurs temporairement remplacées sont affichées avec un remplacement de source.

28.2.3.2 Affichage des paramètres d'exécution

Pour afficher toutes les options définies pour un daemon :

```
cephuser@adm > ceph config show-with-defaults osd.0
```

Pour afficher toutes les options non définies par défaut pour un daemon :

```
cephuser@adm > ceph config show osd.0
```

Pour inspecter une option spécifique :

```
cephuser@adm > ceph config show osd.0 debug_osd
```

Vous pouvez également vous connecter à un daemon en cours d'exécution à partir du noeud sur lequel son processus est exécuté et observer sa configuration :

```
cephuser@adm > cephadm enter --name osd.0 -- ceph daemon osd.0 config show
```

Pour afficher uniquement les paramètres non définis par défaut :

```
cephuser@adm > cephadm enter --name osd.0 -- ceph daemon osd.0 config diff
```

Pour inspecter une option spécifique :

```
cephuser@adm > cephadm enter --name osd.0 -- ceph daemon osd.0 config get debug_osd
```

28.3 config-key stocker

config-key est un service général proposé par les instances Ceph Monitor. Il simplifie la gestion des clés de configuration en stockant de façon permanente des paires clé/valeur. config-key est principalement utilisé par les outils et les daemons Ceph.



Astuce

Après avoir ajouté une nouvelle clé ou modifié une clé existante, redémarrez le service concerné pour que les modifications prennent effet. Pour plus d'informations sur le fonctionnement des services Ceph, reportez-vous au [Chapitre 14, Exécution des services Ceph](#).

Utilisez la commande `ceph config-key` pour utiliser la zone de stockage config-key. La commande **config-key** utilise les sous-commandes suivantes :

ceph config-key rm KEY

Supprime la clé spécifiée.

ceph config-key exists KEY

Vérifie l'existence de la clé spécifiée.

ceph config-key get KEY

Récupère la valeur de la clé spécifiée.

ceph config-key ls

Répertorie toutes les clés.

ceph config-key dump

Vide toutes les clés et leurs valeurs.

ceph config-key set KEY VALUE

Stocke la clé spécifiée avec la valeur donnée.

28.3.1 Passerelle iSCSI

La passerelle iSCSI utilise la zone de stockage config-key pour enregistrer ou lire ses options de configuration. Toutes les clés liées à la passerelle iSCSI sont précédées de la chaîne iscsi, par exemple :

```
iscsi/trusted_ip_list
iscsi/api_port
iscsi/api_user
iscsi/api_password
iscsi/api_secure
```

Si vous avez besoin, par exemple, de deux ensembles d'options de configuration, étendez le préfixe avec un autre mot-clé descriptif, par exemple datacenterA et datacenterB :

```
iscsi/datacenterA/trusted_ip_list
iscsi/datacenterA/api_port
[...]
iscsi/datacenterB/trusted_ip_list
iscsi/datacenterB/api_port
[...]
```

28.4 Ceph OSD et BlueStore

28.4.1 Configuration du dimensionnement automatique du cache

BlueStore peut être configuré pour redimensionner automatiquement ses caches lorsque tc_malloc est configuré comme option d'allocation de mémoire et que le paramètre bluestore_cache_autotune est activé. Cette option est actuellement activée par défaut. BlueStore tente alors de maintenir l'utilisation de la mémoire de segment OSD sous une taille cible désignée via l'option de configuration osd_memory_target. Il s'agit d'un algorithme du meilleur d'effort

(« Best Effort ») ; les caches ne passeront pas en dessous de la quantité minimale spécifiée par la valeur `osd_memory_cache_min`. Les ratios de cache seront choisis en fonction d'une hiérarchie de priorités. Si l'information de priorité n'est pas disponible, les options `bluestore_cache_meta_ratio` et `bluestore_cache_kv_ratio` sont utilisées à la place.

`bluestore_cache_autotune`

Aligne automatiquement les ratios assignés à différents caches BlueStore tout en respectant les valeurs minimales. La valeur par défaut est `True`.

`osd_memory_target`

Lorsque `tc_malloc` et `bluestore_cache_autotune` sont activés, essayez de garder ces nombreux octets assignés en mémoire.



Note

Cela peut ne pas correspondre exactement à l'utilisation de la mémoire RSS du processus. Bien que la quantité totale de mémoire de segment assignée par le processus devrait généralement rester proche de cette cible, il n'y a aucune garantie que le kernel va effectivement récupérer la mémoire dont l'assignation a été annulée.

`osd_memory_cache_min`

Lorsque `tc_malloc` et `bluestore_cache_autotune` sont activés, définissez la quantité minimale de mémoire utilisée pour les caches.



Note

Si vous définissez cette option sur une valeur trop faible, cela peut entraîner un débordement considérable des caches.

28.5 Ceph Object Gateway

Vous pouvez influencer le comportement d'Object Gateway avec plusieurs options. Si une option n'est pas spécifiée, sa valeur par défaut est utilisée. Vous trouverez ci-dessous une liste complète des options Object Gateway.

28.5.1 Paramètres généraux

rgw_frontends

Configure l'interface client HTTP. Pour en spécifier plusieurs, définissez-les dans une liste séparée par des virgules. Chaque configuration d'interface client peut inclure une liste d'options séparées par des espaces, dans laquelle chaque option se présente sous la forme « clé = valeur » ou « clé ». La valeur par défaut est beast port=7480.

rgw_data

Définit l'emplacement des fichiers de données pour Object Gateway. La valeur par défaut est /var/lib/ceph/radosgw/ID_GRAPPE.

rgw_enable_apis

Active les API spécifiées. Par défaut, il s'agit de « s3, swift, swift_auth, admin All APIs ».

rgw_cache_enabled

Active ou désactive le cache Object Gateway. La valeur par défaut est true.

rgw_cache_lru_size

Nombre d'entrées dans le cache Object Gateway. La valeur par défaut est 10000.

rgw_socket_path

Chemin du socket du domaine. FastCgiExternalServer utilise ce socket. Si vous ne spécifiez pas de chemin de socket, Object Gateway ne s'exécutera pas en tant que serveur externe. Le chemin que vous spécifiez ici doit être le même que celui indiqué dans le fichier rgw.conf.

rgw_fcgi_socket_backlog

Backlog de socket pour fcgi. La valeur par défaut est 1024.

rgw_host

Hôte de l'instance Object Gateway. Il peut s'agir d'une adresse IP ou d'un nom hôte. La valeur par défaut est 0.0.0.0.

rgw_port

Numéro de port sur lequel l'instance écoute les requêtes. S'il n'est pas spécifié, Object Gateway exécute la fonction externe FastCGI.

rgw_dns_name

Nom DNS du domaine desservi.

rgw_script_uri

Valeur alternative pour le SCRIPT_URI si elle n'est pas définie dans la requête.

rgw_request_uri

Valeur alternative pour REQUEST_URI si elle n'est pas définie dans la requête.

rgw_print_continue

Permet d'activer le code 100-continue s'il est opérationnel. La valeur par défaut est true.

rgw_remote_addr_param

Paramètre de l'adresse distante. Par exemple, le champ HTTP contenant l'adresse distante ou l'adresse X-Forwarded-For si un proxy inverse est opérationnel. La valeur par défaut est REMOTE_ADDR.

rgw_op_thread_timeout

Timeout en secondes pour les threads ouverts. La valeur par défaut est 600.

rgw_op_thread_suicide_timeout

Timeout en secondes avant la mort du processus Object Gateway. Il est désactivé s'il est défini sur 0 (valeur par défaut).

rgw_thread_pool_size

Nombre de threads pour le serveur Beast. Définissez une valeur plus élevée si vous devez servir plus de demandes. La valeur par défaut est de 100 threads.

rgw_num_rados_handles

Nombre d'identificateurs de grappe RADOS pour Object Gateway. Désormais, chaque thread de travail Object Gateway sélectionne un identificateur RADOS pour toute sa durée de vie. Cette option est à présent obsolète et pourra être supprimée dans les versions futures. La valeur par défaut est 1.

rgw_num_control_oids

Nombre d'objets de notification utilisés pour la synchronisation des caches entre différentes instances d'Object Gateway. La valeur par défaut est 8.

rgw_init_timeout

Nombre de secondes avant qu'Object Gateway renonce à l'initialisation. La valeur par défaut est 30.

rgw_mime_types_file

Chemin et emplacement des types MIME. Utilisés pour détection automatique des types d'objets par Swift. La valeur par défaut est /etc/mime.types.

rgw_gc_max_objs

Nombre maximal d'objets pouvant être gérés par la récupération d'espace mémoire au cours d'un cycle de traitement. La valeur par défaut est 32.

rgw_gc_obj_min_wait

Délai d'attente minimal avant que l'objet puisse être retiré et géré par le processus de récupération d'espace mémoire. La valeur par défaut est $2 * 3600$.

rgw_gc_processor_max_time

Durée maximale entre le début de deux cycles consécutifs de récupération d'espace mémoire. La valeur par défaut est 3600.

rgw_gc_processor_period

Temps de cycle pour la récupération d'espace mémoire. La valeur par défaut est 3600.

rgw_s3_success_create_obj_status

Autre réponse de statut de réussite pour create-obj. La valeur par défaut est 0.

rgw_resolve_cname

Indique si Object Gateway doit utiliser l'enregistrement DNS CNAME du champ de nom d'hôte de la requête (si le nom d'hôte est différent du nom DNS d'Object Gateway). La valeur par défaut est false.

rgw_obj_stripe_size

Taille d'un segment d'objet pour les objets Object Gateway. La valeur par défaut est 4 << 20.

rgw_extended_http_attrs

Ajoutez un nouvel ensemble d'attributs qui peuvent être définis sur une entité (par exemple, un utilisateur, un compartiment ou un objet). Ces attributs supplémentaires peuvent être configurés via les champs d'en-tête HTTP lors de la spécification de l'entité ou de sa modification à l'aide de la méthode POST. S'ils sont définis, ces attributs sont renvoyés sous forme de champs HTTP en cas de demande GET/HEAD sur l'entité. La valeur par défaut est content_foo, content_bar, x-foo-bar.

rgw_exit_timeout_secs

Nombre de secondes à attendre un processus avant de quitter sans condition. La valeur par défaut est 120.

`rgw_get_obj_window_size`

Taille de la fenêtre en octets pour une seule requête d'objet. La valeur par défaut est 16 << 20.

`rgw_get_obj_max_req_size`

Taille maximale de requête pour une seule opération GET envoyée à la grappe de stockage Ceph. La valeur par défaut est 4 << 20.

`rgw_relaxed_s3_bucket_names`

Active les règles souples de nom de compartiment S3 pour les compartiments de la région des États-Unis. La valeur par défaut est false.

`rgw_list_buckets_max_chunk`

Nombre maximal de compartiments à récupérer en une seule opération lors du listage des compartiments utilisateur. La valeur par défaut est 1000.

`rgw_override_bucket_index_max_shards`

Représente le nombre de partitions pour l'objet d'index de compartiment. Définir ce paramètre sur 0 (valeur par défaut) indique qu'il n'y a pas de partitionnement. Il est déconseillé de définir une valeur trop élevée (par exemple 1000), car cela augmente le coût du listage des compartiments. Cette variable doit être définie dans les sections de client ou globales afin qu'elle soit automatiquement appliquée aux commandes **radosgw-admin**.

`rgw_curl_wait_timeout_ms`

Timeout en millisecondes pour certains appels **curl**. La valeur par défaut est 1000.

`rgw_copy_obj_progress`

Active la sortie de la progression d'objet lors des longues opérations de copie. La valeur par défaut est true.

`rgw_copy_obj_progress_every_bytes`

Nombre minimal d'octets entre la sortie de progression de copie. La valeur par défaut est $1024 * 1024$.

`rgw_admin_entry`

Point d'entrée pour une URL de requête d'administration. La valeur par défaut est admin.

`rgw_content_length_compat`

Permet d'activer le traitement de la compatibilité des requêtes FCGI avec les options `CONTENT_LENGTH` et `HTTP_CONTENT_LENGTH` définies. La valeur par défaut est false.

`rgw_bucket_quota_ttl`

Délai en secondes pour lequel les informations de quota mises en cache sont approuvées. Une fois ce temps écoulé, les informations de quota sont à nouveau récupérées auprès de la grappe. La valeur par défaut est 600.

`rgw_user_quota_bucket_sync_interval`

Délai en secondes pendant lequel les informations de quota de compartiment sont accumulées avant d'effectuer une synchronisation avec la grappe. Pendant ce temps, les autres instances Object Gateway ne voient pas les modifications dans les statistiques de quota de compartiment associées aux opérations sur cette instance. La valeur par défaut est 180.

`rgw_user_quota_sync_interval`

Délai en secondes pendant lequel les informations de quota d'utilisateur sont accumulées avant d'effectuer une synchronisation avec la grappe. Pendant ce temps, les autres instances Object Gateway ne voient pas les modifications dans les statistiques de quota d'utilisateur associées aux opérations sur cette instance. La valeur par défaut est 180.

`rgw_bucket_default_quota_max_objects`

Nombre maximal d'objets par compartiment par défaut. Ce paramètre est défini pour les nouveaux utilisateurs si aucun autre quota n'est spécifié, et n'a aucun effet sur les utilisateurs existants. Cette variable doit être définie dans les sections de client ou globales afin qu'elle soit automatiquement appliquée aux commandes **`radosgw-admin`**. La valeur par défaut est -1.

`rgw_bucket_default_quota_max_size`

Capacité maximale par défaut par compartiment (en octets). Ce paramètre est défini pour les nouveaux utilisateurs si aucun autre quota n'est spécifié, et n'a aucun effet sur les utilisateurs existants. La valeur par défaut est -1.

`rgw_user_default_quota_max_objects`

Nombre maximal d'objets par défaut pour un utilisateur. Il inclut tous les objets dans tous les compartiments appartenant à l'utilisateur. Ce paramètre est défini pour les nouveaux utilisateurs si aucun autre quota n'est spécifié, et n'a aucun effet sur les utilisateurs existants. La valeur par défaut est -1.

`rgw_user_default_quota_max_size`

Valeur du quota de taille maximale de l'utilisateur (en octets) défini pour les nouveaux utilisateurs si aucun autre quota n'est spécifié. Ce paramètre n'a aucun effet sur les utilisateurs existants. La valeur par défaut est -1.

`rgw_verify_ssl`

Permet de vérifier les certificats SSL lors de requêtes. La valeur par défaut est true.

`rgw_max_chunk_size`

Taille maximale d'une tranche de données qui sera lue en une seule opération. L'augmentation de la valeur à 4 Mo (4194304) permet d'améliorer les performances de traitement des objets volumineux. La valeur par défaut est 128 Ko (131072).

PARAMÈTRES MULTISITES

`rgw_zone`

Nom de la zone pour l'instance de passerelle. Si aucune zone n'est définie, il est possible de configurer une valeur par défaut à l'échelle de la grappe à l'aide de la commande **radosgw-admin zone default**.

`rgw_zonegroup`

Nom du groupe de zones pour l'instance de passerelle. Si aucune zone de groupes n'est définie, il est possible de configurer une valeur par défaut à l'échelle de la grappe à l'aide de la commande **radosgw-admin zonegroup default**.

`rgw_realm`

Nom du domaine pour l'instance de passerelle. Si aucun domaine n'est défini, il est possible de configurer une valeur par défaut à l'échelle de la grappe à l'aide de la commande **radosgw-admin realm default**.

`rgw_run_sync_thread`

S'il y a d'autres zones dans le domaine à partir desquelles il faut effectuer une synchronisation, ce paramètre permet de générer des threads pour gérer la synchronisation des données et des métadonnées. La valeur par défaut est true.

`rgw_data_log_window`

Fenêtre des entrées de journal de données en secondes. La valeur par défaut est 30.

`rgw_data_log_changes_size`

Nombre d'entrées dans la mémoire à conserver pour le journal des modifications de données. La valeur par défaut est 1000.

`rgw_data_log_obj_prefix`

Préfixe du nom d'objet pour le journal des données. Il s'agit par défaut de « data_log ».

`rgw_data_log_num_shards`

Nombre de partitions (objets) sur lesquelles conserver le journal des modifications de données. La valeur par défaut est 128.

`rgw_md_log_max_shards`

Nombre maximal de partitions pour le journal des métadonnées. La valeur par défaut est 64.

PARAMÈTRES SWIFT

`rgw_enforce_swift_acls`

Applique les paramètres de la liste de contrôle d'accès (ACL) Swift. La valeur par défaut est true.

`rgw_swift_token_expiration`

Délai d'expiration d'un jeton Swift en secondes. La valeur par défaut est $24 * 3600$.

`rgw_swift_url`

URL de l'API Swift de Ceph Object Gateway.

`rgw_swift_url_prefix`

Préfixe de l'URL du stockage Swift qui précède la partie `/v1`. Il permet d'exécuter plusieurs instances de passerelle sur le même hôte. Pour des raisons de compatibilité, si vous définissez cette variable de configuration sur une valeur vide, le système utilise la valeur par défaut `/swift`. Utilisez le préfixe explicite `/` pour démarrer l'URL de stockage à la racine.



Avertissement

La définition de cette option sur `/` ne fonctionnera pas si l'API S3 est activée. N'oubliez pas que la désactivation de S3 rendra impossible le déploiement d'Object Gateway dans la configuration multisite.

`rgw_swift_auth_url`

URL par défaut pour vérifier les jetons d'authentification v1 lorsque l'authentification Swift interne n'est pas utilisée.

`rgw_swift_auth_entry`

Point d'entrée d'une URL d'authentification Swift. La valeur par défaut est auth.

rgw_swift_versioning_enabled

Active le contrôle de version des objets de l'API de stockage d'objets OpenStack. Ce paramètre permet aux clients de définir l'attribut X-Versions-Location sur les conteneurs devant être soumis au contrôle de version. L'attribut spécifie le nom du conteneur stockant les versions archivées. Son propriétaire doit être identique à celui du conteneur soumis au contrôle de version pour des raisons de vérification du contrôle d'accès ; les ACL ne sont *pas* prises en considération. Le contrôle de version de ces conteneurs ne peut pas être assuré par le mécanisme de contrôle de version des objets S3. La valeur par défaut est false.

PARAMÈTRES DE CONSIGNATION

rgw_log_nonexistent_bucket

Permet à Object Gateway de consigner une requête pour un compartiment inexistant. La valeur par défaut est false.

rgw_log_object_name

Format de consignation d'un nom d'objet. Reportez-vous à la page de manuel man 1 date pour plus de détails sur les spécificateurs de format. La valeur par défaut est %Y-%m-%d-%H-%i-%n.

rgw_log_object_name_utc

Indique si un nom d'objet consigné inclut une heure UTC. Si ce paramètre est défini sur false (valeur par défaut), il utilise l'heure locale.

rgw_usage_max_shards

Nombre maximal de partitions pour la consignation de l'utilisation. La valeur par défaut est 32.

rgw_usage_max_user_shards

Nombre maximal de partitions utilisées pour la consignation de l'utilisation d'un seul utilisateur. La valeur par défaut est 1.

rgw_enable_ops_log

Permet d'activer la consignation pour chaque opération Object Gateway réussie. La valeur par défaut est false.

rgw_enable_usage_log

Permet d'activer le journal d'utilisation. La valeur par défaut est false.

`rgw_ops_log_rados`

Définit si le journal des opérations doit être écrit dans l'interface dorsale de la grappe de stockage Ceph. La valeur par défaut est `true`.

`rgw_ops_log_socket_path`

Socket du domaine Unix pour l'écriture des journaux d'opérations.

`rgw_ops_log_data_backlog`

Taille maximale des données du backlog de données pour les journaux d'opérations écrits dans un socket du domaine Unix. La valeur par défaut est 5 < 20.

`rgw_usage_log_flush_threshold`

Nombre d'entrées fusionnées altérées dans le journal d'utilisation avant le vidage synchrone. La valeur par défaut est 1024.

`rgw_usage_log_tick_interval`

Permet de vider les données de journal d'utilisation en attente toutes les « n » secondes. La valeur par défaut est 30.

`rgw_log_http_headers`

Liste (séparée par des virgules) des en-têtes HTTP à inclure dans les entrées de journal. Les noms d'en-tête ne sont pas sensibles à la casse et utilisent le nom d'en-tête complet avec les mots séparés par des traits de soulignement. Par exemple : « `http_x_forwarded_for` », « `http_x_special_k` ».

`rgw_intent_log_object_name`

Format de consignation du nom d'objet du journal d'intentions (« intent log »). Reportez-vous à la page de manuel [`man 1 date`](#) pour plus de détails sur les spécificateurs de format. Il s'agit par défaut de « `%Y-%m-%d-%i-%n` ».

`rgw_intent_log_object_name_utc`

Indique si le nom d'objet du journal d'intentions inclut une heure UTC. Si ce paramètre est défini sur `false` (valeur par défaut), il utilise l'heure locale.

PARAMÈTRES KEYSTONE

`rgw_keystone_url`

URL du serveur Keystone.

`rgw_keystone_api_version`

Version (2 ou 3) de l'API Identity OpenStack qui doit être utilisée pour la communication avec le serveur Keystone. La valeur par défaut est 2.

`rgw_keystone_admin_domain`

Nom du domaine OpenStack avec le privilège d'administration lors de l'utilisation de l'API Identity OpenStack v3.

`rgw_keystone_admin_project`

Nom du projet OpenStack avec le privilège d'administration lors de l'utilisation de l'API Identity OpenStack v3. Si ce paramètre n'est pas défini, le système utilise la valeur de `rgw_keystone_admin_tenant`.

`rgw_keystone_admin_token`

Jeton de l'administrateur Keystone (secret partagé). Dans Object Gateway, l'authentification avec le jeton d'administrateur prime sur l'authentification avec les informations d'identification de l'administrateur (options `rgw_keystone_admin_user`, `rgw_keystone_admin_password`, `rgw_keystone_admin_tenant`, `rgw_keystone_admin_project` et `rgw_keystone_admin_domain`). La fonction de jeton d'administrateur est considérée comme obsolète.

`rgw_keystone_admin_tenant`

Nom du locataire OpenStack avec le privilège d'administration (locataire du service) en cas d'utilisation de l'API Identity OpenStack v2.

`rgw_keystone_admin_user`

Nom de l'utilisateur OpenStack avec le privilège d'administration pour l'authentification Keystone (utilisateur du service) en cas d'utilisation de l'API Identity OpenStack v2.

`rgw_keystone_admin_password`

Mot de passe de l'administrateur OpenStack en cas d'utilisation de l'API Identity OpenStack v2.

`rgw_keystone_accepted_roles`

Rôles nécessaires pour répondre aux requêtes. Il s'agit par défaut de « Member, admin ».

`rgw_keystone_token_cache_size`

Nombre maximal d'entrées dans chaque cache de jetons Keystone. La valeur par défaut est 10000.

`rgw_keystone_revocation_interval`

Nombre de secondes entre les vérifications de révocation de jeton. La valeur par défaut est $15 * 60$.

rgw_keystone_verify_ssl

Permet de vérifier les certificats SSL lors de requêtes de jeton auprès de Keystone. La valeur par défaut est true.

28.5.1.1 Remarques supplémentaires

rgw_dns_name

Permet aux clients d'utiliser des compartiments de type vhost.

L'accès de type vhost fait référence à l'utilisation de bucketname.s3-endpoint/object-path. En comparaison à l'accès de type path : s3-endpoint/bucket/object

Si le nom rgw dns name est défini, vérifiez que le client S3 est configuré pour diriger les requêtes vers le noeud d'extrémité spécifiée par le nom rgw dns name.

28.5.2 Configuration des interfaces clients HTTP

28.5.2.1 Beast

port, ssl_port

Numéros de port d'écoute IPv4 et IPv6. Vous pouvez spécifier plusieurs numéros de port :

```
port=80 port=8000 ssl_port=8080
```

La valeur par défaut est 80.

endpoint, ssl_endpoint

Adresses d'écoute au format « adresse[:port] », où l'adresse est une chaîne d'adresse IPv4 au format décimal avec points ou une adresse IPv6 en notation hexadécimale entourée de crochets. Si vous spécifiez un noeud d'extrémité IPv6, le système écoutera uniquement le protocole IPv6. Le numéro de port facultatif est défini par défaut sur 80 pour endpoint et sur 443 pour ssl_endpoint. Vous pouvez spécifier plusieurs adresses :

```
endpoint=[::1] endpoint=192.168.0.100:8000 ssl_endpoint=192.168.0.100:8080
```

ssl_private_key

Chemin facultatif du fichier de clé privée utilisé pour les noeuds d'extrémité compatibles SSL. Si ce paramètre n'est pas spécifié, le fichier ssl_certificate est utilisé comme clé privée.

tcp_nodelay

Si ce paramètre est spécifié, l'option de socket désactive l'algorithme de Nagle sur la connexion. Cela signifie que les paquets seront envoyés dès que possible au lieu d'attendre un timeout ou une saturation du tampon.

« 1 » désactive l'algorithme de Nagle pour tous les sockets.

« 0 » garde l'algorithme de Nagle activé (valeur par défaut).

EXEMPLE 28.1 : EXEMPLE DE CONFIGURATION BEAST

```
cephuser@adm > ceph config set rgw.myrealm.myzone.ses-min1.kwwazo \  
rgw_frontends beast port=8000 ssl_port=443 \  
ssl_certificate=/etc/ssl/ssl.crt \  
error_log_file=/var/log/radosgw/beast.error.log
```

28.5.2.2 CivetWeb

port

Numéro du port d'écoute. Pour les ports compatibles SSL, ajoutez un suffixe « s » (par exemple, « 443s »). Pour lier une adresse IPv4 ou IPv6 spécifique, utilisez le format « adresse:port ». Vous pouvez spécifier plusieurs noeuds d'extrémité en les joignant avec le caractère « + » ou en fournissant différentes options :

```
port=127.0.0.1:8000+443s  
port=8000 port=443s
```

La valeur par défaut est 7480.

num_threads

Nombre de threads générés par CivetWeb pour gérer les connexions HTTP entrantes. Cela limite efficacement le nombre de connexions simultanées que l'interface client peut desservir.

La valeur par défaut est celle spécifiée par l'option rgw_thread_pool_size.

request_timeout_ms

Délai en millisecondes pendant lequel CivetWeb attend davantage de données entrantes avant d'abandonner.

La valeur par défaut est de 30000 millisecondes.

access_log_file

Chemin du fichier journal des accès. Vous pouvez spécifier un chemin complet ou un chemin par rapport au répertoire de travail actuel. Si ce paramètre n'est pas spécifié (situation par défaut), les accès ne sont pas consignés.

error_log_file

Chemin du fichier journal des erreurs. Vous pouvez spécifier un chemin complet ou un chemin par rapport au répertoire de travail actuel. Si ce paramètre n'est pas spécifié (situation par défaut), les erreurs ne sont pas consignées.

EXEMPLE 28.2 : EXEMPLE DE CONFIGURATION CIVETWEB DANS `/etc/ceph/ceph.conf`

```
cephuser@adm > ceph config set rgw.myrealm.myzone.ses-min2.ingabw \  
rgw_frontends civetweb port=8000+443s request_timeout_ms=30000 \  
error_log_file=/var/log/radosgw/civetweb.error.log
```

28.5.2.3 Options communes

ssl_certificate

Chemin du fichier de certificat SSL utilisé pour les noeuds d'extrémité compatibles SSL.

prefix

Chaîne de préfixe insérée dans l'URI de toutes les requêtes. Par exemple, une interface client Swift uniquement pourrait fournir un préfixe URI /swift.

29 Modules Ceph Manager

L'architecture de Ceph Manager (voir *Manuel « Guide de déploiement », Chapitre 1 « SES et Ceph », Section 1.2.3 « Noeuds et daemons Ceph »* pour une brève présentation) permet d'étendre ses fonctionnalités par le biais de *modules*, tels que le tableau de bord (« dashboard » - voir *Partie I, « Ceph Dashboard »*), Prometheus (voir *Chapitre 16, Surveillance et alertes*) ou l'équilibreur (« balancer »). Pour répertorier tous les modules disponibles, exécutez :

```
cephuser@adm > ceph mgr module ls
{
  "enabled_modules": [
    "restful",
    "status"
  ],
  "disabled_modules": [
    "dashboard"
  ]
}
```

Pour activer ou désactiver un module spécifique, exécutez :

```
cephuser@adm > ceph mgr module enable MODULE-NAME
```

Par exemple :

```
cephuser@adm > ceph mgr module disable dashboard
```

Pour répertorier les services fournis par les modules activés, exécutez :

```
cephuser@adm > ceph mgr services
{
  "dashboard": "http://myserver.com:7789/",
  "restful": "https://myserver.com:8789/"
}
```

29.1 Équilibreur

Le module de l'équilibreur optimise la distribution des groupes de placement (PG) sur les OSD pour assurer un déploiement plus équilibré. Bien que le module soit activé par défaut, il est inactif. Il prend en charge les deux modes suivants : crush-compat et upmap.



Astuce : État et configuration actuels de l'équilibreur

Pour afficher l'état actuel de l'équilibreur et les informations de configuration, exécutez :

```
cephuser@adm > ceph balancer status
```

29.1.1 Mode « crush-compat »

En mode « crush-compat », l'équilibreur ajuste les définitions de repondération des OSD pour améliorer la distribution des données. Il déplace les groupes de placement entre les OSD, ce qui entraîne temporairement un état de grappe `HEALTH_WARN` en raison de groupes de placement mal placés.



Astuce : activation du mode

Bien que le mode « crush-compat » soit le mode par défaut, nous vous recommandons de l'activer explicitement :

```
cephuser@adm > ceph balancer mode crush-compat
```

29.1.2 Planification et exécution de l'équilibrage des données

À l'aide du module de l'équilibreur, vous pouvez créer un plan d'équilibrage des données. Vous pouvez ensuite exécuter le plan manuellement ou laisser en permanence l'équilibreur répartir les groupes de placement.

Pour décider d'exécuter l'équilibreur en mode manuel ou automatique, vous devez tenir compte de plusieurs facteurs, tels que le déséquilibre actuel des données, la taille de la grappe, le nombre de groupes de placement ou l'activité E/S. Nous vous recommandons de créer un plan initial et de l'exécuter à un moment de faible charge E/S dans la grappe. La raison est que le déséquilibre initial sera probablement considérable et qu'il est bon de limiter l'impact sur les clients. Après une première exécution manuelle, vous pouvez envisager d'activer le mode automatique et veiller à ce que le trafic de rééquilibrage reste sous une charge normale d'E/S. Les améliorations apportées à la distribution des groupes de placement doivent être évaluées par rapport au trafic de rééquilibrage causé par l'équilibreur.



Astuce : fraction mobile des groupes de placement

Pendant le processus d'équilibrage, le module de l'équilibreur limite les mouvements de groupes de placement, de sorte que seule une fraction configurable de groupes de placement est déplacée. La valeur par défaut est de 5 %, mais vous pouvez ajuster la fraction, à 9 % par exemple, en exécutant la commande suivante :

```
cephuser@adm > ceph config set mgr target_max_misplaced_ratio .09
```

Pour créer et exécuter un plan d'équilibrage, procédez comme suit :

1. Vérifiez le score actuel de la grappe :

```
cephuser@adm > ceph balancer eval
```

2. Créez un plan. Par exemple, « great_plan » :

```
cephuser@adm > ceph balancer optimize great_plan
```

3. Vérifiez les changements que « great_plan » entraînera :

```
cephuser@adm > ceph balancer show great_plan
```

4. Vérifiez le score potentiel de la grappe si vous décidez d'appliquer « great_plan » :

```
cephuser@adm > ceph balancer eval great_plan
```

5. Exécutez « great_plan » une seule fois :

```
cephuser@adm > ceph balancer execute great_plan
```

6. Observez l'équilibrage de la grappe avec la commande **ceph -s**. Si vous êtes satisfait du résultat, activez l'équilibrage automatique :

```
cephuser@adm > ceph balancer on
```

Si vous décidez par la suite de désactiver l'équilibrage automatique, exécutez la commande suivante :

```
cephuser@adm > ceph balancer off
```



Astuce : équilibrage automatique sans plan initial

Vous pouvez activer l'équilibrage automatique sans exécuter de plan initial. Dans ce cas, attendez-vous à un rééquilibrage potentiellement long des groupes de placement.

29.2 Activation du module de télémétrie

Le plug-in de télémétrie envoie au projet Ceph des données anonymes sur la grappe dans laquelle le plug-in est en cours d'exécution.

Ce composant (en option) contient des compteurs et des statistiques sur la façon dont la grappe a été déployée, la version de Ceph, la distribution des hôtes et d'autres paramètres qui aident le projet à mieux comprendre la façon dont Ceph est utilisé. Il ne reprend pas de données sensibles comme les noms de réserves, d'objets ou d'hôtes, ni le contenu des objets.

Le but du module de télémétrie est de fournir une boucle de rétroaction automatisée pour les développeurs afin d'aider à quantifier les taux d'adoption et le suivi, ou de repérer les points à expliquer davantage ou à mieux valider lors de la configuration pour éviter des résultats indésirables.



Note

Le module de télémétrie nécessite que les noeuds Ceph Manager puissent transmettre des données aux serveurs en amont via HTTPS. Assurez-vous que les pare-feu de votre entreprise permettent cette opération.

1. Pour activer le module de télémétrie :

```
cephuser@adm > ceph mgr module enable telemetry
```



Note

Cette commande vous permet uniquement de visualiser vos données localement. Cette commande ne partage pas vos données avec la communauté Ceph.

2. Pour permettre au module de télémétrie de commencer à partager des données :

```
cephuser@adm > ceph telemetry on
```


3. Pour désactiver le partage de données de télémétrie :

```
cephuser@adm > ceph telemetry off
```

4. Pour générer un rapport JSON pouvant être imprimé :

```
cephuser@adm > ceph telemetry show
```

5. Pour ajouter un contact et une description au rapport :

```
cephuser@adm > ceph config set mgr mgr/telemetry/contact John Doe  
john.doe@example.com  
cephuser@adm > ceph config set mgr mgr/telemetry/description 'My first Ceph cluster'
```

6. Le module compile et envoie un nouveau rapport toutes les 24 heures par défaut. Pour ajuster cet intervalle :

```
cephuser@adm > ceph config set mgr mgr/telemetry/interval HOURS
```

30 Authentification avec cephx

Pour identifier les clients et les protéger contre les attaques de l'homme du milieu, Ceph fournit son système d'authentification `cephx`. Dans ce contexte, les *clients* désignent soit des humains (comme l'administrateur), soit des services/daemons liés à Ceph, par exemple des OSD, des moniteurs ou des instances Object Gateway.



Note

Le protocole `cephx` ne gère pas le chiffrement dans le transport de données, tel que TLS/SSL.

30.1 Architecture d'authentification

`cephx` utilise des clés secrètes partagées pour l'authentification, ce qui signifie que le client et les moniteurs Ceph possèdent une copie de la clé secrète du client. Le protocole d'authentification permet aux deux parties de se prouver qu'elles possèdent une copie de la clé sans vraiment la révéler. Cela permet une authentification mutuelle, ce qui signifie que la grappe est « certaine » que l'utilisateur possède la clé secrète et que l'utilisateur est certain que la grappe possède également une copie de la clé secrète.

La fonctionnalité d'évolutivité des clés de Ceph évite d'avoir une interface centralisée vers le magasin d'objets Ceph. Cela signifie que les clients Ceph peuvent interagir directement avec les OSD. Pour protéger les données, Ceph fournit son système d'authentification `cephx`, qui authentifie les clients Ceph.

Chaque moniteur peut authentifier les clients et distribuer les clés, il n'y a donc pas de point de défaillance unique ou de goulot d'étranglement lors de l'utilisation de `cephx`. Le moniteur renvoie une structure de données d'authentification qui contient une clé de session permettant d'obtenir des services Ceph. Cette clé de session est elle-même chiffrée avec la clé secrète permanente du client de sorte que seul celui-ci peut demander des services aux moniteurs Ceph. Le client utilise ensuite la clé de session pour demander les services souhaités au moniteur, et celui-ci fournit au client un ticket qui authentifiera le client auprès des OSD qui traitent réellement les données. Les moniteurs Ceph et les OSD partagent un secret de sorte que le client peut utiliser le ticket fourni par le moniteur avec n'importe quel serveur OSD ou de métadonnées de la grappe. Les tickets `cephx` expirent de sorte qu'un attaquant ne peut pas utiliser un ticket périmé ou une clé de session périmée obtenus de façon illégitime.

Pour utiliser `cephx`, un administrateur doit configurer tout d'abord les clients/utilisateurs. Dans le diagramme suivant, l'utilisateur `client.admin` appelle `ceph auth get-or-create-key` à partir de la ligne de commande afin de générer un nom d'utilisateur et une clé secrète. Le sous-système `auth` de Ceph génère le nom et la clé de l'utilisateur, stocke une copie avec le ou les moniteurs et transmet le secret de l'utilisateur à l'utilisateur `client.admin`. Cela signifie que le client et le moniteur partagent une clé secrète.

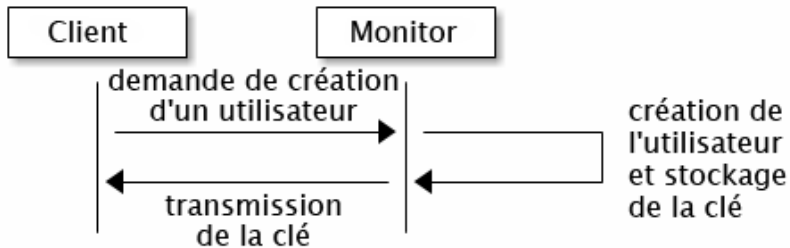


FIGURE 30.1 : AUTHENTIFICATION `cephx` DE BASE

Pour s'authentifier auprès du moniteur, le client transmet le nom d'utilisateur au moniteur. Le moniteur génère une clé de session et la chiffre avec la clé secrète associée au nom d'utilisateur, puis transmet le ticket chiffré au client. Le client déchiffre ensuite les données avec la clé secrète partagée pour récupérer la clé de session. La clé de session identifie l'utilisateur pour la session en cours. Le client demande alors un ticket lié à l'utilisateur et signé par la clé de session. Le moniteur génère un ticket, le chiffre avec la clé secrète de l'utilisateur et le transmet au client. Le client déchiffre le ticket et l'utilise pour signer les requêtes envoyées aux OSD et aux serveurs de métadonnées dans toute la grappe.

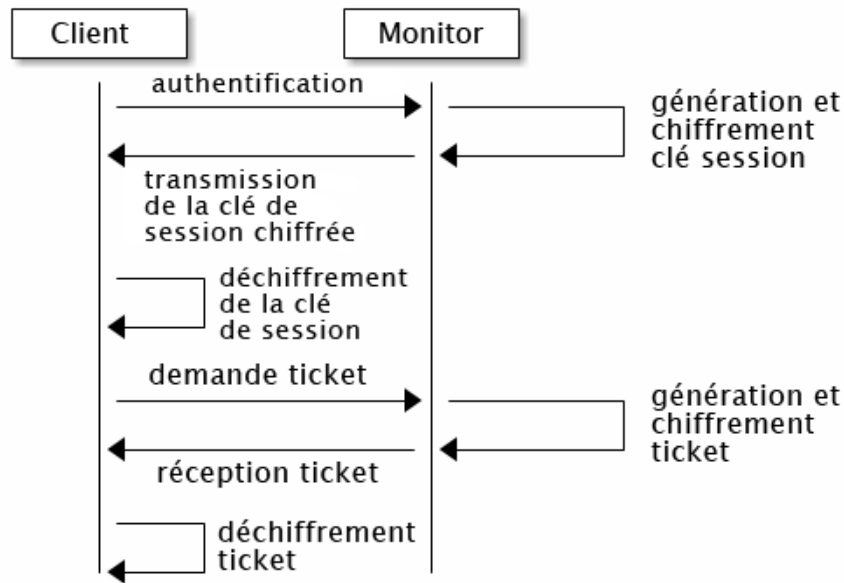


FIGURE 30.2 : **cephx D'AUTHENTIFICATION**

Le protocole cephx authentifie les communications en cours entre la machine cliente et les serveurs Ceph. Chaque message envoyé entre un client et un serveur après l'authentification initiale est signé à l'aide d'un ticket que les moniteurs, les OSD et les serveurs de métadonnées peuvent vérifier avec leur secret partagé.

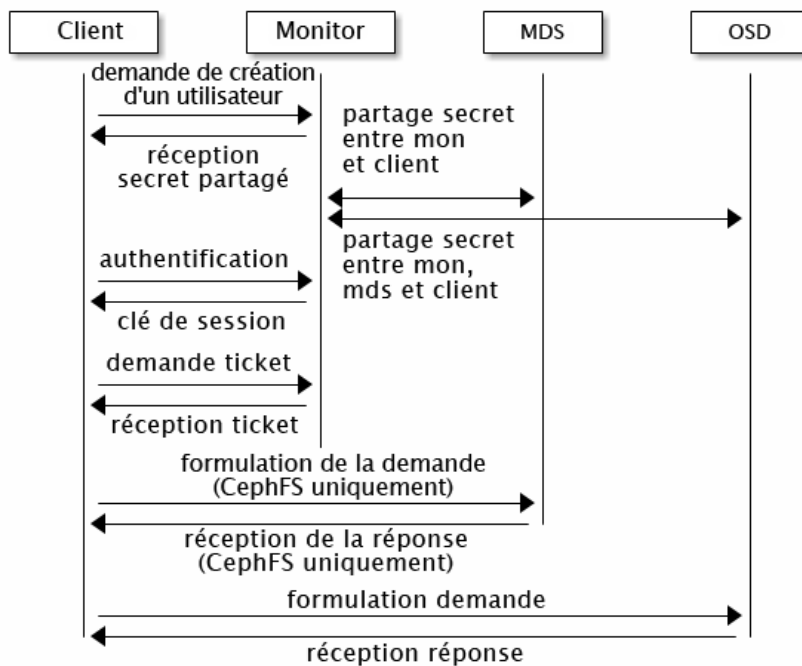


FIGURE 30.3 : AUTHENTIFICATION cephx - MDS ET OSD

! Important

La protection offerte par cette authentification est comprise entre le client Ceph et les hôtes de la grappe Ceph. L'authentification n'est pas étendue au-delà du client Ceph. Si l'utilisateur accède au client Ceph à partir d'un hôte distant, l'authentification Ceph n'est pas appliquée à la connexion entre l'hôte de l'utilisateur et l'hôte client.

30.2 Les zones de gestion principales

Cette section décrit les utilisateurs du client Ceph ainsi que leurs procédures d'authentification et d'autorisation auprès de la grappe de stockage Ceph. Les *utilisateurs* sont des individus ou des acteurs du système, tels que des applications qui s'appuient sur les clients Ceph pour interagir avec les daemons de la grappe de stockage Ceph.

Lorsque Ceph s'exécute avec l'authentification et l'autorisation activées (ce qui est le cas par défaut), vous devez indiquer un nom d'utilisateur et un trousseau contenant la clé secrète de l'utilisateur indiqué (généralement par le biais de la ligne de commande). Si vous n'indiquez pas de nom d'utilisateur, Ceph emploie `client.admin` comme nom d'utilisateur par défaut. Si vous

n'indiquez pas de trousseau de clés, Ceph recherche le paramètre de trousseau de clés dans le fichier de configuration Ceph. Par exemple, si vous exécutez la commande `ceph health` sans indiquer de nom d'utilisateur ou de trousseau de clés, Ceph l'interprète comme suit :

```
cephuser@adm > ceph -n client.admin --keyring=/etc/ceph/ceph.client.admin.keyring health
```

Vous pouvez également utiliser la variable d'environnement `CEPH_ARGS` pour ne pas avoir à saisir une nouvelle fois le nom d'utilisateur et le secret.

30.2.1 Informations de base

Quel que soit le type de client Ceph (par exemple, périphérique de bloc, stockage d'objets, système de fichiers, API native), Ceph stocke toutes les données en tant qu'objets dans des *réserves*. Les utilisateurs Ceph doivent avoir accès aux réserves pour lire et écrire des données. De plus, les utilisateurs Ceph doivent posséder des autorisations d'exécution pour utiliser les commandes d'administration de Ceph. Les concepts suivants vous aideront à comprendre la gestion des utilisateurs Ceph.

30.2.1.1 Utilisateur

Un utilisateur est un acteur individuel ou un acteur système, tel qu'une application. La création d'utilisateurs vous permet de contrôler qui (ou quoi) peut accéder à votre grappe de stockage Ceph, à ses réserves et aux données des réserves.

Ceph utilise des *types* d'utilisateurs. Pour les besoins de gestion des utilisateurs, le type est toujours `client`. Ceph identifie les utilisateurs dans le format point (.) délimité, constitué du type d'utilisateur et de l'ID utilisateur. Par exemple, `TYPE.ID`, `client.admin` ou `client.user1`. Le typage de l'utilisateur se justifie, car les moniteurs, les OSD et les serveurs de métadonnées Ceph emploient également le protocole cephx, mais ne sont pas clients. La distinction du type d'utilisateur permet de distinguer les utilisateurs clients des autres utilisateurs, en rationalisant le contrôle d'accès, la surveillance des utilisateurs et la traçabilité.

Parfois, le type d'utilisateur de Ceph peut sembler ambigu, car la ligne de commande Ceph vous permet de spécifier un utilisateur avec ou sans le type, en fonction de votre syntaxe de la ligne de commande. Si vous spécifiez `--user` ou `--id`, vous pouvez omettre le type. Ainsi, `client.user1` peut être entré simplement en tant que `user1`. Si vous spécifiez `--name` ou `-n`, vous devez spécifier le type et le nom, tels que `client.user1`. Nous vous recommandons d'utiliser le type et le nom comme meilleure pratique à chaque fois que c'est possible.



Note

Un utilisateur de grappe de stockage Ceph n'est pas identique à un utilisateur de stockage d'objets Ceph ou à un utilisateur de système de fichiers Ceph. La passerelle Ceph Object Gateway s'appuie sur un utilisateur de grappe de stockage Ceph pour communiquer entre le daemon passerelle et la grappe de stockage, mais la passerelle dispose de sa propre fonctionnalité de gestion des utilisateurs pour les utilisateurs finaux. Le système de fichiers Ceph utilise la sémantique POSIX. L'espace utilisateur qui lui est associé est distinct d'un utilisateur de grappe de stockage Ceph.

30.2.1.2 Autorisation et fonctions

Ceph utilise le terme « fonctions » (« capabilities » ou « caps » en anglais) pour décrire l'autorisation d'un utilisateur authentifié à exploiter les moniteurs, les OSD et les serveurs de métadonnées. Les fonctions peuvent aussi restreindre l'accès aux données au sein d'une réserve ou d'un espace de noms de réserve. Un administrateur Ceph définit les fonctions d'un utilisateur lors de la création ou de la mise à jour de celui-ci.

La syntaxe d'une fonction obéit au format suivant :

```
daemon-type 'allow capability' [...]
```

Voici une liste des fonctions pour chaque type de service :

Fonctions du moniteur

Inclut r, w, x et allow profile fonction.

```
mon 'allow rwx'  
mon 'allow profile osd'
```

Fonctions OSD

Inclut r, w, x, class-read, class-write et profile osd. En outre, les fonctions OSD autorisent également les paramètres de réserve et d'espace de noms.

```
osd 'allow capability' [pool=poolname] [namespace=namespace-name]
```

Fonction MDS

Requiert seulement allow ou zéro paramètre.

```
mds 'allow'
```

Les entrées suivantes décrivent chaque fonction :

allow

Précède les paramètres d'accès d'un daemon. Implique rw pour MDS uniquement.

r

Donne à l'utilisateur un accès en lecture. Requis avec les moniteurs pour récupérer la carte CRUSH.

w

Donne à l'utilisateur un accès en écriture aux objets.

x

Donne à l'utilisateur les moyens d'appeler des méthodes de classe (lire et écrire) et d'effectuer des opérations auth sur les moniteurs.

class-read

Donne à l'utilisateur la possibilité d'appeler des méthodes de lecture de classe. Sous-ensemble de x.

class-write

Donne à l'utilisateur la possibilité d'appeler des méthodes d'écriture de classe. Sous-ensemble de x.

Fournit à l'utilisateur des autorisations de lecture, d'écriture et d'exécution pour un daemon/une réserve en particulier, et autorise l'utilisateur à exécuter des commandes d'administration.

profile osd

Donne à un utilisateur les autorisations de connexion à d'autres OSD ou moniteurs en tant qu'OSD. Conféré aux OSD pour leur permettre de gérer le trafic de réplication et le rapport d'état.

profile mds

Donne à un utilisateur les autorisations de connexion à d'autres MDS ou moniteurs en tant que MDS.

profile bootstrap-osd

Donne à un utilisateur les autorisations de démarrage d'un OSD. Délégue aux outils de déploiement afin qu'ils disposent des autorisations d'ajout de clés lors du démarrage d'un OSD.

profile bootstrap-mds

Donne à un utilisateur les autorisations de démarrage d'un serveur de métadonnées. Délégué aux outils de déploiement afin qu'ils disposent des autorisations d'ajout de clés lors du démarrage d'un serveur de métadonnées.

30.2.1.3 Réserves

Une réserve est une partition logique dans laquelle les utilisateurs stockent des données. Dans les déploiements Ceph, il est courant de créer une réserve en tant que partition logique pour des types de données similaires. Par exemple, lors du déploiement de Ceph en tant que serveur dorsal pour OpenStack, un déploiement type comporterait des réserves pour les volumes, les images, les sauvegardes et les machines virtuelles, ainsi que des utilisateurs, tels que `client.glance` ou `client.cinder`.

30.2.2 Gestion des utilisateurs

La fonctionnalité de gestion des utilisateurs permet aux administrateurs de grappe Ceph de créer, mettre à jour et supprimer des utilisateurs directement dans la grappe Ceph.

Lorsque vous créez ou supprimez des utilisateurs dans la grappe Ceph, vous devrez distribuer les clés aux clients pour pouvoir les ajouter aux trousseaux de clés. Reportez-vous à la [Section 30.2.3, « Gestion des trousseaux »](#) pour plus d'informations.

30.2.2.1 Liste des utilisateurs

Pour dresser la liste des utilisateurs de votre grappe, exécutez la commande suivante :

```
cephuser@adm > ceph auth list
```

Ceph dresse la liste de tous les utilisateurs de votre grappe. Par exemple, dans une grappe de deux noeuds, la commande `ceph auth list` produit une liste similaire à celle-ci :

```
installed auth entries:

osd.0
    key: AQCvCbtToC6MDhAATtuT70Sl+DymPCfDSsyV4w==
    caps: [mon] allow profile osd
    caps: [osd] allow *
osd.1
    key: AQC4CbtTCFJBChAAVq5spj0ff4eHZICxIOVZeA==
```

```

    caps: [mon] allow profile osd
    caps: [osd] allow *
client.admin
    key: AQBHCbtT6APDHhAA5W00cBchwKQjh3dkKsyPjw==
    caps: [mds] allow
    caps: [mon] allow *
    caps: [osd] allow *
client.bootstrap-mds
    key: AQBICbtT0K9uGBAAdbE5zcIGHZL3T/u2g6EBww==
    caps: [mon] allow profile bootstrap-mds
client.bootstrap-osd
    key: AQBHCbtT4Gxq0RAADE5u7RkpCN/oo4e5W0uBtw==
    caps: [mon] allow profile bootstrap-osd

```



Note : notation TYPE.ID

La notation TYPE.ID s'applique aux utilisateurs ; par exemple, osd.0 définit un utilisateur du type osd et avec l'ID 0. client.admin est un utilisateur du type client et avec l'ID admin. Notez également que chaque entrée possède une entrée key: valeur et une ou plusieurs entrées caps :.

L'option -o nom_fichier avec ceph auth list vous permet d'enregistrer la sortie dans un fichier.

30.2.2.2 Obtention d'informations sur les utilisateurs

Pour récupérer un utilisateur, une clé et des fonctions spécifiques, exécutez la commande suivante :

```
cephuser@adm > ceph auth get TYPE.ID
```

Par exemple :

```

cephuser@adm > ceph auth get client.admin
exported keyring for client.admin
[client.admin]
key = AQA19uZUqIwkHxAAFuUwvq0eJD4S173oFRxe0g==
caps mds = "allow"
caps mon = "allow *"
caps osd = "allow *"

```

Les développeurs peuvent également exécuter la commande suivante :

```
cephuser@adm > ceph auth export TYPE.ID
```

La commande **auth export** ressemble beaucoup à la commande **auth get**, mais permet également d'afficher l'ID d'authentification interne.

30.2.2.3 Ajout d'utilisateurs

L'ajout d'un utilisateur crée un nom d'utilisateur (TYPE.ID), une clé secrète et toutes les fonctionnalités incluses dans la commande destinée à créer l'utilisateur.

La clé d'un utilisateur permet à l'utilisateur de s'authentifier auprès de la grappe de stockage Ceph. Les fonctions de l'utilisateur autorisent celui-ci à lire, écrire ou exécuter sur les moniteurs Ceph (mon), les OSD Ceph (osd) ou les serveurs de métadonnées Ceph (mds).

Quelques commandes permettent d'ajouter un utilisateur :

ceph auth add

Cette commande est la méthode canonique pour ajouter un utilisateur. Elle crée l'utilisateur, génère une clé et ajoute les éventuelles fonctions indiquées.

ceph auth get-or-create

Cette commande est souvent la façon la plus pratique de créer un utilisateur, car elle renvoie un keyfile avec le nom d'utilisateur (entre parenthèses) et la clé. Si l'utilisateur existe déjà, cette commande renvoie simplement le nom d'utilisateur et la clé dans le format du keyfile. L'option -o nom_fichier vous permet d'enregistrer la sortie dans un fichier.

ceph auth get-or-create-key

Cette commande est un moyen pratique de créer un utilisateur et de renvoyer la clé de l'utilisateur (uniquement). Cela est utile pour les clients qui n'ont besoin que de la clé (libvirt, par exemple). Si l'utilisateur existe déjà, cette commande renvoie simplement la clé. L'option -o nom_fichier vous permet d'enregistrer la sortie dans un fichier.

Lors de la création d'utilisateurs clients, vous pouvez définir un utilisateur sans fonction. Un utilisateur sans fonction peut s'authentifier, mais rien de plus. Un tel client ne peut pas extraire une assignation de grappe depuis le moniteur. Cependant, vous pouvez créer un utilisateur sans fonctions si vous souhaitez différer l'ajout de fonctions à une date ultérieure en utilisant la commande **ceph auth caps**.

Un utilisateur normal possède au moins des autorisations de lecture sur le moniteur Ceph et des autorisations de lecture et écriture sur les OSD Ceph. De plus, les autorisations OSD d'un utilisateur sont souvent limitées à l'accès à une réserve particulière.

```
cephuser@adm > ceph auth add client.john mon 'allow r' osd \
```

```
'allow rw pool=liverpool'
cephuser@adm > ceph auth get-or-create client.paul mon 'allow r' osd \
'allow rw pool=liverpool'
cephuser@adm > ceph auth get-or-create client.george mon 'allow r' osd \
'allow rw pool=liverpool' -o george.keyring
cephuser@adm > ceph auth get-or-create-key client.ringo mon 'allow r' osd \
'allow rw pool=liverpool' -o ringo.key
```



Important

Si vous fournissez à un utilisateur des autorisations d'accès OSD *sans* limiter l'accès à des réserves en particulier, l'utilisateur aura accès à *toutes* les réserves de la grappe.

30.2.2.4 Modification des fonctions utilisateur

La commande **ceph auth caps** vous permet d'indiquer un utilisateur et de modifier ses fonctions. La définition de nouvelles fonctionnalités écrase les fonctions en cours. Pour afficher les fonctions actuelles, exécutez **ceph auth get TYPE D'UTILISATEUR.USERID**. Pour ajouter des fonctions, vous devez également indiquer les fonctions existants lors de l'utilisation du formulaire suivant :

```
cephuser@adm > ceph auth caps USERTYPE.USERID daemon 'allow [r|w|x|*|...] \
[pool=pool-name] [namespace=namespace-name]' [daemon 'allow [r|w|x|*|...] \
[pool=pool-name] [namespace=namespace-name]']
```

Par exemple :

```
cephuser@adm > ceph auth get client.john
cephuser@adm > ceph auth caps client.john mon 'allow r' osd 'allow rw pool=prague'
cephuser@adm > ceph auth caps client.paul mon 'allow rw' osd 'allow r pool=prague'
cephuser@adm > ceph auth caps client.brian-manager mon 'allow *' osd 'allow *'
```

Pour supprimer une fonction, vous avez la possibilité de la réinitialiser. Si vous souhaitez que l'utilisateur n'ait pas accès à un daemon particulier précédemment défini, indiquez une chaîne vide :

```
cephuser@adm > ceph auth caps client.ringo mon ' ' osd ' '
```

30.2.2.5 Suppression d'utilisateurs

Pour supprimer un utilisateur, exécutez **auth ceph del** :

```
cephuser@adm > ceph auth del TYPE.ID
```

où TYPE correspond à client, osd, LUN ou mds, et ID correspond au nom d'utilisateur ou à l'ID du daemon.

Si vous avez créé des utilisateurs avec des autorisations strictement pour une réserve qui n'existe plus, vous devez également envisager de supprimer ces utilisateurs.

30.2.2.6 Impression de la clé d'un utilisateur

Pour imprimer la clé d'authentification de l'utilisateur vers la sortie standard, exécutez la commande suivante :

```
cephuser@adm > ceph auth print-key TYPE.ID
```

où TYPE correspond à client, osd, LUN ou mds, et ID correspond au nom d'utilisateur ou à l'ID du daemon.

L'impression de la clé d'un utilisateur est utile lorsque vous devez renseigner le logiciel client avec la clé d'un utilisateur (telle que libvirt), comme dans l'exemple suivant :

```
# mount -t ceph host:/ mount_point \  
-o name=client.user,secret=`ceph auth print-key client.user`
```

30.2.2.7 Importation d'utilisateurs

Pour importer un ou plusieurs utilisateurs, exécutez **ceph auth import** et indiquez un trousseau de clés :

```
cephuser@adm > ceph auth import -i /etc/ceph/ceph.keyring
```



Note

La grappe de stockage Ceph ajoute de nouveaux utilisateurs, leurs clés et leurs fonctions, puis les met à jour.

30.2.3 Gestion des trousseaux

Lorsque vous accédez à Ceph via un client Ceph, le client recherche un trousseau de clés local. Ceph prédéfinit le paramètre de trousseau de clés avec les quatre noms de trousseau de clés suivants par défaut de sorte que vous n'avez pas besoin de les définir dans votre fichier de configuration Ceph, sauf si vous souhaitez remplacer les valeurs par défaut :

```
/etc/ceph/cluster.name.keyring  
/etc/ceph/cluster.keyring  
/etc/ceph/keyring  
/etc/ceph/keyring.bin
```

La métavariable *cluster* correspond à votre nom de grappe Ceph tel qu'il est défini par le nom du fichier de configuration Ceph. *ceph.conf* signifie que le nom de la grappe est *ceph*, donc *ceph.keyring*. La métavariable *name* correspond au type d'utilisateur et à l'ID utilisateur, par exemple *client.admin*, par conséquent *ceph.client.admin.keyring*.

Après avoir créé un utilisateur (par exemple *client.ringo*), vous devez obtenir la clé et l'ajouter à un trousseau de clés sur un client Ceph afin que l'utilisateur puisse accéder à la grappe de stockage Ceph.

La [Section 30.2, « Les zones de gestion principales »](#) explique comment lister, obtenir, ajouter, modifier et supprimer des utilisateurs directement dans la grappe de stockage Ceph. Toutefois, Ceph fournit également l'utilitaire **ceph authtool** pour vous permettre de gérer des trousseaux de clés à partir d'un client Ceph.

30.2.3.1 Création d'un trousseau de clés

Les procédures de la [Section 30.2, « Les zones de gestion principales »](#) qui permettent de créer des utilisateurs vous obligent à fournir des clés d'utilisateur aux clients Ceph afin qu'ils puissent récupérer la clé de l'utilisateur indiqué et s'authentifier avec la grappe de stockage Ceph. Les clients Ceph accèdent aux trousseaux de clés pour rechercher un nom d'utilisateur et récupérer la clé correspondante :

```
cephuser@adm > ceph-authtool --create-keyring /path/to/keyring
```

Lorsque vous créez un trousseau de clés avec plusieurs utilisateurs, nous vous recommandons de reprendre le nom de la grappe (par exemple, *grappe.keyring*) pour le nom de fichier du trousseau de clés et l'enregistrer dans le répertoire */etc/ceph* de sorte que le paramètre par

défaut de la configuration du trousseau de clés récupère le nom du fichier sans que vous ayez à l'indiquer dans la copie locale de votre fichier de configuration Ceph. Par exemple, créez `ceph.keyring` en exécutant la commande suivante :

```
cephuser@adm > ceph-authtool -C /etc/ceph/ceph.keyring
```

Lorsque vous créez un trousseau de clés avec un seul utilisateur, nous vous recommandons de sélectionner le nom de la grappe, le type d'utilisateur et le nom d'utilisateur afin de l'enregistrer dans l'annuaire `/etc/ceph`. Par exemple, `ceph.client.admin.keyring` pour l'utilisateur `client.admin`.

30.2.3.2 Ajout d'un utilisateur à un trousseau de clés

Lorsque vous ajoutez un utilisateur à la grappe de stockage Ceph (voir [Section 30.2.2.3, « Ajout d'utilisateurs »](#)), vous pouvez récupérer l'utilisateur, la clé et les fonctions, et enregistrer l'utilisateur dans un trousseau de clés.

Si vous souhaitez associer un utilisateur par trousseau de clés, la commande `ceph auth get` suivie de l'option `-o` enregistrera la sortie dans le format du fichier de trousseau de clés. Par exemple, pour créer un trousseau de clés pour l'utilisateur `client.admin`, exécutez la commande suivante :

```
cephuser@adm > ceph auth get client.admin -o /etc/ceph/ceph.client.admin.keyring
```

Lorsque vous souhaitez importer des utilisateurs vers un trousseau de clés, vous pouvez utiliser `ceph-authtool` afin d'indiquer le trousseau de destination et le trousseau source :

```
cephuser@adm > ceph-authtool /etc/ceph/ceph.keyring \  
--import-keyring /etc/ceph/ceph.client.admin.keyring
```



Important

Si votre trousseau de clés est compromis, supprimez votre clé du répertoire `/etc/ceph` et recréez une nouvelle clé en suivant les instructions de la [Section 30.2.3.1, « Création d'un trousseau de clés »](#).

30.2.3.3 Création d'un utilisateur

La commande **ceph auth add** de Ceph permet de créer un utilisateur directement dans la grappe de stockage Ceph. Cependant, vous pouvez également créer un utilisateur, des clés et des fonctionnalités directement dans un trousseau de clés du client Ceph. Vous pouvez ensuite importer l'utilisateur vers la grappe de stockage Ceph :

```
cephuser@adm > ceph-authtool -n client.ringo --cap osd 'allow rwx' \
--cap mon 'allow rwx' /etc/ceph/ceph.keyring
```

Vous pouvez également créer un trousseau de clés et lui ajouter un nouvel utilisateur simultanément :

```
cephuser@adm > ceph-authtool -C /etc/ceph/ceph.keyring -n client.ringo \
--cap osd 'allow rwx' --cap mon 'allow rwx' --gen-key
```

Dans les scénarios précédents, le nouvel utilisateur `client.ringo` réside uniquement dans le trousseau de clés. Pour ajouter le nouvel utilisateur à la grappe de stockage Ceph, vous devez encore ajouter celui-ci à la grappe :

```
cephuser@adm > ceph auth add client.ringo -i /etc/ceph/ceph.keyring
```

30.2.3.4 Modification d'utilisateurs

Pour modifier les fonctionnalités d'un enregistrement utilisateur dans un trousseau de clés, indiquez le trousseau de clés et l'utilisateur suivis des fonctionnalités :

```
cephuser@adm > ceph-authtool /etc/ceph/ceph.keyring -n client.ringo \
--cap osd 'allow rwx' --cap mon 'allow rwx'
```

Pour mettre à jour l'utilisateur modifié dans l'environnement de grappe Ceph, vous devez importer les modifications du trousseau de clés dans l'entrée utilisateur de la grappe Ceph :

```
cephuser@adm > ceph auth import -i /etc/ceph/ceph.keyring
```

Reportez-vous à la [Section 30.2.2.7, « Importation d'utilisateurs »](#) pour plus d'informations sur la mise à jour d'un utilisateur de grappe de stockage Ceph à partir d'un trousseau de clés.

30.2.4 Utilisation de la ligne de commande

La commande **ceph** prend en charge les options suivantes liées à la manipulation du nom et du secret de l'utilisateur :

--id ou --user

Ceph identifie les utilisateurs avec un type et un ID (*TYPE.ID*, tel que `client.admin` ou `client.user1`). Les options `id`, `name` et `-n` vous permettent d'indiquer la partie ID du nom d'utilisateur (par exemple, `admin` ou `user1`). Vous pouvez indiquer l'utilisateur avec `--id` et omettre le type. Par exemple, pour indiquer l'utilisateur `client.foo`, entrez la commande suivante :

```
cephuser@adm > ceph --id foo --keyring /path/to/keyring health
cephuser@adm > ceph --user foo --keyring /path/to/keyring health
```

--name ou -n

Ceph identifie les utilisateurs avec un type et un ID (*TYPE.ID*, tel que `client.admin` ou `client.user1`). Les options `--name` et `-n` vous donnent les moyens d'indiquer le nom d'utilisateur complet. Vous devez indiquer le type d'utilisateur (généralement `client`) avec l'ID utilisateur :

```
cephuser@adm > ceph --name client.foo --keyring /path/to/keyring health
cephuser@adm > ceph -n client.foo --keyring /path/to/keyring health
```

--keyring

Chemin d'accès au trousseau de clés contenant un ou plusieurs noms et secrets d'utilisateur. L'option `--secret` fournit la même fonctionnalité, mais elle est inopérante avec Object Gateway, qui utilise `--secret` dans un but différent. Vous pouvez récupérer un trousseau de clés avec **ceph auth get-or-create** et le stocker localement. Il s'agit d'une approche privilégiée, car vous pouvez changer de noms d'utilisateur indépendamment du chemin d'accès aux trousseaux de clés :

```
cephuser@adm > rbd map --id foo --keyring /path/to/keyring mypool/myimage
```

A Mises à jour de la maintenance de Ceph basées sur les versions intermédiaires de « Pacific » en amont

Plusieurs paquetages clés de SUSE Enterprise Storage 7.1 sont basés sur la série de versions Pacific de Ceph. Lorsque le projet Ceph (<https://github.com/ceph/ceph>) publie de nouvelles versions intermédiaires dans la série Pacific, SUSE Enterprise Storage 7.1 est mis à jour pour garantir que le produit bénéficie des derniers rétroports de fonctionnalités et correctifs de bogues en amont.

Ce chapitre contient des résumés des changements notables contenus dans chaque version intermédiaire en amont qui a été ou devrait être incluse dans le produit.

Glossaire

Général

Alertmanager

Binaire unique qui traite les alertes envoyées par le serveur Prometheus et avertit l'utilisateur final.

Arborescence de routage

Terme donné à tout diagramme qui montre les différentes routes qu'un récepteur peut exécuter.

Ceph Dashboard

Application intégrée de gestion et de surveillance Ceph basée sur le Web pour gérer divers aspects et objets de la grappe. Le tableau de bord est implémenté en tant que module Ceph Manager.

Ceph Manager

Ceph Manager ou MGR est le logiciel du gestionnaire Ceph qui centralise tous les états de l'ensemble de la grappe au même endroit.

Ceph Monitor

Ceph Monitor ou MON est le logiciel du moniteur Ceph.

ceph-salt

Fournit des outils pour le déploiement de grappes Ceph gérées par cephadm à l'aide de Salt.

cephadm

cephadm déploie et gère une grappe Ceph en se connectant aux hôtes à partir du daemon du gestionnaire via SSH pour ajouter, supprimer ou mettre à jour les conteneurs du daemon Ceph.

CephFS

Système de fichiers Ceph.

CephX

Protocole d'authentification Ceph. Cephx fonctionne comme Kerberos, mais sans point d'échec unique.

Client Ceph

Collection de composants Ceph qui peuvent accéder à une grappe de stockage Ceph. Il s'agit notamment de l'instance Object Gateway, du périphérique de bloc Ceph, de CephFS et des bibliothèques, modules de kernel et clients FUSE correspondants.

Compartiment

Point qui regroupe d'autres noeuds dans une hiérarchie d'emplacements physiques.

CRUSH, carte CRUSH

Controlled Replication Under Scalable Hashing (Réplication contrôlée sous hachage évolutif) : algorithme qui détermine comment stocker et récupérer des données en calculant les emplacements de stockage de données. CRUSH requiert une assignation de votre grappe pour stocker et récupérer de façon pseudo-aléatoire des données dans les OSD avec une distribution uniforme des données à travers la grappe.

Daemon Ceph OSD

Le daemon **ceph-osd** est le composant de Ceph chargé de stocker les objets sur un système de fichiers local et de fournir un accès à ces objets via le réseau.

Ensemble de règles

Règles qui déterminent le placement des données dans une réserve.

Espace de stockage d'objets Ceph

« Produit », service ou fonctionnalités du stockage d'objets, qui se compose d'une grappe de stockage Ceph et d'une instance Ceph Object Gateway.

Grafana

Solution de surveillance et d'analyse de base de données.

Grappe de stockage Ceph

Ensemble de base du logiciel de stockage qui conserve les données de l'utilisateur. Un tel ensemble comprend des moniteurs Ceph et des OSD.

Groupes d'unités

Les groupes d'unités sont une déclaration d'une ou de plusieurs dispositions OSD pouvant être assignées à des unités physiques. Une disposition OSD définit la manière dont Ceph alloue physiquement l'espace de stockage OSD sur le support en fonction des critères spécifiés.

Module de synchronisation de l'archivage

Module permettant de créer une zone Object Gateway pour conserver l'historique des versions d'objets S3.

Multizone

Noeud

Désigne une machine ou serveur dans une grappe Ceph.

Noeud Admin

Hôte à partir duquel vous exécutez les commandes associées à Ceph pour gérer les hôtes de la grappe.

Noeud OSD

Noeud de grappe qui stocke des données, gère la réplication de données, la récupération, le remplissage, le rééquilibrage et qui fournit des informations de surveillance aux moniteurs Ceph en contrôlant d'autres daemons Ceph OSD.

Object Gateway

Composant de passerelle S3/Swift pour la zone de stockage des objets Ceph. Également appelée passerelle RADOS (RGW).

OSD

Périphérique de stockage d'objets : unité de stockage physique ou logique.

Passerelle Samba

La passerelle Samba assure la liaison avec Active Directory dans le domaine Windows pour authentifier et autoriser les utilisateurs.

Périphérique de bloc RADOS (RBD)

Composant de stockage de blocs pour Ceph. Également appelé périphérique de bloc Ceph.

PG

Placement Group (groupe de placement) : sous-division d'une *réserve* utilisée à des fins d'optimisation des performances.

Prometheus

Toolkit de surveillance et d'alerte des systèmes.

Règle CRUSH

Règle de placement de données CRUSH qui s'applique à une ou plusieurs réserves en particulier.

Réserve

Partitions logiques pour le stockage d'objets, tels que des images de disque.

Samba

Logiciel d'intégration Windows.

Serveur de métadonnées

Le serveur de métadonnées aussi appelé MDS est le logiciel de métadonnées Ceph.

Version ponctuelle

Toute version ad hoc qui inclut uniquement des correctifs de bogues ou de sécurité.

Zone de stockage fiable des objets distribués autonomes fiable (RADOS)

Ensemble de base du logiciel de stockage qui stocke les données de l'utilisateur (MON + OSD).

zonegroup