SUSE Enterprise Storage 7

Guide de déploiement

Guide de déploiement

SUSE Enterprise Storage 7

par Tomáš Bažant, Alexandra Settle, et Liam Proven

Date de publication : 11 déc 2025

https://documentation.suse.com

Table des matières

À propos de ce guide

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 à partir de la version précédente du produit.

SUSE Enterprise Storage 7 est une extension de SUSE Linux Enterprise Server 15 SP2. 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 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

Remarque : 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 :

```
root # 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 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-troubleshooting.html)

Ce guide passe en revue divers problèmes courants lors de l'exécution de SUSE Enterprise Storage 7, 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
- `Utilisateur` : nom de l'utilisateur ou du groupe
- `nom_paquetage` : 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
- `AMD/Intel` Ce paragraphe n'est utile que pour les architectures Intel 64/AMD64. Les flèches marquent le début et la fin du bloc de texte.
  `IBM Z, POWER` 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.

```
root # command
tux > sudo command
```

- Commandes pouvant être exécutées par des utilisateurs non privilégiés.

```
tux > command
```

- Avis

Avertissement : 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.

Remarque : note de remarque

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

Astuce : 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 Cycle de vie et support du produit

Les différents produits SUSE ont des cycles de vie différents. Pour vérifier les dates exactes du cycle de vie de SUSE Enterprise Storage, consultez la page https://www.suse.com/lifecycle/.

4.1 Définitions de support de SUSE

Pour plus d'informations sur nos options et notre politique de support, consultez les adresses https://www.suse.com/support/policy.html et https://www.suse.com/support/programs/long-term-service-pack-support.html.

4.2 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 le nom se termine par `-devel` (contenant les fichiers d'en-tête et les ressources développeurs similaires), qui bénéficient d'un support uniquement conjointement à leur paquetage principal.

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

4.3 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.
- 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. Cela peut être le cas, par exemple, si SUSE découvre qu'une avant-première ne répond pas aux besoins des clients ou du marché, ou ne satisfait pas aux normes des entreprises.

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.

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, reportez-vous au site 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 disposé 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 `cephadm` 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 `root #`, selon les privilèges que la commande concernée nécessite.

6.4 Informations supplémentaires

Pour plus d'informations sur la gestion des clés Ceph, reportez-vous au *Manuel « Guide d'opérations et d'administration », Chapitre 30 « Authentification avec* `cephx` *», Section 30.2 « Les zones de gestion principales »*.

I Présentation de SUSE Enterprise Storage (SES)

1 SES et Ceph

SUSE Enterprise Storage (SES) est un système de stockage distribué conçu pour l'évolutivité, la fiabilité et les performances, et basé sur la technologie Ceph. Une grappe Ceph peut être exécutée sur des serveurs courants dans un réseau standard comme Ethernet. La grappe évolue facilement vers des milliers de serveurs (appelés plus tard noeuds) et dans la plage de pétaoctets. Contrairement aux systèmes conventionnels qui possèdent des tables d'allocation pour stocker et récupérer des données, Ceph utilise un algorithme déterministe pour allouer du stockage des données et ne dispose d'aucune structure centralisée des informations. Ceph suppose que, dans les grappes de stockage, l'ajout ou la suppression de matériel est la règle, et non l'exception. La grappe Ceph automatise les tâches de gestion telles que la distribution et la redistribution des données, la réplication des données, la détection des échecs et la récupération. Ceph peut à la fois s'autoréparer et s'autogérer, ce qui se traduit par une réduction des frais administratifs et budgétaires.

Ce chapitre fournit un aperçu global de SUSE Enterprise Storage 7 et décrit brièvement les éléments les plus importants.

1.1 Fonctions de Ceph

L'environnement Ceph possède les fonctions suivantes :

Évolutivité

Ceph peut s'adapter à des milliers de noeuds et gérer le stockage dans la plage de pétaoctets.

Matériel courant

Aucun matériel spécial n'est requis pour exécuter une grappe Ceph. Pour plus d'informations, reportez-vous au *Chapitre 2, Configuration matérielle requise et recommandations*

Autogestion

La grappe Ceph est autogérée. Lorsque des noeuds sont ajoutés ou supprimés ou échouent, la grappe redistribue automatiquement les données. Elle a également connaissance des disques surchargés.

Aucun point d'échec unique

Aucun noeud d'une grappe ne stocke les informations importantes de manière isolée. Vous pouvez configurer le nombre de redondances.

Logiciels Open Source

Ceph est une solution logicielle Open Source et indépendante du matériel ou de fournisseurs spécifiques.

1.2 Composants de base de Ceph

Pour tirer pleinement parti de la puissance de Ceph, il est nécessaire de comprendre certains des composants et concepts de base. Cette section présente certaines parties de Ceph qui sont souvent référencées dans d'autres chapitres.

1.2.1 RADOS

Le composant de base de Ceph est appelé *RADOS (Reliable Autonomic Distributed Object Store)* (Zone de stockage des objets distribués autonome fiable). Il est chargé de gérer les données stockées dans la grappe. Les données de Ceph sont généralement stockées en tant qu'objets. Chaque objet se compose d'un identificateur et des données.

RADOS fournit les méthodes d'accès suivantes aux objets stockés qui couvrent de nombreux cas d'utilisation :

Object Gateway

Object Gateway est une passerelle HTTP REST pour la zone de stockage des objets RADOS. Elle permet un accès direct aux objets stockés dans la grappe Ceph.

Périphérique de bloc RADOS

Les périphériques de bloc RADOS (RADOS Block Devices, RBD) sont accessibles comme n'importe quel autre périphérique de bloc. Ils peuvent être utilisés, par exemple, en combinaison avec `libvirt` à des fins de virtualisation.

CephFS

Le système de fichiers Ceph est un système de fichiers compatible POSIX.

`librados`

`librados` est une bibliothèque pouvant être utilisée avec de nombreux langages de programmation pour créer une application capable d'interagir directement avec la grappe de stockage.

`librados` est utilisé par Object Gateway et les RBD alors que CephFS s'interface directement avec RADOS *Figure 1.1, « Interfaces avec la zone de stockage des objets Ceph »*.

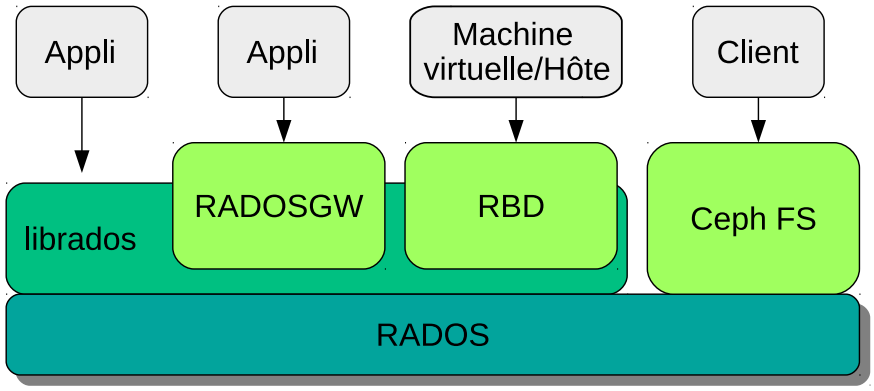

FIGURE 1.1 : INTERFACES AVEC LA ZONE DE STOCKAGE DES OBJETS CEPH

1.2.2 CRUSH

Au coeur d'une grappe Ceph se trouve l'algorithme *CRUSH*. CRUSH est l'acronyme de *Controlled Replication Under Scalable Hashing* (Réplication contrôlée sous hachage évolutif). CRUSH est une fonction qui gère l'allocation de stockage et nécessite relativement peu de paramètres. Cela signifie que seule une petite quantité d'informations est nécessaire pour calculer la position de stockage d'un objet. Les paramètres sont une carte actuelle de la grappe, comprenant l'état de santé, certaines règles de placement définies par l'administrateur et le nom de l'objet qui doit être stocké ou récupéré. Avec ces informations, tous les noeuds de la grappe Ceph sont en mesure de calculer l'emplacement auquel un objet et ses répliques sont stockés. Cela rend l'écriture ou la lecture des données très efficace. CRUSH tente de distribuer équitablement les données sur tous les noeuds de la grappe.

La *carte CRUSH* contient tous les noeuds de stockage et les règles de placement définies par l'administrateur pour le stockage des objets dans la grappe. Elle définit une structure hiérarchique qui correspond généralement à la structure physique de la grappe. Par exemple, les disques contenant des données sont dans des hôtes, les hôtes sont dans des racks, les racks dans des rangées et les rangées dans des centres de données. Cette structure peut être utilisée pour définir des *domaines de défaillance*. Ceph s'assure ensuite que les réplications sont stockées sur différentes branches d'un domaine de défaillance spécifique.

Si le domaine de défaillance est défini sur le niveau rack, les réplications d'objets sont réparties sur différents racks. Cela peut limiter les interruptions de service provoquées par un commutateur défaillant sur un rack. Si une unité de distribution de l'alimentation dessert une rangée

de racks, le domaine de défaillance peut être défini sur rangée. En cas de défaillance de l'unité de distribution de l'alimentation, les données répliquées sont toujours disponibles sur les autres rangées.

1.2.3 Noeuds et daemons Ceph

Dans Ceph, les noeuds sont des serveurs fonctionnant pour la grappe. Ils peuvent exécuter divers types de daemons. Nous recommandons de n'exécuter qu'un seul type de daemon sur chaque noeud, à l'exception des daemons Ceph Manager qui peuvent être colocalisés avec des daemons Ceph Monitor. Chaque grappe nécessite au moins des daemons Ceph Monitor, Ceph Manager, et Ceph OSD :

Noeud Admin

Le *noeud Admin* est un noeud de grappe Ceph à partir duquel vous exécutez des commandes pour gérer la grappe. Le noeud Admin est un point central de la grappe Ceph, car il gère les autres noeuds de la grappe en interrogeant et en instruisant leurs services de Minion Salt.

Ceph Monitor

Les noeuds *Ceph Monitor* (souvent abrégés en *MON*) gèrent des informations sur l'état de santé de la grappe, une carte de tous les noeuds et des règles de distribution des données (reportez-vous à la *Section 1.2.2, « CRUSH »*).

En cas de défaillance ou de conflit, les noeuds Ceph Monitor de la grappe décident, à la majorité, des informations qui sont correctes. Pour former une majorité admissible, il est recommandé de disposer d'un nombre impair de noeuds Ceph Monitor et au minimum de trois d'entre eux.

Si plusieurs sites sont utilisés, les noeuds Ceph Monitor doivent être distribués sur un nombre impair de sites. Le nombre de noeuds Ceph Monitor par site doit être tel que plus de 50 % des noeuds Ceph Monitor restent fonctionnels en cas de défaillance d'un site.

Ceph Manager

Ceph Manager collecte les informations d'état de l'ensemble de la grappe. Le daemon Ceph Manager s'exécute parallèlement aux daemons Ceph Monitor. Il fournit une surveillance supplémentaire et assure l'interface avec les systèmes de surveillance et de gestion externes. Il comprend également d'autres services. Par exemple, l'interface utilisateur Web de Ceph Dashboard s'exécute sur le même noeud que Ceph Manager.

Ceph Manager ne requiert aucune configuration supplémentaire : il suffit de s'assurer qu'il s'exécute.

Ceph OSD

Ceph OSD est un daemon qui gère des périphériques de stockage d'objets (*Object Storage Devices*, OSD) qui sont des unités de stockage physiques ou logiques (disques durs ou partitions). Les périphériques de stockage d'objets peuvent être des disques/partitions physiques ou des volumes logiques. Le daemon prend également en charge la réplication et le rééquilibrage des données en cas d'ajout ou de suppression de noeuds.
Les daemons Ceph OSD communiquent avec les daemons Ceph Monitor et leur fournissent l'état des autres daemons OSD.

Pour utiliser CephFS, Object Gateway, NFS Ganesha ou la passerelle iSCSI, des noeuds supplémentaires sont requis :

Serveur de métadonnées (MDS)

Les métadonnées CephFS sont stockées dans sa propre réserve RADOS (voir *Section 1.3.1, « Réserves »*). Les serveurs de métadonnées agissent comme une couche de mise en cache intelligente pour les métadonnées et sérialisent l'accès en cas de besoin. Cela permet un accès simultané à partir de nombreux clients sans synchronisation explicite.

Object Gateway

Object Gateway est une passerelle HTTP REST pour la zone de stockage des objets RADOS. Elle est compatible avec OpenStack Swift et Amazon S3, et possède sa propre gestion des utilisateurs.

NFS Ganesha

NFS Ganesha fournit un accès NFS à Object Gateway ou à CephFS. NFS Ganesha s'exécute dans l'espace utilisateur au lieu de l'espace kernel, et interagit directement avec Object Gateway ou CephFS.

Passerelle iSCSI

iSCSI est un protocole de réseau de stockage qui permet aux clients d'envoyer des commandes SCSI à des périphériques de stockage SCSI (cibles) sur des serveurs distants.

Passerelle Samba

La passerelle Samba offre un accès Samba aux données stockées sur CephFS.

1.3 Structure de stockage Ceph

1.3.1 Réserves

Les objets qui sont stockés dans une grappe Ceph sont placés dans des *réserves*. Les réserves représentent les partitions logiques de la grappe vers le monde extérieur. Pour chaque réserve, un ensemble de règles peut être défini, par exemple, combien de réplications de chaque objet doivent exister. La configuration standard des réserves est appelée *réserve répliquée*.

Les réserves contiennent généralement des objets, mais peuvent également être configurées pour agir de la même manière qu'un RAID 5. Dans cette configuration, les objets sont stockés sous forme de tranches avec d'autres tranches de codage. Les tranches de codage contiennent les informations redondantes. Le nombre de données et les tranches de codage peuvent être définis par l'administrateur. Dans cette configuration, les réserves sont appelées *réserves codées à effacement* ou *réserves EC*.

1.3.2 Groupes de placement

Les *groupes de placement* (PG) sont utilisés pour la distribution des données au sein d'une réserve. Lorsque vous créez une réserve, un certain nombre de groupes de placement sont définis. Les groupes de placement sont utilisés en interne pour grouper des objets et sont un facteur important en termes de performances d'une grappe Ceph. Le groupe de placement d'un objet est déterminé par le nom de l'objet.

1.3.3 Exemple

Cette section fournit un exemple simplifié de la façon dont Ceph gère les données (reportez-vous à la *Figure 1.2, « Exemple de Ceph à petite échelle »*). Cet exemple ne représente pas une configuration recommandée pour une grappe Ceph. La configuration matérielle comprend trois noeuds de stockage ou noeuds Ceph OSD (`Hôte 1`, `Hôte 2`, `Hôte 3`). Chaque noeud comporte trois disques durs qui sont utilisés comme OSD (`osd.1` à `osd.9`). Les noeuds Ceph Monitor sont ignorés dans cet exemple.

Remarque : différence entre Ceph OSD et OSD

Alors que *Ceph OSD* ou *daemon Ceph OSD* fait référence à un daemon exécuté sur un noeud, le terme *OSD* fait référence au disque logique avec lequel le daemon interagit.

La grappe comporte deux réserves, `Réserve A` et `Réserve B`. Alors que Réserve A ne réplique les objets que deux fois, la résilience pour Réserve B est plus importante et comporte trois réplications pour chaque objet.

Lorsqu'une application place un objet dans une réserve, par exemple via l'API REST, un groupe de placement (`PG1` à `PG4`) est sélectionné en fonction de la réserve et du nom de l'objet. L'algorithme CRUSH calcule ensuite les OSD sur lesquels l'objet est stocké, en fonction du groupe de placement qui contient l'objet.

Dans cet exemple, le domaine de défaillance est défini sur le niveau hôte. Cela garantit que les réplications d'objets sont stockées sur des hôtes différents. En fonction du niveau de réplication défini pour une réserve, l'objet est stocké sur deux ou trois OSD utilisés par le groupe de placement.

Une application qui écrit un objet interagit uniquement avec un Ceph OSD, le Ceph OSD primaire. Le Ceph OSD primaire se charge de la réplication et confirme l'achèvement du processus d'écriture une fois que tous les autres OSD ont stocké l'objet.

Si `osd.5` échoue, tous les objets du `PG1` sont toujours disponibles sur `osd.1`. Dès que la grappe identifie un échec d'OSD, un autre OSD prend le relais. Dans cet exemple `osd.4` est utilisé en remplacement d'`osd.5`. Les objets stockés sur `osd.1` sont ensuite répliqués sur `osd.4` pour restaurer le niveau de la réplication.

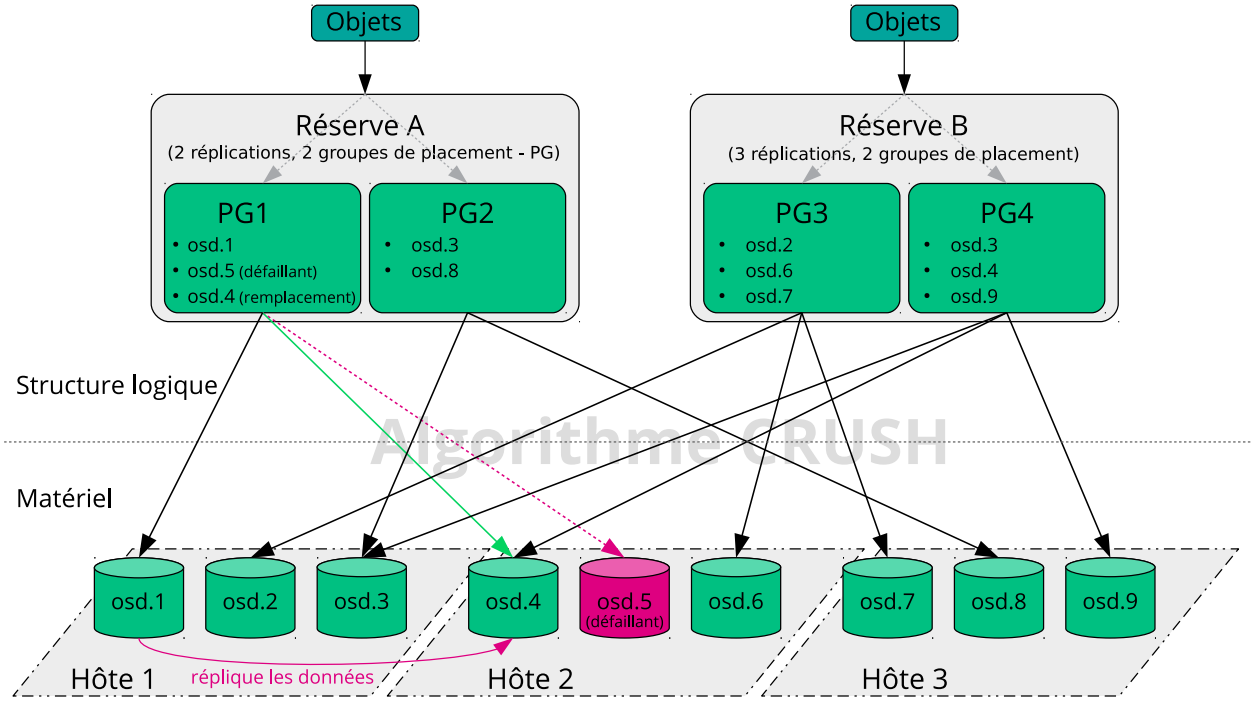

FIGURE 1.2 : EXEMPLE DE CEPH À PETITE ÉCHELLE

Si un nouveau noeud avec de nouveaux OSD est ajouté à la grappe, la carte de la grappe va changer. La fonction CRUSH renvoie alors des emplacements différents pour les objets. Les objets qui reçoivent de nouveaux emplacements seront déplacés. Ce processus aboutit à une utilisation équilibrée de tous les OSD.

1.4 BlueStore

BlueStore est une nouvelle interface dorsale de stockage par défaut pour Ceph depuis la version 5 de SES. Il offre de meilleures performances que FileStore, une somme de contrôle de données complète et une compression intégrée.

BlueStore gère un, deux ou trois périphériques de stockage. Dans le cas le plus simple, BlueStore utilise un seul périphérique de stockage primaire. Le périphérique de stockage est normalement partitionné en deux parties :

1. Une petite partition nommée BlueFS qui met en oeuvre des fonctionnalités de type système de fichiers requises par RocksDB.
2. Le reste du périphérique est généralement une grande partition occupant le reste du périphérique. Il est géré directement par BlueStore et contient toutes les données réelles. Ce périphérique primaire est normalement identifié par un lien symbolique de bloc dans le répertoire de données.

Il est également possible de déployer BlueStore sur deux périphériques supplémentaires :

Un *périphérique WAL* peut être utilisé pour le journal interne ou le journal d'écriture anticipée de BlueStore. Il est identifié par le lien symbolique `block.wal` dans le répertoire de données. Il n'est utile d'utiliser un périphérique WAL distinct que si le périphérique est plus rapide que le périphérique primaire ou le périphérique DB, par exemple dans les cas suivants :

- Le périphérique WAL est une interface NVMe, le périphérique DB est un disque SSD et le périphérique de données est un disque SSD ou un disque HDD.
- Les périphériques WAL et DB sont des disques SSD distincts et le périphérique de données est un disque SSD ou un disque HDD.

Un *périphérique DB* peut être utilisé pour stocker les métadonnées internes de BlueStore. BlueStore (ou plutôt, la base de données RocksDB intégrée) mettra autant de métadonnées que possible sur le périphérique DB pour améliorer les performances. Encore une fois, il n'est utile de déployer un périphérique DB partagé que s'il est plus rapide que le périphérique primaire.

Astuce : planification de la taille du périphérique DB

Planifiez minutieusement la taille du périphérique DB pour qu'elle soit suffisante. Si le périphérique DB sature, les métadonnées débordent sur le périphérique primaire, ce qui dégrade considérablement les performances de l'OSD.

Vous pouvez vérifier si une partition WAL/DB est pleine et déborde avec la commande **`ceph daemon osd.`*`ID`*` perf dump`**. La valeur `slow_used_bytes` indique la quantité de données qui déborde :

```
cephuser@adm > ceph daemon osd.ID perf dump | jq '.bluefs'
```

```
"db_total_bytes": 1073741824,
"db_used_bytes": 33554432,
"wal_total_bytes": 0,
"wal_used_bytes": 0,
"slow_total_bytes": 554432,
"slow_used_bytes": 554432,
```

1.5 Informations supplémentaires

- Ceph, en tant que projet communautaire, dispose de sa propre documentation en ligne. Pour les rubriques non trouvées dans ce manuel, reportez-vous au site https://docs.ceph.com/en/octopus/.
- La publication d'origine *CRUSH: Controlled, Scalable, Decentralized Placement of Replicated Data* (CRUSH : placement contrôlé, évolutif et décentralisé des données répliquées) par *S.A. Weil, S.A. Brandt, E.L. Miller, C. Maltzahn* fournit des informations utiles sur le fonctionnement interne de Ceph. Il est surtout recommandé de le lire lors du déploiement de grappes à grande échelle. La publication peut être consultée à l'adresse http://www.ssrc.ucsc.edu/papers/weil-sc06.pdf.
- SUSE Enterprise Storage peut être utilisé avec des distributions non-SUSE OpenStack. Les clients Ceph doivent être à un niveau compatible avec SUSE Enterprise Storage.

Remarque

SUSE prend en charge le composant serveur du déploiement Ceph et le client est pris en charge par le fournisseur de la distribution OpenStack.

2 Configuration matérielle requise et recommandations

La configuration matérielle requise de Ceph dépend fortement du workload des E/S. La configuration matérielle requise et les recommandations suivantes doivent être considérées comme un point de départ de la planification détaillée.

En général, les recommandations données dans cette section dépendent du processus. Si plusieurs processus sont situés sur la même machine, les besoins en UC, RAM, disque et réseau doivent être additionnés.

2.1 Aperçu du réseau

Ceph compte plusieurs réseaux logiques :

- Un réseau d'interface client appelé `réseau public`.
- Un réseau interne approuvé, le réseau back-end, appelé `réseau de grappes`. Cette option est facultative.
- Un ou plusieurs réseaux client pour les passerelles. Ceci est facultatif et sort du cadre de ce chapitre.

Le réseau public est le réseau sur lequel les daemons Ceph communiquent entre eux et avec leurs clients. Cela signifie que tout le trafic de la grappe Ceph passe par ce réseau, sauf dans le cas où un réseau de grappe est configuré.

Le réseau de grappe est le réseau back-end entre les noeuds OSD, pour la réplication, le rééquilibrage et la récupération. Lorsqu'il est configuré, ce réseau facultatif devrait idéalement fournir deux fois la bande passante du réseau public avec la réplication tripartite par défaut, étant donné que l'OSD primaire envoie deux copies aux autres OSD par le biais de ce réseau. Le réseau public est situé entre les clients et les passerelles d'un côté pour communiquer avec les moniteurs, les gestionnaires ainsi qu'avec les noeuds MDS et OSD. Il est également utilisé par les moniteurs, les gestionnaires et les noeuds MDS pour communiquer avec les noeuds OSD.

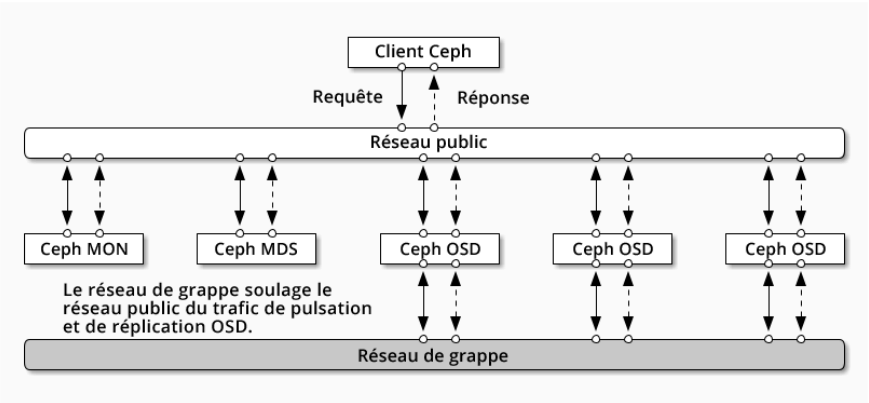

FIGURE 2.1 : APERÇU DU RÉSEAU

2.1.1 Recommandations concernant le réseau

Nous vous recommandons d'utiliser un seul réseau tolérant aux pannes avec une bande passante suffisante pour répondre à vos besoins. Pour l'environnement de réseau public Ceph, nous recommandons deux interfaces réseau 25 GbE (ou plus rapides) liées à l'aide de 802.3ad (LACP). Cette recommandation est considérée comme la configuration minimale pour Ceph. Si vous utilisez également un réseau en grappe, nous recommandons quatre interfaces réseau 25 GbE liées. La liaison de deux ou plusieurs interfaces réseau offre un meilleur débit grâce à l'agrégation de liens et, compte tenu des liens et commutateurs redondants, une meilleure tolérance aux pannes et une meilleure maintenabilité.

Vous pouvez également créer des VLAN pour isoler différents types de trafic sur une liaison. Par exemple, vous pouvez créer une liaison pour fournir deux interfaces VLAN, l'une pour le réseau public et l'autre pour le réseau en grappe. Ce n'est toutefois *pas* nécessaire lors de la configuration du réseau Ceph. Vous trouverez des détails sur la liaison des interfaces sur le site https://documentation.suse.com/sles/15-SP2/html/SLES-all/cha-network.html#sec-network-iface-bonding.

La tolérance aux pannes peut être améliorée en isolant les composants dans des domaines de défaillance. Pour améliorer la tolérance aux pannes du réseau, la liaison d'une interface à partir de deux cartes d'interface réseau (NIC) distinctes offre une protection contre la défaillance d'une seule carte réseau. De même, la création d'une liaison entre deux commutateurs protège contre la défaillance d'un commutateur. Nous vous recommandons de vous adresser au fournisseur de l'équipement réseau afin de définir le niveau de tolérance aux pannes requis.

Important : réseau d'administration non pris en charge

La configuration réseau d'administration supplémentaire (qui permet, par exemple, de séparer les réseaux SSH, Salt ou DNS) n'est ni testée ni prise en charge.

Astuce : noeuds configurés via DHCP

Si les noeuds de stockage sont configurés via DHCP, les timeouts par défaut peuvent ne pas être suffisants pour que le réseau soit configuré correctement avant le démarrage des différents daemons Ceph. Si cela se produit, les instances MON et OSD de Ceph ne démarreront pas correctement (l'exécution de **`systemctl status ceph\*`** entraînera des erreurs de type « unable to bind » [liaison impossible]). Pour éviter ce problème, nous vous recommandons d'augmenter le timeout du client DHCP sur au moins 30 secondes pour chaque noeud de votre grappe de stockage. Pour ce faire, vous pouvez modifier les paramètres suivants sur chaque noeud :

Dans `/etc/sysconfig/network/dhcp`, définissez

```
DHCLIENT_WAIT_AT_BOOT="30"
```

Dans `/etc/sysconfig/network/config`, définissez

```
WAIT_FOR_INTERFACES="60"
```

2.1.1.1 Ajout d'un réseau privé à une grappe en cours d'exécution

Si vous ne spécifiez pas de réseau pour la grappe lors du déploiement de Ceph, il suppose un environnement de réseau public unique. Bien que Ceph fonctionne parfaitement avec un réseau public, ses performances et sa sécurité s'améliorent lorsque vous définissez un second réseau de grappe privé. Pour prendre en charge deux réseaux, chaque noeud Ceph doit disposer d'au moins deux cartes réseau.

Vous devez apporter les modifications suivantes à chaque noeud Ceph. Ces modifications sont relativement rapides à apporter à une petite grappe, mais peuvent prendre beaucoup de temps si votre grappe est composée de centaines ou de milliers de noeuds.

1. Définissez le réseau de la grappe à l'aide de la commande suivante :

```
root # ceph config set global cluster_network MY_NETWORK
```

Redémarrez les OSD pour qu'ils se lient au réseau de grappe spécifié :

```
root # systemctl restart ceph-*@osd.*.service
```

2. Vérifiez que le réseau privé de la grappe fonctionne comme prévu au niveau du système d'exploitation.

2.1.1.2 Noeuds de moniteur sur des sous-réseaux distincts

Si les noeuds de moniteur se trouvent sur plusieurs sous-réseaux, par exemple s'ils se trouvent dans des pièces différentes et sont desservis par différents commutateurs, vous devez spécifier leur adresse de réseau public en notation CIDR :

```
cephuser@adm > ceph config set mon public_network
 "MON_NETWORK_1, MON_NETWORK_2, MON_NETWORK_N
```

Par exemple :

```
cephuser@adm > ceph config set mon public_network "192.168.1.0/24, 10.10.0.0/16"
```

Avertissement

Si vous spécifiez plusieurs segments de réseau pour le réseau public (ou en grappe) comme décrit dans cette section, chacun de ces sous-réseaux doit pouvoir être routé vers tous les autres. Dans le cas contraire, les daemons MON et les autres daemons Ceph sur différents segments de réseau ne peuvent pas communiquer et il en résulte une grappe divisée. De plus, si vous utilisez un pare-feu, veillez à inclure chaque sous-réseau ou adresse IP dans vos iptables et ouvrez leur les ports nécessaires sur tous les noeuds.

2.2 Configurations d'architecture multiples

SUSE Enterprise Storage prend en charge les architectures x86 et Arm. Si nous considérons chaque architecture, il est important de noter que d'un point de vue du nombre de coeurs par OSD, de la fréquence et de la mémoire RAM, il n'y a pas de différence réelle entre les architectures de processeurs en termes de dimensionnement.

Comme pour les processeurs x86 plus petits (non-serveurs), les coeurs basés sur Arm moins performants peuvent ne pas fournir une expérience optimale, en particulier lorsqu'ils sont utilisés pour des réserves codées à effacement.

Remarque

Dans toute la documentation, *SYSTEM-ARCH* est utilisé à la place de x86 ou Arm.

2.3 Configuration du matériel

Pour bénéficier d'une expérience optimale avec le produit, nous vous recommandons de commencer par configurer la grappe recommandée. Pour une grappe de test ou une grappe ne nécessitant pas de performances très importantes, nous documentons une configuration minimale prise en charge pour la grappe.

2.3.1 Configuration minimale de la grappe

Une configuration de grappe minimale pour le produit se compose des éléments suivants :

- Au moins quatre noeuds physiques (noeuds OSD) avec cohabitation des services
- Ethernet Dual-10 Go en tant que réseau lié
- Un noeud Admin distinct (peut être virtualisé sur un noeud externe)

Une configuration détaillée est la suivante :

- Un noeud Admin distinct avec 4 Go de RAM, quatre coeurs et 1 To de capacité de stockage. Il s'agit généralement du noeud Salt Master. Les passerelles et services Ceph et de passerelles, tels que Ceph Monitor, Metadata Server, Ceph OSD, Object Gateway ou NFS Ganesha ne sont pas pris en charge sur le noeud Admin, car il doit orchestrer indépendamment les processus de mise à jour et de mise à niveau de la grappe.
- Au moins quatre noeuds OSD physiques, avec huit disques OSD chacun, reportez-vous à la *Section 2.4.1, « Configuration minimale requise »* pour consulter la configuration requise. La capacité totale de la grappe doit être dimensionnée de sorte que, même si un noeud est indisponible, la capacité totale utilisée (y compris la redondance) ne dépasse pas 80 %.
- Trois instances de Ceph Monitor. Pour des raisons de latence, les moniteurs doivent être exécutés à partir d'un espace de stockage SSD/NVMe et non à partir de disques durs.

- Les moniteurs, le serveur de métadonnées et les passerelles peuvent cohabiter sur les noeuds OSD. Reportez-vous à la *Section 2.12, « OSD et Monitor partageant un même serveur »* concernant la cohabitation des moniteurs. Si vous faites cohabiter des services, les exigences de mémoire et d'UC doivent être additionnées.

- La passerelle iSCSI, Object Gateway et le serveur de métadonnées ont au minimum besoin de 4 Go de RAM incrémentielle et de quatre coeurs.

- Si vous utilisez CephFS, S3/Swift, iSCSI, au moins deux instances des rôles respectifs (serveur de métadonnées, Object Gateway, iSCSI) sont requises à des fins de redondance et de disponibilité.

- Les noeuds doivent être dédiés à SUSE Enterprise Storage et ne doivent pas être utilisés pour d'autres workloads physiques, conteneurisés ou virtualisés.

- Si l'une des passerelles (iSCSI, Object Gateway, NFS Ganesha, serveur de métadonnées, etc.) est déployée au sein de machines virtuelles, ces dernières ne doivent pas être hébergées sur les machines physiques servant d'autres rôles de grappe. (Ce n'est pas nécessaire puisqu'elles sont prises en charge en tant que services cohabitants.)

- Lors du déploiement de services en tant que machines virtuelles sur des hyperviseurs en dehors de la grappe physique principale, les domaines de défaillance doivent être respectés afin de garantir la redondance.
  Par exemple, ne déployez pas plusieurs rôles du même type sur le même hyperviseur, tels que plusieurs instances MON ou MDS.

- Lors d'un déploiement sur des machines virtuelles, il est particulièrement important de s'assurer que les noeuds disposent d'une solide connectivité réseau et d'une synchronisation horaire fonctionnant convenablement.

- Les noeuds d'hyperviseur doivent avoir la bonne taille pour éviter les interférences occasionnées par d'autres workloads consommant des ressources de processeur, de RAM, de réseau et de stockage.

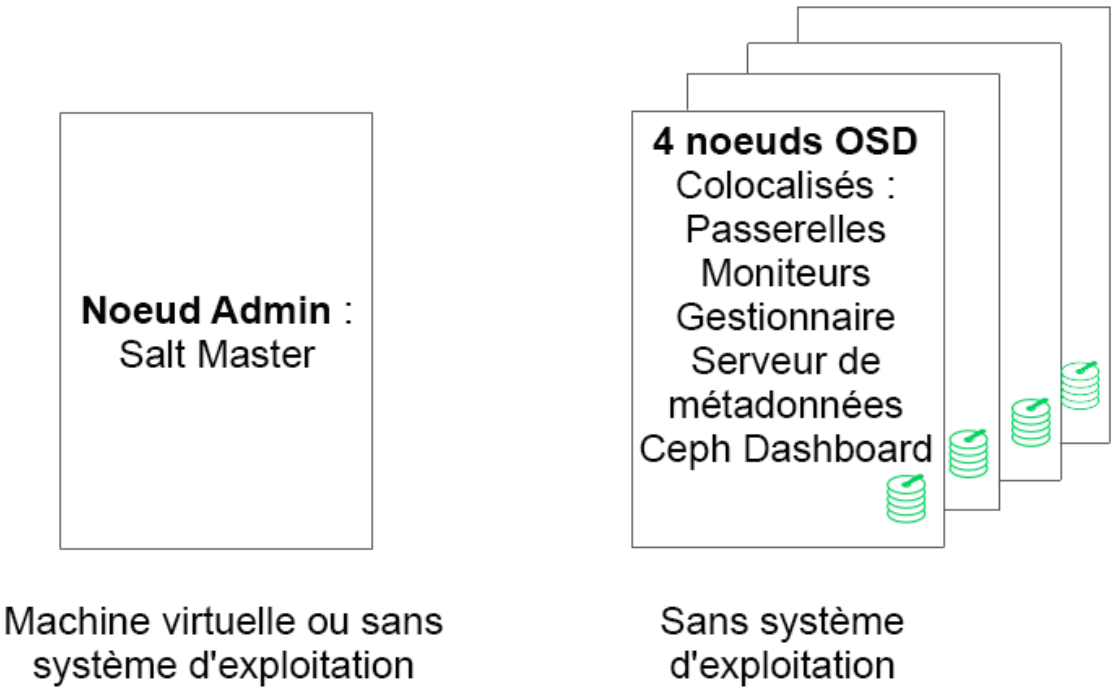

FIGURE 2.2 : CONFIGURATION MINIMALE DE LA GRAPPE

2.3.2 Configuration recommandée pour une grappe en production

Une fois votre grappe développée, nous vous recommandons de déplacer les instances Ceph Monitor, les serveurs de métadonnées et les passerelles vers des noeuds distincts afin d'améliorer la tolérance aux pannes.

- Sept noeuds de stockage des objets
  - Aucun noeud n'excède environ 15 % du stockage total.
  - La capacité totale de la grappe doit être dimensionnée de sorte que, même si un noeud est indisponible, la capacité totale utilisée (y compris la redondance) ne dépasse pas 80 %.
  - Ethernet de 25 Go ou plus chacun lié à la grappe interne et au réseau public externe.

- Plus de 56 OSD par grappe de stockage.
- Reportez-vous à la *Section 2.4.1, « Configuration minimale requise »* pour d'autres recommandations.

- Noeuds d'infrastructure physique dédiés.
  - Trois noeuds de moniteur Ceph : 4 Go de RAM, processeur à 4 coeurs, disques SSD RAID 1.
    Reportez-vous à la *Section 2.5, « Noeuds de moniteur »* pour d'autres recommandations.
  - Noeuds Object Gateway : 32 Go de RAM, processeur à 8 coeurs, disques SSD RAID 1.
    Reportez-vous à la *Section 2.6, « Noeuds Object Gateway »* pour d'autres recommandations.
  - Noeuds Passerelle iSCSI : 16 Go de RAM, processeur à 8 coeurs, disques SSD RAID 1.
    Reportez-vous à la *Section 2.9, « Noeuds de passerelle iSCSI »* pour d'autres recommandations.
  - Noeuds MDS (un actif/un de secours) : 32 Go de RAM, processeur à 8 coeurs, disques SSD RAID 1.
    Reportez-vous à la *Section 2.7, « Noeuds de serveur de métadonnées »* pour d'autres recommandations.
  - Un noeud Admin SES : 4 Go de RAM, processeur à 4 coeurs, disques SSD RAID 1.

2.3.3 Configuration multipath

Si vous souhaitez utiliser du matériel multipath, assurez-vous que LVM voit `multipath_component_detection = 1` dans le fichier de configuration sous la section `devices`. Vous pouvez vérifier à l'aide de la commande **`lvm config`**.

Vous pouvez également vous assurer que LVM filtre les composants mpath d'un périphérique via la configuration du filtre LVM. Cela dépendra de l'hôte.

Remarque

Ce n'est pas recommandé et ne doit être envisagé que si `multipath_component_detection = 1` ne peut pas être défini.

Pour plus d'informations sur la configuration multipath, consultez l'adresse https://documentation.suse.com/sles/15-SP2/html/SLES-all/cha-multipath.html#sec-multipath-lvm.

2.4 Noeuds de stockage des objets

2.4.1 Configuration minimale requise

- Les recommandations d'UC suivantes tiennent compte des périphériques indépendamment de l'utilisation par Ceph :
  - 1x thread d'UC 2 GHz par disque rotatif.
  - 2x threads d'UC 2 GHz par disque SSD.
  - 4x threads d'UC 2 GHz par disque NVMe.
- Réseaux 10 GbE séparés (public/client et interne), 4x 10 GbE requis, 2x 25 GbE recommandés.
- Mémoire RAM totale requise = nombre d'OSD x (1 Go + `osd_memory_target`) + 16 Go
  Reportez-vous au *Manuel « Guide d'opérations et d'administration », Chapitre 28 « Configuration de la grappe Ceph », Section 28.4.1 « Configuration du dimensionnement automatique du cache »* pour plus de détails sur `osd_memory_target`.
- Disques OSD dans les configurations JBOD ou les configurations RAID-0 individuelles.
- Le journal OSD peut résider sur le disque OSD.
- Les disques OSD doivent être exclusivement utilisés par SUSE Enterprise Storage.
- Disque et SSD dédiés au système d'exploitation, de préférence dans une configuration RAID 1.
- Allouez au moins 4 Go de RAM supplémentaires si cet hôte OSD doit héberger une partie d'une réserve en cache utilisée pour la hiérarchisation du cache.
- Les moniteurs Ceph, la passerelle et les serveurs de métadonnées peuvent résider sur les noeuds de stockage des objets.

- Pour des raisons de performances de disque, les noeuds OSD sont des noeuds sans système d'exploitation. Aucun autre workload ne doit s'exécuter sur un noeud OSD, sauf s'il s'agit d'une configuration minimale d'instances Ceph Monitor et Ceph Manager.
- Disques SSD pour journal avec un ratio journal SSD à OSD de 6:1.

Remarque

Assurez-vous qu'aucun périphérique de bloc en réseau n'est assigné aux noeuds OSD, tels que des images de périphérique de bloc RADOS ou iSCSI.

2.4.2 Taille minimale du disque

Deux types d'espace disque sont nécessaires pour l'exécution sur un OSD : l'espace pour le périphérique WAL/DB et l'espace primaire pour les données stockées. La valeur minimale (et par défaut) pour le périphérique WAL/DB est de 6 Go. L'espace minimum pour les données est de 5 Go, car les partitions inférieures à 5 Go sont automatiquement assignées à une pondération de 0.

Ainsi, bien que l'espace disque minimal pour un OSD soit de 11 Go, il est recommandé de ne pas utiliser un disque inférieur à 20 Go, même à des fins de test.

2.4.3 Taille recommandée pour les périphériques WAL et DB de BlueStore

Astuce : supplément d'informations

Reportez-vous à la *Section 1.4, « BlueStore »* pour plus d'informations sur BlueStore.

- Nous vous recommandons de réserver 4 Go pour le périphérique WAL. La taille recommandée pour DB est de 64 Go pour la plupart des workloads.

Important

Nous recommandons des volumes de base de données plus importants pour les déploiements à charge élevée, en particulier si l'utilisation de RGW ou de CephFS est importante. Réservez une certaine capacité (emplacements) pour installer plus de matériel et plus d'espace DB si nécessaire.

- Si vous envisagez de placer les périphériques WAL et DB sur le même disque, il est recommandé d'utiliser une partition unique pour les deux périphériques, plutôt que d'utiliser une partition distincte pour chacun. Cela permet à Ceph d'utiliser également le périphérique DB pour les opérations WAL. La gestion de l'espace disque est donc plus efficace, car Ceph n'utilise la partition DB pour le WAL qu'en cas de nécessité. Un autre avantage est que la probabilité que la partition WAL soit pleine est très faible, et lorsqu'elle n'est pas entièrement utilisée, son espace n'est pas gaspillé, mais utilisé pour les opérations de DB.
  Pour partager le périphérique DB avec le WAL, ne spécifiez *pas* le périphérique WAL et indiquez uniquement le périphérique DB.
  Pour plus d'informations sur la spécification d'une disposition OSD, reportez-vous à la *Manuel « Guide d'opérations et d'administration », Chapitre 13 « Tâches opérationnelles », Section 13.4.3 « Ajout d'OSD à l'aide de la spécification DriveGroups »*.

2.4.4 SSD pour les partitions WAL/DB

Les disques SSD n'ont pas de pièces mobiles. Cela réduit le temps d'accès aléatoire et la latence de lecture tout en accélérant le débit de données. Comme leur prix par Mo est significativement plus élevé que le prix des disques durs tournants, les disques SSD ne conviennent que pour un stockage plus petit.

Les OSD peuvent constater une amélioration significative de leurs performances en stockant leurs partitions sur un disque SSD et les données d'objet sur un disque dur distinct.

Astuce : partage d'un disque SSD pour plusieurs partitions WAL/DB

Étant donné que les partitions WAL/DB occupent relativement peu d'espace, vous pouvez partager un disque SSD avec plusieurs partitions WAL/DB. N'oubliez pas qu'avec chaque partition WAL/DB, les performances du disque SSD se dégradent. Il est déconseillé de partager plus de six partitions WAL/DB sur le même disque SSD et 12 sur des disques NVMe.

2.4.5 Nombre maximal recommandé de disques

Vous pouvez avoir autant de disques sur un serveur que ce dernier l'autorise. Il existe quelques aspects à considérer lorsque vous planifiez le nombre de disques par serveur :

- *Bande passante du réseau.* Plus vous avez de disques sur un serveur, plus il y a de données à transférer via la ou les cartes réseau pour les opérations d'écriture sur disque.
- *Mémoire.* Plus de 2 Go de RAM sont utilisés pour le cache BlueStore. Avec la valeur par défaut de l'option `osd_memory_target`, à savoir 4 Go, le système dispose d'une taille de cache de départ raisonnable pour les supports rotatifs. Si vous utilisez des disques SSD ou NVMe, pensez à augmenter la taille du cache et l'allocation de RAM par OSD afin d'optimiser les performances.
- *Tolérance aux pannes.* Si le serveur complet tombe en panne, plus il comporte de disques, plus le nombre d'OSD perdus temporairement par la grappe est important. En outre, pour maintenir les règles de réplication en cours d'exécution, vous devez copier toutes les données du serveur en panne vers les autres noeuds de la grappe.

2.5 Noeuds de moniteur

- Au moins trois noeuds MON sont requis. Le nombre de moniteurs doit toujours être impair ($1 + 2n$).
- 4 Go de RAM.
- Processeur à quatre coeurs logiques.

- Un SSD ou un autre type de support de stockage suffisamment rapide est vivement recommandé pour les moniteurs, en particulier pour le chemin `/var/lib/ceph` sur chaque noeud de moniteur, car le quorum peut être instable avec des latences de disque élevées. Deux disques en configuration RAID 1 sont recommandés pour la redondance. Il est recommandé d'utiliser des disques distincts ou au moins des partitions de disque distinctes pour les processus Monitor, afin de protéger l'espace disque disponible du moniteur contre des phénomènes tels que le grossissement excessif du fichier journal.
- Il ne doit y avoir qu'un seul processus Monitor par noeud.
- La combinaison des noeuds OSD, Monitor ou Object Gateway n'est prise en charge que s'il y a suffisamment de ressources matérielles disponibles. Cela signifie que les besoins de tous les services doivent être additionnés.
- Deux interfaces réseau liées à plusieurs commutateurs.

2.6 Noeuds Object Gateway

Les noeuds Object Gateway doivent avoir au moins six coeurs de processeur et 32 Go de RAM. Lorsque d'autres processus sont colocalisés sur la même machine, leurs besoins doivent être additionnés.

2.7 Noeuds de serveur de métadonnées

Le dimensionnement approprié des noeuds MDS (Metadata Server, Serveur de métadonnées) dépend du cas d'utilisation spécifique. En règle générale, plus le nombre de fichiers ouverts que le serveur de métadonnées doit traiter est important, plus l'UC et la RAM requises sont importantes. La configuration minimale requise est la suivante :

- 4 Go de RAM pour chaque daemon MDS.
- Interface réseau liée.
- 2,5 GHz d'UC avec au moins 2 coeurs.

2.8 Noeud Admin

Au moins 4 Go de RAM et une UC quadruple coeur sont requis. Cela inclut l'exécution de Salt Master sur le noeud Admin. Pour les grappes de grande taille avec des centaines de noeuds, 6 Go de RAM sont conseillés.

2.9 Noeuds de passerelle iSCSI

Les noeuds de passerelle iSCSI doivent avoir au moins six coeurs de processeur et 16 Go de RAM.

2.10 SES et autres produits SUSE

Cette section contient des informations importantes concernant l'intégration de SES avec d'autres produits SUSE.

2.10.1 SUSE Manager

SUSE Manager et SUSE Enterprise Storage n'étant pas intégrés, SUSE Manager ne peut actuellement pas gérer une grappe SES.

2.11 Limitations concernant les noms

Ceph ne prend généralement pas en charge les caractères non-ASCII dans les fichiers de configuration, les noms de réserve, les noms d'utilisateur, etc. Lors de la configuration d'une grappe Ceph, il est recommandé d'utiliser uniquement des caractères alphanumériques simples (A-Z, a-z, 0-9) et des ponctuations minimales (« . », « - », « _ ») dans tous les noms d'objet/de configuration Ceph.

2.12 OSD et Monitor partageant un même serveur

Bien qu'il soit techniquement possible d'exécuter des OSD et des instances Monitor sur le même serveur dans des environnements de test, il est vivement recommandé de disposer d'un serveur distinct pour chaque noeud de moniteur en production. Le principal motif concerne la perfor-

mance : plus la grappe contient d'OSD, plus les noeuds MON doivent effectuer un grand nombre d'opérations d'E/S. Et lorsqu'un serveur est partagé entre un noeud MON et plusieurs OSD, les opérations d'E/S des OSD constituent un facteur limitant pour le noeud de moniteur.

Un autre aspect consiste à déterminer s'il convient de partager des disques entre un OSD, un noeud MON et le système d'exploitation sur le serveur. La réponse est simple : si possible, dédiez un disque distinct à l'OSD et un serveur distinct au noeud de moniteur.

Bien que Ceph prenne en charge les OSD basés sur les répertoires, un OSD doit toujours avoir un disque dédié autre que celui du système d'exploitation.

Astuce

S'il est *vraiment* nécessaire d'exécuter l'OSD et le noeud MON sur le même serveur, exécutez MON sur un disque distinct en montant le disque dans le répertoire `/var/lib/ceph/mon` pour obtenir des performances légèrement meilleures.

3 Configuration haute disponibilité du noeud Admin

Le *noeud Admin* est un noeud de la grappe Ceph sur lequel le service Salt Master s'exécute. Il gère le reste des noeuds de la grappe en interrogeant et en indiquant leurs services de minion Salt. En général, il inclut également d'autres services, par exemple, le tableau de bord *Grafana* soutenu par le toolkit de surveillance *Prometheus*.

En cas de défaillance du noeud Admin, vous devez généralement fournir un nouveau matériel pour le noeud et restaurer la pile de configuration complète de la grappe à partir d'une sauvegarde récente. Cette méthode prend énormément de temps et provoque l'interruption de la grappe.

Pour éviter le temps hors service de la grappe Ceph causé par la défaillance du noeud Admin, il est recommandé d'utiliser une grappe haute disponibilité (High Availability, HA) pour le noeud Admin Ceph.

3.1 Aperçu de la grappe HA pour le noeud Admin

Le principe d'une grappe HA est qu'en cas de défaillance d'un noeud de la grappe, un autre noeud prend automatiquement en charge son rôle, y compris le noeud Admin virtualisé. De cette façon, les autres noeuds de la grappe Ceph ne remarquent pas la défaillance du noeud Admin.

La solution HA minimale pour le noeud Admin requiert le matériel suivant :

- Deux serveurs sans système d'exploitation en mesure d'exécuter SUSE Linux Enterprise avec l'extension haute disponibilité et de virtualiser le noeud Admin.
- Deux ou plusieurs chemins de communication réseau redondants, par exemple via la liaison de périphérique réseau.
- Un stockage partagé pour héberger les images de disque de la machine virtuelle du noeud Admin. Le stockage partagé doit être accessible depuis les deux serveurs. Il peut s'agir, par exemple, d'une exportation NFS, d'un partage Samba ou d'une cible iSCSI.

Pour plus de détails sur les conditions requises de la grappe, reportez-vous à la page https://documentation.suse.com/sle-ha/15-SP2/html/SLE-HA-all/art-sleha-install-quick.html#sec-ha-inst-quick-req.

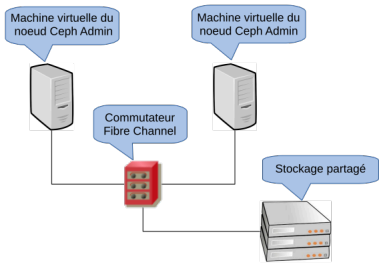

FIGURE 3.1 : GRAPPE HAUTE DISPONIBILITÉ À 2 NOEUDS POUR LE NOEUD ADMIN

3.2 Création d'une grappe haute disponibilité avec le noeud Admin

La procédure suivante résume les étapes les plus importantes de la création d'une grappe haute disponibilité pour la virtualisation du noeud Admin. Pour plus d'informations, reportez-vous aux liens indiqués.

1. Configurez une grappe haute disponibilité à 2 noeuds de base avec un stockage partagé comme décrit dans le manuel https://documentation.suse.com/sle-ha/15-SP2/html/SLE-HA-all/art-sleha-install-quick.html.
2. Sur les deux noeuds de la grappe, installez tous les paquetages requis pour l'exécution de l'hyperviseur KVM et du toolkit `libvirt` comme décrit dans le manuel https://documentation.suse.com/sles/15-SP2/html/SLES-all/cha-vt-installation.html#sec-vt-installation-kvm.
3. Sur le premier noeud de la grappe, créez une machine virtuelle KVM en utilisant `libvirt` comme décrit dans le manuel https://documentation.suse.com/sles/15-SP2/html/SLES-all/cha-kvm-inst.html#sec-libvirt-inst-virt-install. Utilisez le stockage partagé préconfiguré pour stocker les images de disque de la machine virtuelle.
4. Une fois la configuration de la machine virtuelle terminée, exportez sa configuration dans un fichier XML sur le stockage partagé. Utilisez la syntaxe suivante.

```
root # virsh dumpxml VM_NAME > /path/to/shared/vm_name.xml
```

5. Créez une ressource pour la machine virtuelle du noeud Admin. Reportez-vous à l'adresse https://documentation.suse.com/sle-ha/15-SP2/html/SLE-HA-all/cha-conf-hawk2.html pour des informations générales sur la création de ressources haute disponibilité. Des informations détaillées sur la création de ressources pour une machine virtuelle KVM sont décrites à l'adresse http://www.linux-ha.org/wiki/VirtualDomain_%28resource_agent%29.

6. Sur l'invité de la machine virtuelle qui vient d'être créé, déployez le noeud Admin, y compris les services supplémentaires dont vous avez besoin. Suivez les étapes appropriées de la *Section 5.2, « Déploiement de Salt »*. En même temps, déployez les noeuds de la grappe Ceph restants sur les serveurs de grappe non-HA.

II Déploiement de la grappe Ceph

4 Introduction et tâches courantes

Depuis SUSE Enterprise Storage 7, les services Ceph sont déployés en tant que conteneurs à l'aide de l'utilitaire cephadm au lieu des paquetages RPM. Pour plus de détails, reportez-vous au *Chapitre 5, Déploiement avec cephadm*.

4.1 Lisez les notes de version

Les notes de version contiennent des informations supplémentaires sur les modifications apportées depuis la version précédente de SUSE Enterprise Storage. Consultez les notes de version pour vérifier les aspects suivants :

- Votre matériel doit tenir compte de certaines considérations spéciales.
- Les paquetages logiciels utilisés ont été considérablement modifiés.
- Des précautions spéciales sont nécessaires pour votre installation.

Les notes de version incluent également des informations de dernière minute qui, faute de temps, n'ont pas pu être intégrées au manuel. Elles contiennent également des notes concernant les problèmes connus.

Après avoir installé le paquetage `release-notes-ses`, les notes de version sont disponibles en local dans le répertoire `/usr/share/doc/release-notes` ou en ligne à l'adresse https://www.suse.com/releasenotes/.

5 Déploiement avec cephadm

SUSE Enterprise Storage 7 utilise l'outil `ceph-salt` basé sur Salt pour préparer le système d'exploitation sur chaque noeud de la grappe participant en vue d'un déploiement via cephadm. cephadm déploie et gère une grappe Ceph en se connectant aux hôtes à partir du daemon Ceph Manager via SSH. cephadm gère le cycle de vie complet d'une grappe Ceph. Il commence par démarrer une petite grappe sur un seul noeud (un service MON et un service MGR), puis utilise l'interface d'orchestration pour étendre la grappe afin d'inclure tous les hôtes et de provisionner tous les services Ceph. Vous pouvez effectuer cette opération via l'interface de ligne de commande Ceph (CLI) ou partiellement via Ceph Dashboard (GUI).

Important

Notez que la documentation de la communauté Ceph utilise la commande de démarrage **`cephadm bootstrap`** lors du déploiement initial. `ceph-salt` appelle la commande **`cephadm bootstrap`** et ne doit pas être exécuté directement. Tout déploiement manuel de grappes Ceph à l'aide de l'outil de **`cephdm bootstrap`** ne sera pas pris en charge.

Pour déployer une grappe Ceph à l'aide de cephadm, procédez comme suit :

1. Installez et effectuez la configuration de base du système d'exploitation sous-jacent (SUSE Linux Enterprise Server 15 SP2) sur tous les noeuds de la grappe.
2. Déployez l'infrastructure Salt sur tous les noeuds de la grappe pour préparer le déploiement initial via `ceph-salt`.
3. Configurez les propriétés de base de la grappe via `ceph-salt` et déployez-la.
4. Ajoutez de nouveaux noeuds et rôles à la grappe et déployez des services sur ces derniers à l'aide de cephadm.

5.1 Installation et configuration de SUSE Linux Enterprise Server

1. Installez et enregistrez SUSE Linux Enterprise Server 15 SP2 sur chaque noeud de la grappe. Lors de l'installation de SUSE Enterprise Storage, vous devrez pouvoir accéder aux dépôts de mise à jour. L'enregistrement est donc obligatoire. Incluez au moins les modules suivants :

   - Module Basesystem (Système de base)
   - Module Server Applications (Applications serveur)

   Pour plus de détails sur l'installation de SUSE Linux Enterprise Server, reportez-vous à la documentation https://documentation.suse.com/sles/15-SP2/html/SLES-all/cha-install.html.

2. Installez l'extension *SUSE Enterprise Storage 7* sur chaque noeud de la grappe.

   

   ### Astuce : installez SUSE Enterprise Storage en même temps que SUSE Linux Enterprise Server

   Vous pouvez installer l'extension SUSE Enterprise Storage 7 séparément après avoir installé SUSE Linux Enterprise Server 15 SP2 ou l'ajouter au cours de la procédure d'installation de SUSE Linux Enterprise Server 15 SP2.

   Pour plus de détails sur l'installation des extensions, reportez-vous à la documentation https://documentation.suse.com/sles/15-SP2/html/SLES-all/cha-register-sle.html.

3. Configurez les paramètres réseau, y compris la résolution de nom DNS appropriée sur chaque noeud. Pour plus d'informations sur la configuration d'un réseau, reportez-vous à l'adresse https://documentation.suse.com/sles/15-SP2/html/SLES-all/cha-network.html#sec-network-yast. Pour plus d'informations sur la configuration d'un serveur DNS, reportez-vous à l'adresse https://documentation.suse.com/sles/15-SP2/html/SLES-all/cha-dns.html.

5.2 Déploiement de Salt

SUSE Enterprise Storage utilise Salt et `ceph-salt` pour la préparation initiale de la grappe. Salt vous aide à configurer et à exécuter des commandes sur plusieurs noeuds de grappe simultanément à partir d'un hôte dédié appelé *Salt Master*. Avant de déployer Salt, tenez compte des points importants suivants :

- Les *minions Salt* sont les noeuds contrôlés par un noeud dédié appelé Salt Master.
- Si l'hôte Salt Master doit faire partie de la grappe Ceph, il doit exécuter son propre minion Salt, mais ce n'est pas obligatoire.

Astuce : partage de plusieurs rôles par serveur

Vous obtiendrez les meilleures performances de votre grappe Ceph lorsque chaque rôle est déployé sur un noeud distinct. Toutefois, les déploiements réels nécessitent parfois le partage d'un noeud entre plusieurs rôles. Pour éviter les problèmes liés aux performances et à la procédure de mise à niveau, ne déployez pas le rôle Ceph OSD, Serveur de métadonnées ou Ceph Monitor sur le noeud Admin.

- Les minions Salt doivent résoudre correctement le nom d'hôte de Salt Master sur le réseau. Par défaut, ils recherchent le nom d'hôte `salt`, mais vous pouvez spécifier tout autre nom d'hôte accessible par le réseau dans le fichier `/etc/salt/minion`.

1. Installez `salt-master` sur le noeud Salt Master :

```
root@master # zypper in salt-master
```

2. Vérifiez si le service `salt-master` est activé et démarré, puis activez-le et démarrez-le, le cas échéant :

```
root@master # systemctl enable salt-master.service
root@master # systemctl start salt-master.service
```

3. Si vous envisagez d'utiliser le pare-feu, vérifiez si les ports 4505 et 4506 du noeud Salt Master sont ouverts pour tous les noeuds minions Salt. Si les ports sont fermés, vous pouvez les ouvrir à l'aide de la commande **`yast2 firewall`** en autorisant le service *salt-master* à accéder à la zone appropriée. Par exemple, `public`.
4. Installez le paquetage `salt-minion` sur tous les noeuds des minions.

```
root@minion > zypper in salt-minion
```

5. Modifiez `/etc/salt/minion` et supprimez les commentaires de la ligne suivante :

```
#log_level_logfile: warning
```

Remplacez le niveau de consignation `warning` par `info`.

Remarque : `log_level_logfile` et `log_level`

Tandis que `log_level` contrôle quels messages du journal seront affichés à l'écran, `log_level_logfile` contrôle quels messages du journal seront écrits dans `/var/log/salt/minion`.

Remarque

Veillez à modifier le niveau de consignation sur *tous* les noeuds de grappe (minion).

6. Assurez-vous que le *nom de domaine complet* de chaque noeud peut être résolu sur une adresse IP du réseau de grappes public par tous les autres noeuds.

7. Configurez tous les minions pour qu'ils se connectent au maître. Si le nom d'hôte `salt` ne parvient pas à joindre votre Salt Master, modifiez le fichier `/etc/salt/minion` ou créez un fichier `/etc/salt/minion.d/master.conf` avec le contenu suivant :

```
master: host_name_of_salt_master
```

Si vous avez apporté des modifications aux fichiers de configuration mentionnés ci-dessus, redémarrez le service Salt sur tous les minions Salt qui y sont associés :

```
root@minion > systemctl restart salt-minion.service
```

8. Vérifiez que le service `salt-minion` est activé et démarré sur tous les noeuds. Activez et démarrez-le, le cas échéant :

```
root # systemctl enable salt-minion.service
root # systemctl start salt-minion.service
```

9. Vérifiez l'empreinte digitale de chaque minion Salt et acceptez toutes les clés Salt sur Salt Master si les empreintes digitales correspondent.

Remarque

Si l'empreinte de minion Salt revient vide, vérifiez que le minion Salt a une configuration Salt Master et qu'il peut communiquer avec Salt Master.

Affichez l'empreinte digitale de chaque minion :

```
root@minion > salt-call --local key.finger
local:
3f:a3:2f:3f:b4:d3:d9:24:49:ca:6b:2c:e1:6c:3f:c3:83:37:f0:aa:87:42:e8:ff...
```

Après avoir collecté les empreintes digitales de tous les minions Salt, répertoriez les empreintes de toutes les clés de minions non acceptées sur Salt Master :

```
root@master # salt-key -F
[...]
Unaccepted Keys:
minion1:
3f:a3:2f:3f:b4:d3:d9:24:49:ca:6b:2c:e1:6c:3f:c3:83:37:f0:aa:87:42:e8:ff...
```

Si les empreintes des minions correspondent, acceptez-les :

```
root@master # salt-key --accept-all
```

10. Vérifiez que les clés ont été acceptées :

```
root@master # salt-key --list-all
```

11. Testez si tous les minions Salt répondent :

```
root@master # salt-run manage.status
```

5.3 Déploiement de la grappe Ceph

Cette section vous guide tout au long du processus de déploiement d'une grappe Ceph de base. Lisez attentivement les sous-sections suivantes et exécutez les commandes incluses dans l'ordre indiqué.

5.3.1 Installation de `ceph-salt`

`ceph-salt` fournit des outils pour le déploiement de grappes Ceph gérées par cephdm. `ceph-salt` utilise l'infrastructure Salt pour gérer le système d'exploitation (par exemple, les mises à jour logicielles ou la synchronisation horaire) et définir les rôles pour les minions Salt.

Sur Salt Master, installez le paquetage `ceph-salt` :

```
root@master # zypper install ceph-salt
```

La commande ci-dessus a installé `ceph-salt-formula` en tant que dépendance qui a modifié la configuration de Salt Master en insérant des fichiers supplémentaires dans le répertoire `/etc/salt/master.d`. Pour appliquer les modifications, redémarrez `salt-master.service` et synchronisez les modules Salt :

```
root@master # systemctl restart salt-master.service
root@master # salt \* saltutil.sync_all
```

5.3.2 Configuration des propriétés de grappe

Utilisez la commande **`ceph-salt config`** pour configurer les propriétés de base de la grappe.

Important

Le fichier `/etc/ceph/ceph.conf` est géré par cephadm et les utilisateurs ne doivent *pas* le modifier. Les paramètres de configuration Ceph doivent être définis à l'aide de la nouvelle commande **`ceph config`**. Pour plus d'informations, reportez-vous au *Manuel « Guide d'opérations et d'administration », Chapitre 28 « Configuration de la grappe Ceph », Section 28.2 « Base de données de configuration »*.

5.3.2.1 Utilisation du shell `ceph-salt`

Si vous exécutez **`ceph-salt config`** sans chemin ni sous-commande, vous accédez à un shell `ceph-salt` interactif. Le shell est pratique si vous devez configurer plusieurs propriétés dans un seul lot sans devoir saisir la syntaxe complète de la commande.

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

```
o- / ..................................................... [...]
  o- ceph_cluster ....................................... [...]
  | o- minions ..................................... [no minions]
  | o- roles ....................................... [...]
  |   o- admin ..................................... [no minions]
  |   o- bootstrap ..................................... [no minion]
  |   o- cephadm ..................................... [no minions]
  |   o- tuned ....................................... [...]
  |     o- latency ..................................... [no minions]
  |     o- throughput ..................................... [no minions]
  o- cephadm_bootstrap ..................................... [...]
  | o- advanced ....................................... [...]
  | o- ceph_conf ....................................... [...]
  | o- ceph_image_path ............................... [ no image path]
  | o- dashboard ....................................... [...]
  | | o- force_password_update ............................... [enabled]
  | | o- password ..................................... [admin]
  | | o- ssl_certificate ..................................... [not set]
  | | o- ssl_certificate_key ..................................... [not set]
  | | o- username ..................................... [admin]
  | o- mon_ip ..................................... [not set]
  o- containers ..................................... [...]
  | o- registries_conf ..................................... [enabled]
  | | o- registries ..................................... [empty]
  | o- registry_auth ..................................... [...]
  |   o- password ..................................... [not set]
  |   o- registry ..................................... [not set]
  |   o- username ..................................... [not set]
  o- ssh ..................................... [no key pair set]
  | o- private_key ............................... [no private key set]
  | o- public_key ................................ [no public key set]
  o- time_server ......................... [enabled, no server host set]
    o- external_servers ..................................... [empty]
    o- servers ..................................... [empty]
    o- subnet ..................................... [not set]
```

Comme vous pouvez le constater dans la sortie de la commande `ls` de **`ceph-salt`**, la configuration de la grappe est organisée en arborescence. Pour configurer une propriété spécifique de la grappe dans le shell `ceph-salt`, deux options s'offrent à vous :

- Exécutez la commande à partir de la position actuelle et entrez le chemin absolu de la propriété comme premier argument :

```
/> /cephadm_bootstrap/dashboard ls
o- dashboard ..................................................... [...]
  o- force_password_update ................................... [enabled]
  o- password ..................................................... [admin]
```

```
  o- ssl_certificate ..................................... [not set]
  o- ssl_certificate_key ................................. [not set]
  o- username ................................................ [admin]
/> /cephadm_bootstrap/dashboard/username set ceph-admin
Value set.
```

- Accédez au chemin dont vous devez configurer la propriété et exécutez la commande :

```
/> cd /cephadm_bootstrap/dashboard/
/ceph_cluster/minions> ls
o- dashboard ..................................................... [...]
  o- force_password_update ..................................... [enabled]
  o- password ................................................... [admin]
  o- ssl_certificate ..................................... [not set]
  o- ssl_certificate_key ................................. [not set]
  o- username ..............................................[ceph-admin]
```

Astuce : saisie semi-automatique des extraits de configuration

Lorsque vous êtes dans un shell `ceph-salt`, vous pouvez utiliser la fonction de saisie semi-automatique similaire à celle d'un shell Linux normal (bash). Il complète les chemins de configuration, les sous-commandes ou les noms des minions Salt. Lors de la saisie semi-automatique d'un chemin de configuration, deux options s'offrent à vous :

- Pour laisser le shell terminer un chemin par rapport à votre position actuelle, appuyez deux fois sur la touche TAB. `→|`
- Pour laisser le shell terminer un chemin absolu, entrez `/` et appuyez deux fois sur la touche TAB `→|` .

Astuce : navigation avec les flèches du clavier

Si vous entrez **`cd`** à partir du shell `ceph-salt` sans aucun chemin, la commande imprime une structure d'arborescence de la configuration de grappe avec la ligne du chemin actuellement actif. Vous pouvez utiliser les flèches Haut et Bas pour parcourir les différentes lignes. Après avoir confirmé avec la touche `Entrée`, le chemin de configuration est remplacé par le dernier chemin actif.

Important : convention

Pour assurer la cohérence de la documentation, nous utiliserons une syntaxe de commande unique sans accéder au shell `ceph-salt`. Par exemple, vous pouvez répertorier l'arborescence de configuration de la grappe à l'aide de la commande suivante :

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

5.3.2.2 Ajout de minions Salt

Incluez tout ou partie des minions Salt que vous avez déployés et acceptés à la *Section 5.2, « Déploiement de Salt »* dans la configuration de la grappe Ceph. Vous pouvez spécifier les minions Salt à l'aide de leur nom complet ou utiliser une expression globale « * » et « ? » pour inclure plusieurs minions Salt simultanément. Utilisez la sous-commande **add** sous le chemin `/ceph_cluster/minions`. La commande suivante inclut tous les minions Salt acceptés :

```
root@master # ceph-salt config /ceph_cluster/minions add '*'
```

Vérifiez que les minions Salt spécifiés ont été ajoutés :

```
root@master # ceph-salt config /ceph_cluster/minions ls
o- minions ..................................................... [Minions: 5]
  o- ses-master.example.com ..................................... [no roles]
  o- ses-min1.example.com ....................................... [no roles]
  o- ses-min2.example.com ....................................... [no roles]
  o- ses-min3.example.com ....................................... [no roles]
  o- ses-min4.example.com ....................................... [no roles]
```

5.3.2.3 Spécification des minions Salt gérés par cephadm

Spécifiez les noeuds qui appartiendront à la grappe Ceph et qui seront gérés par cephadm. Incluez tous les noeuds qui exécuteront les services Ceph ainsi que le noeud Admin :

```
root@master # ceph-salt config /ceph_cluster/roles/cephadm add '*'
```

5.3.2.4 Spécification du noeud Admin

Le noeud Admin est le noeud sur lequel le fichier de configuration `ceph.conf` et le trousseau de clés Ceph admin sont installés. les commandes liées à Ceph sont généralement exécutées sur le noeud Admin.

Astuce : noeud Admin et Salt Master sur le même noeud

Dans un environnement homogène dans lequel tous les hôtes ou la plupart d'entre eux appartiennent à SUSE Enterprise Storage, il est recommandé de placer le noeud Admin sur le même hôte que Salt Master.

Dans un environnement hétérogène dans lequel une infrastructure Salt héberge plusieurs grappes, par exemple SUSE Enterprise Storage avec SUSE Manager, ne placez *pas* le noeud Admin sur le même hôte que Salt Master.

Pour spécifier le noeud Admin, exécutez la commande suivante :

```
root@master # ceph-salt config /ceph_cluster/roles/admin add ses-master.example.com
1 minion added.
root@master # ceph-salt config /ceph_cluster/roles/admin ls
o- admin ..................................................... [Minions: 1]
  o- ses-master.example.com ..................... [Other roles: cephadm]
```

Astuce : installation de `ceph.conf` et du trousseau de clés admin sur plusieurs noeuds

Vous pouvez installer le fichier de configuration Ceph et le trousseau de clés admin sur plusieurs noeuds si votre déploiement l'exige. Pour des raisons de sécurité, évitez de les installer sur tous les noeuds de la grappe.

5.3.2.5 Spécification du premier noeud MON/MGR

Vous devez spécifier quel minion Salt de la grappe va démarrer la grappe. Ce minion sera alors le premier à exécuter les services Ceph Monitor et Ceph Manager.

```
root@master # ceph-salt config /ceph_cluster/roles/bootstrap set ses-min1.example.com
Value set.
root@master # ceph-salt config /ceph_cluster/roles/bootstrap ls
```

```
o- bootstrap ..................................... [ses-min1.example.com]
```

En outre, vous devez spécifier l'adresse IP de démarrage du service MON sur le réseau public pour vous assurer que le paramètre `public_network` est correctement défini, par exemple :

```
root@master # ceph-salt config /cephadm_bootstrap/mon_ip set 192.168.10.20
```

5.3.2.6 Spécification de profils ajustés

Vous devez spécifier quels minions de la grappe ont des profils activement ajustés. Pour ce faire, ajoutez ces rôles explicitement à l'aide des commandes suivantes :

Remarque

Un minion ne peut pas cumuler les rôles de `latency` (latence) et `throughput` (débit).

```
root@master # ceph-salt config /ceph_cluster/roles/tuned/latency add ses-min1.example.com
Adding ses-min1.example.com...
1 minion added.
root@master # ceph-salt config /ceph_cluster/roles/tuned/throughput add ses-
min2.example.com
Adding ses-min2.example.com...
1 minion added.
```

5.3.2.7 Génération d'une paire de clés SSH

cephadm utilise le protocole SSH pour communiquer avec les noeuds de la grappe. Un compte utilisateur nommé `cephadm` est automatiquement créé et utilisé pour la communication SSH.

Vous devez générer la partie privée et la partie publique de la paire de clés SSH :

```
root@master # ceph-salt config /ssh generate
Key pair generated.
root@master # ceph-salt config /ssh ls
o- ssh ................................................ [Key Pair set]
  o- private_key ..... [53:b1:eb:65:d2:3a:ff:51:6c:e2:1b:ca:84:8e:0e:83]
  o- public_key ...... [53:b1:eb:65:d2:3a:ff:51:6c:e2:1b:ca:84:8e:0e:83]
```

5.3.2.8 Configuration du serveur horaire

L'heure de tous les noeuds de la grappe doit être synchronisée avec une source horaire fiable. Plusieurs scénarios existent pour aborder la synchronisation horaire :

- Si tous les noeuds de la grappe sont déjà configurés pour synchroniser leur heure à l'aide du service NTP de votre choix, désactivez complètement la gestion du serveur horaire :

```
root@master # ceph-salt config /time_server disable
```

- Si votre site possède déjà une seule source horaire, spécifiez le nom d'hôte de la source horaire :

```
root@master # ceph-salt config /time_server/servers add time-server.example.com
```

- Dans le cas contraire, `ceph-salt` a la possibilité de configurer l'un des minions Salt pour qu'il fasse office de serveur horaire pour le reste de la grappe. Il est parfois appelé « serveur horaire interne ». Dans ce scénario, `ceph-salt` configure le serveur horaire interne (qui doit être l'un des minions Salt) pour synchroniser son heure avec celle d'un serveur horaire externe, tel que `pool.ntp.org` et configure tous les autres minions pour obtenir leur heure à partir du serveur horaire interne. Pour ce faire, procédez comme suit :

```
root@master # ceph-salt config /time_server/servers add ses-master.example.com
root@master # ceph-salt config /time_server/external_servers add pool.ntp.org
```

  L'option `/time_server/subnet` spécifie le sous-réseau à partir duquel les clients NTP sont autorisés à accéder au serveur NTP. Elle est définie automatiquement lorsque vous spécifiez `/time_server/servers`. Si vous devez la modifier ou la spécifier manuellement, exécutez :

```
root@master # ceph-salt config /time_server/subnet set 10.20.6.0/24
```

Vérifiez les paramètres du serveur horaire :

```
root@master # ceph-salt config /time_server ls
o- time_server ............................................. [enabled]
  o- external_servers ........................................... [1]
  | o- pool.ntp.org ............................................ [...]
  o- servers .................................................... [1]
  | o- ses-master.example.com ................................... [...]
  o- subnet ............................................ [10.20.6.0/24]
```

Pour plus d'informations sur la configuration de la synchronisation horaire, reportez-vous à l'adresse https://documentation.suse.com/sles/15-SP2/html/SLES-all/cha-ntp.html#sec-ntp-yast.

5.3.2.9 Configuration des informations d'identification de connexion de Ceph Dashboard

Ceph Dashboard sera disponible après le déploiement de la grappe de base. Pour y accéder, vous devez définir un nom d'utilisateur et un mot de passe valides, par exemple :

```
root@master # ceph-salt config /cephadm_bootstrap/dashboard/username set admin
root@master # ceph-salt config /cephadm_bootstrap/dashboard/password set PWD
```

Astuce : mise à jour forcée du mot de passe

Par défaut, le premier utilisateur du tableau de bord est obligé de modifier son mot de passe lors de la première connexion au tableau de bord. Pour désactiver cette fonction, exécutez la commande suivante :

```
root@master # ceph-salt config /cephadm_bootstrap/dashboard/force_password_update
 disable
```

5.3.2.10 Configuration du chemin d'accès aux images du conteneur

cephadm doit avoir connaissance d'un chemin URI valide vers les images du conteneur qui sera utilisé au cours de l'étape de déploiement. Vérifiez si le chemin par défaut a été défini :

```
root@master # ceph-salt config /cephadm_bootstrap/ceph_image_path ls
```

Si aucun chemin par défaut n'a été défini ou si votre déploiement nécessite un chemin spécifique, ajoutez-le comme suit :

```
root@master # ceph-salt config /cephadm_bootstrap/ceph_image_path set registry.suse.com/
ses/7/ceph/ceph
```

Remarque

Pour la pile de surveillance, d'autres images de conteneur sont nécessaires. Pour un déploiement à vide ou à partir d'un registre local, vous souhaiterez peut-être obtenir ces images à ce stade afin de préparer votre registre local comme il se doit.

Notez que `ceph-salt` n'utilisera pas ces images de conteneur pour effectuer le déploiement. Il s'agit de préparatifs en vue d'une étape ultérieure lors de laquelle cephadm sera utilisé pour le déploiement ou la migration des composants de surveillance.

Pour plus d'informations sur les images utilisées par la pile de surveillance et sur la façon de les personnaliser, consultez la page *Manuel « Guide d'opérations et d'administration », Chapitre 16 « Surveillance et alertes », Section 16.1 « Configuration d'images personnalisées ou locales »*.

5.3.2.11 Configuration du registre de conteneurs

Vous pouvez éventuellement définir un registre local de conteneurs. Il fera office de miroir pour le registre `registration.suse.com`. N'oubliez pas que vous devez resynchroniser le registre local chaque fois que de nouveaux conteneurs mis à jour sont disponibles à partir du site `registration.suse.com`.

La création d'un registre local est utile dans les scénarios suivants :

- Vous disposez d'un grand nombre de noeuds de grappe et souhaitez réduire le temps de téléchargement et économiser de la bande passante en créant un miroir local des images de conteneur.
- Votre grappe n'a pas accès au registre en ligne (déploiement à vide) et vous avez besoin d'un miroir local pour extraire les images de conteneur.
- Si des problèmes de configuration ou de réseau empêchent votre grappe d'accéder aux registres distants par le biais d'un lien sécurisé, vous aurez alors besoin d'un registre local non chiffré.

Important

Pour déployer des correctifs temporaires de programme (PTF) sur un système pris en charge, vous devez déployer un registre local de conteneurs.

Pour configurer une URL de registre local avec des informations d'identification d'accès, procédez comme suit :

1. Configurez l'URL du registre local :

```
cephuser@adm > ceph-salt config /containers/registry_auth/registry set REGISTRY_URL
```

2. Configurez le nom d'utilisateur et le mot de passe pour accéder au registre local :

```
cephuser@adm > ceph-salt config /containers/registry_auth/username
 set REGISTRY_USERNAME
```

```
cephuser@adm > ceph-salt config /containers/registry_auth/password
 set REGISTRY_PASSWORD
```

3. Exécutez **`ceph-salt apply`** pour mettre à jour Salt Pillar sur tous les minions.

Astuce : cache de registre

Pour éviter de resynchroniser le registre local lorsque de nouveaux conteneurs mis à jour apparaissent, vous pouvez configurer un *cache de registre*.

Les méthodes de développement et de distribution des applications natives dans le cloud nécessitent un registre et une instance CI/CD (intégration/distribution continue) pour le développement et la production d'images de conteneur. Vous pouvez utiliser un registre privé dans cette instance.

5.3.2.12 Activation du chiffrement des données à la volée (msgr2)

Le protocole Messenger v2 (MSGR2) est le protocole on-wire de Ceph. Il offre un mode de sécurité qui chiffre toutes les données passant sur le réseau, l'encapsulation des charges utiles d'authentification et l'activation de l'intégration future de nouveaux modes d'authentification (tels que Kerberos).

Important

msgr2 n'est actuellement pas pris en charge par les clients Ceph du kernel Linux, tels que le périphérique de bloc RADOS et CephFS.

Les daemons Ceph peuvent se lier à plusieurs ports, ce qui permet aux clients Ceph hérités et aux nouveaux clients compatibles v2 de se connecter à la même grappe. Par défaut, les instances MON se lient désormais au nouveau port 3300 assigné par l'IANA (CE4h ou 0xCE4) pour le nouveau protocole v2, tout en se liant également à l'ancien port par défaut 6789 pour le protocole v1 hérité.

Le protocole v2 (MSGR2) prend en charge deux modes de connexion :

crc mode
: Authentification initiale forte lorsque la connexion est établie et contrôle d'intégrité CRC32C.

secure mode
: Authentification initiale forte lorsque la connexion est établie et un chiffrement complet de tout le trafic post-authentification, y compris une vérification de l'intégrité cryptographique.

Pour la plupart des connexions, il existe des options qui contrôlent les modes utilisés :

ms_cluster_mode
: Mode de connexion (ou modes autorisés) utilisé(s) pour la communication entre les daemons Ceph au sein de la grappe. Si plusieurs modes sont répertoriés, ceux s'affichant en haut de la liste sont les préférés.

ms_service_mode
: Liste des modes que les clients sont autorisés à utiliser lors de la connexion à la grappe.

ms_client_mode
: Liste des modes de connexion, par ordre de préférence, que les clients peuvent ou sont autorisés à utiliser lorsqu'ils communiquent avec une grappe Ceph.

Il existe un ensemble parallèle d'options qui s'appliquent spécifiquement aux moniteurs, ce qui permet aux administrateurs de définir d'autres exigences (généralement plus sécurisées) en matière de communication avec les moniteurs.

ms_mon_cluster_mode
: Mode de connexion (ou modes autorisés) à utiliser entre les moniteurs.

ms_mon_service_mode
: Liste des modes autorisés que les clients ou les autres daemons Ceph sont autorisés à utiliser lors de la connexion aux moniteurs.

ms_mon_client_mode
: Liste des modes de connexion, par ordre de préférence, que les clients ou les daemons non-moniteurs peuvent utiliser pour se connecter à des moniteurs.

Pour activer le mode de chiffrement MSGR2 pendant le déploiement, vous devez ajouter certaines options de configuration à la configuration de `ceph-salt` avant d'exécuter **`ceph-salt apply`**.

Pour utiliser le mode `secure`, exécutez les commandes suivantes.

Ajoutez la section globale à `ceph_conf` dans l'outil de configuration `ceph-salt` :

```
root@master # ceph-salt config /cephadm_bootstrap/ceph_conf add global
```

Paramétrez les options suivantes :

```
root@master # ceph-salt config /cephadm_bootstrap/ceph_conf/global set ms_cluster_mode
 "secure crc"
root@master # ceph-salt config /cephadm_bootstrap/ceph_conf/global set ms_service_mode
 "secure crc"
root@master # ceph-salt config /cephadm_bootstrap/ceph_conf/global set ms_client_mode
 "secure crc"
```

Remarque

Vérifiez que `secure` se trouve devant `crc`.

Pour *forcer le mode* `secure`, exécutez les commandes suivantes :

```
root@master # ceph-salt config /cephadm_bootstrap/ceph_conf/global set ms_cluster_mode
 secure
root@master # ceph-salt config /cephadm_bootstrap/ceph_conf/global set ms_service_mode
 secure
root@master # ceph-salt config /cephadm_bootstrap/ceph_conf/global set ms_client_mode
 secure
```

Astuce : mise à jour des paramètres

Si vous souhaitez modifier l'un des paramètres ci-dessus, définissez les modifications à apporter à la configuration dans la zone de stockage de la configuration du moniteur. Pour ce faire, utilisez la commande **`ceph config set`**.

```
root@master # ceph config set global CONNECTION_OPTION CONNECTION_MODE [--force]
```

Par exemple :

```
root@master # ceph config set global ms_cluster_mode "secure crc"
```

Si vous souhaitez vérifier la valeur actuelle, y compris la valeur par défaut, exécutez la commande suivante :

```
root@master # ceph config get CEPH_COMPONENT CONNECTION_OPTION
```

Par exemple, pour obtenir le mode `ms_cluster_mode` pour les OSD, exécutez :

```
root@master # ceph config get osd ms_cluster_mode
```

5.3.2.13 Configuration du réseau de grappes

Éventuellement, si vous exécutez un réseau de grappe distinct, vous devrez peut-être définir l'adresse IP réseau de la grappe suivie de la partie masque de sous-réseau après la barre oblique, par exemple `192.168.10.22/24`.

Exécutez les commandes suivantes pour activer `cluster_network` :

```
root@master # ceph-salt config /cephadm_bootstrap/ceph_conf add global
root@master # ceph-salt config /cephadm_bootstrap/ceph_conf/global set
 cluster_network NETWORK_ADDR
```

5.3.2.14 Vérification de la configuration de la grappe

La configuration de grappe minimale est terminée. Inspectez-la pour voir si elle contient des d'erreurs flagrantes :

```
root@master # ceph-salt config ls
o- / ..................................................................... [...]
  o- ceph_cluster ........................................................ [...]
  | o- minions ..................................... [Minions: 5]
  | | o- ses-master.example.com ..................... [admin]
  | | o- ses-min1.example.com ..................... [bootstrap, admin]
  | | o- ses-min2.example.com ..................... [no roles]
  | | o- ses-min3.example.com ..................... [no roles]
  | | o- ses-min4.example.com ..................... [no roles]
  | o- roles ..................................... [...]
  |   o- admin ..................................... [Minions: 2]
  |   | o- ses-master.example.com ..... [no other roles]
  |   | o- ses-min1.example.com ..... [other roles: bootstrap]
  |   o- bootstrap ..... [ses-min1.example.com]
```

```
|   o- cephadm ..................................... [Minions: 5]
|   o- tuned ....................................................... [...]
|     o- latency ..................................... [no minions]
|     o- throughput .................................. [no minions]
o- cephadm_bootstrap ..................................... [...]
| o- advanced ..................................................... [...]
| o- ceph_conf .................................................... [...]
| o- ceph_image_path ............... [registry.suse.com/ses/7/ceph/ceph]
| o- dashboard .................................................... [...]
|   o- force_password_update ............................... [enabled]
|   o- password ................................ [randomly generated]
|   o- username ................................................ [admin]
| o- mon_ip ....................................... [192.168.10.20]
o- containers ................................................... [...]
| o- registries_conf ..................................... [enabled]
| | o- registries ........................................... [empty]
| o- registry_auth ........................................... [...]
|   o- password ............................................ [not set]
|   o- registry ............................................ [not set]
|   o- username ............................................ [not set]
o- ssh ..................................................... [Key Pair set]
| o- private_key ..... [53:b1:eb:65:d2:3a:ff:51:6c:e2:1b:ca:84:8e:0e:83]
| o- public_key ...... [53:b1:eb:65:d2:3a:ff:51:6c:e2:1b:ca:84:8e:0e:83]
o- time_server ..................................... [enabled]
  o- external_servers ..................................... [1]
  | o- 0.pt.pool.ntp.org ..................................... [...]
  o- servers ..................................................... [1]
  | o- ses-master.example.com ................................ [...]
  o- subnet ..................................... [10.20.6.0/24]
```

Astuce : état de la configuration de la grappe

Vous pouvez vérifier si la configuration de la grappe est valide en exécutant la commande suivante :

```
root@master # ceph-salt status
cluster: 5 minions, 0 hosts managed by cephadm
config: OK
```

5.3.2.15 Exportation des configurations de grappe

Une fois que vous avez configuré la grappe de base et que sa configuration est valide, il est recommandé d'exporter sa configuration dans un fichier :

```
root@master # ceph-salt export > cluster.json
```

Avertissement

La sortie de l'exportation **`ceph-salt export`** inclut la clé privée SSH. Si les implications au niveau de la sécurité vous inquiètent, n'exécutez pas cette commande sans prendre les précautions appropriées.

Si vous interrompez la configuration de la grappe et que vous devez rétablir un état de sauvegarde, exécutez :

```
root@master # ceph-salt import cluster.json
```

5.3.3 Mise à jour des noeuds et de la grappe minimale de démarrage

Avant de déployer la grappe, mettez à jour tous les paquetages logiciels sur l'ensemble des noeuds :

```
root@master # ceph-salt update
```

Si un noeud signale `Reboot is needed` (Redémarrage requis) au cours de la mise à jour, cela signifie que des paquetages de système d'exploitation importants, tels que le kernel, ont été mis à jour vers une version plus récente. Vous devrez alors redémarrer le noeud pour que les modifications soient prises en compte.

Pour redémarrer tous les noeuds requis, ajoutez l'option `--reboot`.

```
root@master # ceph-salt update --reboot
```

Vous pouvez également les redémarrer dans le cadre d'une étape distincte :

```
root@master # ceph-salt reboot
```

Important

Salt Master n'est jamais redémarré par les commandes **`ceph-salt update --reboot`** ou **`ceph-salt reboot`**. Si Salt Master doit être redémarré, vous devez le redémarrer manuellement.

Une fois les noeuds mis à jour, démarrez la grappe minimale :

```
root@master # ceph-salt apply
```

Remarque

Une fois démarrée, la grappe dispose d'une instance de Ceph Monitor et d'une instance de Ceph Manager.

La commande ci-dessus ouvre une interface utilisateur interactive qui affiche la progression actuelle de chaque minion.

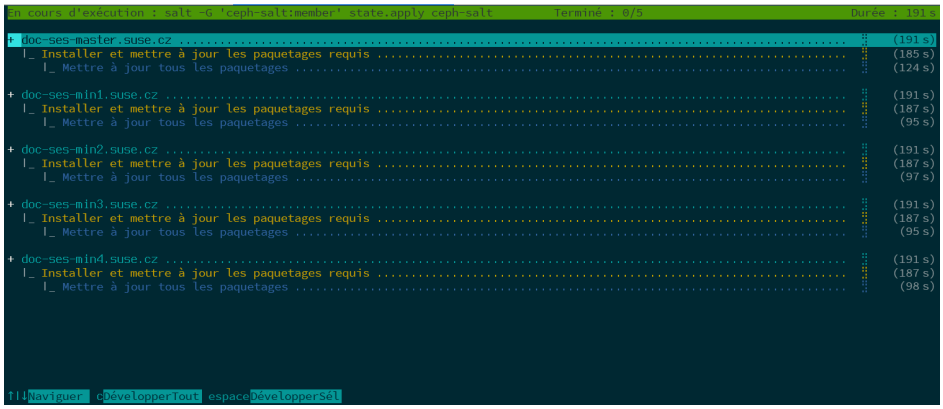

FIGURE 5.1 : DÉPLOIEMENT D'UNE GRAPPE MINIMALE

Astuce : mode non interactif

Si vous devez appliquer la configuration à partir d'un script, il existe également un mode de déploiement non interactif. Ce mode est aussi utile lors du déploiement de la grappe à partir d'une machine distante, car la mise à jour constante des informations de progression à l'écran sur le réseau risque de vous distraire :

```
root@master # ceph-salt apply --non-interactive
```

5.3.4 Examen des dernières étapes

Une fois la commande **`ceph-salt apply`** exécutée, vous devez disposer d'une instance de Ceph Monitor et d'une instance de Ceph Manager. Vous devriez pouvoir exécuter la commande **`ceph status`** sur n'importe quel minion ayant reçu le rôle `admin` en tant que `root` ou sur l'utilisateur de `cephadm` à l'aide de `sudo`.

Pour les étapes suivantes, vous devrez utiliser cephadm pour déployer des instances Ceph Monitor ou Ceph Manager, des OSD, la pile de surveillance et des passerelles supplémentaires.

Avant de continuer, passez en revue les paramètres réseau de votre nouvelle grappe. À ce stade, le paramètre `public_network` a été renseigné en fonction de ce qui a été entré pour `/cephadm_bootstrap/mon_ip` dans la configuration `ceph-salt`. Toutefois, ce paramètre était uniquement appliqué à Ceph Monitor. Vous pouvez vérifier ce paramètre à l'aide de la commande suivante :

```
root@master # ceph config get mon public_network
```

Il s'agit du paramètre minimal pour que Ceph puisse fonctionner, mais nous vous recommandons de définir le paramètre `public_network` sur `global`, ce qui signifie qu'il s'appliquera à tous les types de daemons Ceph, et pas uniquement aux instances MON :

```
root@master # ceph config set global public_network "$(ceph config get mon
 public_network)"
```

Remarque

Cette étape n'est pas obligatoire. Toutefois, si vous n'utilisez pas ce paramètre, les OSD Ceph et les autres daemons (à l'exception de Ceph Monitor) écouteront *toutes les adresses*.

Si vous souhaitez que vos OSD communiquent entre eux au moyen d'un réseau totalement distinct, exécutez la commande suivante :

```
root@master # ceph config set global cluster_network
 "cluster_network_in_cidr_notation"
```

L'exécution de cette commande garantit que les OSD créés dans votre déploiement utiliseront le réseau de grappe prévu dès le début.

Si votre grappe est configurée pour contenir des noeuds denses (plus de 62 OSD par hôte), veillez à assigner suffisamment de ports aux OSD Ceph. La plage par défaut (6800-7300) n'autorise pas plus de 62 OSD par hôte. Pour une grappe contenant des noeuds denses, configurez le

paramètre `ms_bind_port_max` sur une valeur appropriée. Chaque OSD consomme huit ports supplémentaires. Par exemple, si un hôte est configuré pour exécuter 96 OSD, 768 ports seront nécessaires. `ms_bind_port_max` doit au moins être défini sur 7568 en exécutant la commande suivante :

```
root@master # ceph config set osd.* ms_bind_port_max 7568
```

Vous devrez ajuster vos paramètres de pare-feu en conséquence pour que cela fonctionne. Pour plus d'informations, reportez-vous au *Manuel « Troubleshooting Guide », Chapitre 13 « Hints and tips », Section 13.7 « Firewall settings for Ceph »*.

5.4 Déploiement de services et de passerelles

Après avoir déployé la grappe Ceph de base, déployez les services principaux sur d'autres noeuds de la grappe. Pour que les clients puissent accéder aux données de la grappe, déployez aussi des services supplémentaires.

Actuellement, nous prenons en charge le déploiement des services Ceph sur la ligne de commande à l'aide de l'orchestrateur Ceph (sous-commandes **`ceph orch`**).

5.4.1 Commande **`ceph orch`**

La commande de l'orchestrateur Ceph **`ceph orch`**, qui sert d'interface avec le module cephadm, se charge d'établir la liste les composants de la grappe et de déployer les services Ceph sur les nouveaux noeuds de la grappe.

5.4.1.1 Affichage de l'état de l'orchestrateur

La commande suivante affiche le mode et l'état actuels de l'orchestrateur Ceph.

```
cephuser@adm > ceph orch status
```

5.4.1.2 Liste des périphériques, services et daemons

Pour obtenir la liste de tous les périphériques de disque, exécutez la commande suivante :

```
cephuser@adm > ceph orch device ls
Hostname   Path      Type  Serial  Size   Health   Ident  Fault  Available
ses-master /dev/vdb  hdd   0d8a... 10.7G  Unknown  N/A    N/A    No
```

```
ses-min1   /dev/vdc  hdd   8304... 10.7G  Unknown  N/A    N/A    No
ses-min1   /dev/vdd  hdd   7b81... 10.7G  Unknown  N/A    N/A    No
[...]
```

Astuce : services et daemons

Service est un terme général désignant un service Ceph d'un type spécifique, par exemple Ceph Manager.

Daemon est une instance spécifique d'un service, par exemple un processus `mgr.ses-min1.gdlcik` exécuté sur un noeud `appelé ses-min1`.

Pour obtenir la liste de tous les services connus de cephadm, exécutez :

```
cephuser@adm > ceph orch ls
NAME  RUNNING  REFRESHED  AGE  PLACEMENT  IMAGE NAME                  IMAGE ID
mgr       1/0  5m ago     -    <no spec>  registry.example.com/[...]  5bf12403d0bd
mon       1/0  5m ago     -    <no spec>  registry.example.com/[...]  5bf12403d0bd
```

Astuce

Vous pouvez limiter la liste aux services d'un noeud en particulier avec le paramètre facultatif `--host` et aux services d'un type particulier avec le paramètre facultatif `--service-type` (les types acceptables sont `mon`, `osd`, `mgr`, `mds` et `rgw`).

Pour obtenir la liste de tous les daemons en cours d'exécution déployés par cephadm, exécutez :

```
cephuser@adm > ceph orch ps
NAME            HOST     STATUS   REFRESHED AGE VERSION    IMAGE ID     CONTAINER ID
mgr.ses-min1.gd ses-min1 running) 8m ago    12d 15.2.0.108 5bf12403d0bd b8104e09814c
mon.ses-min1    ses-min1 running) 8m ago    12d 15.2.0.108 5bf12403d0bd a719e0087369
```

Astuce

Pour interroger l'état d'un daemon en particulier, utilisez `--daemon_type` et `--daemon_id`. Pour les OSD, l'ID est l'ID numérique de l'OSD. Pour MDS, l'ID est le nom du système de fichiers :

```
cephuser@adm > ceph orch ps --daemon_type osd --daemon_id 0
cephuser@adm > ceph orch ps --daemon_type mds --daemon_id my_cephfs
```

5.4.2 Spécification de service et de placement

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.

5.4.2.1 Création de spécifications de service

Vous pouvez créer un fichier de spécifications distinct pour chaque type de service, par exemple :

```
root@master # cat nfs.yml
service_type: nfs
service_id: EXAMPLE_NFS
placement:
  hosts:
  - ses-min1
  - ses-min2
spec:
  pool: EXAMPLE_POOL
  namespace: EXAMPLE_NAMESPACE
```

Vous pouvez également spécifier plusieurs (ou tous les) types de service dans un fichier (par exemple, `cluster.yml`) pour décrire les noeuds qui exécuteront des services spécifiques. N'oubliez pas de séparer les différents types de services par trois tirets (`---`) :

```
cephuser@adm > cat cluster.yml
service_type: nfs
service_id: EXAMPLE_NFS
placement:
  hosts:
  - ses-min1
  - ses-min2
spec:
  pool: EXAMPLE_POOL
  namespace: EXAMPLE_NAMESPACE
---
service_type: rgw
service_id: REALM_NAME.ZONE_NAME
placement:
  hosts:
  - ses-min1
  - ses-min2
  - ses-min3
---
[...]
```

Les propriétés susmentionnées ont la signification suivante :

`service_type`
: Type de service. Il peut s'agir d'un service Ceph (`mon`, `mgr`, `mds`, `crash`, `osd` ou `rbd-mirror`), d'une passerelle (`nfs` ou `rgw`) ou d'une partie de la pile de surveillance (`alertmanager`, `grafana`, `node-exporter` ou `prometheus`).

`service_id`
: Nom du service. Les spécifications de type `mon`, `mgr`, `alertmanager`, `grafana`, `node-exporter` et `prometheus` ne nécessitent pas la propriété `service_id`.

`placement`
: Spécifie les noeuds qui exécuteront le service. Reportez-vous à la *Section 5.4.2.2, « Création d'une spécification de placement »* pour plus de détails.

`spec`
: Spécification supplémentaire pertinente pour le type de service.

Astuce : application de services spécifiques

Les services de grappe Ceph ont généralement un certain nombre de propriétés intrinsèques. Pour obtenir des exemples et des détails sur la spécification des différents services, reportez-vous à la *Section 5.4.3, « Déploiement des services Ceph »*.

5.4.2.2 Création d'une spécification de placement

Pour déployer les services Ceph, cephadm doit savoir sur quels noeuds les déployer. Utilisez la propriété `placement` et répertoriez les noms d'hôte courts des noeuds auxquels le service s'applique :

```
cephuser@adm > cat cluster.yml
[...]
placement:
  hosts:
  - host1
  - host2
  - host3
[...]
```

5.4.2.3 Application de la spécification de grappe

Après avoir créé un fichier `cluster.yml` complet avec les spécifications de tous les services et leur placement, vous pouvez appliquer la grappe en exécutant la commande suivante :

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

Pour afficher l'état de la grappe, exécutez la commande **`ceph orch status`**. Pour plus de détails, reportez-vous à la *Section 5.4.1.1, « Affichage de l'état de l'orchestrateur »*.

5.4.2.4 Exportation de la spécification d'une grappe en cours d'exécution

Bien que vous ayez déployé des services sur la grappe Ceph à l'aide des fichiers de spécifications comme décrit à la *Section 5.4.2, « Spécification de service et de placement »*, la configuration de la grappe peut être différente de la spécification d'origine lors de son fonctionnement. Il se peut également que vous ayez accidentellement supprimé les fichiers de spécifications.

Pour récupérer une spécification complète d'une grappe en cours d'exécution, exécutez :

```
cephuser@adm > ceph orch ls --export
placement:
  hosts:
  - hostname: ses-min1
    name: ''
    network: ''
service_id: my_cephfs
service_name: mds.my_cephfs
service_type: mds
---
placement:
  count: 2
service_name: mgr
service_type: mgr
---
[...]
```

Astuce

Vous pouvez ajouter l'option `--format` pour modifier le format de sortie `yaml` par défaut. Vous pouvez choisir entre `json`, `json-attractive` ou `yaml`. Par exemple :

```
ceph orch ls --export --format json
```

5.4.3 Déploiement des services Ceph

Une fois la grappe de base en cours d'exécution, vous pouvez déployer des services Ceph sur d'autres noeuds.

5.4.3.1 Déploiement d'instances de Monitor et Ceph Manager

La grappe Ceph comporte trois ou cinq instances MON déployées sur différents noeuds. Si la grappe contient au moins cinq noeuds, il est recommandé de déployer cinq instances MON. Une bonne pratique consiste à déployer les instances MGR sur les mêmes noeuds que les instances MON.

Important : inclusion de l'instance MON de démarrage

Lorsque vous déployez des instances MON et MGR, n'oubliez pas d'inclure la première instance MON que vous avez ajoutée lors de la configuration de la grappe de base à la *Section 5.3.2.5, « Spécification du premier noeud MON/MGR »*.

Pour déployer des instances MON, appliquez la spécification suivante :

```
service_type: mon
placement:
  hosts:
  - ses-min1
  - ses-min2
  - ses-min3
```

Remarque

Si vous devez ajouter un autre noeud, ajoutez le nom d'hôte à la même liste YAML. Par exemple :

```
service_type: mon
placement:
 hosts:
 - ses-min1
 - ses-min2
 - ses-min3
 - ses-min4
```

De la même manière, pour déployer des instances MGR, appliquez la spécification suivante :

Important

Assurez-vous que votre déploiement contient au moins trois instances Ceph Manager dans chaque déploiement.

```
service_type: mgr
placement:
  hosts:
  - ses-min1
  - ses-min2
  - ses-min3
```

Astuce

Si les instances MON ou MGR ne se trouvent *pas* dans le même sous-réseau, vous devez ajouter les adresses de sous-réseau. Par exemple :

```
service_type: mon
placement:
  hosts:
  - ses-min1:10.1.2.0/24
  - ses-min2:10.1.5.0/24
  - ses-min3:10.1.10.0/24
```

5.4.3.2 Déploiement des OSD Ceph

Important : lorsque le périphérique de stockage est disponible

Un périphérique de stockage est considéré comme *disponible* si toutes les conditions suivantes sont remplies :

- Le périphérique n'a pas de partitions.
- Le périphérique n'a pas d'état LVM.
- Le périphérique n'est pas monté.
- Le périphérique ne contient pas de système de fichiers.

- Le périphérique ne contient pas d'OSD BlueStore.
- La taille du périphérique est supérieure à 5 Go.

Si les conditions ci-dessus ne sont pas remplies, Ceph refuse de provisionner ces OSD.

Deux méthodes permettent de déployer des OSD :

- Indiquez à Ceph de consommer tous les périphériques de stockage disponibles et inutilisés :

```
cephuser@adm > ceph orch apply osd --all-available-devices
```

- Utilisez DriveGroups (reportez-vous au *Manuel « Guide d'opérations et d'administration », Chapitre 13 « Tâches opérationnelles », Section 13.4.3 « Ajout d'OSD à l'aide de la spécification DriveGroups »*) pour créer une spécification OSD décrivant les périphériques qui seront déployés sur la base de leurs propriétés, telles que le type de périphérique (SSD ou HDD), les noms de modèle de périphérique, la taille ou les noeuds sur lesquels les périphériques existent. Appliquez ensuite la spécification en exécutant la commande suivante :

```
cephuser@adm > ceph orch apply osd -i drive_groups.yml
```

5.4.3.3 Déploiement de serveurs de métadonnées

CephFS nécessite un ou plusieurs services de serveur de métadonnées (MDS). Pour créer un CephFS, commencez par créer des serveurs MDS en appliquant la spécification suivante :

Remarque

Assurez-vous d'avoir créé au moins deux réserves, l'une pour les données CephFS et l'autre pour les métadonnées CephFS, avant d'appliquer la spécification suivante.

```
service_type: mds
service_id: CEPHFS_NAME
placement:
  hosts:
  - ses-min1
  - ses-min2
  - ses-min3
```

Une fois que les MDS sont fonctionnels, créez le CephFS :

```
ceph fs new CEPHFS_NAME metadata_pool data_pool
```

5.4.3.4 Déploiement d'instances Object Gateway

cephadm déploie Object Gateway en tant qu'ensemble de daemons qui gèrent un *domaine* et une *zone* spécifique.

Vous pouvez soit associer un service Object Gateway à une zone et un domaine préexistants (reportez-vous au *Manuel « Guide d'opérations et d'administration », Chapitre 21 « Ceph Object Gateway », Section 21.13 « Passerelles Object Gateway multisites »* pour plus de détails) ou vous pouvez spécifier des noms *NOM_DOMAINE* et *NOM_ZONE* n'existant pas encore. Ils seront créés automatiquement après avoir appliqué la configuration suivante :

```
service_type: rgw
service_id: REALM_NAME.ZONE_NAME
placement:
  hosts:
  - ses-min1
  - ses-min2
  - ses-min3
spec:
  rgw_realm: RGW_REALM
  rgw_zone: RGW_ZONE
```

5.4.3.4.1 Utilisation d'un accès SSL sécurisé

Pour utiliser une connexion SSL sécurisée à l'instance Object Gateway, vous avez besoin d'une paire de certificats SSL et de fichiers de clé valides (reportez-vous au *Manuel « Guide d'opérations et d'administration », Chapitre 21 « Ceph Object Gateway », Section 21.7 « Activation de HTTPS/SSL pour les passerelles Object Gateway »* pour plus de détails). Vous devez activer SSL, spécifier un numéro de port pour les connexions SSL, ainsi que le certificat SSL et les fichiers de clé.

Pour activer SSL et spécifier le numéro de port, incluez les éléments suivants dans votre spécification :

```
spec:
  ssl: true
  rgw_frontend_port: 443
```

Pour spécifier le certificat et la clé SSL, vous pouvez coller leur contenu directement dans le fichier de spécifications YAML. Le signe de barre verticale (|) à la fin de la ligne indique à l'analyseur de s'attendre à une valeur contenant une chaîne s'étendant sur plusieurs lignes. Par exemple :

```
spec:
  ssl: true
  rgw_frontend_port: 443
  rgw_frontend_ssl_certificate: |
   -----BEGIN CERTIFICATE-----
   MIIFmjCCA4KgAwIBAgIJAIZ2n35bmwXTMA0GCSqGSIb3DQEBCwUAMGIxCzAJBgNV
   BAYTAkFVMQwwCgYDVQQIDANOU1cxHTAbBgNVBAoMFEV4YW1wbGUgUkdXIFNTTCBp
   [...]
   -----END CERTIFICATE-----
   rgw_frontend_ssl_key: |
   -----BEGIN PRIVATE KEY-----
   MIIJRAIBADANBgkqhkiG9w0BAQEFAASCCS4wggkqAgEAAoICAQDLtFwg6LLl2j4Z
   BDV+iL4AO7VZ9KbmWIt37Ml2W6y2YeKX3Qwf+3eBz7TVHR1dm6iPpCpqpQjXUsT9
   [...]
   -----END PRIVATE KEY-----
```

Astuce

Au lieu de coller le contenu du certificat SSL et des fichiers de clé, vous pouvez omettre les mots-clés `rgw_frontend_ssl_certificate:` et `rgw_frontend_ssl_key:` et les télécharger dans la base de données de configuration :

```
cephuser@adm > ceph config-key set rgw/cert/REALM_NAME/ZONE_NAME.crt \
 -i SSL_CERT_FILE
cephuser@adm > ceph config-key set rgw/cert/REALM_NAME/ZONE_NAME.key \
 -i SSL_KEY_FILE
```

5.4.3.4.2 Déploiement avec une sous-grappe

Les *sous-grappes* vous aident à organiser les noeuds de vos grappes afin d'isoler les workloads et de gagner en flexibilité. Si vous effectuez un déploiement avec une sous-grappe, appliquez la configuration suivante :

```
service_type: rgw
service_id: REALM_NAME.ZONE_NAME.SUBCLUSTER
placement:
```

```
  hosts:
  - ses-min1
  - ses-min2
  - ses-min3
spec:
  rgw_realm: RGW_REALM
  rgw_zone: RGW_ZONE
  subcluster: SUBCLUSTER
```

5.4.3.5 Déploiement de passerelles iSCSI

cephadm déploie une passerelle iSCSI, à savoir un protocole SAN (Storage area network) qui permet aux clients (appelés initiateurs) d'envoyer des commandes SCSI aux périphériques de stockage SCSI (cibles) sur des serveurs distants.

Appliquez la configuration suivante pour effectuer le déploiement. Assurez-vous que la liste `trusted_ip_list` contient les adresses IP de tous les noeuds de passerelle iSCSI et Ceph Manager (comme dans l'exemple de sortie ci-dessous).

Remarque

Vérifiez que la réserve a été créée avant d'appliquer la spécification suivante.

```
service_type: iscsi
service_id: EXAMPLE_ISCSI
placement:
  hosts:
  - ses-min1
  - ses-min2
  - ses-min3
spec:
  pool: EXAMPLE_POOL
  api_user: EXAMPLE_USER
  api_password: EXAMPLE_PASSWORD
  trusted_ip_list: "IP_ADDRESS_1,IP_ADDRESS_2"
```

Remarque

Assurez-vous que les adresses IP répertoriées pour `trusted_ip_list` ne contiennent *pas* d'espace après la séparation par une virgule.

5.4.3.5.1 Configuration SSL sécurisée

Pour utiliser une connexion SSL sécurisée entre Ceph Dashboard et l'API cible iSCSI, vous avez besoin d'une paire de fichiers de clés et de certificats SSL valides. Ceux-ci peuvent être émis par une autorité de certification ou auto-signés (reportez-vous au *Manuel « Guide d'opérations et d'administration », Chapitre 10 « Configuration manuelle », Section 10.1.1 « Création de certificats auto-signés »*). Pour activer SSL, ajoutez le paramètre `api_secure: true` à votre fichier de spécifications :

```
spec:
  api_secure: true
```

Pour spécifier le certificat et la clé SSL, vous pouvez coller le contenu directement dans le fichier de spécifications YAML. Le signe de barre verticale (`|`) à la fin de la ligne indique à l'analyseur de s'attendre à une valeur contenant une chaîne s'étendant sur plusieurs lignes. Par exemple :

```
spec:
  pool: EXAMPLE_POOL
  api_user: EXAMPLE_USER
  api_password: EXAMPLE_PASSWORD
  trusted_ip_list: "IP_ADDRESS_1,IP_ADDRESS_2"
  api_secure: true
  ssl_cert: |
    -----BEGIN CERTIFICATE-----
    MIIDtTCCAp2gAwIBAgIYMC4xNzc1NDQxNjEzMzc2MjMyXzxvQ7EcMA0GCSqGSIb3
    DQEBCwUAMG0xCzAJBgNVBAYTAlVTMQ0wCwYDVQQIDARVdGFoMRcwFQYDVQQHDA5T
    [...]
    -----END CERTIFICATE-----
  ssl_key: |
    -----BEGIN PRIVATE KEY-----
    MIIEvQIBADANBgkqhkiG9w0BAQEFAASCBKcwggSjAgEAAoIBAQC5jdYbjtNTAKW4
    /CwQr/7wOiLGzVxChn3mmCIF3DwbL/qvTFTX2d8bDf6LjGwLYloXHscRfxszX/4h
    [...]
    -----END PRIVATE KEY-----
```

5.4.3.6 Déploiement de NFS Ganesha

cephadm déploie NFS Ganesha à l'aide d'une réserve RADOS prédéfinie et d'un espace de nom facultatif. Pour déployer NFS Ganesha, appliquez la spécification suivante :

Remarque

Vous devez disposer d'une réserve RADOS prédéfinie, sinon l'opération **`ceph orch apply`** échouera. Pour plus d'informations sur la création d'une, reportez-vous au *Manuel « Guide d'opérations et d'administration », Chapitre 18 « Gestion des réserves de stockage », Section 18.1 « Création d'une réserve »*.

```
service_type: nfs
service_id: EXAMPLE_NFS
placement:
  hosts:
  - ses-min1
  - ses-min2
spec:
  pool: EXAMPLE_POOL
  namespace: EXAMPLE_NAMESPACE
```

- *EXEMPLE_NFS* avec une chaîne arbitraire qui identifie l'exportation NFS.
- *EXEMPLE_RÉSERVE* avec le nom de la réserve dans laquelle l'objet de configuration NFS Ganesha RADOS sera stocké.
- *EXEMPLE_ESPACE_DE_NOM* (facultatif) avec l'espace de nom NFS Object Gateway souhaité (par exemple, `ganesha`).

5.4.3.7 Déploiement de `rbd-mirror`

Le service `rbd-mirror` prend en charge la synchronisation des images de périphériques de bloc RADOS entre deux grappes Ceph (pour plus de détails, reportez-vous au *Manuel « Guide d'opérations et d'administration », Chapitre 20 « Périphérique de bloc RADOS », Section 20.4 « Miroirs d'image RBD »*). Pour déployer `rbd-mirror`, utilisez la spécification suivante :

```
service_type: rbd-mirror
service_id: EXAMPLE_RBD_MIRROR
placement:
  hosts:
```

```
  - ses-min3
```

5.4.3.8 Déploiement de la pile de surveillance

La pile de surveillance se compose de Prometheus, d'exportateurs Prometheus, de Prometheus Alertmanager et de Grafana. Ceph Dashboard utilise ces composants pour stocker et visualiser des mesures détaillées concernant l'utilisation et les performances de la grappe.

Astuce

Si votre déploiement nécessite des images de conteneur personnalisées ou mises à disposition localement pour les services de pile de surveillance, reportez-vous au *Manuel « Guide d'opérations et d'administration », Chapitre 16 « Surveillance et alertes », Section 16.1 « Configuration d'images personnalisées ou locales »*.

Pour déployer la pile de surveillance, procédez comme suit :

1. Activez le module `prometheus` dans le daemon Ceph Manager. Cela permet de rendre visibles les mesures Ceph internes afin que Prometheus puisse les lire :

```
cephuser@adm > ceph mgr module enable prometheus
```

Remarque

Assurez-vous que cette commande est exécutée avant le déploiement de Prometheus. Si la commande n'a pas été exécutée avant le déploiement, vous devez redéployer Prometheus pour mettre à jour la configuration de Prometheus :

```
cephuser@adm > ceph orch redeploy prometheus
```

2. Créez un fichier de spécifications (par exemple, `monitoring.yaml`) dont le contenu ressemble à ce qui suit :

```
service_type: prometheus
placement:
  hosts:
  - ses-min2
---
```

```
service_type: node-exporter
---
service_type: alertmanager
placement:
  hosts:
  - ses-min4
---
service_type: grafana
placement:
  hosts:
  - ses-min3
```

3. Appliquez les services de surveillance en exécutant :

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

Le déploiement des services de surveillance peut prendre une à deux minutes.

Important

Prometheus, Grafana et Ceph Dashboard sont tous automatiquement configurés pour pouvoir communiquer entre eux, ce qui permet une intégration Grafana entièrement fonctionnelle dans Ceph Dashboard lorsque le déploiement a été effectué comme décrit ci-dessus.

La seule exception à cette règle est la surveillance avec des images RBD. Pour plus d'informations, reportez-vous au *Manuel « Guide d'opérations et d'administration », Chapitre 16 « Surveillance et alertes », Section 16.5.4 « Activation de la surveillance des images RBD »*.

III Installation de services supplémentaires

6 Installation d'une passerelle iSCSI

iSCSI est un protocole SAN (Storage area network) qui permet aux clients (appelés *initiateurs*) d'envoyer des commandes SCSI aux périphériques de stockage SCSI (*cibles*) sur des serveurs distants. SUSE Enterprise Storage 7 comprend une interface qui ouvre la gestion du stockage Ceph aux clients hétérogènes, tels que Microsoft Windows* et VMware* vSphere, via le protocole iSCSI. L'accès iSCSI multipath assure la disponibilité et l'évolutivité à ces clients, et le protocole iSCSI standard fournit également une couche supplémentaire d'isolation de sécurité entre les clients et la grappe SUSE Enterprise Storage 7. La fonction de configuration est nommée `ceph-iscsi`. À l'aide de la fonction `ceph-iscsi`, les administrateurs du stockage Ceph peuvent définir des volumes à allocation dynamique, hautement disponibles et répliqués, prenant en charge les instantanés en lecture seule, les clones en lecture-écriture et le redimensionnement automatique avec le périphérique de bloc RADOS (RADOS Block Device, RBD) de Ceph. Les administrateurs peuvent ensuite exporter des volumes via un seul hôte de passerelle `ceph-iscsi` ou via plusieurs hôtes de passerelle prenant en charge le basculement multipath. Les hôtes Linux, Microsoft Windows et VMware peuvent se connecter aux volumes à l'aide du protocole iSCSI, ce qui les rend disponibles comme tout autre périphérique de bloc SCSI. Cela signifie que les clients de SUSE Enterprise Storage 7 peuvent exécuter efficacement un sous-système d'infrastructure de stockage de blocs complet sur Ceph, qui offre toutes les fonctionnalités et avantages d'un SAN conventionnel, ce qui permet une croissance future.

Ce chapitre présente des informations détaillées permettant de configurer une infrastructure de grappe Ceph avec une passerelle iSCSI afin que les hôtes clients puissent utiliser des données stockées à distance en tant que périphériques de stockage locaux à l'aide du protocole iSCSI.

6.1 Stockage de blocs iSCSI

iSCSI est une implémentation de la commande SCSI (Small Computer System Interface) définie en utilisant le protocole Internet (IP), spécifié dans la norme RFC 3720. iSCSI est implémenté comme un service dans lequel un client (initiateur) se connecte à un serveur (cible) via une session sur le port TCP 3260. L'adresse IP et le port d'une cible iSCSI sont appelés *portail iSCSI*, où une cible peut être exposée par le bais d'un ou plusieurs portails. La combinaison d'une cible et d'un ou de plusieurs portails est appelée *groupe de portails cible* (TPG).

Le protocole de couche de liaison de données sous-jacent pour iSCSI est le plus souvent Ethernet. Plus spécifiquement, les infrastructures iSCSI modernes utilisent des réseaux 10 GigE Ethernet ou plus rapides pour un débit optimal. Une connectivité 10 Gigabit Ethernet entre la passerelle iSCSI et la grappe Ceph de l'interface dorsale est fortement recommandée.

6.1.1 Cible iSCSI du kernel Linux

La cible iSCSI du kernel Linux s'appelait à l'origine LIO pour `linux-iscsi.org`, le domaine et le site Web d'origine du projet. Pendant un certain temps, pas moins de quatre implémentations de cibles iSCSI concurrentes étaient disponibles pour la plate-forme Linux, mais LIO a finalement prévalu en tant que cible de référence iSCSI unique. Le code du kernel principal de LIO utilise le nom simple, mais quelque peu ambigu « target », en faisant la distinction entre « target core » et une variété de modules cibles d'interface client et d'interface dorsale.

Le module d'interface client le plus couramment utilisé est sans doute iSCSI. Toutefois, LIO prend également en charge Fibre Channel (FC), Fibre Channel sur Ethernet (FCoE) et plusieurs autres protocoles d'interface client. Pour l'instant, seul le protocole iSCSI est pris en charge par SUSE Enterprise Storage.

Le module d'interface dorsale cible le plus fréquemment utilisé est celui qui est capable de réexporter simplement n'importe quel périphérique de bloc disponible sur l'hôte cible. Ce module est nommé *iblock*. Cependant, LIO dispose également d'un module d'interface dorsale spécifique à RBD prenant en charge l'accès E/S multipath parallélisé aux images RBD.

6.1.2 Initiateurs iSCSI

Cette section présente des informations succinctes sur les initiateurs iSCSI utilisés sur les plates-formes Linux, Microsoft Windows et VMware.

6.1.2.1 Linux

L'initiateur standard de la plate-forme Linux est `open-iscsi`. `open-iscsi` lance le daemon `iscsid`, que l'utilisateur peut ensuite utiliser pour découvrir des cibles iSCSI sur n'importe quel portail donné, se connecter à des cibles et assigner des volumes iSCSI. `iscsid` communique avec la mi couche SCSI pour créer des périphériques de bloc du kernel que ce dernier peut ensuite

traiter comme n'importe quel autre périphérique de bloc SCSI du système. L'initiateur `open-iscsi` peut être déployé en association avec l'outil Device Mapper Multipath (`dm-multipath`) capable de fournir un périphérique de bloc iSCSI hautement disponible.

6.1.2.2 Microsoft Windows et Hyper-V

L'initiateur iSCSI par défaut pour le système d'exploitation Microsoft Windows est l'initiateur iSCSI de Microsoft. Le service iSCSI peut être configuré via une interface graphique (GUI) et prend en charge les E/S multipath pour une haute disponibilité.

6.1.2.3 VMware

L'initiateur iSCSI par défaut pour VMware vSphere et ESX est l'initiateur iSCSI du logiciel VMware ESX, `vmkiscsi`. Lorsque qu'il est activé, il peut être configuré à partir du client vSphere ou à l'aide de la commande **`vmkiscsi-tool`**. Vous pouvez ensuite formater les volumes de stockage connectés via l'adaptateur de disque vSphere iSCSI avec VMFS et les utiliser comme tout autre périphérique de stockage VM. L'initiateur VMware prend également en charge les E/S multipath pour une haute disponibilité.

6.2 Informations générales concernant `ceph-iscsi`

La fonction `ceph-iscsi` associe les avantages des périphériques de bloc RADOS à la polyvalence omniprésente du protocole iSCSI. En utilisant `ceph-iscsi` sur un hôte de la cible iSCSI (connu sous le nom de passerelle iSCSI), toute application qui a besoin d'utiliser le stockage de blocs peut bénéficier de Ceph, même si elle n'utilise pas un protocole de client Ceph. Au lieu de cela, les utilisateurs peuvent utiliser le protocole iSCSI ou tout autre protocole d'interface client cible pour se connecter à une cible LIO, ce qui traduit toutes les E/S cibles en opérations de stockage RBD.

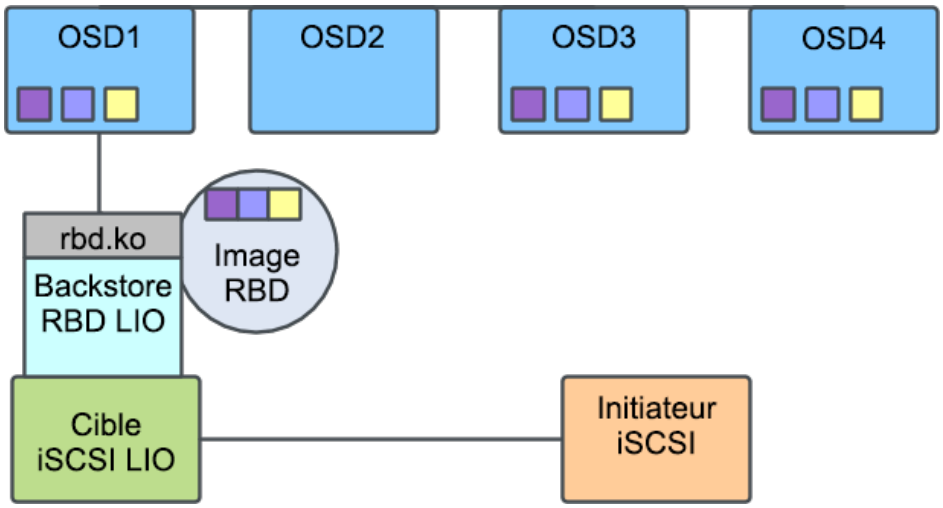

FIGURE 6.1 : GRAPPE CEPH AVEC UNE PASSERELLE ISCSI UNIQUE

`ceph-iscsi` est intrinsèquement hautement disponible et prend en charge les opérations multi-path. Ainsi, les hôtes initiateurs en aval peuvent utiliser plusieurs passerelles iSCSI pour la haute disponibilité et l'évolutivité. Lors de la communication avec une configuration iSCSI avec plus d'une passerelle, les initiateurs peuvent équilibrer la charge des requêtes iSCSI sur plusieurs passerelles. En cas d'échec d'une passerelle, d'inaccessibilité temporaire ou de désactivation pour maintenance, les E/S continuent de manière transparente via une autre passerelle.

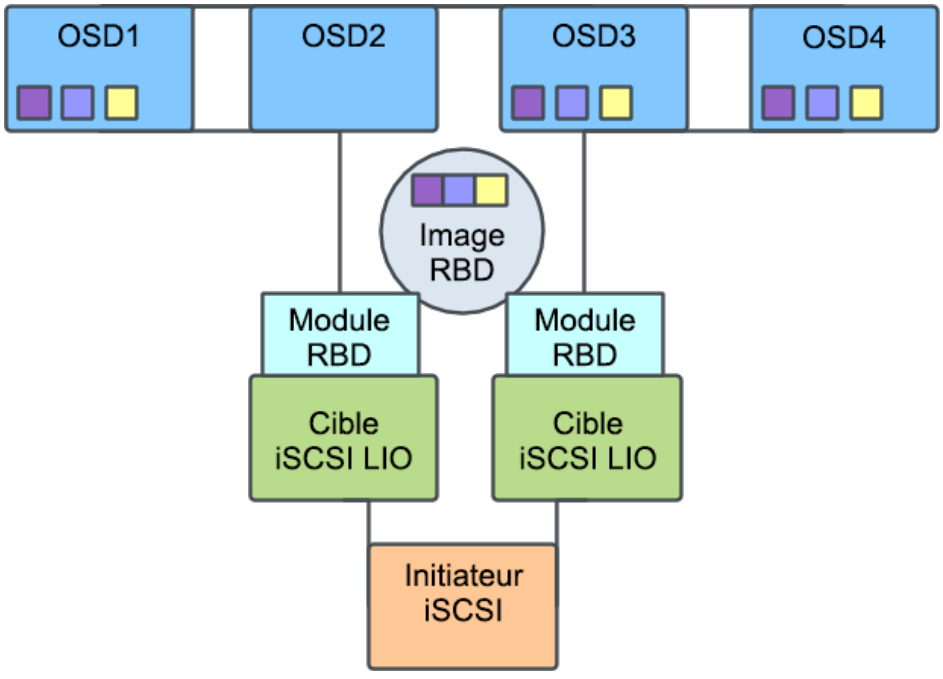

FIGURE 6.2 : GRAPPE CEPH AVEC PLUSIEURS PASSERELLES ISCSI

6.3 Aspects à prendre en considération pour le déploiement

Une configuration minimale de SUSE Enterprise Storage 7 avec `ceph-iscsi` comprend les composants suivants :

- Une grappe de stockage Ceph. La grappe Ceph consiste en un minimum de quatre serveurs physiques hébergeant au moins huit daemons de stockage des objets (Object Storage Daemon, OSD) chacun. Dans une telle configuration, les trois noeuds OSD doublent également en tant qu'hôte Monitor (MON).
- Un serveur cible iSCSI exécutant la cible iSCSI LIO, configurée via `ceph-iscsi`.
- Un hôte initiateur iSCSI, exécutant `open-iscsi` (Linux), l'initiateur Microsoft iSCSI (Microsoft Windows) ou toute autre implémentation d'initiateur iSCSI compatible.

La configuration recommandée en production pour SUSE Enterprise Storage 7 avec `ceph-iscsi` est constituée des éléments suivants :

- Une grappe de stockage Ceph. Une grappe Ceph de production est constituée de n'importe quel nombre de noeuds OSD (généralement plus de 10), chacun exécutant généralement 10 à 12 OSD, avec pas moins de trois hôtes MON dédiés.
- Plusieurs serveurs cibles iSCSI exécutant la cible iSCSI LIO, configurés via `ceph-iscsi`. Pour un basculement et un équilibrage de charge iSCSI, ces serveurs doivent exécuter un kernel prenant en charge le module `target_core_rbd`. Des paquetages de mise à jour sont disponibles à partir du canal de maintenance SUSE Linux Enterprise Server.
- N'importe quel nombre d'hôtes initiateurs iSCSI exécutant `open-iscsi` (Linux), l'initiateur Microsoft iSCSI (Microsoft Windows) ou toute autre implémentation d'initiateur iSCSI compatible.

6.4 Installation et configuration

Cette section décrit les étapes d'installation et de configuration d'une passerelle iSCSI en plus de SUSE Enterprise Storage.

6.4.1 Déploiement de la passerelle iSCSI sur une grappe Ceph

Le déploiement de la passerelle iSCSI Ceph suit la même procédure que le déploiement d'autres services Ceph, à l'aide de cephdm. Pour plus de détails, reportez-vous à la *Section 5.4.3.5, « Déploiement de passerelles iSCSI »*.

6.4.2 Création d'images RBD

Les images RBD sont créées dans la zone de stockage Ceph et ensuite exportées vers iSCSI. Il est recommandé d'utiliser une réserve RADOS dédiée à cette fin. Vous pouvez créer un volume à partir de n'importe quel hôte pouvant se connecter à votre grappe de stockage à l'aide de l'utilitaire de ligne de commande Ceph **`rbd`**. Pour cela, le client doit avoir au moins un fichier de configuration `ceph.conf` minimal et des informations d'identification d'authentification CephX appropriées.

Pour créer un volume pour l'exportation ultérieure via iSCSI, utilisez la commande **`rbd create`** en spécifiant la taille du volume en mégaoctets. Par exemple, pour créer un volume de 100 Go nommé `testvol` dans la réserve intitulée `iscsi-images`, exécutez la commande suivante :

```
cephuser@adm > rbd --pool iscsi-images create --size=102400 testvol
```

6.4.3 Exportation d'images RBD via iSCSI

Pour exporter des images RBD via iSCSI, vous pouvez utiliser l'interface Web Ceph Dashboard ou l'utilitaire gwcli `ceph-iscsi`. Dans cette section, nous nous concentrons sur gwcli uniquement et indiquons comment créer une cible iSCSI qui exporte une image RBD à l'aide de la ligne de commande.

Remarque

Les images RBD avec les propriétés suivantes ne peuvent pas être exportées via iSCSI :

- images avec la fonction `journaling` activée
- images avec une unité `stripe unit` inférieure à 4 096 octets

En tant qu'utilisateur `root`, entrez le conteneur de la passerelle iSCSI :

```
root # cephadm enter --name CONTAINER_NAME
```

En tant qu'utilisateur `root`, démarrez l'interface de ligne de commande de la passerelle iSCSI :

```
root # gwcli
```

Accédez à `iscsi-targets` et créez une cible nommée `iqn.2003-01.org.linux-iscsi.iscsi.`*`ARCH-SYSTÈME`*`:testvol` :

```
gwcli >  /> cd /iscsi-targets
gwcli >  /iscsi-targets> create iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol
```

Créez les passerelles iSCSI en spécifiant le nom de la passerelle (`name`) et l'adresse IP (`ip`) :

```
gwcli >  /iscsi-targets> cd iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol/
gateways
gwcli >  /iscsi-target...tvol/gateways> create iscsi1 192.168.124.104
gwcli >  /iscsi-target...tvol/gateways> create iscsi2 192.168.124.105
```

Astuce

Utilisez la commande `help` pour afficher la liste des commandes disponibles sur le noeud de configuration actuel.

Ajoutez l'image RBD avec le nom `testvol` dans la réserve `iscsi-images` :

```
gwcli >  /iscsi-target...tvol/gateways> cd /disks
gwcli >  /disks> attach iscsi-images/testvol
```

Assignez l'image RBD à la cible :

```
gwcli >  /disks> cd /iscsi-targets/iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol/
disks
gwcli >  /iscsi-target...testvol/disks> add iscsi-images/testvol
```

Remarque

Vous pouvez utiliser des outils de niveau inférieur, tels que **`targetcli`**, pour interroger la configuration locale, mais pas pour la modifier.

Astuce

Vous pouvez utiliser la commande **`ls`** pour examiner la configuration. Certains noeuds de configuration prennent également en charge la commande **`info`**, qui permet d'afficher des informations plus détaillées.

Notez que, par défaut, l'authentification ACL est activée de sorte que cette cible n'est pas encore accessible. Reportez-vous à la *Section 6.4.4, « Authentification et contrôle d'accès »* pour plus d'informations sur l'authentification et le contrôle d'accès.

6.4.4 Authentification et contrôle d'accès

L'authentification iSCSI est flexible et couvre de nombreuses possibilités d'authentification.

6.4.4.1 Désactivation de l'authentification ACL

No Authentication (Pas d'authentification) signifie que tout initiateur sera en mesure d'accéder à n'importe quelle unité logique sur la cible correspondante. Vous pouvez activer l'option *No authentication* en désactivant l'authentification ACL :

```
gwcli >  /> cd /iscsi-targets/iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol/hosts
gwcli >  /iscsi-target...testvol/hosts> auth disable_acl
```

6.4.4.2 Utilisation de l'authentification ACL

Lors de l'utilisation de l'authentification ACL basée sur le nom de l'initiateur, seuls les initiateurs définis sont autorisés à se connecter. Vous pouvez définir un initiateur comme suit :

```
gwcli >  /> cd /iscsi-targets/iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol/hosts
gwcli >  /iscsi-target...testvol/hosts> create iqn.1996-04.de.suse:01:e6ca28cc9f20
```

Les initiateurs définis pourront se connecter, mais n'auront accès qu'aux images RBD qui leur ont été explicitement ajoutées :

```
gwcli >  /iscsi-target...:e6ca28cc9f20> disk add rbd/testvol
```

6.4.4.3 Activation de l'authentification CHAP

En plus de l'ACL, vous pouvez activer une authentification CHAP en spécifiant un nom d'utilisateur et un mot de passe pour chaque initiateur :

```
gwcli >  /> cd /iscsi-targets/iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol/
hosts/iqn.1996-04.de.suse:01:e6ca28cc9f20
gwcli >  /iscsi-target...:e6ca28cc9f20> auth username=common12 password=pass12345678
```

Remarque

Les noms d'utilisateur doivent comporter entre 8 et 64 caractères et peuvent contenir des caractères alphanumériques, ., `@`, `-`, _ ou `:`.

Les mots de passe doivent comporter entre 12 et 16 caractères et peuvent contenir des caractères alphanumériques, `@`, `-`, _ ou `/`.

Éventuellement, vous pouvez aussi activer une authentification mutuelle CHAP en spécifiant les paramètres `mutual_username` et `mutual_password` dans la commande **`auth`**.

6.4.4.4 Configuration de l'authentification mutuelle et de la découverte

L'*authentification de la découverte* est indépendante des méthodes d'authentification précédentes. Elle nécessite des informations d'identification pour la navigation, est facultative et peut être configurée comme suit :

```
gwcli >  /> cd /iscsi-targets
gwcli >  /iscsi-targets> discovery_auth username=du123456 password=dp1234567890
```

Remarque

Les noms d'utilisateur doivent comporter entre 8 et 64 caractères et ne peuvent contenir que des lettres, `.`, `@`, `-`, _ ou `:`.

Les mots de passe doivent comporter entre 12 et 16 caractères, et ne peuvent contenir que des lettres, `@`, `-`, _ ou `/`.

Éventuellement, vous pouvez aussi spécifier les paramètres `mutual_username` et `mutual_password` dans la commande **`discovery_auth`**.

L'authentification de découverte peut être désactivée à l'aide de la commande suivante :

```
gwcli >  /iscsi-targets> discovery_auth nochap
```

6.4.5 Configuration des paramètres avancés

`ceph-iscsi` peut être configuré avec des paramètres avancés qui sont ensuite transmis à la cible des E/S LIO. Les paramètres sont divisés en paramètres `target` (cible) et `disk` (disque).

Avertissement

Sauf indication contraire, il est déconseillé de modifier ces paramètres par rapport à leur valeur par défaut.

6.4.5.1 Affichage des paramètres cibles

Vous pouvez afficher la valeur de ces paramètres à l'aide de la commande **`info`** :

```
gwcli >  /> cd /iscsi-targets/iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol
gwcli >  /iscsi-target...i.SYSTEM-ARCH:testvol> info
```

Vous pouvez modifier un paramètre à l'aide de la commande **`reconfigure`** :

```
gwcli >  /iscsi-target...i.SYSTEM-ARCH:testvol> reconfigure login_timeout 20
```

Les paramètres `target` disponibles sont les suivants :

default_cmdsn_depth

Profondeur de CmdSN (numéro de séquence de commande) par défaut. Limite le nombre de requêtes qu'un initiateur iSCSI peut avoir en attente à tout moment.

default_erl

Niveau de la récupération d'erreur par défaut.

login_timeout

Valeur de timeout de connexion en secondes.

netif_timeout

Timeout d'échec de la carte d'interface réseau en secondes.

prod_mode_write_protect

Une valeur définie sur `1` empêche les écritures sur les LUN (numéros d'unité logique).

6.4.5.2 Affichage des paramètres du disque

Vous pouvez afficher la valeur de ces paramètres à l'aide de la commande **`info`** :

```
gwcli >  /> cd /disks/rbd/testvol
gwcli >  /disks/rbd/testvol> info
```

Vous pouvez modifier un paramètre à l'aide de la commande **`reconfigure`** :

```
gwcli >  /disks/rbd/testvol> reconfigure rbd/testvol emulate_pr 0
```

Les paramètres `disk` disponibles sont les suivants :

block_size

Taille du bloc du périphérique sous-jacent.

emulate_3pc

Une valeur définie sur `1` active l'option Third Party Copy (Copie tierce).

emulate_caw

Une valeur définie sur `1` active l'option Compare and Write (Comparer et écrire).

emulate_dpo

Si définie sur 1, active l'option Disable Page Out (Désactiver la sortie de page).

emulate_fua_read

Une valeur définie sur `1` active la lecture pour l'option Force Unit Access (Forcer l'accès aux unités).

emulate_fua_write

Une valeur définie sur `1` active l'écriture pour l'option Force Unit Access (Forcer l'accès aux unités).

emulate_model_alias

Une valeur définie sur `1` utilise le nom du périphérique d'interface dorsale pour l'alias du modèle.

emulate_pr

Si définie sur 0, la prise en charge des réservations SCSI, y compris les réservations de groupe persistantes, est désactivée. La passerelle iSCSI SES peut alors ignorer l'état de réservation, ce qui améliore la latence des requêtes.

Astuce

Il est recommandé de définir `backstore_emulate_pr` sur `0` si les initiateurs iSCSI n'ont pas besoin de la prise en charge des réservations SCSI.

emulate_rest_reord

Si la valeur est définie sur `0`, la réorganisation est restreinte dans le modificateur d'algorithme de file d'attente.

emulate_tas

Une valeur définie sur `1` active l'option Task Aborted Status (État Tâche abandonnée).

emulate_tpu

Une valeur définie sur `1` active l'option Thin Provisioning Unmap (Annulation d'assignation du provisioning léger).

emulate_tpws

Une valeur définie sur `1` active l'option Thin Provisioning Write Same (Écriture identique provisioning léger).

emulate_ua_intlck_ctrl

Une valeur définie sur `1` active l'option Unit Attention Interlock (Interverrouillage attention unité).

emulate_write_cache

Une valeur définie sur `1` active l'option Write Cache Enable (Écriture sur le cache activée).

enforce_pr_isids

Une valeur définie sur `1` applique les ISID de réservation persistante.

is_nonrot

Si la valeur est définie sur `1`, le backstore est un périphérique non rotationnel.

max_unmap_block_desc_count

Nombre maximal de descripteurs de bloc pour UNMAP (Annuler l'assignation).

max_unmap_lba_count:

Nombre maximal de LBA pour UNMAP (Annuler l'assignation).

max_write_same_len

Longueur maximale pour WRITE_SAME (Écriture identique).

optimal_sectors

Taille de la requête optimale dans les secteurs.

pi_prot_type

Type de protection DIF.

queue_depth

Profondeur de la file d'attente.

unmap_granularity

Granularité UNMAP (Annuler l'assignation).

unmap_granularity_alignment

Alignement de la granularité UNMAP (Annuler l'assignation).

force_pr_aptpl

Lorsque ce paramètre est activé, LIO écrira toujours l'état *persistent reservation* (réservation persistante) dans l'espace de stockage persistant, que le client l'ait demandé ou non via `aptpl=1`. Cela n'a aucun effet avec l'interface dorsale RBD de kernel pour LIO, qui conserve toujours l'état PR. Idéalement, l'option `target_core_rbd` devrait imposer la valeur « 1 » et renvoyer une erreur si un utilisateur tente de la désactiver via la configuration.

unmap_zeroes_data

Détermine si LIO annoncera LBPRZ aux initiateurs SCSI, indiquant que les zéros seront relus à partir d'une région suivant UNMAP ou WRITE SAME avec un bit d'annulation d'assignation (« unmap »).

6.5 Exportation d'images de périphérique de bloc RADOS à l'aide de `tcmu-runner`

`ceph-iscsi` prend en charge les backstores `rbd` (basé sur le kernel) et `user:rbd` (tcmu-runner), ce qui rend toute la gestion transparente et indépendante du backstore.

Avertissement : avant-première technologique

Les déploiements de passerelles iSCSI basés sur `tcmu-runner` représentent actuellement un aperçu de la technologie.

Contrairement aux déploiements de passerelles iSCSI basés sur le kernel, les passerelles iSCSI basées sur `tcmu-runner` ne prennent pas en charge les E/S multipath ou les réservations persistantes SCSI.

Pour exporter une image RBD à l'aide de `tcmu-runner`, vous devez juste spécifier le backstore `user:rbd` lorsque vous attachez le disque :

```
gwcli >  /disks> attach rbd/testvol backstore=user:rbd
```

Remarque

Lors de l'utilisation de `tcmu-runner`, l'image RBD exportée doit avoir la fonction `exclusive-lock` activée.

IV Mise à jour à partir de versions précédentes

7 Mise à niveau à partir d'une version précédente

Ce chapitre présente les étapes requises pour mettre à niveau SUSE Enterprise Storage 6 vers la version 7.

La mise à niveau inclut les tâches suivantes :

- mise à niveau de Ceph Nautilus vers Octopus ;
- passage de l'installation et de l'exécution de Ceph via des paquetages RPM à l'exécution dans des conteneurs ;
- suppression complète de DeepSea et remplacement par `ceph-salt` et cephadm.

Avertissement

Les informations de mise à niveau de ce chapitre s'appliquent *uniquement* aux mises à niveau de DeepSea vers cephadm. N'essayez pas de suivre ces instructions si vous souhaitez déployer SUSE Enterprise Storage sur la plate-forme SUSE CaaS.

Important

La mise à niveau des versions SUSE Enterprise Storage antérieures à la version 6 n'est pas prise en charge. Vous devez d'abord passer à la dernière version SUSE Enterprise Storage 6, puis suivre les étapes de ce chapitre.

7.1 Avant la mise à niveau

Les tâches suivantes *doivent* être effectuées avant de commencer la mise à niveau. Cette opération peut être effectuée à tout moment pendant la durée de vie de SUSE Enterprise Storage 6.

- La migration OSD de FileStore vers BlueStore *doit* avoir lieu avant la mise à niveau car FileStore n'est pas pris en charge dans SUSE Enterprise Storage 7. Pour plus d'informations sur BlueStore et sur la procédure de migration à partir de FileStore, consultez le site https://documentation.suse.com/ses/6/html/ses-all/cha-ceph-upgrade.html#filestore2bluestore.
- Si vous exécutez une grappe plus ancienne qui utilise toujours des OSD `ceph-disk`, vous *devez* passer à `ceph-volume` avant la mise à niveau. Pour plus de détails, consultez le site https://documentation.suse.com/ses/6/html/ses-all/cha-ceph-upgrade.html#upgrade-osd-deployment.

7.1.1 Considérations

Avant la mise à niveau, veillez à lire les sections suivantes pour vous assurer que vous comprenez toutes les tâches à exécuter.

- *Lisez les notes de version* : elles contiennent des informations supplémentaires sur les modifications apportées depuis la version précédente de SUSE Enterprise Storage. Consultez les notes de version pour vérifier les aspects suivants :
  - votre matériel doit tenir compte de certaines considérations spéciales ;
  - les paquetages logiciels utilisés ont été considérablement modifiés ;
  - des précautions spéciales sont nécessaires pour votre installation.

  Les notes de version incluent également des informations de dernière minute qui, faute de temps, n'ont pas pu être intégrées au manuel. Elles contiennent également des notes concernant les problèmes connus.
  Les notes de version de SES 7 sont disponibles en ligne à l'adresse https://www.suse.com/releasenotes/.
  En outre, après avoir installé le paquetage `release-notes-ses` à partir de l'espace de stockage SES 7, vous trouverez les notes de version en local dans le répertoire `/usr/share/doc/release-notes` ou en ligne à l'adresse https://www.suse.com/releasenotes/.
- Lisez le *Chapitre 5, Déploiement avec cephadm* pour vous familiariser avec `ceph-salt` et l'orchestrateur Ceph, et en particulier prendre connaissance des informations sur les spécifications de service.
- La mise à niveau de la grappe peut prendre beaucoup de temps : environ le temps nécessaire pour mettre à niveau une machine multiplié par le nombre de noeuds de la grappe.

- Vous devez d'abord mettre à niveau Salt Master, puis remplacer DeepSea par `ceph-salt` et cephadm ensuite. Vous ne pourrez *pas* commencer à utiliser le module orchestrateur cephedm tant que vous n'aurez pas mis à niveau toutes les instances de Ceph Manager.

- La mise à niveau de l'utilisation des RPM Nautilus vers les conteneurs Octopus doit s'effectuer en une seule étape. Cela signifie mettre à niveau un noeud entier à la fois, et non un daemon à la fois.

- La mise à niveau des services principaux (MON, MGR, OSD) s'effectue de façon ordonnée. Chaque service est disponible pendant la mise à niveau. Les services de passerelle (Metadata Server, Object Gateway, NFS Ganesha, iSCSI Gateway) doivent être redéployés après la mise à niveau des services principaux. Un certain temps hors service existe pour chacun des services suivants :

  - 

    ### Important

    Les instances de Metadata Server et d'Object Gateway sont hors service à partir du moment où les noeuds sont mis à niveau de SUSE Linux Enterprise Server 15 SP1 vers SUSE Linux Enterprise Server 15 SP2 jusqu'au redéploiement des services à la fin de la procédure de mise à niveau. Il est particulièrement important de garder cela à l'esprit si ces services cohabitent avec des MON, des MGR ou des OSD, car dans ce cas, ils peuvent être hors service pendant toute la durée de la mise à niveau de la grappe. En cas de problème, envisagez de déployer ces services séparément sur des noeuds supplémentaires avant de procéder à la mise à niveau, afin qu'ils soient hors service le moins longtemps possible. Il s'agit de la durée de la mise à niveau des noeuds de passerelle, et non de la durée de la mise à niveau de l'ensemble de la grappe.

  - Les instances NFS Ganesha and iSCSI Gateway ne sont hors service que pendant le redémarrage des noeuds lors de la mise à niveau de SUSE Linux Enterprise Server 15 SP1 vers SUSE Linux Enterprise Server 15 SP2, puis brièvement lors du redéploiement de chaque service en mode conteneurisé.

7.1.2 Sauvegarde de la configuration et des données de grappe

Nous vous recommandons vivement de sauvegarder l'ensemble de la configuration et des données de la grappe avant la mise à niveau vers SUSE Enterprise Storage 7. Pour savoir comment sauvegarder toutes vos données, consultez https://documentation.suse.com/ses/6/html/ses-all/cha-deployment-backup.html.

7.1.3 Vérification des étapes de la mise à niveau précédente

Si vous avez au préalable effectué une mise à niveau à partir de la version 5, vérifiez que la mise à niveau vers la version 6 a bien abouti :

Vérifiez l'existence du fichier `/srv/salt/ceph/configuration/files/ceph.conf.import`.

Ce fichier est créé par le processus engulf lors de la mise à niveau de SUSE Enterprise Storage 5 vers la version 6. L'option `configuration_init: default-import` est définie dans le répertoire `/srv/pillar/ceph/proposals/config/stack/default/ceph/cluster.yml`.

Si l'option `configuration_init` reste définie sur `default-import`, la grappe utilise `ceph.conf.import` comme fichier de configuration, au lieu du fichier `ceph.conf` par défaut de DeepSea, qui est compilé à partir des fichiers dans le répertoire `/srv/salt/ceph/configuration/files/ceph.conf.d/`.

Par conséquent, vous devez rechercher les configurations personnalisées potentielles dans `ceph.conf.import` et, éventuellement, les déplacer dans l'un des fichiers du répertoire `/srv/salt/ceph/configuration/files/ceph.conf.d/`.

Supprimez ensuite la ligne `configuration_init: default-import` du répertoire `/srv/pilleur/ceph/proposals/config/stack/default/ceph/cluster.yml`.

7.1.4 Mise à jour des noeuds de la grappe et vérification de l'état de santé de la grappe

Vérifiez que les dernières mises à jour de SUSE Linux Enterprise Server 15 SP1 et SUSE Enterprise Storage 6 sont toutes appliquées à tous les noeuds de la grappe :

```
root # zypper refresh && zypper patch
```

Une fois les mises à jour appliquées, vérifiez l'état de santé de la grappe :

```
cephuser@adm > ceph -s
```

7.1.5 Vérification de l'accès aux dépôts logiciels et aux images de conteneur

Vérifiez que chaque noeud de la grappe a accès aux dépôts logiciels SUSE Linux Enterprise Server 15 SP2 et SUSE Enterprise Storage 7, ainsi qu'au registre des images de conteneur.

7.1.5.1 Dépôts logiciels

Si tous les noeuds sont enregistrés auprès de SCC, vous pouvez utiliser la commande **`zypper migration`** pour effectuer la mise à niveau. Pour plus d'informations, consultez https://documentation.suse.com/sles/15-SP2/html/SLES-all/cha-upgrade-online.html#sec-upgrade-online-zypper.

Si les noeuds ne sont **pas** enregistrés auprès de SCC, désactivez tous les dépôts logiciels existants et ajoutez les dépôts `Pool` et `Updates` pour chacune des extensions suivantes :

- SLE-Product-SLES/15-SP2
- SLE-Module-Basesystem/15-SP2
- SLE-Module-Server-Applications/15-SP2
- SUSE-Enterprise-Storage-7

7.1.5.2 Images de conteneur

Tous les noeuds de la grappe doivent accéder au registre des images de conteneur. Dans la plupart des cas, vous utiliserez le registre SUSE public à l'adresse `registration.suse.com`. Les images suivantes sont nécessaires :

- registry.suse.com/ses/7/ceph/ceph
- registry.suse.com/ses/7/ceph/grafana
- registry.suse.com/caasp/v4.5/prometheus-server
- registry.suse.com/caasp/v4.5/prometheus-node-exporter
- registry.suse.com/caasp/v4.5/prometheus-alertmanager

Vous pouvez également, par exemple pour les déploiements à vide, configurer un registre local et vérifier que vous disposez de l'ensemble approprié d'images de conteneur. Pour plus d'informations sur la configuration d'un registre d'images de conteneur local, reportez-vous à la *Section 5.3.2.11, « Configuration du registre de conteneurs »*.

7.2 Mise à niveau de Salt Master

La procédure suivante décrit le processus de mise à niveau de Salt Master :

1. Mettez à niveau le système d'exploitation sous-jacent vers SUSE Linux Enterprise Server 15 SP2 :

   - Pour les grappes dont l'ensemble des noeuds sont enregistrés auprès de SCC, exécutez **`zypper migration`**.
   - Pour les grappes dont les noeuds ont des dépôts logiciels assignés manuellement, exécutez **`zypper dup`** suivi de **`reboot`**.

2. Désactivez les phases DeepSea pour éviter toute utilisation accidentelle. Ajoutez le contenu suivant à `/srv/pillar/ceph/stack/global.yml` :

   ```
   stage_prep: disabled
   stage_discovery: disabled
   stage_configure: disabled
   stage_deploy: disabled
   stage_services: disabled
   stage_remove: disabled
   ```

   Enregistrez le fichier et appliquez les modifications :

   ```
   root@master # salt '*' saltutil.pillar_refresh
   ```

3. Si vous n'utilisez **pas** d'images de conteneur de `registry.suse.com`, mais plutôt le registre configuré localement, modifiez `/srv/pillar/ceph/stack/global.yml` pour indiquer à DeepSea l'image de conteneur Ceph et le registre utiliser. Par exemple, pour utiliser `192.168.121.1:5000/my/ceph/image`, ajoutez les lignes suivantes :

   ```
   ses7_container_image: 192.168.121.1:5000/my/ceph/image
   ses7_container_registries:
     - location: 192.168.121.1:5000
   ```

   Enregistrez le fichier et appliquez les modifications :

   ```
   root@master # salt '*' saltutil.refresh_pillar
   ```

4. Assimilez la configuration existante :

   ```
   cephuser@adm > ceph config assimilate-conf -i /etc/ceph/ceph.conf
   ```

5. Vérifiez l'état de la mise à niveau. Il est possible que votre sortie soit différente en fonction de la configuration de votre grappe :

```
root@master # salt-run upgrade.status
The newest installed software versions are:
 ceph: ceph version 15.2.2-60-gf5864377ab (f5864377abb5549f843784c93577980aa264b9bc)
 octopus (stable)
 os: SUSE Linux Enterprise Server 15 SP2
Nodes running these software versions:
 admin.ceph (assigned roles: master, prometheus, grafana)
Nodes running older software versions must be upgraded in the following order:
 1: mon1.ceph (assigned roles: admin, mon, mgr)
 2: mon2.ceph (assigned roles: admin, mon, mgr)
 3: mon3.ceph (assigned roles: admin, mon, mgr)
 4: data4.ceph (assigned roles: storage, mds)
 5: data1.ceph (assigned roles: storage)
 6: data2.ceph (assigned roles: storage)
 7: data3.ceph (assigned roles: storage)
 8: data5.ceph (assigned roles: storage, rgw)
```

7.3 Mise à niveau des noeuds MON, MGR et OSD

Mettez à niveau les noeuds Ceph Monitor, Ceph Manager et OSD un par un. Pour chaque service, procédez comme suit :

1. Si vous mettez à niveau un noeud OSD, évitez de marquer l'OSD comme `sorti` pendant la mise à niveau en exécutant la commande suivante :

```
cephuser@adm > ceph osd add-noout SHORT_NODE_NAME
```

   Remplacez *SHORT_NODE_NAME* par le nom abrégé du noeud tel qu'il apparaît dans la sortie de la commande **`ceph osd tree`**. Dans l'entrée suivante, les noms d'hôte abrégés sont `ses-min1` et `ses-min2`

```
root@master # ceph osd tree
ID   CLASS  WEIGHT   TYPE NAME       STATUS  REWEIGHT  PRI-AFF
 -1         0.60405  root default
-11         0.11691      host ses-min1
  4    hdd  0.01949          osd.4       up   1.00000  1.00000
  9    hdd  0.01949          osd.9       up   1.00000  1.00000
 13    hdd  0.01949          osd.13      up   1.00000  1.00000
[...]
 -5         0.11691      host ses-min2
```

```
  2    hdd  0.01949          osd.2       up   1.00000  1.00000
  5    hdd  0.01949          osd.5       up   1.00000  1.00000
[...]
```

2. Mettez à niveau le système d'exploitation sous-jacent vers SUSE Linux Enterprise Server 15 SP2 :

   - Si les noeuds de la grappe sont tous enregistrés auprès de SCC, exécutez **`zypper migration`**.
   - Si les noeuds de la grappe ont des dépôts logiciels assignés manuellement, exécutez **`zypper dup`** suivi de **`reboot`**.

3. Une fois le noeud redémarré, conteneurisez tous les daemons MON, MGR et OSD existants sur ce noeud en exécutant la commande suivante sur Salt Master :

   ```
   root@master # salt MINION_ID state.apply ceph.upgrade.ses7.adopt
   ```

   Remplacez *`MINION_ID`* par l'ID du minion que vous mettez à niveau. Vous pouvez obtenir la liste des ID de minion en exécutant la commande **`salt-key -L`** sur Salt Master.

   

   ### Astuce

   Pour voir l'état et la progression de l'*adoption*, consultez Ceph Dashboard ou exécutez l'une des commandes suivantes sur Salt Master :

   ```
   root@master # ceph status
   root@master # ceph versions
   root@master # salt-run upgrade.status
   ```

4. Une fois l'adoption terminée, désélectionnez l'indicateur `noout` si vous mettez à niveau un noeud OSD :

   ```
   cephuser@adm > ceph osd rm-noout SHORT_NODE_NAME
   ```

7.4 Mise à niveau des noeuds de passerelle

Mettez ensuite à niveau vos noeuds de passerelle distincts (Metadata Server, Object Gateway, NFS Ganesha ou iSCSI Gateway). Mettez à niveau le système d'exploitation sous-jacent vers SUSE Linux Enterprise Server 15 SP2 pour chaque noeud :

- Si les noeuds de la grappe sont tous enregistrés auprès de SUSE Customer Center, exécutez la commande **`zypper migration`**.
- Si les noeuds de la grappe ont des dépôts logiciels assignés manuellement, exécutez la commande **`zypper dup`** suivie de **`reboot`**.

Cette étape s'applique également à tous les noeuds qui font partie de la grappe, mais qui n'ont pas encore de rôle assigné (en cas de doute, consultez la liste des hôtes sur Salt Master fournie par la commande **`salt-key -L`** et comparez-la à la sortie de la commande **`salt-run upgrade.status`**).

Une fois le système d'exploitation mis à niveau sur tous les noeuds de la grappe, l'étape suivante consiste à installer le paquetage `ceph-salt` et à appliquer la configuration de la grappe. Les services de passerelle proprement dits sont redéployés en mode conteneurisé à la fin de la procédure de mise à niveau.

Remarque

Les services Metadata Server et Object Gateway sont indisponibles à partir de la mise à niveau vers SUSE Linux Enterprise Server 15 SP2 jusqu'à leur redéploiement à la fin de la procédure de mise à niveau.

7.5 Installation de `ceph-salt` et application de la configuration de la grappe

Avant de lancer la procédure d'installation de `ceph-salt` et d'appliquer la configuration de la grappe, vérifiez l'état de la grappe et de la mise à niveau en exécutant les commandes suivantes :

```
root@master # ceph status
root@master # ceph versions
root@master # salt-run upgrade.status
```

1. Supprimez les tâches cron `rbd_exporter` et `rgw_exporter` créées par DeepSea. Sur Salt Master en tant que `root`, exécutez la commande **`crontab -e`** pour modifier le fichier crontab. Supprimez les éléments suivants, le cas échéant :

```
# SALT_CRON_IDENTIFIER:deepsea rbd_exporter cron job
*/5 * * * * /var/lib/prometheus/node-exporter/rbd.sh > \
 /var/lib/prometheus/node-exporter/rbd.prom 2> /dev/null
# SALT_CRON_IDENTIFIER:Prometheus rgw_exporter cron job
*/5 * * * * /var/lib/prometheus/node-exporter/ceph_rgw.py > \
 /var/lib/prometheus/node-exporter/ceph_rgw.prom 2> /dev/null
```

2. Exportez la configuration de la grappe de DeepSea en exécutant les commandes suivantes :

```
root@master # salt-run upgrade.ceph_salt_config > ceph-salt-config.json
root@master # salt-run upgrade.generate_service_specs > specs.yaml
```

3. Désinstallez DeepSea et installez `ceph-salt` sur Salt Master :

```
root@master # zypper remove 'deepsea*'
root@master # zypper install ceph-salt
```

4. Redémarrez Salt Master et synchronisez les modules Salt :

```
root@master # systemctl restart salt-master.service
root@master # salt \* saltutil.sync_all
```

5. Importez la configuration de la grappe de DeepSea dans `ceph-salt` :

```
root@master # ceph-salt import ceph-salt-config.json
```

6. Générez des clés SSH pour la communication des noeuds de la grappe :

```
root@master # ceph-salt config /ssh generate
```

Astuce

Vérifiez que la configuration de la grappe a bien été importée de DeepSea et spécifiez les options potentiellement manquantes :

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

Pour une description complète de la configuration de la grappe, reportez-vous à la *Section 5.3.2, « Configuration des propriétés de grappe »*.

7. Appliquez la configuration et activez cephamp :

```
root@master # ceph-salt apply
```

8. Si vous devez fournir une URL de registre local de conteneurs et des informations d'identification d'accès, suivez les étapes décrites à la *Section 5.3.2.11, « Configuration du registre de conteneurs »*.

9. Si vous n'utilisez **pas** d'images de conteneur de `registry.suse.com`, mais plutôt le registre configuré localement, indiquez à Ceph quelle image de conteneur utiliser en exécutant

```
root@master # ceph config set global container_image IMAGE_NAME
```

   Par exemple :

```
root@master # ceph config set global container_image 192.168.121.1:5000/my/ceph/
image
```

10. Arrêtez et désactivez les daemons `ceph-crash` de SUSE Enterprise Storage 6. Les nouvelles formes conteneurisées de ces daemons seront lancées automatiquement plus tard.

```
root@master # salt '*' service.stop ceph-crash
root@master # salt '*' service.disable ceph-crash
```

7.6 Mise à niveau et adoption de la pile de surveillance

La procédure suivante adopte tous les composants de la pile de surveillance (voir *Manuel « Guide d'opérations et d'administration », Chapitre 16 « Surveillance et alertes »* pour plus de détails).

1. Mettez l'orchestrateur en pause :

```
cephuser@adm > ceph orch pause
```

2. Sur le noeud qui exécute Prometheus, Grafana et Alertmanager (l'instance Salt Master par défaut), exécutez les commandes suivantes :

```
cephuser@adm > cephadm adopt --style=legacy --name prometheus.$(hostname)
cephuser@adm > cephadm adopt --style=legacy --name alertmanager.$(hostname)
cephuser@adm > cephadm adopt --style=legacy --name grafana.$(hostname)
```

Astuce

Si vous n'exécutez **pas** le registre d'images de conteneur par défaut `registration.suse.com`, vous devez spécifier l'image à utiliser, par exemple :

```
cephuser@adm > cephadm --image 192.168.121.1:5000/caasp/v4.5/prometheus-
server:2.18.0 \
  adopt --style=legacy --name prometheus.$(hostname)
cephuser@adm > cephadm --image 192.168.121.1:5000/caasp/v4.5/prometheus-
alertmanager:0.16.2 \
  adopt --style=legacy --name alertmanager.$(hostname)
cephuser@adm > cephadm --image 192.168.121.1:5000/ses/7/ceph/grafana:7.0.3 \
 adopt --style=legacy --name grafana.$(hostname)
```

Pour plus d'informations sur l'utilisation des images de conteneur personnalisées ou locales, voir *Manuel « Guide d'opérations et d'administration », Chapitre 16 « Surveillance et alertes », Section 16.1 « Configuration d'images personnalisées ou locales »*.

3. Supprimez le service Node-Exporter. Il ne doit pas nécessairement être migré et il sera réinstallé en tant que conteneur lorsque le fichier `specs.yaml` sera appliqué.

```
tux > sudo zypper rm golang-github-prometheus-node_exporter
```

4. Appliquez les spécifications de service que vous avez précédemment exportées de DeepSea :

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

5. Relancez l'orchestrateur :

```
cephuser@adm > ceph orch resume
```

7.7 Redéploiement du service de passerelle

7.7.1 Mise à niveau de la passerelle Object Gateway

Dans SUSE Enterprise Storage 7, les passerelles Object Gateway sont toujours configurées avec un domaine, ce qui permet une utilisation sur plusieurs sites (voir *Manuel « Guide d'opérations et d'administration », Chapitre 21 « Ceph Object Gateway », Section 21.13 « Passerelles Object Gateway multisites »* pour plus de détails) à l'avenir. Si vous avez utilisé une configuration Object Gateway monosite dans SUSE Enterprise Storage 6, procédez comme suit pour ajouter un domaine. Si vous ne prévoyez pas d'utiliser à proprement parler la fonctionnalité multisite, vous pouvez utiliser la `valeur par défaut` pour les noms de domaine, de groupe de zones et de zone.

1. Créez un domaine :

```
cephuser@adm > radosgw-admin realm create --rgw-realm=REALM_NAME --default
```

2. Vous pouvez éventuellement renommer la zone et le groupe de zones par défaut.

```
cephuser@adm > radosgw-admin zonegroup rename \
 --rgw-zonegroup default \
 --zonegroup-new-name=ZONEGROUP_NAME
cephuser@adm > radosgw-admin zone rename \
 --rgw-zone default \
 --zone-new-name ZONE_NAME \
 --rgw-zonegroup=ZONEGROUP_NAME
```

3. Configurez le groupe de zones maître :

```
cephuser@adm > radosgw-admin zonegroup modify \
 --rgw-realm=REALM_NAME \
 --rgw-zonegroup=ZONEGROUP_NAME \
 --endpoints http://RGW.EXAMPLE.COM:80 \
 --master --default
```

4. Configurez la zone maître. Pour ce faire, vous aurez besoin des clés ACCESS_KEY et SECRET_KEY d'un utilisateur Object Gateway avec l'indicateur `system` activé. Il s'agit généralement de l'utilisateur `admin`. Pour obtenir les clés ACCESS_KEY et SECRET_KEY, exécutez **`radosgw-admin user info --uid admin`**.

```
cephuser@adm > radosgw-admin zone modify \
 --rgw-realm=REALM_NAME \
```

```
--rgw-zonegroup=ZONEGROUP_NAME \
--rgw-zone=ZONE_NAME \
--endpoints http://RGW.EXAMPLE.COM:80 \
--access-key=ACCESS_KEY \
--secret=SECRET_KEY \
--master --default
```

5. Validez la configuration mise à jour :

```
cephuser@adm > radosgw-admin period update --commit
```

Pour conteneuriser le service Object Gateway, créez son fichier de spécification comme décrit à la *Section 5.4.3.4, « Déploiement d'instances Object Gateway »* et appliquez-le.

```
cephuser@adm > ceph orch apply -i RGW.yml
```

7.7.2 Mise à niveau de NFS Ganesha

Les paragraphes qui suivent expliquent comment migrer un service NFS Ganesha existant qui exécute Ceph Nautilus vers un conteneur NFS Ganesha qui exécute Ceph Octopus.

Avertissement

La documentation suivante implique que vous ayez déjà mis à niveau les services Ceph principaux.

NFS Ganesha stocke la configuration par daemon supplémentaire et l'exporte dans une réserve RADOS. La réserve RADOS configurée se trouve sur la ligne `watch_url` du bloc `RADOS_URLS` dans le fichier `ganesha.conf`. Par défaut, cette réserve est nommée `ganesha_config`

Avant de tenter une migration, nous vous recommandons vivement de copier les objets de configuration d'exportation et de daemon situés dans la réserve RADOS. Pour localiser la réserve RADOS configurée, exécutez la commande suivante :

```
cephuser@adm > grep -A5 RADOS_URLS /etc/ganesha/ganesha.conf
```

Pour répertorier le contenu de la réserve RADOS :

```
cephuser@adm > rados --pool ganesha_config --namespace ganesha ls | sort
  conf-node3
```

```
  export-1
  export-2
  export-3
  export-4
```

Pour copier les objets RADOS :

```
cephuser@adm > RADOS_ARGS="--pool ganesha_config --namespace ganesha"
cephuser@adm > OBJS=$(rados $RADOS_ARGS ls)
cephuser@adm > for obj in $OBJS; do rados $RADOS_ARGS get $obj $obj; done
cephuser@adm > ls -lah
total 40K
drwxr-xr-x 2 root root 4.0K Sep 8 03:30 .
drwx------ 9 root root 4.0K Sep 8 03:23 ..
-rw-r--r-- 1 root root 90 Sep 8 03:30 conf-node2
-rw-r--r-- 1 root root 90 Sep 8 03:30 conf-node3
-rw-r--r-- 1 root root 350 Sep 8 03:30 export-1
-rw-r--r-- 1 root root 350 Sep 8 03:30 export-2
-rw-r--r-- 1 root root 350 Sep 8 03:30 export-3
-rw-r--r-- 1 root root 358 Sep 8 03:30 export-4
```

Pour chaque noeud, un service NFS Ganesha existant doit être arrêté, puis remplacé par un conteneur géré par cephadm.

1. Arrêtez et désactivez le service NFS Ganesha existant :

   ```
   cephuser@adm > systemctl stop nfs-ganesha
   cephuser@adm > systemctl disable nfs-ganesha
   ```

2. Une fois le service NFS Ganesha existant arrêté, vous pouvez en déployer un nouveau dans un conteneur à l'aide de cephadm. Pour ce faire, vous devez créer une spécification de service contenant un `service_id` qui sera utilisé pour identifier cette nouvelle grappe NFS, le nom d'hôte du noeud que nous migrons répertorié comme hôte dans la spécification de placement, ainsi que la réserve RADOS et l'espace de noms qui contient les objets d'exportation NFS configurés. Par exemple :

   ```
   service_type: nfs
   service_id: SERVICE_ID
   placement:
     hosts:
     - node2
     pool: ganesha_config
     namespace: ganesha
   ```

Pour plus d'informations sur la création d'une spécification de placement, voir *Section 5.4.2, « Spécification de service et de placement »*.

3. Appliquez la spécification de placement :

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

4. Confirmez que le daemon NFS Ganesha est en cours d'exécution sur l'hôte :

```
cephuser@adm > ceph orch ps --daemon_type nfs
NAME           HOST   STATUS         REFRESHED  AGE  VERSION  IMAGE NAME
                    IMAGE ID      CONTAINER ID
nfs.foo.node2  node2  running (26m)  8m ago     27m  3.3      registry.suse.com/
ses/7/ceph/ceph:latest  8b4be7c42abd  c8b75d7c8f0d
```

5. Répétez ces étapes pour chaque noeud NFS Ganesha. Il n'est pas nécessaire de créer une spécification de service distincte pour chaque noeud. Il suffit d'ajouter le nom d'hôte de chaque noeud à la spécification de service NFS existante et de l'appliquer à nouveau.

Les exportations existantes peuvent être migrées de deux manières différentes :

- recréation ou réassignation manuelle à l'aide de Ceph Dashboard ;
- copie manuelle du contenu de chaque objet RADOS par daemon dans la configuration commune NFS Ganesha nouvellement créée.

PROCÉDURE 7.1 : COPIE MANUELLE DES EXPORTATIONS DANS LE FICHIER DE CONFIGURATION COMMUN NFS GANESHA

1. Déterminez la liste des objets RADOS par daemon :

```
cephuser@adm > RADOS_ARGS="--pool ganesha_config --namespace ganesha"
cephuser@adm > DAEMON_OBJS=$(rados $RADOS_ARGS ls | grep 'conf-')
```

2. Copiez les objets RADOS par daemon :

```
cephuser@adm > for obj in $DAEMON_OBJS; do rados $RADOS_ARGS get $obj $obj; done
cephuser@adm > ls -lah
total 20K
drwxr-xr-x 2 root root 4.0K Sep 8 16:51 .
drwxr-xr-x 3 root root 4.0K Sep 8 16:47 ..
-rw-r--r-- 1 root root 90 Sep 8 16:51 conf-nfs.SERVICE_ID
-rw-r--r-- 1 root root 90 Sep 8 16:51 conf-node2
-rw-r--r-- 1 root root 90 Sep 8 16:51 conf-node3
```

3. Triez et fusionnez en une seule liste d'exportations :

```
cephuser@adm > cat conf-* | sort -u > conf-nfs.SERVICE_ID
cephuser@adm > cat conf-nfs.foo
%url "rados://ganesha_config/ganesha/export-1"
%url "rados://ganesha_config/ganesha/export-2"
%url "rados://ganesha_config/ganesha/export-3"
%url "rados://ganesha_config/ganesha/export-4"
```

4. Écrivez le nouveau fichier de configuration commun NFS Ganesha :

```
cephuser@adm > rados $RADOS_ARGS put conf-nfs.SERVICE_ID conf-nfs.SERVICE_ID
```

5. Notifiez le daemon NFS Ganesha :

```
cephuser@adm > rados $RADOS_ARGS notify conf-nfs.SERVICE_ID conf-nfs.SERVICE_ID
```

Remarque

Cette opération entraîne le rechargement de la configuration par le daemon.

Une fois le service migré, le service NFS Ganesha basé sur Nautilus peut être supprimé.

1. Supprimez NFS Ganesha :

```
cephuser@adm > zypper rm nfs-ganesha
Reading installed packages...
Resolving package dependencies...
The following 5 packages are going to be REMOVED:
  nfs-ganesha nfs-ganesha-ceph nfs-ganesha-rados-grace nfs-ganesha-rados-urls nfs-
ganesha-rgw
5 packages to remove.
After the operation, 308.9 KiB will be freed.
Continue? [y/n/v/...? shows all options] (y): y
(1/5) Removing nfs-ganesha-
ceph-2.8.3+git0.d504d374e-3.3.1.x86_64 ..........................................
[done]
(2/5) Removing nfs-ganesha-
rgw-2.8.3+git0.d504d374e-3.3.1.x86_64 ...........................................
[done]
(3/5) Removing nfs-ganesha-rados-
urls-2.8.3+git0.d504d374e-3.3.1.x86_64 ..........................................
[done]
```

```
(4/5) Removing nfs-ganesha-rados-
grace-2.8.3+git0.d504d374e-3.3.1.x86_64 ..........................................
[done]
(5/5) Removing nfs-
ganesha-2.8.3+git0.d504d374e-3.3.1.x86_64 ..........................................
[done]
Additional rpm output:
warning: /etc/ganesha/ganesha.conf saved as /etc/ganesha/ganesha.conf.rpmsave
```

2. Supprimez les paramètres hérités de la grappe de l'instance Ceph Dashboard :

```
cephuser@adm > ceph dashboard reset-ganesha-clusters-rados-pool-namespace
```

7.7.3 Mise à niveau du serveur de métadonnées

Contrairement aux instances MON, MGR et OSD, le serveur de métadonnées ne peut pas être adopté sur place. Vous devez le redéployer dans des conteneurs à l'aide de l'orchestrateur Ceph.

1. Exécutez la commande **`ceph fs ls`** pour obtenir le nom de votre système de fichiers, par exemple :

```
cephuser@adm > ceph fs ls
name: cephfs, metadata pool: cephfs_metadata, data pools: [cephfs_data ]
```

2. Créez un nouveau fichier de spécification de service `mds.yml` comme décrit à la *Section 5.4.3.3, « Déploiement de serveurs de métadonnées »* en utilisant le nom du système de fichiers comme `service_id` et en spécifiant les hôtes qui exécuteront les daemons MDS. Par exemple :

```
service_type: mds
service_id: cephfs
placement:
  hosts:
  - ses-min1
  - ses-min2
  - ses-min3
```

3. Exécutez la commande **`ceph orch apply -i mds.yml`** pour appliquer la spécification de service et démarrer les daemons MDS.

7.7.4 Mise à niveau de la passerelle iSCSI

Pour mettre à niveau la passerelle iSCSI, vous devez la redéployer dans des conteneurs à l'aide de l'orchestrateur Ceph. Si vous disposez de plusieurs passerelles iSCSI, vous devez les redéployer une par une pour réduire le temps d'indisponibilité du service.

1. Arrêtez et désactivez les daemons iSCSI existants sur chaque noeud de passerelle iSCSI :

```
tux > sudo systemctl stop rbd-target-gw
tux > sudo systemctl disable rbd-target-gw
tux > sudo systemctl stop rbd-target-api
tux > sudo systemctl disable rbd-target-api
```

2. Créez une spécification de service pour la passerelle iSCSI comme décrit à la *Section 5.4.3.5, « Déploiement de passerelles iSCSI »*. Pour ce faire, vous avez besoin des paramètres `pool`, `trusted_ip_list` et `api_*` du fichier `/etc/ceph/iscsi-gateway.cfg` existant. Si la prise en charge SSL est activée (`api_secure = true`), vous avez également besoin du certificat SSL (`/etc/ceph/iscsi-gateway.crt`) et de la clé (`/etc/ceph/iscsi-gateway.key`).
   Par exemple, si le fichier `/etc/ceph/iscsi-gateway.cfg` contient ce qui suit :

```
[config]
cluster_client_name = client.igw.ses-min5
pool = iscsi-images
trusted_ip_list = 10.20.179.203,10.20.179.201,10.20.179.205,10.20.179.202
api_port = 5000
api_user = admin
api_password = admin
api_secure = true
```

   Vous devez créer le fichier de spécification de service `iscsi.yml` suivant :

```
service_type: iscsi
service_id: igw
placement:
  hosts:
  - ses-min5
spec:
  pool: iscsi-images
  trusted_ip_list: "10.20.179.203,10.20.179.201,10.20.179.205,10.20.179.202"
  api_port: 5000
  api_user: admin
  api_password: admin
  api_secure: true
```

```
  ssl_cert: |
    -----BEGIN CERTIFICATE-----
    MIIDtTCCAp2gAwIBAgIYMC4xNzc1NDQxNjEzMzc2MjMyXzxvQ7EcMA0GCSqGSIb3
    DQEBCwUAMG0xCzAJBgNVBAYTAlVTMQ0wCwYDVQQIDARVdGFoMRcwFQYDVQQHDA5T
    [...]
    -----END CERTIFICATE-----
  ssl_key: |
    -----BEGIN PRIVATE KEY-----
    MIIEvQIBADANBgkqhkiG9w0BAQEFAASCBKcwggSjAgEAAoIBAQC5jdYbjtNTAKW4
    /CwQr/7wOiLGzVxChn3mmCIF3DwbL/qvTFTX2d8bDf6LjGwLYloXHscRfxszX/4h
    [...]
    -----END PRIVATE KEY-----
```

Remarque

Les paramètres `pool`, `trusted_ip_list`, `api_port`, `api_user`, `api_password`, `api_secure` sont identiques à ceux du fichier `/etc/ceph/iscsi-gateway.cfg`. Les valeurs `ssl_cert` et `ssl_key` peuvent être copiées à partir du certificat SSL et des fichiers de clé existants. Vérifiez qu'elles sont correctement indentées et que le caractère *barre verticale* (`|`) apparaît à la fin des lignes `ssl_cert:` et `ssl_key:` (voir le contenu du fichier `iscsi.yml` ci-dessus).

3. Exécutez la commande **`ceph orch apply -i iscsi.yml`** pour appliquer la spécification de service et démarrer les daemons de la passerelle iSCSI.

4. Supprimez l'ancien paquetage `ceph-iscsi` de chacun des noeuds existants de la passerelle iSCSI :

```
cephuser@adm > zypper rm -u ceph-iscsi
```

7.8 Nettoyage après mise à niveau

Après la mise à niveau, effectuez les opérations de nettoyage suivantes :

1. Vérifiez que la grappe a bien été mise à niveau en vérifiant la version actuelle de Ceph :

```
cephuser@adm > ceph versions
```

2. Assurez-vous qu'aucun ancien OSD ne rejoindra la grappe :

```
cephuser@adm > ceph osd require-osd-release octopus
```

3. Activez le module de mise à l'échelle automatique :

```
cephuser@adm > ceph mgr module enable pg_autoscaler
```

Important

Dans SUSE Enterprise Storage 6, les réserves avaient le paramètre `pg_autoscale_mode` défini sur `warn` par défaut. Cela entraînait un message d'avertissement si le nombre de groupes de placement était sous-optimal, mais la mise à l'échelle automatique ne se produisait pas. Dans SUSE Enterprise Storage 7, l'option `pg_autoscale_mode` est réglée sur `on` par défaut pour les nouvelles réserves et les groupes de placement seront mis à l'échelle automatiquement. Le processus de mise à niveau ne modifie pas automatiquement le paramètre `pg_autoscale_mode` des réserves existantes. Si vous souhaitez le régler sur `on` pour tirer pleinement parti de la mise à l'échelle automatique, reportez-vous aux instructions figurant dans le *Manuel « Guide d'opérations et d'administration », Chapitre 17 « Gestion des données stockées », Section 17.4.12 « Activation de la mise à l'échelle automatique des groupes de placement »*.

Pour plus de détails, reportez-vous au *Manuel « Guide d'opérations et d'administration », Chapitre 17 « Gestion des données stockées », Section 17.4.12 « Activation de la mise à l'échelle automatique des groupes de placement »*.

4. Évitez les clients de versions antérieures à Luminous :

```
cephuser@adm > ceph osd set-require-min-compat-client luminous
```

5. Activez le module de l'équilibreur :

```
cephuser@adm > ceph balancer mode upmap
cephuser@adm > ceph balancer on
```

Pour plus de détails, reportez-vous au *Manuel « Guide d'opérations et d'administration », Chapitre 29 « Modules Ceph Manager », Section 29.1 « Équilibreur »*.

6. Vous pouvez éventuellement activer le module de télémétrie :

```
cephuser@adm > ceph mgr module enable telemetry
cephuser@adm > ceph telemetry on
```

Pour plus de détails, reportez-vous au *Manuel « Guide d'opérations et d'administration », Chapitre 29 « Modules Ceph Manager », Section 29.2 « Activation du module de télémétrie »*.

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

Plusieurs paquetages clés de SUSE Enterprise Storage 7 sont basés sur la série de versions Octopus de Ceph. Lorsque le projet Ceph (https://github.com/ceph/ceph) publie de nouvelles versions intermédiaires dans la série Octopus, SUSE Enterprise Storage 7 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.

Version intermédiaire 15.2.5 d'Octopus

La version intermédiaire 15.2.5 d'Octopus a apporté les correctifs et autres modifications suivants :

- CephFS : des stratégies de partitionnement statique automatique en sous-arborescences peuvent désormais être configurées à l'aide des nouveaux attributs étendus d'épinglage éphémère distribué et aléatoire sur les répertoires. Pour plus d'informations, reportez-vous à la documentation suivante : https://docs.ceph.com/docs/master/cephfs/multimds/
- Les moniteurs disposent désormais d'une option de configuration `mon_osd_warn_num_repaired`, définie sur 10 par défaut. Si un OSD a réparé plus que ces nombreuses erreurs d'E/S dans les données stockées, un avertissement d'état de santé `OSD_TOO_MANY_REPAIRS` est généré.
- Désormais, lorsque des indicateurs `no scrub` et/ou `no deep-scrub` sont définis globalement ou par réserve, les nettoyages planifiés du type désactivé sont abandonnés. Tous les nettoyages lancés par l'utilisateur ne sont PAS interrompus.
- Résolution d'un problème d'optimisation des cartes OSD dans une grappe saine.

Version intermédiaire 15.2.4 d'Octopus

La version intermédiaire 15.2.4 d'Octopus a apporté les correctifs et autres modifications suivants :

- CVE-2020-10753 : rgw: assainissement des nouvelles lignes dans la propriété Expose-Header de s3 CORSConfiguration
- Object Gateway : les sous-commandes **`radosgw-admin`** relatives aux orphelins (**`radosgw-admin orphans find`**, **`radosgw-admin orphans finish`** et **`radosgw-admin orphans list-jobs`**) ont été abandonnées. Elles n'avaient pas fait l'objet d'une maintenance active, et comme elles stockent les résultats intermédiaires sur la grappe, elles pouvaient potentiellement remplir une grappe presque pleine. Elles ont été remplacées par **`rgw-orphelin-list`**, un outil actuellement considéré comme expérimental.
- RBD : l'objet de réserve RBD utilisé pour stocker la planification de la purge de la corbeille RBD s'appelle désormais `rbd_trash_purge_schedule` au lieu de `rbd_trash_trash_purge_schedule`. Les utilisateurs qui ont déjà commencé à utiliser la fonctionnalité de planification de purge de la corbeille RBD et qui ont configuré des planifications par réserve ou espace de noms doivent copier l'objet `rbd_trash_trash_purge_schedule` dans `rbd_trash_purge_schedule` avant la mise à niveau et supprimer `rbd_trash_purge_schedule` à l'aide des commandes suivantes dans chaque réserve et espace de noms RBD où une planification de purge de la corbeille avait été précédemment configurée :

```
rados -p pool-name [-N namespace] cp rbd_trash_trash_purge_schedule
 rbd_trash_purge_schedule
rados -p pool-name [-N namespace] rm rbd_trash_trash_purge_schedule
```

Vous pouvez également utiliser tout autre moyen pratique permettant de restaurer la planification après la mise à niveau.

Version intermédiaire 15.2.3 d'Octopus

- La version intermédiaire 15.2.3 d'Octopus était une version corrective destinée à résoudre un problème d'altération de WAL qui survenait lorsque les options `bluefs_preextend_wal_files` et `bluefs_buffered_io` étaient activées en même temps. Le correctif de la version 15.2.3 n'est qu'une mesure temporaire (définition de la valeur par défaut de

`bluefs_preextend_wal_files` sur `false`). Le correctif permanent consistera à supprimer complètement l'option `bluefs_preextend_wal_files` : ce correctif arrivera probablement dans la version intermédiaire 15.2.6.

Version intermédiaire 15.2.2 d'Octopus

La version intermédiaire 15.2.2 d'Octopus a corrigé une faille de sécurité :

- CVE-2020-10736 : correction d'un contournement d'autorisation dans les MON et les MGR

Version intermédiaire 15.2.1 d'Octopus

La version intermédiaire 15.2.1 d'Octopus a résolu un problème qui entraînait le blocage des OSD en cas de mise à niveau rapide de Luminous (SES5.5) vers Nautilus (SES6), puis vers Octopus (SES7). Elle a également corrigé deux failles de sécurité qui étaient présentes dans la version initiale d'Octopus (15.2.0) :

- CVE-2020-1759 : correction du problème de réutilisation du nonce dans le mode sécurisé de msgr V2
- CVE-2020-1760 : correction du problème de XSS en raison du fractionnement des en-têtes dans RGW GetObject

B Mises à jour de la documentation

Ce chapitre répertorie les modifications apportées au contenu de ce document qui s'appliquent à la version actuelle de SUSE Enterprise Storage.

- Déplacement du chapitre Ceph Dashboard (*Manuel « Guide d'opérations et d'administration »*) d'un niveau vers le haut dans la hiérarchie du livre afin que ses rubriques détaillées puissent faire l'objet de recherches directement à partir de la table des matières.
- La structure de *Manuel « Guide d'opérations et d'administration »* a été mise à jour pour correspondre à la plage de guides actuelle. Certains chapitres ont été déplacés vers d'autres guides (jsc#SES-1397).

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
: cephdm 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