



SUSE Enterprise Storage 7.1

# Guia de Administração e Operações

# Guia de Administração e Operações


SUSE Enterprise Storage 7.1


por Tomáš Bažant, Alexandra Settle, e Liam Proven

Data de Publicação: 20/03/2025

<https://documentation.suse.com> 

Copyright © 2020–2025 SUSE LLC e colaboradores. Todos os direitos reservados.

Exceto quando especificado de outra forma, este documento está licenciado nos termos da Creative Commons Attribution-ShareAlike 4.0 International (CC-BY-SA 4.0): <https://creativecommons.org/licenses/by-sa/4.0/legalcode> .

Para ver as marcas registradas da SUSE, visite <http://www.suse.com/company/legal/>. Todas as marcas registradas de terceiros pertencem aos seus respectivos proprietários. Os símbolos de marca registrada (®, <sup>TM</sup> etc.) indicam as marcas registradas da SUSE e de suas afiliadas. Os asteriscos (\*) indicam marcas registradas de terceiros.

Todas as informações deste manual foram compiladas com a maior atenção possível aos detalhes. Entretanto, isso não garante uma precisão absoluta. A SUSE LLC, suas afiliadas, os autores ou tradutores não serão responsáveis por possíveis erros nem pelas consequências resultantes de tais erros.

# Conteúdo

## Sobre este guia **xviii**

- 1 Documentação disponível **xviii**
- 2 Inserindo comentários **xix**
- 3 Convenções da documentação **xx**
- 4 Suporte **xxii**  
Declaração de suporte do SUSE Enterprise Storage **xxii** • Prévias de tecnologia **xxiii**
- 5 Colaboradores do Ceph **xxiv**
- 6 Comandos e prompts de comando usados neste guia **xxiv**  
Comandos relacionados ao Salt **xxiv** • Comandos relacionados ao Ceph **xxv** • Comandos gerais do Linux **xxvi** • Informações adicionais **xxvi**

## I **CEPH DASHBOARD 1**

### 1 Sobre o Ceph Dashboard **2**

### 2 Interface do usuário da Web do painel de controle **3**

- 2.1 Efetuando login **3**
- 2.2 Menu de utilitários **5**
- 2.3 Menu principal **6**
- 2.4 Painel de conteúdo **7**
- 2.5 Recursos comuns da IU da Web **7**

- 2.6 Widgets do painel de controle 7
  - Widgets de status 7 • Widgets de capacidade 8 • Widgets de desempenho 9

### **3 Gerenciar usuários e funções do Ceph Dashboard 11**

- 3.1 Listando usuários 11
- 3.2 Adicionando novos usuários 11
- 3.3 Editando usuários 12
- 3.4 Apagando usuários 12
- 3.5 Listando funções de usuário 13
- 3.6 Adicionando funções personalizadas 13
- 3.7 Editando funções personalizadas 15
- 3.8 Apagando funções personalizadas 15

### **4 Ver detalhes internos do cluster 16**

- 4.1 Visualizando nós do cluster 16
- 4.2 Acessando o inventário do cluster 16
- 4.3 Visualizando Ceph Monitors 17
- 4.4 Exibindo serviços 18
- 4.5 Exibindo Ceph OSDs 19
  - Adicionando OSDs 22
- 4.6 Visualizando a configuração do cluster 25
- 4.7 Vendo o mapa CRUSH do 25
- 4.8 Visualizando módulos do gerenciador 26
- 4.9 Visualizando registros 27
- 4.10 Visualizando o monitoramento 27

## 5 Gerenciar pools 28

- 5.1 Adicionando um novo pool 29
- 5.2 Apagando pools 29
- 5.3 Editando as opções de um pool 30

## 6 Gerenciar dispositivos de blocos RADOS 31

- 6.1 Visualizando detalhes sobre RBDs 32
- 6.2 Visualizando a configuração do RBD 33
- 6.3 Criando RBDs 34
- 6.4 Apagando RBDs 35
- 6.5 Criando instantâneos de dispositivo de blocos RADOS 35
- 6.6 Espelhamento do RBD 36
  - Configurando clusters principais e secundários 37 • Habilitando o daemon rbd-mirror 38 • Desabilitando o espelhamento 39 • Inicializando peers 39 • Removendo o peer do cluster 40 • Configurando a replicação de pool no Ceph Dashboard 40 • Verificando se a replicação de imagens RBD funciona 45
- 6.7 Gerenciando iSCSI Gateways 48
  - Adicionando destinos iSCSI 49 • Editando destinos iSCSI 51 • Apagando destinos iSCSI 51
- 6.8 Qualidade do Serviço (QoS) do RBD 51
  - Configurando opções globalmente 52 • Configurando opções em um novo pool 52 • Configurando opções em pool existente 53 • Opções de configuração 53 • Criando opções de QoS do RBD com uma nova imagem RBD 54 • Editando opções de QoS do RBD em imagens existentes 54 • Mudando as opções de configuração ao copiar ou clonar imagens 54

## 7 Gerenciar o NFS Ganesha 55

- 7.1 Criando exportações do NFS 56

7.2 Apagando exportações do NFS 58

7.3 Editando exportações do NFS 58

## **8 Gerenciar o CephFS 60**

8.1 Acessando a visão geral do CephFS 60

## **9 Gerenciar o Gateway de Objetos 62**

9.1 Vendo Gateways de Objetos 62

9.2 Gerenciando usuários do Gateway de Objetos 63

Adicionando um novo usuário do gateway 64 • Apagando usuários do gateway 66 • Editando detalhes do usuário do gateway 66

9.3 Gerenciando compartimentos de memória do Gateway de Objetos 66

Adicionando um novo compartimento de memória 66 • Visualizando detalhes do compartimento de memória 67 • Editando o compartimento de memória 68 • Apagando um compartimento de memória 69

## **10 Configuração manual 70**

10.1 Configurando o suporte a TLS/SSL 70

Criando certificados autoassinados 71 • Usando certificados assinados por CA 71

10.2 Mudando nome de host e número de porta 72

10.3 Ajustando nomes de usuário e senhas 73

10.4 Habilitando o front end de gerenciamento do Gateway de Objetos 73

10.5 Habilitando o gerenciamento de iSCSI 75

10.6 Habilitando o login único 75

## **11 Gerenciar usuários e funções na linha de comando 78**

11.1 Gerenciando a política de senha 78

11.2 Gerenciando contas dos usuários 79

- 11.3 Funções e permissões de usuário 80
  - Definindo escopos de segurança 80 • Especificando funções de usuário 81
- 11.4 Configuração de proxy 84
  - Acessando o painel de controle com proxies reversos 84 • Desabilitando redirecionamentos 84 • Configurando códigos de status de erro 85 • Exemplo de configuração do HAProxy 85
- 11.5 Fazendo auditoria das solicitações de API 86
- 11.6 Configurando o NFS Ganesha no Ceph Dashboard 87
  - Configurando vários clusters do NFS Ganesha 87
- 11.7 Plug-ins de depuração 88

## II OPERAÇÃO DO CLUSTER 89

### 12 Determinar o estado do cluster 90

- 12.1 Verificando o status de um cluster 90
- 12.2 Verificando a saúde do cluster 92
- 12.3 Verificando as estatísticas de uso de um cluster 101
- 12.4 Verificando o status do OSD 103
- 12.5 Verificando se há OSDs cheios 104
- 12.6 Verificando o status do monitor 105
- 12.7 Verificando estados de grupos de posicionamento 106
- 12.8 Capacidade de armazenamento 106
- 12.9 Monitorando OSDs e grupos de posicionamento 108
  - Monitorando OSDs 109 • Atribuindo conjuntos de grupos de posicionamento 110 • Emparelhamento 112 • Monitorando estados de grupos de posicionamento 112 • Encontrando o local de um objeto 118

### 13 Tarefas operacionais 119

- 13.1 Modificando a configuração do cluster 119



- 13.2 Adicionando nós 119
- 13.3 Removendo nós 120
- 13.4 Gerenciamento de OSD 122
  - Listando dispositivos de disco 122 • Apagando dispositivos de disco 123 • Adicionando OSDs por meio da especificação DriveGroups 123 • Removendo OSDs 133 • Substituindo OSDs 134
- 13.5 Movendo o Master Salt para um novo nó 135
- 13.6 Atualizando os nós do cluster 137
  - Repositórios do software 137 • Propagação em fases do repositório 137 • Tempo de espera dos serviços do Ceph 138 • Executando a atualização 138
- 13.7 Atualizando o Ceph 138
  - Iniciando a atualização 139 • Monitorando a atualização 139 • Cancelando uma atualização 139
- 13.8 Parando ou reiniciando o cluster 140
- 13.9 Removendo um cluster inteiro do Ceph 141
- 14 Operação de serviços do Ceph 142**
  - 14.1 Operando serviços individuais 142
  - 14.2 Operando tipos de serviço 143
  - 14.3 Operando serviços em um único nó 143
    - Identificando serviços e destinos 143 • Operando todos os serviços em um nó 144 • Operando um serviço individual em um nó 144 • Consultando o status do serviço 145
  - 14.4 Encerrando e reiniciando todo o cluster do Ceph 145
- 15 Backup e restauração 147**
  - 15.1 Fazer backup da configuração e dos dados do cluster 147
    - Fazer backup da configuração do ceph-salt 147 • Fazer backup da configuração do Ceph 147 • Fazer backup da configuração do Salt 147 • Fazer backup das configurações personalizadas 148

15.2 Restaurando um nó do Ceph 148

## **16 Monitoramento e alerta 150**

16.1 Configurando imagens personalizadas ou locais 151

16.2 Atualizando os serviços de monitoramento 153

16.3 Desabilitando o monitoramento 153

16.4 Configurando o Grafana 154

16.5 Configurando o módulo do gerenciador do Prometheus 155

Configurando a interface de rede 155 • Configurando o  
scrape\_interval 155 • Configurando o cache 156 • Habilitando o  
monitoramento de imagens RBD 156

16.6 Modelo de segurança do Prometheus 157

16.7 Gateway SNMP do Alertmanager do Prometheus 157

## **III ARMAZENANDO DADOS EM UM CLUSTER 159**

## **17 Gerenciamento de dados armazenados 160**

17.1 Dispositivos OSD 161

Classes de dispositivo 161

17.2 Compartimentos de memória 169

17.3 Conjuntos de regras 172

Iterando a árvore de nós 174 • firstn e indep 176

17.4 Grupos de posicionamento 177

Usando grupos de posicionamento 177 • Determinando o  
valor de *PG\_NUM* 179 • Definindo o número de grupos de  
posicionamento 181 • Encontrando o número de grupos  
de posicionamento 181 • Encontrando as estatísticas de  
PG de um cluster 182 • Encontrando as estatísticas de  
PGs travados 182 • Pesquisando o mapa de um grupo  
de posicionamento 182 • Recuperando as estatísticas de  
grupos de posicionamento 183 • Depurando um grupo de

- posicionamento 183 • Priorizando o provisionamento e a recuperação de grupos de posicionamento 183 • Revertendo objetos perdidos 184 • Habilitando o dimensionador automático de PG 184
- 17.5 Manipulação de mapa CRUSH 185
  - Editando um mapa CRUSH 185 • Adicionando ou movendo um OSD 187 • Diferença entre **ceph osd reweight** e **ceph osd crush reweight** 187 • Removendo um OSD 188 • Adicionando um compartimento de memória 188 • Movendo um compartimento de memória 189 • Removendo um compartimento de memória 189
- 17.6 Depurando grupos de posicionamento 189
- 18 Gerenciar pools de armazenamento 192**
- 18.1 Criando um pool 193
- 18.2 Listando os pools 194
- 18.3 Renomeando um pool 194
- 18.4 Apagando um pool 195
- 18.5 Outras operações 196
  - Associando pools a um aplicativo 196 • Definindo cotas de pool 196 • Mostrando as estatísticas do pool 197 • Obtendo valores do pool 198 • Definindo valores de um pool 199 • Definindo o número de réplicas do objeto 203
- 18.6 Migração de pool 204
  - Limitações 205 • Migração usando a camada de cache 205 • Migrando imagens RBD 207
- 18.7 Instantâneos de pool 208
  - Criando um instantâneo de um pool 208 • Listando instantâneos de um pool 209 • Removendo um instantâneo de um pool 209
- 18.8 Compactação de dados 209
  - Habilitando a compactação 210 • Opções de compactação do pool 210 • Opções globais de compactação 211

## 19 Pools codificados para eliminação 213

- 19.1 Pré-requisito para pools codificados para eliminação 213
- 19.2 Criando um pool codificado para eliminação de exemplo 214
- 19.3 Perfis de código de eliminação 214
  - Criando um novo perfil de código de eliminação 217 • Removendo um perfil de código de eliminação 218 • Exibindo detalhes do perfil de código de eliminação 219 • Listando perfis de código de eliminação 219
- 19.4 Marcando pools codificados para eliminação com dispositivo de blocos RADOS 219

## 20 Dispositivo de blocos RADOS 220

- 20.1 Comandos do dispositivo de blocos 220
  - Criando uma imagem de dispositivo de blocos em um pool replicado 221 • Criando uma imagem de dispositivo de blocos em um pool codificado para eliminação 221 • Listando imagens de dispositivo de blocos 222 • Recuperando informações da imagem 222 • Redimensionando uma imagem de dispositivo de blocos 222 • Removendo uma imagem de dispositivo de blocos 222
- 20.2 Montando e desmontando 223
  - Criando uma conta do usuário do Ceph 223 • Autenticação de usuário 224 • Preparando um Dispositivo de Blocos RADOS para uso 224 • **rbdmap** Mapear dispositivos RBD no momento da inicialização 226 • Aumentando o tamanho dos dispositivos RBD 227
- 20.3 Instantâneos 228
  - Habilitando e configurando o cephx 228 • Aspectos básicos do instantâneo 229 • Camadas de instantâneo 231
- 20.4 Espelhos de imagens RBD 235
  - Configuração do pool 236 • Configuração de imagens RBD 240 • Verificando o status do espelho 245
- 20.5 Configurações de cache 245
- 20.6 Configurações de QoS 247

20.7	Configurações de leitura com ajuda	248
20.8	Recursos avançados	249
20.9	Mapeando o RBD por meio de clientes antigos do kernel	251
20.10	Habilitando dispositivos de blocos e Kubernetes	253
	Usando dispositivos de blocos do Ceph no Kubernetes	255
IV	ACESSANDO OS DADOS DO CLUSTER	259
21	Gateway de Objetos do Ceph	260
21.1	Restrições e limitações de nomeação do Gateway de Objetos	260
	Limitações de compartimento de memória	260
	Limitações de objetos armazenados	260
	Limitações de cabeçalho HTTP	261
21.2	Implantando o Gateway de Objetos	261
21.3	Operando o serviço Gateway de Objetos	261
21.4	Opções de configuração	261
21.5	Gerenciando o acesso ao Gateway de Objetos	262
	Acessando o Gateway de Objetos	262
	Gerenciar contas do S3 e do Swift	264
21.6	Front ends HTTP	268
21.7	Habilitar HTTPS/SSL para Gateways de Objetos	268
	Criando um certificado autoassinado	268
	Configurando o Gateway de Objetos com SSL	269
21.8	Módulos de sincronização	270
	Configurando módulos de sincronização	270
	Sincronizando zonas	271
	Módulo de sincronização Elasticsearch	273
	Módulo de sincronização de nuvem	276
	Módulo de sincronização de arquivo	281
21.9	Autenticação LDAP	281
	Mecanismo de autenticação	282
	Requisitos	282
	Configurando o Gateway de Objetos para usar a autenticação LDAP	283
	Usando um filtro de	

- pesquisa personalizado para limitar o acesso do usuário 283 • Gerando um token de acesso para autenticação LDAP 284
- 21.10 Fragmentação de índice do compartimento de memória 285  
Refragmentação de índice do compartimento de memória 285 • Fragmentação de índice para novos compartimentos de memória 289
- 21.11 Integração do OpenStack Keystone 290  
Configurando o OpenStack 290 • Configurando o Gateway de Objetos do Ceph 291
- 21.12 Posicionamento do pool e classes de armazenamento 293  
Exibindo destinos de posicionamento 293 • Classes de armazenamento 294 • Configurando grupos de zonas e zonas 294 • Personalização de posicionamento 296 • Usando classes de armazenamento 298
- 21.13 Gateways de Objetos multissite 298  
Requisitos e considerações 299 • Configurando uma zona master 300 • Configurar zonas secundárias 306 • Manutenção geral do Gateway de Objetos 312 • Executando failover e recuperação de desastre 314

## 22 Ceph iSCSI Gateway 316

- 22.1 Destinos gerenciados pelo ceph-iscsi 316  
Conectando-se ao open-iscsi 316 • Conectando-se ao Microsoft Windows (iniciador iSCSI para Microsoft) 320 • Conectando-se ao VMware 327
- 22.2 Conclusão 333

## 23 Sistema de arquivos em cluster 334

- 23.1 Montando o CephFS 334  
Preparando o cliente 334 • Criando um arquivo secreto 335 • Montando o CephFS 335
- 23.2 Desmontando o CephFS 337
- 23.3 Montando o CephFS em /etc/fstab 337

- 23.4 Vários daemons MDS ativos (MDS ativo-ativo) **337**  
Usando o MDS ativo-ativo **337** • Aumentando o tamanho do cluster MDS ativo **338** • Diminuindo o número de classificações **339** • Fixando manualmente árvores de diretório em uma classificação **339**
- 23.5 Gerenciando o failover **340**  
Configurando a reprodução de standby **340**
- 23.6 Definindo cotas do CephFS **341**  
Limitações de cota do CephFS **341** • Configurando cotas do CephFS **342**
- 23.7 Gerenciando instantâneos do CephFS **343**  
Criando instantâneos **343** • Apagando instantâneos **344**
- 24 Exportar dados do Ceph por meio do Samba 345**
- 24.1 Exportar o CephFS por meio do compartilhamento do Samba **345**  
Configurando e exportando pacotes do Samba **345** • Exemplo de gateway único **346** • Configurando a alta disponibilidade **349**
- 24.2 Ingressando no Gateway do Samba e no Active Directory **355**  
Preparando a instalação do Samba **355** • Verificando o DNS **355** • Resolvendo registros SRV **356** • Configurando o kerberos **357** • Resolvendo o nome de host local **357** • Configurando o Samba **358** • Ingressando no domínio do Active Directory **361** • Configurando o Name Service Switch **361** • Iniciando os serviços **362** • Testar a conectividade de winbindd **362**
- 25 NFS Ganesha 363**
- 25.1 Criando um serviço do NFS **364**
- 25.2 Iniciando ou reiniciando o NFS Ganesha **365**
- 25.3 Listando objetos no pool de recuperação do NFS **365**
- 25.4 Criando uma exportação do NFS **365**
- 25.5 Verificando a exportação do NFS **366**
- 25.6 Montando a exportação do NFS **367**
- 25.7 Vários clusters do NFS Ganesha **367**

## V INTEGRAÇÃO COM FERRAMENTAS DE VIRTUALIZAÇÃO 368

### 26 libvirt e Ceph 369

- 26.1 Configurando o Ceph com libvirt 369
- 26.2 Preparando o gerenciador de VM 370
- 26.3 Criando uma VM 371
- 26.4 Configurando a VM 371
- 26.5 Resumo 374

### 27 Ceph como back end para instância de KVM QEMU 375

- 27.1 Instalando qemu-block-rbd 375
- 27.2 Usando o QEMU 375
- 27.3 Criando imagens com o QEMU 376
- 27.4 Redimensionando imagens com o QEMU 376
- 27.5 Recuperando informações da imagem com o QEMU 376
- 27.6 Executando o QEMU com o RBD 377
- 27.7 Habilitando descarte e TRIM 377
- 27.8 Definindo as opções de cache do QEMU 378

## VI CONFIGURANDO UM CLUSTER 380

### 28 Configuração do cluster do Ceph 381

- 28.1 Configurar o arquivo `ceph.conf` 381
  - Acessando o `ceph.conf` dentro das imagens de container 381
- 28.2 Banco de dados de configuração 382
  - Configurando seções e máscaras 382
  - Definindo e lendo as opções de configuração 383
  - Configurando daemons em tempo de execução 383
- 28.3 `config-key` armazenar 386
  - iSCSI Gateway 387



28.4	Ceph OSD e BlueStore	387
	Configurando o dimensionamento automático do cache	387
28.5	Gateway de Objetos do Ceph	389
	Configurações gerais	389
	Configurando front ends HTTP	399
<b>29</b>	<b>Módulos do Ceph Manager</b>	<b>402</b>
29.1	Balanceador	402
	O modo “crush-compatible”	403
	Planejando e executando o balanceamento de dados	403
29.2	Habilitando o módulo de telemetria	405
<b>30</b>	<b>Autenticação com cephx</b>	<b>407</b>
30.1	Arquitetura de autenticação	407
30.2	As principais áreas de	410
	Informações de referência	411
	Gerenciando usuários	414
	Gerenciando chaveiros	418
	Uso da linha de comando	421
<b>A</b>	<b>Atualizações de manutenção do Ceph baseadas nos point releases de upstream do “Pacific”</b>	<b>423</b>
	<b>Glossário</b>	<b>424</b>

# Sobre este guia

O foco deste guia são as tarefas de rotina que você, como administrador, precisa realizar após a implantação do cluster básico do Ceph (operações do dia 2). Ele também descreve todos os meios possíveis para acessar os dados armazenados em um cluster do Ceph.

SUSE Enterprise Storage 7.1 é uma extensão do SUSE Linux Enterprise Server 15 SP3. Ele combina os recursos do projeto de armazenamento do Ceph (<http://ceph.com/>) com a engenharia corporativa e o suporte da SUSE. O SUSE Enterprise Storage 7.1 oferece às organizações de TI o recurso para implantar uma arquitetura de armazenamento distribuído capaz de suportar inúmeros casos de uso por meio de plataformas de hardware convencional.

## 1 Documentação disponível



### Nota: Documentação online e atualizações mais recentes

A documentação para nossos produtos está disponível no site <https://documentation.suse.com/>, onde você também pode encontrar as atualizações mais recentes e procurar ou fazer download da documentação em vários formatos. As atualizações mais recentes da documentação estão disponíveis em inglês.

Além disso, a documentação do produto está disponível no seu sistema instalado em `/usr/share/doc/manual`. Ela está incluída em um pacote RPM chamado `ses-manual_LANG_CODE`. Instale esse pacote caso ainda não esteja em seu sistema, por exemplo:

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

A seguinte documentação está disponível para este produto:

*Guia de Implantação* (<https://documentation.suse.com/ses/html/ses-all/book-storage-deployment.html>)

O foco deste guia é a implantação de um cluster básico do Ceph e de serviços adicionais. Ele também abrange as etapas de upgrade para o SUSE Enterprise Storage 7.1 a partir da versão anterior do produto.

*Guia de Administração e Operações* (<https://documentation.suse.com/ses/html/ses-all/book-storage-admin.html>) ↗

O foco deste guia são as tarefas de rotina que você, como administrador, precisa realizar após a implantação do cluster básico do Ceph (operações do dia 2). Ele também descreve todos os meios possíveis para acessar os dados armazenados em um cluster do Ceph.

*Guia de Reforço da Segurança* (<https://documentation.suse.com/ses/html/ses-all/book-storage-security.html>) ↗

O foco deste guia é a garantia da segurança do cluster.

*Troubleshooting Guide (Guia de Solução de Problemas)* (<https://documentation.suse.com/ses/html/ses-all/book-storage-troubleshooting.html>) ↗

Este guia aborda os vários problemas comuns durante a execução do SUSE Enterprise Storage 7.1 e outras questões relacionadas a componentes relevantes, como o Ceph ou o Gateway de Objetos.

*Guia do SUSE Enterprise Storage para Windows* (<https://documentation.suse.com/ses/html/ses-all/book-storage-windows.html>) ↗

Este guia descreve a integração, a instalação e a configuração dos ambientes Microsoft Windows e SUSE Enterprise Storage que usam o Driver do Windows.

## 2 Inserindo comentários

Ficamos satisfeitos com seus comentários e contribuições para esta documentação. E você pode contribuir por meio de vários canais:

### Solicitações de serviço e suporte

Para conhecer os serviços e as opções de suporte disponíveis para o seu produto, consulte <http://www.suse.com/support/> ↗.

Para abrir uma solicitação de serviço, você precisa de uma assinatura do SUSE registrada no SUSE Customer Center. Vá para <https://scc.suse.com/support/requests> ↗, efetue login e clique em *Criar Novo*.


### Relatórios de bugs

Relate os problemas com a documentação em <https://bugzilla.suse.com/> ↗. A geração de relatórios de problemas requer uma conta do Bugzilla.

Para simplificar esse processo, você pode usar os links *Relatar Bug na Documentação* ao lado dos cabeçalhos na versão HTML deste documento. Isso pré-seleciona o produto e a categoria certos no Bugzilla e adiciona um link à seção atual. Você pode começar a digitar o relatório do bug imediatamente.

### Contribuições

Para contribuir com esta documentação, use os links *Editar Fonte* ao lado dos cabeçalhos na versão HTML deste documento. Eles levarão você até o código-fonte no GitHub, onde é possível abrir uma solicitação pull. A contribuição requer uma conta do GitHub.


Para obter mais informações sobre o ambiente da documentação usado para este documento, consulte o README do repositório em <https://github.com/SUSE/doc-ses> .

### E-mail

É possível também relatar erros e enviar comentários sobre a documentação para [doc-team@suse.com](mailto:doc-team@suse.com). Inclua o título do documento, a versão do produto e a data de publicação do documento. Mencione também o número e o título da seção relevante (ou inclua o URL) e insira uma breve descrição do problema.

## 3 Convenções da documentação

Os seguintes avisos e convenções tipográficas são usados nesta documentação:

- `/etc/passwd`: Nomes de diretório e arquivo
- `PLACEHOLDER`: Substitua `PLACEHOLDER` pelo valor real
- `PATH`: Uma variável de ambiente
- `ls, --help`: Comandos, opções e parâmetros
- `user`: O nome do usuário ou grupo
- `package_name`: O nome de um pacote de software
- `Alt`, `Alt - F1`: Uma tecla para pressionar ou uma combinação de teclas. As teclas são mostradas em maiúsculas como no teclado.
- *Arquivo*, *Arquivo* > *Gravar Como*: itens de menu, botões
- `AMD/Intel` Este parágrafo é relevante apenas para as arquiteturas AMD64/Intel 64. As setas marcam o início e o fim do bloco de texto. 

**IBM Z, POWER** Este parágrafo é relevante apenas para as arquiteturas IBM Z e POWER. As setas marcam o início e o fim do bloco de texto. ◀

- *Capítulo 1, “Capítulo de exemplo”*: Uma referência cruzada a outro capítulo deste guia.
- Comandos que devem ser executados com privilégios root. Geralmente, você também pode usar o comando sudo como prefixo nesses comandos para executá-los como usuário não privilegiado.

```
# command  
> sudo command
```

- Comandos que podem ser executados por usuários sem privilégios.

```
> command
```

- Avisos



### Atenção: Mensagem de aviso

Informações vitais que você deve saber antes de continuar. Avisa sobre problemas de segurança, potencial perda de dados, danos no hardware ou perigos físicos.



### Importante: Aviso importante

Informações importantes que você deve saber antes de continuar.



### Nota: Nota

Informações adicionais, por exemplo, sobre diferenças nas versões do software.



### Dica: Aviso de dica

Informações úteis, como uma diretriz ou informação prática.

- Avisos compactos



Informações adicionais, por exemplo, sobre diferenças nas versões do software.



Informações úteis, como uma diretriz ou informação prática.

## 4 Suporte

Encontre a declaração de suporte do SUSE Enterprise Storage e as informações gerais sobre as prévias de tecnologia a seguir. Para obter detalhes sobre o ciclo de vida do produto, consulte <https://www.suse.com/lifecycle>.

Se você tiver direito a suporte, encontre os detalhes de como coletar informações para um ticket de suporte em <https://documentation.suse.com/sles-15/html/SLES-all/cha-adm-support.html>.

### 4.1 Declaração de suporte do SUSE Enterprise Storage

Para receber suporte, você precisa de uma inscrição apropriada na SUSE. Para ver as ofertas de suporte específicas que estão disponíveis para você, acesse <https://www.suse.com/support/> e selecione seu produto.

Os níveis de suporte são definidos da seguinte forma:

#### L1

Determinação do problema, que significa suporte técnico designado para fornecer informações de compatibilidade, suporte ao uso, manutenção contínua, coleta de informações e solução básica de problemas usando a documentação disponível.

#### L2

Isolamento do problema, que significa suporte técnico designado para analisar os dados, reproduzir os problemas dos clientes, isolar a área problemática e resolver os problemas não resolvidos no Nível 1 ou preparar-se para o Nível 3.

#### L3

Resolução do problema, que significa suporte técnico designado para resolver os problemas com a participação da engenharia para solucionar defeitos nos produtos que foram identificados pelo Suporte de Nível 2.

Para clientes e parceiros contratados, o SUSE Enterprise Storage é entregue com suporte L3 para todos os pacotes, com as seguintes exceções:

- Prévias de tecnologia.
- Som, gráficos, fontes e arte.
- Pacotes que requerem um contrato de cliente adicional.
- Alguns pacotes enviados como parte do módulo *Workstation Extension* contam apenas com o suporte L2.
- Os pacotes com nomes que terminam em `-devel` (contendo arquivos de cabeçalho e recursos de desenvolvedor semelhantes) serão suportados apenas junto com seus pacotes principais.

A SUSE apenas oferecerá suporte ao uso dos pacotes originais. Isto é, pacotes que não foram modificados nem recompilados.

## 4.2 Prévias de tecnologia

As Prévias de tecnologia são pacotes, pilhas ou recursos fornecidos pela SUSE como amostras de inovações futuras. As prévias de tecnologia foram incluídas para sua conveniência e para que você possa testar as novas tecnologias em seu ambiente. Agradecemos seus comentários! Se você testar uma prévia de tecnologia, contate seu representante SUSE e conte sobre sua experiência e seus casos de uso. Suas informações são úteis para o desenvolvimento futuro.

As prévias de tecnologia têm as seguintes limitações:

- As prévias de tecnologia ainda estão em desenvolvimento. Portanto, elas podem ter funcionalidades incompletas, instáveis ou, de alguma outra maneira, *inadequadas* para uso em produção.
- As prévias de tecnologia *não* contam com suporte.
- As prévias de tecnologia talvez estejam disponíveis apenas para arquiteturas de hardware específicas.

- Os detalhes e as funcionalidades das prévias de tecnologia estão sujeitos a mudanças. Consequentemente, o upgrade para as versões subsequentes de uma prévia de tecnologia pode ser impossível e exigir uma instalação nova.
- A SUSE pode descobrir que uma prévia não atende às necessidades do cliente ou do mercado, ou não está em conformidade com os padrões da empresa. As prévias de tecnologia podem ser removidas de um produto a qualquer momento. A SUSE não se compromete em oferecer uma versão com suporte desse tipo de tecnologia no futuro.

Para obter uma visão geral das prévias de tecnologia fornecidas com seu produto, consulte as notas de lançamento em [https://www.suse.com/releasenotes/x86\\_64/SUSE-Enterprise-Storage/7.1](https://www.suse.com/releasenotes/x86_64/SUSE-Enterprise-Storage/7.1).

## 5 Colaboradores do Ceph

O projeto do Ceph e a respectiva documentação são resultados do trabalho de centenas de colaboradores e organizações. Visite <https://ceph.com/contributors/> para obter mais detalhes.

## 6 Comandos e prompts de comando usados neste guia

Como administrador de cluster do Ceph, você configura e ajusta o comportamento do cluster executando comandos específicos. Há vários tipos de comandos que serão necessários:

### 6.1 Comandos relacionados ao Salt

Esses comandos ajudam você a implantar nós do cluster do Ceph, executar comandos em vários (ou todos) nós do cluster ao mesmo tempo ou a adicionar ou remover nós do cluster. Os comandos usados com mais frequência são **ceph-salt** e **ceph-salt config**. Você precisa executar os comandos do Salt no nó do Master Salt como root. Esses comandos são introduzidos com o seguinte prompt:

```
root@master #
```

Por exemplo:

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



## 6.2 Comandos relacionados ao Ceph

Esses são comandos de nível inferior para configurar e ajustar todos os aspectos do cluster e seus gateways na linha de comando, por exemplo, `ceph`, `cephadm`, `rbd` ou `radosgw-admin`.

Para executar os comandos relacionados ao Ceph, você precisa ter acesso de leitura a uma chave do Ceph. Os recursos da chave definem seus privilégios no ambiente do Ceph. Uma opção é executar os comandos do Ceph como `root` (ou por meio do `sudo`) e usar o chaveiro padrão irrestrito “ceph.client.admin.key”.

A opção mais segura e recomendada é criar uma chave individual mais restritiva para cada usuário administrador e colocá-la em um diretório onde os usuários possam lê-la, por exemplo:

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



### Dica: Caminho para as chaves do Ceph

Para utilizar um usuário admin e um chaveiro personalizados, você precisa especificar o nome de usuário e o caminho para a chave toda vez que executar o comando `ceph` com as opções `-n client.USER_NAME` e `--keyring PATH/TO/KEYRING`.

Para evitar isso, inclua essas opções na variável `CEPH_ARGS` nos arquivos `~/bashrc` dos usuários individuais.

É possível executar os comandos relacionados ao Ceph em qualquer nó do cluster, mas a recomendação é executá-los no Nó de Admin. Esta documentação utiliza o usuário `cephuser` para executar os comandos, portanto, eles são introduzidos com o seguinte prompt:

```
cephuser@adm >
```

Por exemplo:

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



### Dica: Comandos para nós específicos

Se a documentação orientar você a executar um comando em um nó do cluster com uma função específica, isso será processado pelo prompt. Por exemplo:

```
cephuser@mon >
```

### 6.2.1 Executando o **ceph-volume**

A partir do SUSE Enterprise Storage 7, os serviços do Ceph são executados em containers. Se você precisar executar o **ceph-volume** em um nó OSD, anteceda-o com o comando **cephadm**, por exemplo:

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

## 6.3 Comandos gerais do Linux

Os comandos do Linux não relacionados ao Ceph, como **mount**, **cat** ou **openssl**, são introduzidos com os prompts `cephuser@adm >` ou `#`, dependendo dos privilégios exigidos pelo comando relacionado.

## 6.4 Informações adicionais

Para obter mais informações sobre o gerenciamento de chaves do Ceph, consulte o [Seção 30.2, “As principais áreas de”](#).

# I Ceph Dashboard

- 1 Sobre o Ceph Dashboard 2
- 2 Interface do usuário da Web do painel de controle 3
- 3 Gerenciar usuários e funções do Ceph Dashboard 11
- 4 Ver detalhes internos do cluster 16
- 5 Gerenciar pools 28
- 6 Gerenciar dispositivos de blocos RADOS 31
- 7 Gerenciar o NFS Ganesha 55
- 8 Gerenciar o CephFS 60
- 9 Gerenciar o Gateway de Objetos 62
- 10 Configuração manual 70
- 11 Gerenciar usuários e funções na linha de comando 78

# 1 Sobre o Ceph Dashboard

O Ceph Dashboard é um aplicativo de monitoramento e gerenciamento do Ceph baseado na Web que administra vários aspectos e objetos do cluster. O painel de controle será habilitado automaticamente após a implantação do cluster básico descrita no *Livro “Guia de Implantação”, Capítulo 7 “Implantando o cluster de boot usando ceph-salt”*.

O Ceph Dashboard para SUSE Enterprise Storage 7.1 adicionou mais recursos de gerenciamento baseados na Web para facilitar a administração do Ceph, incluindo o monitoramento e a administração de aplicativos no Ceph Manager. Você não precisa mais saber comandos complexos relacionados ao Ceph para gerenciar e monitorar o cluster do Ceph. Você pode usar a interface intuitiva da Ceph Dashboard ou a API REST incorporada.

O módulo Ceph Dashboard visualiza informações e estatísticas sobre o cluster do Ceph usando um servidor Web hospedado pelo `ceph-mgr`. Consulte o *Livro “Guia de Implantação”, Capítulo 1 “SES e Ceph”, Seção 1.2.3 “Nós e daemons do Ceph”* para obter mais detalhes sobre o Ceph Manager.

## 2 Interface do usuário da Web do painel de controle

### 2.1 Efetuando login

Para efetuar login no Ceph Dashboard, aponte seu navegador para o URL dele, incluindo o número da porta. Execute o seguinte comando para encontrar o endereço:

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

O comando retorna o URL em que o Ceph Dashboard está localizado. Se você tiver problemas com esse comando, consulte o *Livro “Troubleshooting Guide”, Capítulo 10 “Troubleshooting the Ceph Dashboard”, Seção 10.1 “Locating the Ceph Dashboard”*.

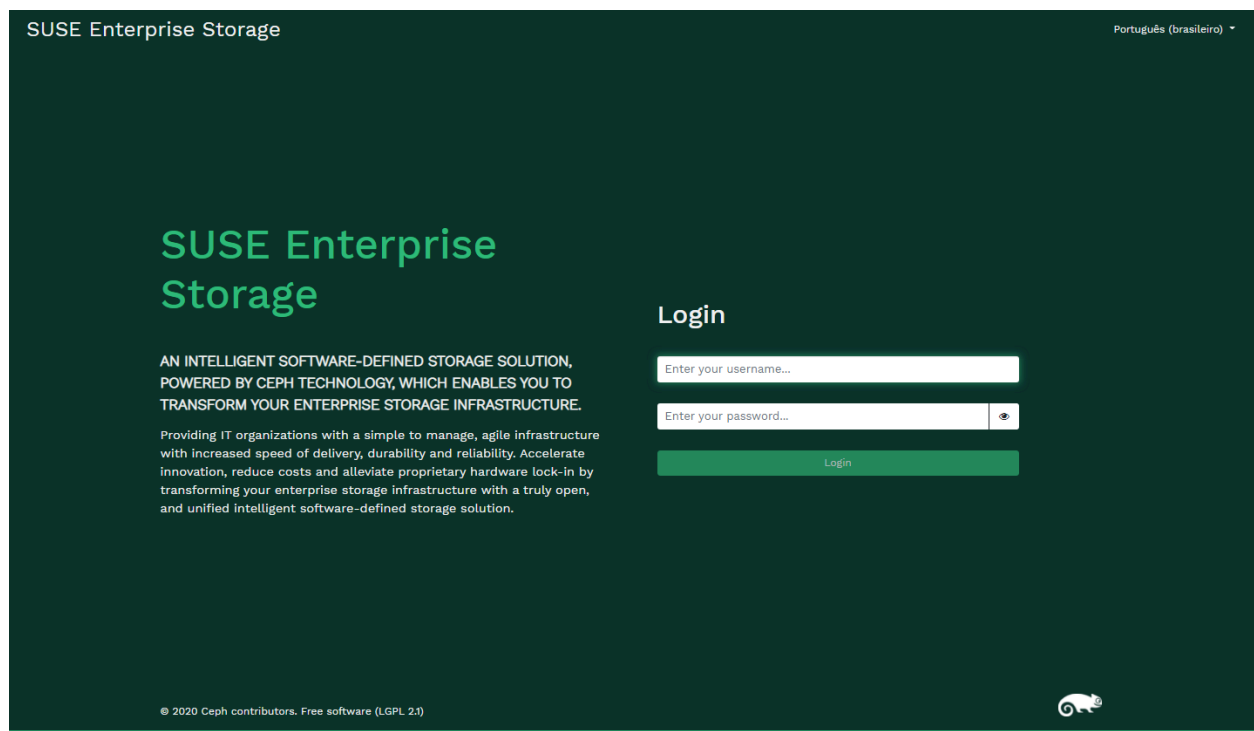


FIGURA 2.1: TELA DE LOGIN DO CEPH DASHBOARD

Efetue login usando as credenciais que você criou durante a implantação do cluster (consulte o *Livro “Guia de Implantação”, Capítulo 7 “Implantando o cluster de boot usando ceph-salt”, Seção 7.2.9 “Configurando as credenciais de login do Ceph Dashboard”*).



## Dica: Conta do usuário personalizada

Se você não deseja usar a conta padrão *admin* para acessar o Ceph Dashboard, crie uma conta do usuário personalizada com privilégios de administrador. Consulte [Capítulo 11, Gerenciar usuários e funções na linha de comando](#) para obter mais detalhes.



## Importante

Assim que um upgrade para uma nova versão principal do Ceph (codinome: Pacific) estiver disponível, o Ceph Dashboard exibirá uma mensagem relevante na área de notificação superior. Para fazer o upgrade, siga as instruções no Livro “*Guia de Implantação*”, Capítulo 11 “*Upgrade do SUSE Enterprise Storage 7 para 7.1*”.

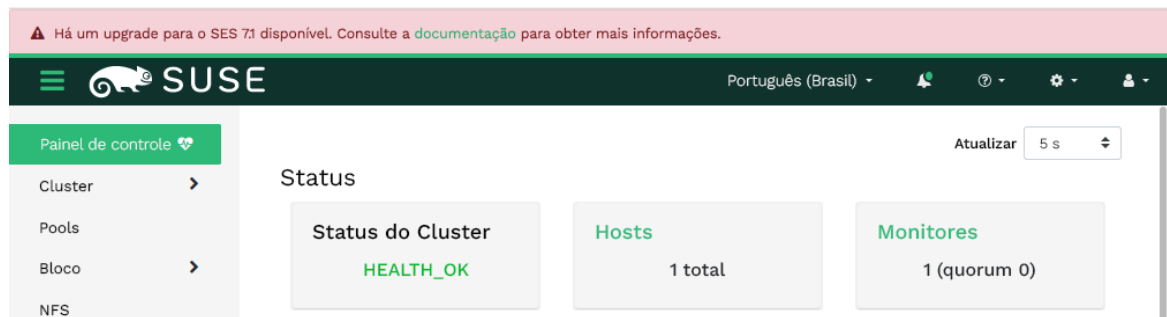


FIGURA 2.2: NOTIFICAÇÃO SOBRE UMA NOVA VERSÃO DO SUSE ENTERPRISE STORAGE

A interface do usuário do painel de controle é dividida graficamente em vários *blocos*: o *menu de utilitários* na lateral superior direita da tela, o *menu principal* na lateral esquerda e o *painel de conteúdo principal*.

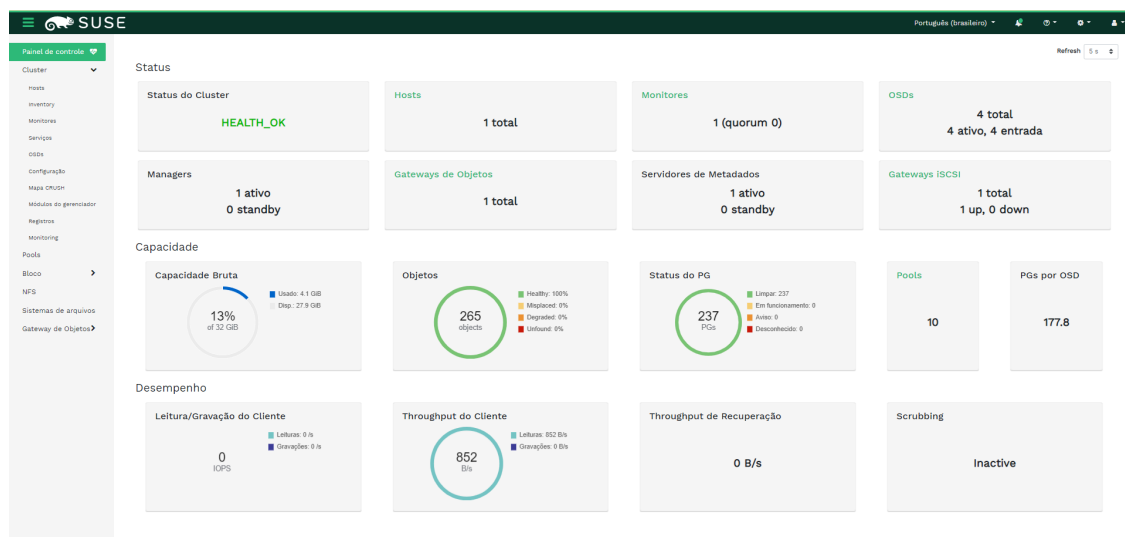


FIGURA 2.3: HOME PAGE DO CEPH DASHBOARD

## 2.2 Menu de utilitários

A lateral superior direita da tela contém um menu de utilitários. Ele inclui tarefas gerais relacionadas mais ao painel de controle do que ao cluster do Ceph. Ao clicar nas opções, você pode acessar os seguintes tópicos:

- Mudar a interface de idioma do painel de controle para: tcheco, alemão, inglês, espanhol, francês, indonésio, italiano, japonês, coreano, polonês, português (Brasil) e chinês
- Tarefas e notificações
- Consultar a documentação, as informações sobre a API REST ou outras informações sobre o painel de controle
- Gerenciamento de usuários e configuração de telemetria



### Nota

Para ver as descrições de linha de comando mais detalhadas das funções de usuário, consulte o [Capítulo 11, Gerenciar usuários e funções na linha de comando](#).

- Configuração de login; mudar a senha ou efetuar logout

## 2.3 Menu principal

O menu principal do painel de controle ocupa a lateral esquerda da tela. Ela abrange os seguintes tópicos:

### *Painel de controle*

Retornar à home page do Ceph Dashboard.

### *Cluster*

Veja informações detalhadas sobre hosts, inventário, Ceph Monitors, serviços, Ceph OSDs, configuração de cluster, Mapa CRUSH, módulos do Ceph Manager, registros e monitoramento.

### *Pools*

Ver e gerenciar pools de cluster.

### *Bloquear*

Veja informações detalhadas e gerencie imagens de dispositivos de blocos RADOS, espelhamento e iSCSI.

### *NFS*

Ver e gerenciar implantações do NFS Ganesha.



### Nota

Se o NFS Ganesha não foi implantado, um aviso informativo é exibido. Consulte a [Seção 11.6, “Configurando o NFS Ganesha no Ceph Dashboard”](#).

### *Sistemas de arquivos*

Ver e gerenciar CephFSs.

### *Gateway de Objetos*

Ver e gerenciar daemons, usuários e compartimentos de memória do Gateway de Objetos.



### Nota

Se o Gateway de Objetos não foi implantado, um aviso informativo é exibido. Consulte a [Seção 10.4, “Habilitando o front end de gerenciamento do Gateway de Objetos”](#).





## 2.4 Painel de conteúdo

O painel de conteúdo ocupa a parte principal da tela do painel de controle. A home page do painel de controle mostra muitos widgets úteis para informar você brevemente sobre as informações de status atual, capacidade e desempenho do cluster.


## 2.5 Recursos comuns da IU da Web

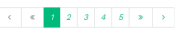
No Ceph Dashboard, você costuma trabalhar com *listas*. Por exemplo, listas de pools, nós OSD ou dispositivos RBD. Todas as listas se atualizarão automaticamente por padrão a cada cinco segundos. Os seguintes widgets comuns ajudam você a gerenciar ou ajustar essas listas:

Clique em  para acionar uma atualização manual da lista.

Clique em  para exibir ou ocultar colunas individuais da tabela.

Clique em  e insira (ou selecione) quantas linhas serão exibidas em uma única página.

Clique dentro de  e filtre as linhas digitando a string de pesquisa.

Use  para mudar a página exibida no momento, se a lista ocupar várias páginas.

## 2.6 Widgets do painel de controle

Cada widget do painel de controle mostra informações de status específicas relacionadas a um determinado aspecto de um cluster do Ceph em execução. Alguns widgets são links ativos e, depois de clicar neles, redirecionarão você para uma página detalhada relacionada do tópico que eles representam.



### Dica: Mais detalhes ao mover o cursor do mouse

Alguns widgets gráficos mostram mais detalhes quando você move o cursor do mouse sobre eles.

### 2.6.1 Widgets de status

Os widgets de *status* apresentam uma breve visão geral do status atual do cluster.

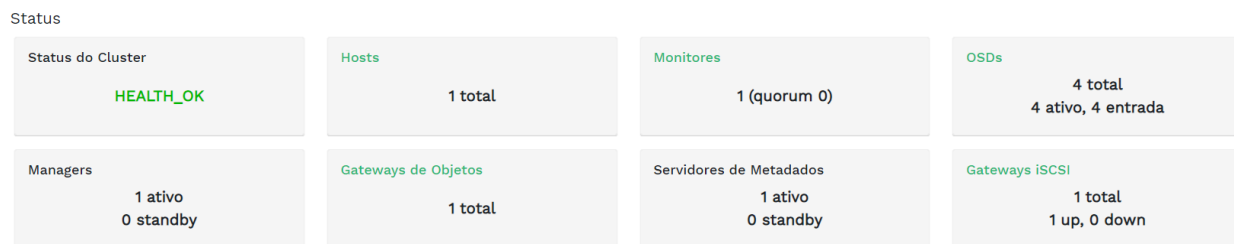


FIGURA 2.4: WIDGETS DE STATUS

### **Status do Cluster**

Apresenta informações básicas sobre a saúde do cluster.

### **Hosts**

Mostra o número total de nós do cluster.

### **Monitores**

Mostra o número de monitores em execução e o quorum deles.

### **OSDs**

Mostra o número total de OSDs e o número de OSDs *ativos* e *incluídos*.

### **Gerentes**

Mostra o número de daemons Ceph Manager ativos e em standby.

### **Gateways de Objetos**

Mostra o número de Gateways de Objetos em execução.

### **Servidores de Metadados**

Mostra o número de Servidores de Metadados.

### **iSCSI Gateways**

Mostra o número de gateways iSCSI configurados.

## 2.6.2 Widgets de capacidade

Os widgets de *capacidade* mostram informações resumidas sobre a capacidade de armazenamento.

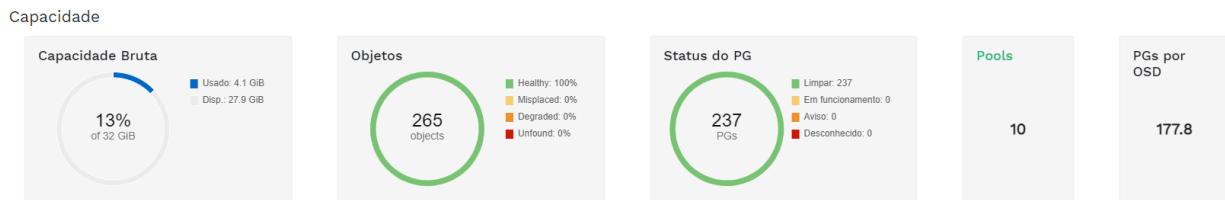


FIGURA 2.5: WIDGETS DE CAPACIDADE

### **Capacidade Bruta**

Mostra a proporção de capacidade de armazenamento usado e bruto disponível.

### **Objetos**

Mostra o número de objetos de dados armazenados no cluster.

### **Status do PG**

Exibe um gráfico dos grupos de posicionamento de acordo com o status deles.

### **Pools**

Mostra o número de pools no cluster.

### **PGs por OSD**

Mostra o número médio de grupos de posicionamento por OSD.

## 2.6.3 Widgets de desempenho

Os widgets de *desempenho* referem-se aos dados básicos de desempenho dos clientes do Ceph.

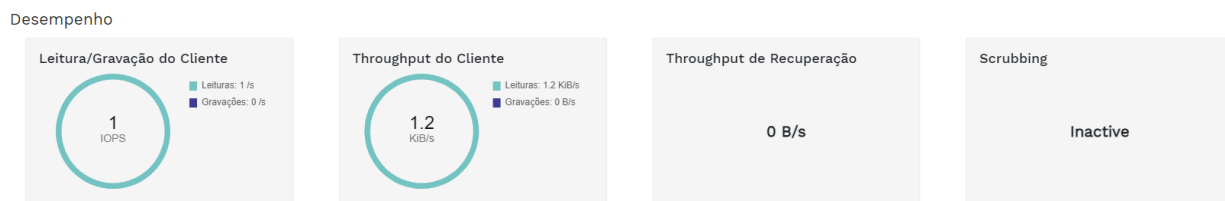


FIGURA 2.6: WIDGETS DE DESEMPENHO

### **Leitura/Gravação do Cliente**

A quantidade de operações de leitura e gravação dos clientes por segundo.

### **Throughput do Cliente**

A quantidade de dados transferidos de e para os clientes do Ceph em bytes por segundo.

### ***Throughput de Recuperação***

O throughput de dados recuperados por segundo.

### ***Depuração***


Mostra o status da remoção (consulte a [Seção 17.4.9, “Depurando um grupo de posicionamento”](#)). Ele é inativo, habilitado ou ativo.

## 3 Gerenciar usuários e funções do Ceph Dashboard

O gerenciamento de usuários do painel de controle realizado pelos comandos do Ceph na linha de comando já foi apresentado na [Capítulo 11, Gerenciar usuários e funções na linha de comando](#).

Esta seção descreve como gerenciar as contas de usuário usando a interface do usuário da Web do painel de controle.

### 3.1 Listando usuários

Clique em  no menu de utilitários e selecione *Gerenciamento de usuários*.

A lista contém o nome de usuário, o nome completo e o e-mail de cada usuário, uma lista de funções atribuídas, se a função está habilitada e a data de vencimento da senha.



Nome de usuário	Nome	E-mail	Funções	Habilitado	Password expiration date
admin			administrator	✓	
Alex	Alexandra Settle	tux@example.com	cluster-manager, pool-manager	✓	
dashboard user 1	Dashboard User1	du1@example.com		✓	
rgw user	RGW User	rgw@example.com	pool-manager, rgw-manager	✓	

0 selecionado(s) / 4 total

FIGURA 3.1: GERENCIAMENTO DE USUÁRIOS

### 3.2 Adicionando novos usuários

Clique em *Criar* na parte superior esquerda do cabeçalho da tabela para adicionar um novo usuário. Insira seu nome de usuário, senha e, opcionalmente, um nome completo e um e-mail.

Criar Usuário

Nome de usuário \*

potato

✓

Senha ?

.....

✓

👁

Confirmar senha

.....

✓

👁

Password expiration date ?

Password expiration date...

✕

Nome completo

Mr Potato

✓

E-mail

potato@example.com

✓

Funções

✎

There are no roles.

☒ Habilitado

☒ User must change password at next logon

Criar Usuário

Cancelar

FIGURA 3.2: ADICIONANDO UM USUÁRIO

Clique no ícone pequeno de caneta para atribuir funções predefinidas ao usuário. Clique em *Criar usuário* para confirmar.


### 3.3 Editando usuários

Clique na linha da tabela de um usuário para realçar a seleção e escolha *Editar* para editar os detalhes sobre o usuário. Clique em *Editar Usuário* para confirmar.

### 3.4 Apagando usuários

Clique na linha da tabela de um usuário para realçar a seleção, escolha o caixa suspensa ao lado de *Editar* e selecione *Excluir* na lista para apagar a conta do usuário. Ative a caixa de seleção *Sim, desejo* e clique em *Excluir usuário* para confirmar.

## 3.5 Listando funções de usuário

Clique em  no menu de utilitários e selecione *Gerenciamento de usuários*. Em seguida, clique na guia *Funções*.

A lista mostra o nome e a descrição de cada função e se é uma função do sistema.



Nome	Descrição	Função do Sistema
administrator	Administrator	✓
block-manager	Block Manager	✓
cephfs-manager	CephFS Manager	✓
cluster-manager	Cluster Manager	✓
ganasha-manager	NFS Ganesha Manager	✓
pool-manager	Pool Manager	✓
read-only	Read-Only	✓
rgw-manager	RGW Manager	✓

FIGURA 3.3: FUNÇÕES DE USUÁRIO

## 3.6 Adicionando funções personalizadas

Clique em *Criar* na parte superior esquerda do cabeçalho da tabela para adicionar uma nova função personalizada. Insira o *Nome* e a *Descrição* e selecione as permissões apropriadas ao lado de *Permissões*.



### Dica: Purgando funções personalizadas

Se você cria funções de usuário personalizadas e depois planeja remover o cluster do Ceph com o comando **ceph-salt purge**, precisa purgar primeiro as funções personalizadas. Encontre mais detalhes na [Seção 13.9, “Removendo um cluster inteiro do Ceph”](#).

Criar Role

Nome \*

ganesha pool user ✓

Descrição

a user that can only manage ganesha and pools ✓

Permissões

<input type="checkbox"/> Tudo	<input type="checkbox"/> Leitura	<input type="checkbox"/> Criar	<input type="checkbox"/> Atualizar	<input type="checkbox"/> Excluir
<input type="checkbox"/> cephfs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> config-opt	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> dashboard-settings	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> grafana	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> hosts	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> iscsi	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> log	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> manager	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> monitor	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/> nfs-ganesha	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input type="checkbox"/> osd	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> pool	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> prometheus	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> rbd-image	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> rbd-mirroring	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> rgw	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> user	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Criar Role

Cancelar

FIGURA 3.4: ADICIONANDO UMA FUNÇÃO



### Dica: Ativação múltipla

Ao ativar a caixa de seleção que antecede o nome do tópico, você ativa todas as permissões para esse tópico. Ao ativar a caixa de seleção *Tudo*, você ativa todas as permissões para todos os tópicos.

Clique em *Criar função* para confirmar.



### 3.7 Editando funções personalizadas

Clique na linha da tabela de um usuário para realçar a seleção e escolha *Editar* na parte superior esquerda do cabeçalho da tabela para editar uma descrição e as permissões da função personalizada. Clique em *Editar Função* para confirmar.

### 3.8 Apagando funções personalizadas

Clique na linha da tabela de uma função para realçar a seleção, escolha o caixa suspensa ao lado de *Editar* e selecione *Excluir* na lista para apagar a função. Ative a caixa de seleção *Sim, desejo* e clique em *Excluir função* para confirmar.

## 4 Ver detalhes internos do cluster

O item de menu *Cluster* permite ver informações detalhadas sobre hosts de cluster do Ceph, inventário, Ceph Monitors, serviços, OSDs, configuração, Mapa CRUSH, Ceph Manager, registros e arquivos de monitoramento.

### 4.1 Visualizando nós do cluster

Clique em *Cluster* > *Hosts* para ver uma lista de nós do cluster.



Nome de host	Serviços	Labels	Versão
> master			
> node1	mgr.node1.wbmq, mon.node1, osd.0, osd.3		15.2.4-557-g4ac763f0b3
> node2	mgr.node2.qcwalx, mon.node2, osd.1, osd.4		15.2.4-557-g4ac763f0b3
> node3	mgr.node3.rhkzzy, mon.node3, osd.2, osd.5		15.2.4-557-g4ac763f0b3

FIGURA 4.1: HOSTS

Clique na seta suspensa ao lado do nome de um nó na coluna *Nome de host* para ver os detalhes do desempenho do nó.

A coluna *Serviços* lista todos os daemons que estão em execução em cada nó relacionado. Clique no nome de um daemon para ver sua configuração detalhada.

### 4.2 Acessando o inventário do cluster

Clique em *Cluster* > *Inventário* para ver uma lista de dispositivos. A lista inclui o caminho, o tipo, a disponibilidade, o fornecedor, o modelo, o tamanho e os OSDs do dispositivo.

Clique para selecionar o nome de um nó na coluna *Nome de host*. Quando selecionado, clique em *Identificar* para identificar o dispositivo no qual o host está sendo executado. Esse procedimento instrui o dispositivo a piscar os LEDs. Selecione a duração dessa ação entre 1, 2, 5, 10 ou 15 minutos. Clique em *Executar*.

Identify							
Nome de host	Device path	Tipo	Available	Vendor	Model	Tamanho	OSDs
master	/dev/vda	HDD		0x1af4		42 GiB	
node1	/dev/vda	HDD		0x1af4		42 GiB	
node1	/dev/vdb	HDD		0x1af4		8 GiB	osd.0
node1	/dev/vdc	HDD		0x1af4		8 GiB	osd.3
node2	/dev/vda	HDD		0x1af4		42 GiB	
node2	/dev/vdb	HDD		0x1af4		8 GiB	osd.1
node2	/dev/vdc	HDD		0x1af4		8 GiB	osd.4
node3	/dev/vda	HDD		0x1af4		42 GiB	
node3	/dev/vdb	HDD		0x1af4		8 GiB	osd.2
node3	/dev/vdc	HDD		0x1af4		8 GiB	osd.5

0 selecionado(s) / 10 total

FIGURA 4.2: SERVIÇOS

## 4.3 Visualizando Ceph Monitors

Clique em *Cluster > Monitores* para ver uma lista de nós do cluster com monitores do Ceph em execução. O painel de conteúdo é dividido em duas telas: Status e No Quorum ou Não está no Quorum.

A tabela *Status* mostra as estatísticas gerais dos Ceph Monitors em execução, incluindo as seguintes:

- ID do cluster
- monmap modificado
- época de monmap
- quorum con
- quorum mon
- con obrigatório
- mon obrigatório

Os painéis No Quorum e Não está no Quorum incluem o nome, o número de classificação, o endereço IP público e o número de sessões abertas de cada monitor.

Clique no nome de um nó na coluna *Nome* para ver a configuração do Ceph Monitor relacionado.

Status	
ID do Cluster	05766fa4-a9a7-11eb-9e46-525400b22828
monmap modificado	2021-04-30T11:27:20.465652Z
época de monmap	1
quorum con	4540138292840890367
quorum mon	kraken,luminous,mimic,osdmap-prune,nautilus,octopus
con obrigatório	2449958747315978244
mon obrigatório	kraken,luminous,mimic,osdmap-prune,nautilus,octopus

No Quorum

Nome	Posição	Endereço Público	Sessões Abertas
node1	0	10.20.156.201:6789/0	.....
node2	2	10.20.156.202:6789/0	.....
node3	1	10.20.156.203:6789/0	.....
3 total			

Não está no Quorum

Nome	Posição	Endereço Público
No data to display		
0 total		

FIGURA 4.3: CEPH MONITORS

## 4.4 Exibindo serviços

Clique em *Cluster* > *Serviços* para ver os detalhes de cada um dos serviços disponíveis: *crash*, *Ceph Manager* e *Ceph Monitors*. A lista inclui o nome da imagem do container, o ID da imagem do container, o status do elemento que está em execução, o tamanho e a data da última atualização.

Clique na seta suspensa ao lado do nome de um serviço na coluna *Serviço* para ver os detalhes do daemon. A lista de detalhes inclui nome de host, tipo de daemon, ID do daemon, ID do container, nome da imagem do container, ID da imagem do container, número da versão, status e data da última atualização.

Cluster > Services							
✕ Excluir							
Serviço	Container image name	Container image ID	Placement	Running	Tamanho	Last Refreshed	
crash	registry.suse.de/devel/storage/7.0/containers/sep/ceph/ceph:latest	6549871c3f67			4	2020-08-14T13:37:34.148847	
Daemons							
Nome de host	Daemon type	Daemon ID	Container ID	Container image name	Container image ID	Versão	Status
master	crash	master	3acfc1fb607e	registry.suse.de/devel/storage/7.0/containers/sep/ceph/ceph:latest	6549871c3f67	15.2.4.557	running
node1	crash	node1	3d56e2a421eb	registry.suse.de/devel/storage/7.0/containers/sep/ceph/ceph:latest	6549871c3f67	15.2.4.557	running
node2	crash	node2	8fa9790b9a51	registry.suse.de/devel/storage/7.0/containers/sep/ceph/ceph:latest	6549871c3f67	15.2.4.557	running
node3	crash	node3	b047531bbf2a	registry.suse.de/devel/storage/7.0/containers/sep/ceph/ceph:latest	6549871c3f67	15.2.4.557	running
4 total							
> mgr	registry.suse.de/devel/storage/7.0/containers/sep/ceph/ceph:latest	dcfacef0831b	master	1	1	2021-06-05T12:57:02.424523Z	
> mon	registry.suse.de/devel/storage/7.0/containers/sep/ceph/ceph:latest	dcfacef0831b	master:10.20.165.200	1	1	2021-06-05T12:57:02.425266Z	
> node-exporter	registry.suse.com/caasp/v4.5/prometheus-node-exporter:0.18.1	a149a78bcd37	master	1	1	2021-06-05T12:57:02.426742Z	
> osd.sesdev_osd_deployer	registry.suse.de/devel/storage/7.0/containers/sep/ceph/ceph:latest	dcfacef0831b	master	4	4	2021-06-05T12:57:02.425374Z	
1 selecionado(s) / 5 total							

FIGURA 4.4: SERVIÇOS

## 4.5 Exibindo Ceph OSDs

Clique em *Cluster* > *OSDs* para ver uma lista de nós com daemons OSD em execução. A lista inclui nome, ID, status, classe do dispositivo, número de grupos de posicionamento, tamanho, uso, cronograma de leituras/gravações e taxa de operações de leitura/gravação por segundo de cada nó.

Lista de OSDs

Desempenho Geral

+ Criar

Cluster-wide configuration

10

	Host	ID	Status	Device class	PGs	Tamanho	Flags	Uso	Bytes de leitura	Bytes de gravação	Op. de leitura	Op. de gravação
<input type="checkbox"/>	> node1	0	<div>in up</div>	hdd	0	8 GiB		<div><div></div>13%</div>	<div><div></div></div>	<div><div></div></div>	0 /s	0 /s
<input type="checkbox"/>	> node2	1	<div>in up</div>	hdd	1	8 GiB		<div><div></div>13%</div>	<div><div></div></div>	<div><div></div></div>	0 /s	0 /s
<input type="checkbox"/>	> node3	2	<div>in up</div>	hdd	1	8 GiB		<div><div></div>13%</div>	<div><div></div></div>	<div><div></div></div>	0 /s	0 /s
<input type="checkbox"/>	> node1	3	<div>in up</div>	hdd	1	8 GiB		<div><div></div>13%</div>	<div><div></div></div>	<div><div></div></div>	0 /s	0 /s
<input type="checkbox"/>	> node2	4	<div>in up</div>	hdd	1	8 GiB		<div><div></div>13%</div>	<div><div></div></div>	<div><div></div></div>	0 /s	0 /s
<input type="checkbox"/>	> node3	5	<div>in up</div>	hdd	0	8 GiB		<div><div></div>13%</div>	<div><div></div></div>	<div><div></div></div>	0 /s	0 /s

0 selecionado(s) / 6 total

FIGURA 4.5: CEPH OSDS

Selecione *Flags* no menu suspenso *Configuração de todo o cluster* no cabeçalho da tabela para abrir uma janela popup. Ela mostra uma lista de flags que se aplicam a todo o cluster. É possível ativar ou desativar flags individuais e clicar em *Enviar* para confirmar.

Flags OSD de todo o Cluster

☐ Não Entrada

Os OSDs que já foram marcados como saída não serão remarcados como entrada ao serem iniciados

☐ Não Saída

Os OSDs não serão automaticamente marcados como saída após o intervalo configurado

☐ Não Ativo

Os OSDs não podem ser iniciados

☐ Não Inativo

Os relatórios de falha de OSD estão sendo ignorados, portanto, os monitores não marcarão os OSDs como inativos

☐ Pausar

Pausa leituras e gravações

☐ Sem Remoção

Remoção desabilitada

☐ Sem Remoção Profunda

Remoção Profunda desabilitada

☐ Sem Provisionamento

Provisionamento de PGs suspenso

☐ No Rebalance

OSD will choose not to backfill unless PG is also degraded

☐ Sem Recuperação

Recuperação de PGs suspensa

☒ Classificação Bit a Bit

Usar classificação bit a bit

☒ Snapdirs Purgados

Enviar

Cancelar

FIGURA 4.6: **FLAGS OSD**

Selecione *Prioridade de Recuperação* no menu suspenso *Configuração de todo o cluster* no cabeçalho da tabela para abrir uma janela popup. Ela mostra uma lista de prioridades de recuperação de OSD que se aplicam a todo o cluster. É possível ativar o perfil de prioridade preferencial e ajustar os valores individuais a seguir. Clique em *Enviar* para confirmar.

Prioridade de Recuperação de OSD

Prioridade \*

Personalizado

Personalizar valores de prioridade

Máx. de Provisionamentos \* ?

1

Máx. Recuperação Ativo \* ?

0

Máx. Recuperação Inicialização Única \*

1

Suspensão de Recuperação \* ?

0

Enviar

Cancelar

FIGURA 4.7: PRIORIDADE DE RECUPERAÇÃO DE OSD

Clique na seta suspensa ao lado do nome de um nó na coluna *Host* para ver uma tabela estendida com os detalhes das configurações e do desempenho do dispositivo. Ao navegar pelas diversas guias, você pode ver listas de *Atributos*, *Metadados*, *Saúde do dispositivo*, *Contador de desempenho*, um *Histograma* gráfico das leituras e gravações e *Detalhes de Desempenho*.

Lista de OSDs

Desempenho Geral

Editar

Cluster-wide configuration

10

	Host	ID	Status	Device class	PGs	Tamanho	Flags	Uso	Bytes de leitura	Bytes de gravação	Op. de leitura	Op. de gravação
<input checked="" type="checkbox"/>	master	0	In Up	hdd	171	8 GiB		12%			1,799587935387338 /s	0 /s

Devices

Atributos (mapa OSD)

Metadados

Device health

Contador de desempenho

Histograma

Detalhes de Desempenho

10

Nome	Descrição	Valor
bluefs.bytes_written_slow	Bytes written to WAL/SSTs at slow device	0
bluefs.bytes_written_sst	Bytes written to SSTs	0
bluefs.bytes_written_wal	Bytes written to WAL	0
bluefs.db_total_bytes	Total bytes (main db device)	1073741824
bluefs.db_used_bytes	Used bytes (main db device)	312868864
bluefs.log_bytes	Size of the metadata log	6750208
bluefs.logged_bytes	Bytes written to the metadata log	0
bluefs.num_files	File count	12
bluefs.read_bytes	Bytes requested in buffered read mode	0
bluefs.read_prefetch_bytes	Bytes requested in prefetch read mode	0

112 total

<

1

2

3

4

5

>

FIGURA 4.8: DETALHES DO OSD

21

Exibindo Ceph OSDs | SES 7.1



## Dica: Executando tarefas específicas nos OSDs

Depois que você clicar no nome de um nó OSD, a linha da tabela será realçada. Isso significa que agora você pode executar uma tarefa no nó. É possível executar qualquer uma das seguintes ações: *Editar*, *Criar*, *Remoção*, *Remoção Profunda*, *Reponderar*, *Marcar como Saída*, *Marcar como Entrada*, *Marcar como Inativo*, *Marcar como Perdido*, *Purgar*, *Destruir* ou *Excluir*.

Clique na seta para baixo na parte superior esquerda do cabeçalho da tabela ao lado do botão *Criar* e selecione a tarefa que deseja executar.

### 4.5.1 Adicionando OSDs

Para adicionar novos OSDs, siga estas etapas:

1. Verifique se alguns nós do cluster têm dispositivos de armazenamento com status disponível. Em seguida, clique na seta para baixo na parte superior esquerda do cabeçalho da tabela e selecione *Criar*. Isso abre a janela *Criar OSDs*.

The screenshot shows a 'Criar OSDs' window. It has a title bar 'Criar OSDs'. Inside, there are sections for 'Primary devices' and 'Shared devices'. 'Primary devices' has an 'Adicionar' button. 'Shared devices' has a sub-section 'WAL devices' and 'DB devices', each with an 'Adicionar' button. Below these is a 'Configuração' section with a 'Recursos' label and an 'Encryption' checkbox. At the bottom right are 'Visualizar' and 'Cancelar' buttons.

FIGURA 4.9: CRIAR OSDS

2. Para adicionar dispositivos de armazenamento principais aos OSDs, clique em *Adicionar*. Antes de adicionar dispositivos de armazenamento, você precisa especificar os critérios de filtragem na parte superior direita da tabela *Dispositivos principais*, por exemplo *Tipo hdd*. Confirme com *Add* (Adicionar).





Visualização da criação de OSD

DriveGroups

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

Criar

Cancelar

FIGURA 4.12:

## 5. Novos dispositivos serão adicionados à lista de OSDs.

	Host	ID	Status	Device class	PGs	Tamanho	Flags	Uso	Bytes de leitura	Bytes de gravação	Op. de leitura	Op. de gravação
<input type="checkbox"/>	> doc-ses-min2	0	<span>in</span> <span>up</span>	hdd	119	10 GiB		<div><div></div></div> 11%			0.7999105934891158 /s	0 /s
<input type="checkbox"/>	> doc-ses-min3	1	<span>in</span> <span>up</span>	hdd	108	10 GiB		<div><div></div></div> 11%			1.5998816768416986 /s	0 /s
<input type="checkbox"/>	> doc-ses-min4	2	<span>in</span> <span>up</span>	hdd	126	10 GiB		<div><div></div></div> 11%			0 /s	0 /s
<input type="checkbox"/>	> doc-ses-min1	3	<span>in</span> <span>up</span>	hdd	96	12 GiB		<div><div></div></div> 9%			0.3999455526088382 /s	0 /s
<input type="checkbox"/>	> doc-ses-min1	4	<span>in</span> <span>up</span>	hdd	76	12 GiB		<div><div></div></div> 9%			1.9995708432976873 /s	0 /s

0 selecionado(s) / 5 total

FIGURA 4.13: OSDS RECÉM-ADICIONADOS

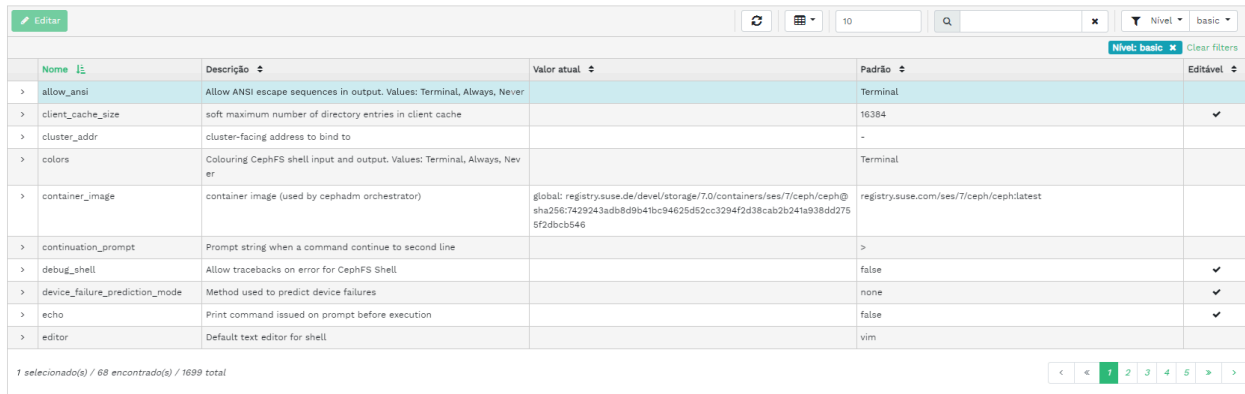


## Nota

Não há visualização do andamento do processo de criação de OSDs. Leva um tempo para eles serem realmente criados. Os OSDs aparecerão na lista quando forem implantados. Para ver os registros e verificar o status da implantação, clique em *Cluster > Registros*.

## 4.6 Visualizando a configuração do cluster

Clique em *Cluster* > *Configuração* para ver uma lista completa de opções de configuração de cluster do Ceph. A lista contém o nome da opção, sua descrição resumida e seus valores atuais e padrão e indica se a opção é editável.



The screenshot shows the Ceph configuration interface. At the top, there is a search bar and a filter dropdown set to 'Nível: basic'. Below this is a table with columns: Nome, Descrição, Valor atual, Padrão, and Editável. The table lists various configuration options like allow\_ansi, client\_cache\_size, cluster\_addr, colors, container\_image, continuation\_prompt, debug\_shell, device\_failure\_prediction\_mode, echo, and editor. Each row has a dropdown arrow next to the 'Nome' column. At the bottom right, there is a pagination bar showing '1 selecionado(s) / 68 encontrado(s) / 1699 total' and a set of numbered links (1, 2, 3, 4, 5).

Nome	Descrição	Valor atual	Padrão	Editável
allow_ansi	Allow ANSI escape sequences in output. Values: Terminal, Always, Never		Terminal	
client_cache_size	soft maximum number of directory entries in client cache		16384	✓
cluster_addr	cluster-facing address to bind to		-	
colors	Colouring CephFS shell input and output. Values: Terminal, Always, Never		Terminal	
container_image	container image (used by cephadm orchestrator)	global: registry.suse.de/devrel/storage/7.0/containers/ses/7/ceph/ceph@sha256:7429243adb8d9b41bc94625d52cc3294f2d38cab2b241a938dd2755f2d0cb546	registry.suse.com/ses/7/ceph/cephlatest	
continuation_prompt	Prompt string when a command continue to second line		>	
debug_shell	Allow tracebacks on error for CephFS Shell		false	✓
device_failure_prediction_mode	Method used to predict device failures		none	✓
echo	Print command issued on prompt before execution		false	✓
editor	Default text editor for shell		vim	

FIGURA 4.14: CONFIGURAÇÃO DO CLUSTER

Clique na seta suspensa ao lado de uma opção de configuração na coluna *Nome* para ver uma tabela estendida com informações detalhadas sobre a opção, como tipo de valor, valores mínimo e máximo permitidos, se é possível atualizá-la em tempo de execução etc.

Após realçar uma opção específica, você poderá editar o(s) valor(s) dela clicando no botão *Editar* na parte superior esquerda do cabeçalho da tabela. Clique em *Salvar* para confirmar as mudanças.

## 4.7 Vendo o mapa CRUSH do

Clique em *Cluster* > *Mapa CRUSH* para ver um Mapa CRUSH do cluster. Para obter mais informações gerais sobre Mapas CRUSH, consulte a [Seção 17.5, "Manipulação de mapa CRUSH"](#).

Clique em cada raiz, nó ou OSD para ver informações mais detalhadas, como peso do CRUSH, profundidade na árvore de mapa, classe de dispositivo do OSD e muitos mais.

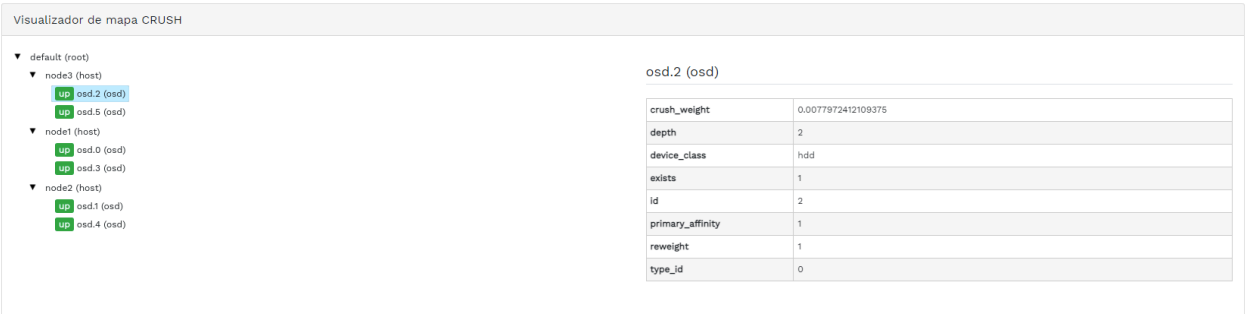


FIGURA 4.15: MAPA CRUSH

## 4.8 Visualizando módulos do gerenciador

Clique em *Cluster > Módulos do gerenciador* para ver uma lista dos módulos disponíveis do Ceph Manager. Cada linha consiste no nome do módulo e nas informações que indicam se ele está ou não habilitado no momento.

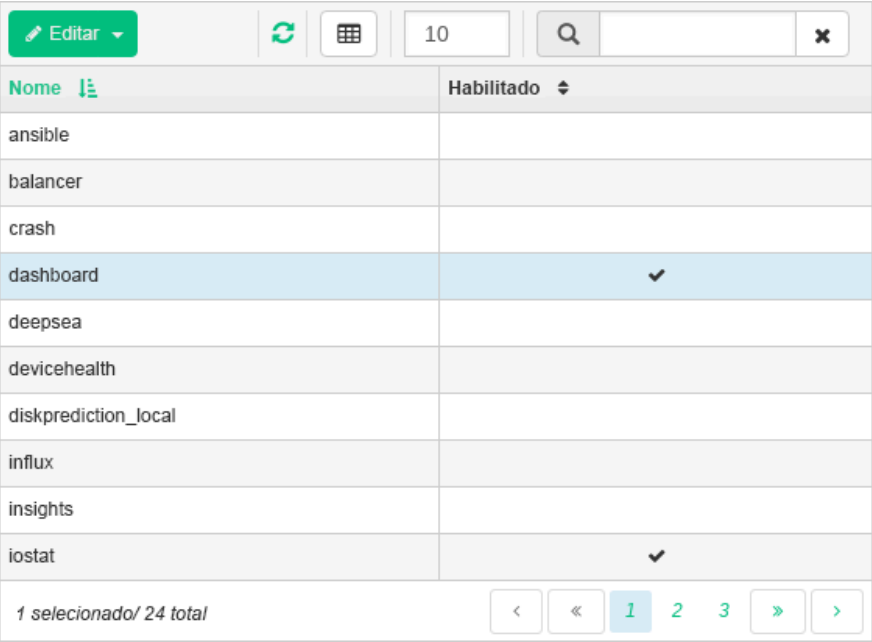


FIGURA 4.16: MÓDULOS DO GERENCIADOR

Clique na seta suspensa ao lado de um módulo na coluna *Nome* para ver uma tabela estendida com as configurações detalhadas na tabela *Detalhes* abaixo. Para editá-las, clique no botão *Editar* na parte superior esquerda do cabeçalho da tabela. Clique em *Atualizar* para confirmar as mudanças.

Clique na seta suspensa ao lado do botão *Editar* na parte superior esquerda do cabeçalho da tabela para *Habilitar* ou *Desabilitar* um módulo.

## 4.9 Visualizando registros

Clique em *Cluster* > *Registros* para ver uma lista das entradas de registro recentes do cluster. Cada linha consiste em uma marcação de horário, no tipo de entrada de registro e na própria mensagem registrada.

Clique na guia *Registros de Auditoria* para ver as entradas de registro do subsistema de auditoria. Consulte a [Seção 11.5, “Fazendo auditoria das solicitações de API”](#) para ver os comandos que habilitam ou desabilitam a auditoria.

Registros do Cluster		Registros de Auditoria
04-04-2019 02:53:02.989912	[INF]	from='mgr.3269933 10.100.24.61:0/27247' entity='mgr.doc-ses6min2' cmd='[{"prefix": "osd in", "format": "json", "ids": ["8"]}]: finished
04-04-2019 02:53:02.240046	[INF]	from='mgr.3269933 10.100.24.61:0/27247' entity='mgr.doc-ses6min2' cmd=[{"prefix": "osd in", "format": "json", "ids": ["8"]}]: dispatch
04-04-2019 02:52:34.282994	[INF]	from='mgr.3269933 10.100.24.61:0/27247' entity='mgr.doc-ses6min2' cmd=[{"prefix": "osd purge-actual", "format": "json", "id": 4, "yes_i_really_mean_it": true}]: finished
04-04-2019 02:52:33.946127	[INF]	from='mgr.3269933 10.100.24.61:0/27247' entity='mgr.doc-ses6min2' cmd=[{"prefix": "osd purge-actual", "format": "json", "id": 4, "yes_i_really_mean_it": true}]: dispatch
04-04-2019 02:52:09.570264	[INF]	from='mgr.3269933 10.100.24.61:0/27247' entity='mgr.doc-ses6min2' cmd='[{"prefix": "osd in", "format": "json", "ids": ["4"]}]: finished

FIGURA 4.17: REGISTROS

## 4.10 Visualizando o monitoramento

Clique em *Cluster* > *Monitoramento* para gerenciar e ver detalhes sobre os alertas do Prometheus. Se o Prometheus estiver ativo, você poderá ver informações detalhadas sobre *Alertas Ativos*, *Todos os Alertas* ou *Silêncios* nesse painel de conteúdo.



### Nota

Se você não tiver o Prometheus implantado, um banner de informações será exibido com um link para a documentação relevante.

## 5 Gerenciar pools



### Dica: Mais informações sobre pools

Para obter mais informações gerais sobre os pools do Ceph, consulte o [Capítulo 18, Gerenciar pools de armazenamento](#). Para obter informações específicas dos pools com código de eliminação, consulte o [Capítulo 19, Pools codificados para eliminação](#).

Para listar todos os pools disponíveis, clique em *Pools* no menu principal.

A lista mostra o nome, o tipo, o aplicativo relacionado, o status do grupo de posicionamento, o tamanho da réplica, a última mudança, o perfil codificado para eliminação, o conjunto de regras crush, o uso e as estatísticas de leitura/gravação de cada pool.

Lista de Pools

Desempenho Geral

<div>+ Adicionar</div> <div> <div>10</div> <div>Q</div> <div>X</div> </div>												
Nome	Tipo	Aplicativos	Status do PG	Tam da R	Últim Alteri	Perfil Codificado para Eliminação	Conjunto de Regras do Crush	Uso	Bytes de leitura	Bytes de gravação	Op. de lei	Op. de gr
.rgw.root	replicado	rgw	8 ativo+limpo	3	22		replicated_rule	0%	<div></div>	<div></div>	0/s	0/s
cephfs_data	replicado	cephfs	256 ativo+limpo	3	209		replicated_rule	0%	<div></div>	<div></div>	0/s	0/s
cephfs_metadata	replicado	cephfs	64 ativo+limpo	3	210		replicated_rule	0%	<div></div>	<div></div>	0/s	0/s
default.rgw.buckets.index	replicado	rgw	8 ativo+limpo	3	75		replicated_rule	0%	<div></div>	<div></div>	0/s	0/s
default.rgw.control	replicado	rgw	8 ativo+limpo	3	25		replicated_rule	0%	<div></div>	<div></div>	0/s	0/s
default.rgw.log	replicado	rgw	8 ativo+limpo	3	30		replicated_rule	0%	<div></div>	<div></div>	0/s	0/s
default.rgw.meta	replicado	rgw	8 ativo+limpo	3	28		replicated_rule	0%	<div></div>	<div></div>	0/s	0/s
family_photos	replicado	cephfs	128 ativo+limpo	3	226		replicated_rule	0%	<div></div>	<div></div>	0/s	0/s
testing_rbd_pool	replicado	cephfs,rbd	128 ativo+limpo	3	76		replicated_rule	0%	<div></div>	<div></div>	0,8/s	0/s
0 selecionado/9 total												

FIGURA 5.1: LISTA DE POOLS

Clique na seta suspensa ao lado do nome de um pool na coluna *Nome* para ver uma tabela estendida com informações detalhadas sobre o pool, como detalhes gerais e de desempenho e configuração.

## 5.1 Adicionando um novo pool

Para adicionar um novo pool, clique em *Criar* na parte superior esquerda da tabela de pools. No formulário do pool, você pode inserir o nome, o tipo, os aplicativos, o modo de compactação e as cotas do pool, incluindo o máximo de bytes e de objetos. O próprio formulário do pool pré-calcula o número de grupos de posicionamento mais adequado a esse pool específico. O cálculo é baseado na quantidade de OSDs no cluster e no tipo de pool selecionado com suas configurações específicas. Assim que um número de grupos de posicionamento for definido manualmente, ele será substituído por um número calculado. Clique em *Criar Pool* para confirmar.

The screenshot shows the 'Criar Pool' form with the following fields and values:

- Nome \***: potato-pool
- Tipo de pool \***: replicated
- PG Autoscale**: on
- Tamanho replicado \***: 3
- Aplicativos**: cephfs
- CRUSH**
  - Conjunto de regras do Crush**: replicated\_rule
- Compactação**
  - Modo**: none
- Quotas**
  - Max bytes**: ex. 10 GiB
  - Max objects**: 0

Buttons at the bottom: Criar Pool, Cancelar.

FIGURA 5.2: ADICIONANDO UM NOVO POOL

## 5.2 Apagando pools

Para apagar um pool, selecione-o na linha da tabela. Clique na seta suspensa ao lado do botão *Criar* e clique em *Excluir*.

## 5.3 Editando as opções de um pool

Para editar as opções de um pool, selecione-o na linha da tabela e clique em *Editar* na parte superior esquerda da tabela de pools.

É possível mudar o nome do pool, aumentar o número de grupos de posicionamento, modificar a lista de aplicativos e as configurações de compactação do pool. Clique em *Editar Pool* para confirmar.



## 6 Gerenciar dispositivos de blocos RADOS

Para listar todos os Dispositivos de Blocos RADOS (RBDs, RADOS Block Devices) disponíveis, clique em *Bloco > Imagens* no menu principal.

A lista mostra informações resumidas sobre o dispositivo, como nome do dispositivo, nome do pool relacionado, namespace, tamanho do dispositivo, número e tamanho dos objetos no dispositivo, dados sobre provisionamento dos detalhes e pai.

Imagens

Namespaces

Lixo

Desempenho Geral

+ Criar

10

Q

X

	Nome	Pool	Namespace	Tamanho	Objetos	Tamanho do objeto	Aprovisionado	Total provisionado	Pai
>	example_rbd_device	rbd		4 MiB	1	4 MiB	0 B	0 B	-
>	potato_rbd	rbd		10 MiB	3	4 MiB	0 B	0 B	-

0 selecionado(s) / 2 total

FIGURA 6.1: LISTA DE IMAGENS RBD

## 6.1 Visualizando detalhes sobre RBDs

Para ver informações mais detalhadas sobre um dispositivo, clique na linha dele na tabela:

Detalhes	Instantâneos	Configuração	Desempenho
Nome	example_rbd_device		
Pool	rbd		
Pool de Dados	-		
Criado	30/04/2021 15:02:38		
Tamanho	4 MiB		
Objetos	1		
Tamanho do objeto	4 MiB		
Recursos	deep-flatten exclusive-lock fast-diff layering object-map		
Aprovisionado	0 B		
Total provisionado	0 B		
Unidade de distribuição	4 MiB		
Total de distribuições	1		
Pai	-		
Prefixo do nome do bloco	rbd_data.37ff2b17a0d1		
Ordem	22		
Format Version	2		

FIGURA 6.2: DETALHES DO RBD

## 6.2 Visualizando a configuração do RBD

Para ver a configuração detalhada de um dispositivo, clique na linha dele na tabela e, em seguida, na guia *Configuração* na tabela inferior:

Detalhes	Instantâneos	Configuração	Desempenho
Nome ↕	Descrição ↕	Chave ↕	Origem ↕
Intermitência de BPS	O limite de bytes de E/S de intermitência desejado.	rbd_qos_bps_burst	Imagem
Limite de BPS	O limite de bytes por segundo de E/S desejado.	rbd_qos_bps_limit	Imagem
Intermitência de IOPS	O limite de operações de E/S de intermitência desejado.	rbd_qos_iops_burst	Imagem
Limite de IOPS	O limite de operações por segundo de E/S desejado.	rbd_qos_iops_limit	Imagem
Intermitência de BPS de Leitura	O limite de bytes de leitura de intermitência desejado.	rbd_qos_read_bps_burst	Imagem
Limite de BPS de Leitura	O limite de bytes por segundo de leitura desejado.	rbd_qos_read_bps_limit	Imagem
Intermitência de IOPS de Leitura	O limite de operações de leitura de intermitência desejado.	rbd_qos_read_iops_burst	Imagem
Limite de IOPS de Leitura	O limite de operações por segundo de leitura desejado.	rbd_qos_read_iops_limit	Imagem
Intermitência de BPS de Gravação	O limite de bytes de gravação de intermitência desejado.	rbd_qos_write_bps_burst	Imagem
Limite de BPS de Gravação	O limite de bytes por segundo de gravação desejado.	rbd_qos_write_bps_limit	Imagem

FIGURA 6.3: CONFIGURAÇÃO DO RBD

## 6.3 Criando RBDs

Para adicionar um novo dispositivo, clique em *Criar* na parte superior esquerda do cabeçalho da tabela e faça o seguinte na tela *Criar RBD*:

FIGURA 6.4: ADICIONANDO UM NOVO RBD

1. Insira o nome do novo dispositivo. Consulte o Livro *“Guia de Implantação”, Capítulo 2 “Requisitos e recomendações de hardware”, Seção 2.11 “Limitações de nome”* para saber as limitações de nomeação.
2. Selecione o pool com o aplicativo `rbd` atribuído que será usado como base para criação do novo dispositivo RBD.

3. Especifique o tamanho do novo dispositivo.
4. Especifique outras opções para o dispositivo. Para ajustar os parâmetros do dispositivo, clique em *Avançado* e insira os valores para tamanho do objeto, unidade de distribuição ou total de distribuições. Para inserir limites de Qualidade do Serviço (QoS), clique em *Qualidade do Serviço*.
5. Clique em *Criar RBD* para confirmar.

## 6.4 Apagando RBDs

Para apagar um dispositivo, selecione-o na linha da tabela. Clique na seta suspensa ao lado do botão *Criar* e clique em *Excluir*. Clique em *Excluir RBD* para confirmar a exclusão.



### Dica: Movendo RBDs para o lixo

A exclusão de um RBD é uma ação irreversível. Em vez disso, se você *Mover para Lixo*, poderá restaurar o dispositivo posteriormente selecionando-o na guia *Lixo* da tabela principal e clicando em *Restaurar* na parte superior esquerda do cabeçalho da tabela.

## 6.5 Criando instantâneos de dispositivo de blocos RADOS

Para criar um instantâneo de Dispositivo de Blocos RADOS, selecione o dispositivo na linha da tabela, e o painel de conteúdo da configuração detalhada é exibido. Selecione a guia *Instantâneos* e clique em *Criar* na parte superior esquerda do cabeçalho da tabela. Insira o nome do instantâneo e clique em *Criar Instantâneo do RBD* para confirmar.

Após selecionar um instantâneo, você poderá executar outras ações no dispositivo, como renomear, proteger, clonar, copiar ou apagar. *Rollback* restaura o estado do dispositivo com base no instantâneo atual.

Detalhes	Instantâneos	Configuração
----------	--------------	--------------

+ Criar		10	Q	X
Nome	Tamanho	Aprovisionado	Estado	Criado
testing_rbd-20190215T095402Z	10 MiB	0 B	NÃO PROTEGIDO	15/02/19 10:54:08 AM
testing_rbd-20190405T074138Z	10 MiB	0 B	NÃO PROTEGIDO	15/04/19 9:41:42 AM
0 selecionado/2 total				

FIGURA 6.5: INSTANTÂNEOS DO RBD

## 6.6 Espelhamento do RBD

É possível espelhar as imagens de Dispositivo de Blocos RADOS de forma assíncrona entre dois clusters do Ceph. Você pode usar o Ceph Dashboard para configurar a replicação de imagens RBD entre dois ou mais clusters. Esse recurso está disponível em dois modos:

### Com base em diário

Esse modo usa o recurso de registro de imagens RBD em diário para garantir a replicação consistente com o ponto no tempo e a falha entre os clusters.

### Com base em instantâneo

Esse modo usa instantâneos de espelho de imagens RBD programados periodicamente ou criados manualmente para replicar imagens RBD consistentes com a falha entre os clusters.

O espelhamento é configurado por pool nos clusters de peer e pode ser configurado em um subconjunto específico de imagens no pool ou configurado para espelhar automaticamente todas as imagens em um pool ao usar apenas o espelhamento com base em diário.

O espelhamento é configurado usando o comando **rbd**, que é instalado por padrão no SUSE Enterprise Storage 7.1. O daemon **rbd-mirror** é responsável por capturar as atualizações da imagem do cluster de peer remoto e aplicá-las à imagem no cluster local. Consulte a [Seção 6.6.2, “Habilitando o daemon rbd-mirror”](#) para obter mais informações sobre como habilitar o daemon **rbd-mirror**.

Dependendo da necessidade de replicação, o espelhamento de Dispositivo de Blocos RADOS pode ser configurado para replicação unidirecional ou bidirecional:

#### Replicação unidirecional

Quando os dados são espelhados apenas de um cluster principal para um cluster secundário, o daemon `rbd-mirror` é executado somente no cluster secundário.

#### Replicação bidirecional

Quando os dados são espelhados de imagens principais em um cluster para imagens não principais em outro cluster (e vice-versa), o daemon `rbd-mirror` é executado nos dois clusters.



### Importante

Cada instância do daemon `rbd-mirror` deve ser capaz de se conectar aos clusters do Ceph local e remoto simultaneamente, por exemplo, todos os hosts de monitor e OSD. Além disso, a rede deve ter largura de banda suficiente entre os dois data centers para processar a carga de trabalho de espelhamento.



### Dica: Informações gerais

Para obter informações gerais e a abordagem de linha de comando para espelhamento do Dispositivo de Blocos RADOS, consulte a [Seção 20.4, “Espelhos de imagens RBD”](#).

## 6.6.1 Configurando clusters principais e secundários

O cluster *principal* é onde o pool original com as imagens é criado. O cluster *secundário* é onde o pool ou as imagens são replicados do cluster *principal*.



### Nota: Nomeação relativa

Os termos *principal* e *secundário* podem ser relativos no contexto de replicação porque estão mais relacionados aos pools individuais do que aos clusters. Por exemplo, na replicação de duas vias, é possível espelhar um pool do cluster *principal* para o *secundário*. É possível também espelhar outro pool do cluster *secundário* para o *principal*.

## 6.6.2 Habilitando o daemon rbd-mirror

Os procedimentos a seguir demonstram como executar as tarefas administrativas básicas para configurar o espelhamento usando o comando `rbd`. O espelhamento é configurado por pool nos clusters do Ceph.

As etapas de configuração do pool devem ser executadas nos dois clusters de peer. Estes procedimentos consideram a existência de dois clusters, chamados “principal” e “secundário”, acessíveis de um único host, por motivos de clareza.

O daemon `rbd-mirror` executa a replicação de dados do cluster real.

1. Renomeie os arquivos `ceph.conf` e de chaveiro e copie-os do host principal para o host secundário:

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

2. Para habilitar o espelhamento em um pool com `rbd`, especifique `mirror pool enable`, o nome do pool e o modo de espelhamento:

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



### Nota

O modo de espelhamento pode ser `image` ou `pool`. Por exemplo:

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

3. No Ceph Dashboard, navegue até *Bloco > Espelhamento*. A tabela *Daemons* à esquerda mostra os daemons `rbd-mirror` em execução ativa e a saúde deles.



## Daemons







  10  <input type="text"/> 				
Instância ▾	ID 	Nome de host ▾	Versão ▾	Saúde ▾
292255	test	doc-ses-min4	14.2.2-354-g8878cf2360	
1 total				

FIGURA 6.6: EXECUTANDO DAEMONS `rbd-mirror`

### 6.6.3 Desabilitando o espelhamento

Para desabilitar o espelhamento em um pool com `rbd`, especifique o comando `mirror pool disable` e o nome do pool:

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

Quando o espelhamento é desabilitado dessa maneira em um pool, ele também é desabilitado em todas as imagens (no pool) para as quais ele foi explicitamente habilitado.

### 6.6.4 Inicializando peers

Para que `rbd-mirror` descubra seu cluster de peer, o peer precisa ser registrado no pool, e uma conta do usuário precisa ser criada. Esse processo pode ser automatizado com o `rbd` e usando os comandos `mirror pool peer bootstrap create` e `mirror pool peer bootstrap import`.

Para criar manualmente um novo token de boot com o `rbd`, especifique o comando `mirror pool peer bootstrap create`, o nome de um pool, junto com o nome de um site opcional, para descrever o cluster local:

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

A saída do `mirror pool peer bootstrap create` será um token que deve ser inserido no comando `mirror pool peer bootstrap import`. Por exemplo, no cluster principal:

```
cephuser@adm > rbd --cluster primary mirror pool peer bootstrap create --site-name
primary
image-pool
eyJmc2lkIjoioWY1MjgyZGI0Yjg5S0S0NTk2LTgwOTgtMzIwYzFmYzYzM5NmYzIiwiaWY2xpZW50X2lkIjoicmJkL \
```

```
WlpcnJvcilwZWVyIiwia2V5IjoiQVFBUnczOWQwdkhvQmhBQVlMM1I4RmR5dHNJQU50bkFTZ0l0TVE9PSIsIm1vbl9ob3N0I
\
joiW3Yy0jE5Mi4xNjguMS4z0jY4MjAsdjE6MTkyLjE2OC4xLjM6NjgyMV0ifQ==
```

Para importar manualmente o token de boot criado por outro cluster com o comando **rbd**, especifique o comando **mirror pool peer bootstrap import**, o nome do pool, um caminho de arquivo para o token criado (ou '-' para ler a entrada padrão), juntamente com um nome de site opcional para descrever o cluster local e uma direção de espelhamento (o padrão é **rx-tx** para espelhamento bidirecional, mas também pode ser definido como **rx-only** para espelhamento unidirecional):

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

Por exemplo, no cluster secundário:

```
cephuser@adm > cat >>EOF < token
eyJmc2lkIjoiOWY1MjgyZGItYjg5OS00NTk2LTgwOTgtMzIwYzFmYzY5MjYzIiwia2V5IjoiQVFBUnczOWQwdkhvQmhBQVlMM1I4RmR5dHNJQU50bkFTZ0l0TVE9PSIsIm1vbl9ob3N0I
\
JvcilwZWVyIiwia2V5IjoiQVFBUnczOWQwdkhvQmhBQVlMM1I4RmR5dHNJQU50bkFTZ0l0TVE9PSIsIm1vbl9ob3N0I
\
joiW3Yy0jE5Mi4xNjguMS4z0jY4MjAsdjE6MTkyLjE2OC4xLjM6NjgyMV0ifQ==
EOF
cephuser@adm > rbd --cluster secondary mirror pool peer bootstrap import --site-name
secondary image-pool token
```

## 6.6.5 Removendo o peer do cluster

Para remover um cluster do Ceph de peer de espelhamento, com o comando **rbd**, especifique o comando **mirror pool peer remove**, o nome do pool e o UUID do peer (disponível ao executar o comando **rbd mirror pool info**):

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

## 6.6.6 Configurando a replicação de pool no Ceph Dashboard

O daemon **rbd-mirror** precisa ter acesso ao cluster principal para poder espelhar imagens RBD. Verifique se você seguiu as etapas na [Seção 6.6.4, “Inicializando peers”](#) antes de continuar.

1. Em ambos os clusters *principal* e *secundário*, crie pools com nomes idênticos e atribua o aplicativo rbd a eles. Consulte a [Seção 5.1, “Adicionando um novo pool”](#) para obter mais detalhes sobre a criação de um novo pool.

Criar Pool

Nome \*

mirrored-pool

✓

Tipo de pool \*

replicated

✓ ⬆ ⬇ ⬆

PG Autoscale

off

✓ ⬆ ⬇ ⬆

Grupos de posicionamento \*

4

✓

Ajuda no cálculo

Tamanho replicado \*

3

Aplicativos

✎

rbd

✕

CRUSH

Conjunto de regras do Crush

replicated\_rule

⬆

?

+

🗑

Compactação

Modo

none

⬆ ⬇ ⬆

Quotas

Max bytes ?

ex. 10 GiB

Max objects ?

0

Configuração de RBD

Qualidade do Serviço +

Criar Pool

Cancelar

2. Em ambos os painéis de controle dos clusters *principal* e *secundário*, navegue até *Bloco > Espelhamento*. Na tabela *Pools* à direita, clique no nome do pool que será replicado e, após clicar em *Modo de Edição*, selecione o modo de replicação. Neste exemplo, trabalharemos com um modo de replicação de *pool*, o que significa que todas as imagens em um determinado pool serão replicadas. Clique em *Atualizar* para confirmar.



Editar modo de espelho do pool

Para editar o modo de espelho do pool `mirrored-pool`, selecione o novo modo na lista e clique em **Atualizar**.

Modo

Pool

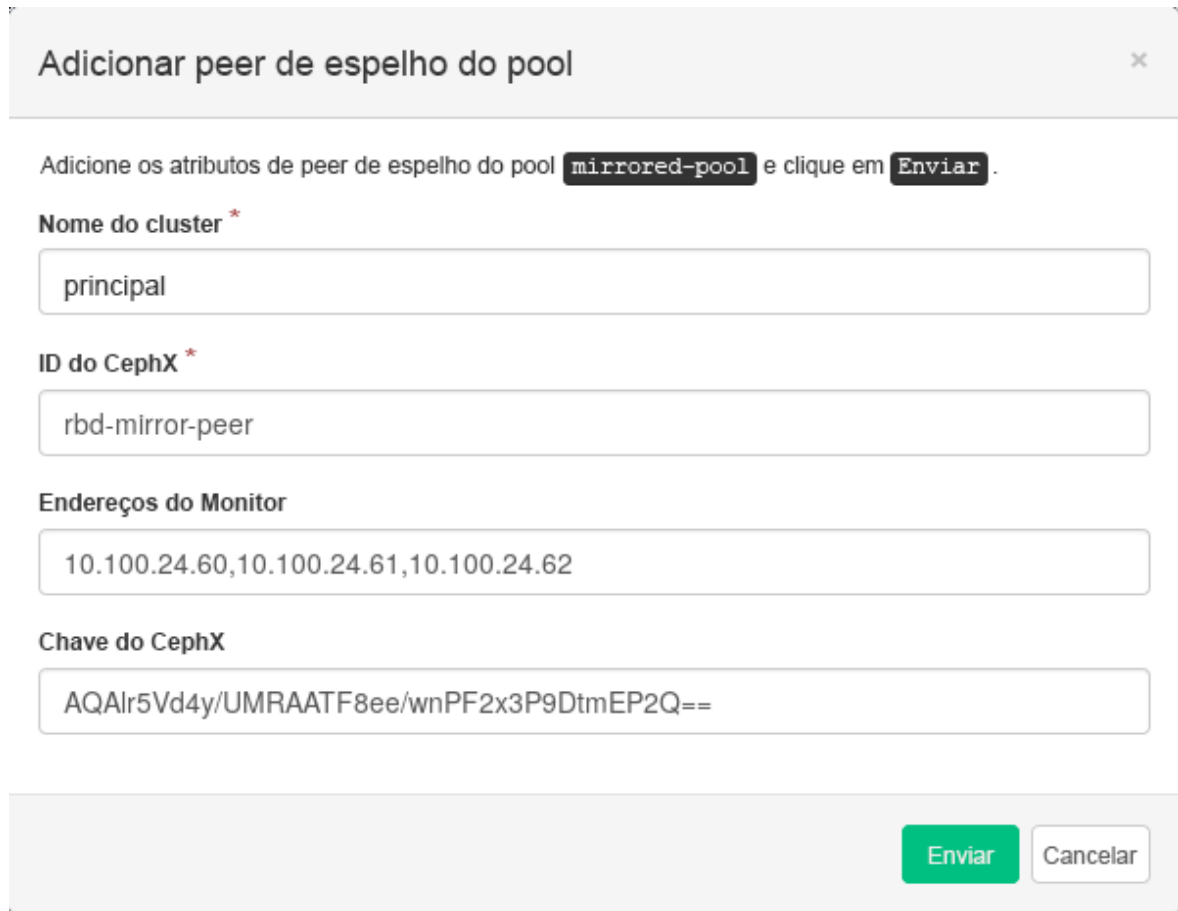
Atualizar Cancelar

FIGURA 6.8: CONFIGURANDO O MODO DE REPLICAÇÃO

### ! Importante: Erro ou aviso sobre o cluster principal

Após a atualização do modo de replicação, um flag de erro ou aviso aparecerá na coluna direita correspondente. Isso acontece porque o pool ainda não tem um usuário peer atribuído para replicação. Ignore esse flag para o cluster *principal* já que atribuímos um usuário peer apenas ao cluster *secundário*.

3. No Painel de Controle do cluster *secundário*, navegue até *Bloco > Espelhamento*. Adicione o peer do espelho do pool selecionando *Adicionar Peer*. Insira os detalhes do cluster *principal*:



**Adicionar peer de espelho do pool** ✕

Adicione os atributos de peer de espelho do pool `mirrored-pool` e clique em **Enviar**.

**Nome do cluster \***

principal

**ID do CephX \***

rbd-mirror-peer

**Endereços do Monitor**

10.100.24.60,10.100.24.61,10.100.24.62

**Chave do CephX**

AQAlr5Vd4y/UMRAATF8ee/wnPF2x3P9DtmEP2Q==

**Enviar** **Cancelar**

FIGURA 6.9: ADICIONANDO CREDENCIAIS DO PEER

#### **Cluster Name**

Uma string exclusiva arbitrária que identifica o cluster principal, como “primary”. O nome do cluster precisa ser diferente do nome do cluster secundário real.

#### **ID do CephX**

O ID de usuário do Ceph que você criou como um peer de espelhamento. Neste exemplo, ele é o “rbd-mirror-peer”.

#### **Endereços do Monitor**

Lista de endereços IP separados por vírgula dos nós do Ceph Monitor do cluster principal.

### Chave do CephX

A chave relacionada ao ID de usuário peer. Você pode recuperá-la executando o seguinte comando de exemplo no cluster principal:

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

Clique em *Enviar* para confirmar.

### Pools

Modo de Edição					
10					
Nome	Modo	Leader	# Local	# Remoto	Saúde
example_rbd_pool	pool	292255	2	2	OK
mirrored-pool	pool	292255	0	0	OK
pool3	imagem	292255	2	2	OK
pool4	pool	292255	1	1	OK
1 selecionado/4 total					

FIGURA 6.10: LISTA DE POOLS REPLICADOS

## 6.6.7 Verificando se a replicação de imagens RBD funciona

Quando o daemon `rbd-mirror` está em execução e a replicação de imagens RBD foi configurada no Ceph Dashboard, é hora de verificar se a replicação está realmente funcionando:

1. No Ceph Dashboard do cluster *principal*, crie uma imagem RBD de modo que seu pool pai seja o pool que você já criou para fins de replicação. Habilite os recursos Bloqueio exclusivo e Registro em diário para a imagem. Consulte a [Seção 6.3, “Criando RBDs”](#) para obter detalhes sobre como criar imagens RBD.

Criar RBD

Nome \*

mirrored-image1

Pool \*

mirrored-pool

☐ Usar pool de dados dedicado

Tamanho \*

60 GiB

Recursos

☐ Nivelamento profundo

☒ Disposição em camadas

☒ Bloqueio exclusivo

☐ Mapa de objetos (requer bloqueio exclusivo)

☒ Registro em diário (requer bloqueio exclusivo)

☐ Comparação rápida (requer mapa de objetos)

Avançado ...

Criar RBD

Cancelar





3. No cluster *principal*, grave os dados na imagem RBD. No Ceph Dashboard do cluster *secundário*, navegue até *Bloco > Imagens* e monitore se o tamanho da imagem correspondente aumenta à medida que os dados no cluster principal são gravados.

## 6.7 Gerenciando iSCSI Gateways



### Dica: Mais informações sobre gateways iSCSI

Para obter mais informações gerais sobre os Gateways iSCSI, consulte o [Capítulo 22, Ceph iSCSI Gateway](#).

Para listar todos os gateways e imagens mapeadas disponíveis, clique em *Bloco > iSCSI* no menu principal. A guia *Visão geral* é aberta com a lista atual dos iSCSI Gateways configurados e das imagens RBD mapeadas.

A tabela *Gateways* lista o estado de cada gateway, o número de destinos iSCSI e a quantidade de sessões. A tabela *Imagens* lista o nome de cada imagem mapeada, o tipo de backstore do nome do pool relacionado e outros detalhes estatísticos.

A guia *Destinos* lista os destinos iSCSI configurados no momento.

Destino	Portais	Imagens	# Sessions
> iqn.2001-07.com.ceph:1619785904397	master.ses7-mini.test:10.20.165.200	rbd/example_rbd_device_potato	0
> iqn.2001-07.com.ceph:1619785974221	master.ses7-mini.test:192.168.121.185	rbd/potato-rbd	0

0 selecionado(s) / 2 total

FIGURA 6.14: LISTA DE DESTINOS iSCSI

Para ver informações mais detalhadas sobre um destino, clique na seta suspensa na linha da tabela de destinos. Um esquema estruturado em árvore é aberto com uma lista de discos, portais, iniciadores e grupos. Clique em um item para expandi-lo e ver seu conteúdo detalhado, opcionalmente com uma configuração relacionada na tabela à direita.

Visão geral
Destinos

+ Criar
Autenticação de descoberta

10

Q

Destino	Portais	Imagens	# Sessions
iqn.2001-07.com.ceph:1597683071527	node1.asettle-dashboards.test:10.20.164.201	rbid/example_rbd_device_potato	0

Topologia iSCSI

iqn.2001-07.com.ceph:1597683071527

Disks

rbid/example\_rbd\_device\_potato

Portais

node1.asettle-dashboards.test:10.20.164.201

Initiators

Groups

rbid/example\_rbd\_device\_potato

0

Q

Nome	Atual	Padrão
backstore	usuário:rbid (tcmu-runner)	rbid
hw_max_sectors	1024	1024
lun	0	
max_data_area_mb	8	8
osd_op_timeout	30	30
qfull_timeout	5	5
wwn	bf60abfd-9159-4098-bc9b-2be4daaefa5c	
7 total		

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

1 selecionado(s) / 2 total

FIGURA 6.15: DETALHES DO DESTINO iSCSI

## 6.7.1 Adicionando destinos iSCSI

Para adicionar um novo destino iSCSI, clique em *Criar* na parte superior esquerda da tabela *Destinos* e insira as informações necessárias.

49

Adicionando destinos iSCSI | SES 7.1

Criar Target

IQN de Destino \*

iqn.2001-07.com.ceph:1620215063543

Portais \*

master.ses7-mini.test:10.20.165.200

+ Adicionar portal

Imagens

rbd/potato\_rbd

lun: 0

Backstore: rbd.

+ Adicionar imagem

☐ Autenticação ACL

Usuário

Senha

Usuário Mútuo

Senha Mútua

Criar Target

Cancelar

FIGURA 6.16: ADICIONANDO UM NOVO DESTINO

1. Digite o endereço de destino do novo gateway.
2. Clique em *Adicionar portal* e selecione um ou vários portais iSCSI na lista.
3. Clique em *Adicionar imagem* e selecione uma ou várias imagens RBD para o gateway.

4. Se você precisa usar a autenticação para acessar o gateway, ative a caixa de seleção *Autenticação ACL* e insira as credenciais. Você poderá encontrar opções de autenticação mais avançadas após ativar *Autenticação mútua* e *Autenticação de descoberta*.
5. Clique em *Criar Destino* para confirmar.

## 6.7.2 Editando destinos iSCSI

Para editar um destino iSCSI existente, clique na linha dele na tabela *Destinos* e clique em *Editar* na parte superior esquerda da tabela.

Em seguida, você pode modificar o destino iSCSI, adicionar ou apagar portais e adicionar ou apagar imagens RBD relacionadas. Você também pode ajustar as informações de autenticação para o gateway.

## 6.7.3 Apagando destinos iSCSI

Para apagar um destino iSCSI, selecione a linha da tabela e clique na seta suspensa ao lado do botão *Editar* e selecione *Excluir*. Ative *Sim, desejo* e clique em *Excluir destino iSCSI* para confirmar.

## 6.8 Qualidade do Serviço (QoS) do RBD



### Dica: Para obter mais informações

Para obter mais informações gerais e uma descrição das opções de configuração de QoS do RBD, consulte a [Seção 20.6, “Configurações de QoS”](#).

É possível configurar as opções de QoS em diferentes níveis.

- Globalmente
- Por pool
- Por imagem

A configuração *global* fica na parte superior da lista e será usada para todas as imagens RBD recém-criadas e para as imagens que não anulam esses valores no pool nem na camada da imagem RBD. Um valor de opção especificado globalmente pode ser anulado por pool ou por

imagem. As opções especificadas em um pool serão aplicadas a todas as imagens RBD desse pool, exceto se forem anuladas por uma opção de configuração definida em uma imagem. As opções especificadas em uma imagem anularão as opções especificadas em um pool e globalmente.

Dessa forma, é possível definir padrões globalmente, adaptá-los a todas as imagens RBD de um pool específico e anular a configuração do pool para imagens RBD individuais.

### 6.8.1 Configurando opções globalmente

Para configurar as opções do Dispositivo de Blocos RADOS globalmente, selecione *Cluster > Configuração* no menu principal.

1. Para listar todas as opções de configuração global disponíveis, ao lado de *Nível*, escolha *Avançado* no menu suspenso.
2. Em seguida, filtre os resultados da tabela por rbd\_qos no campo de pesquisa. Esse procedimento lista todas as opções de configuração disponíveis para QoS.
3. Para mudar um valor, clique na linha da tabela e selecione *Editar* na parte superior esquerda da tabela. A caixa de diálogo *Editar* inclui seus campos diferentes para especificar valores. Os valores da opção de configuração do RBD são obrigatórios na caixa de texto *mgr*.



#### Nota

Ao contrário das outras caixas de diálogo, esta não permite especificar o valor em unidades práticas. É necessário definir esses valores em bytes ou IOPS, dependendo da opção que você está editando.

### 6.8.2 Configurando opções em um novo pool

Para criar um novo pool e definir opções de configuração do RBD nele, clique em *Pools > Criar*. Selecione *replicado* como tipo de pool. Em seguida, você precisa adicionar a tag do aplicativo rbd ao pool para poder configurar as opções de QoS do RBD.



## Nota

Não é possível definir opções de configuração de QoS do RBD em um pool codificado para eliminação. Para definir opções de QoS do RBD para pools codificados para eliminação, você precisa editar o pool replicado de metadados de uma imagem RBD. Na sequência, a configuração será aplicada ao pool de dados codificado para eliminação dessa imagem.

### 6.8.3 Configurando opções em pool existente

Para definir opções de QoS do RBD em um pool existente, clique em *Pools*, clique na linha do pool na tabela e selecione *Editar* na parte superior esquerda da tabela.

Você deve ver a seção *Configuração do RBD* na caixa de diálogo, seguida de uma seção *Qualidade do Serviço*.



## Nota

Se a seção *Configuração do RBD* ou *Qualidade do Serviço* não for exibida, provavelmente você estará editando um pool *codificado para eliminação*, que não pode ser usado para definir opções de configuração do RBD, ou o pool não foi configurado para uso por imagens RBD. Neste último caso, atribua a tag do aplicativo *rbd* ao pool, e as seções de configuração correspondentes serão exibidas.

### 6.8.4 Opções de configuração

Clique em *Qualidade do Serviço* + para expandir as opções de configuração. Uma lista de todas as opções disponíveis será exibida. As unidades das opções de configuração já aparecem nas caixas de texto. No caso de qualquer opção de bytes por segundo (BPS, bytes per second), use atalhos como “1M” ou “5G”. Eles serão automaticamente convertidos em “1 MB/s” e “5 GB/s”, respectivamente.

Ao clicar no botão de redefinição à direita de cada caixa de texto, qualquer valor definido no pool será removido. Isso não remove os valores de configuração das opções definidas globalmente ou em uma imagem RBD.

### 6.8.5 Criando opções de QoS do RBD com uma nova imagem RBD

Para criar uma imagem RBD com opções de QoS do RBD definidas nela, selecione *Bloco > Imagens* e clique em *Criar*. Clique em *Avançado...* para expandir a seção de configuração avançada. Clique em *Qualidade do Serviço* + para abrir todas as opções de configuração disponíveis.

### 6.8.6 Editando opções de QoS do RBD em imagens existentes

Para editar as opções de QoS do RBD em uma imagem existente, selecione *Bloco > Imagens*, clique na linha do pool na tabela e, por último, clique em *Editar*. A caixa de diálogo de edição será exibida. Clique em *Avançado...* para expandir a seção de configuração avançada. Clique em *Qualidade do Serviço* + para abrir todas as opções de configuração disponíveis.

### 6.8.7 Mudando as opções de configuração ao copiar ou clonar imagens

Se uma imagem RBD for clonada ou copiada, os valores definidos nela também serão copiados por padrão. Para mudá-los durante a cópia ou clonagem, você pode especificar os valores de configuração atualizados na caixa de diálogo de cópia/clonagem da mesma forma que na criação ou edição de uma imagem RBD. Esse procedimento apenas definirá (ou redefinirá) os valores para a imagem RBD que é copiada ou clonada. Essa operação não muda a configuração da imagem RBD de origem nem a configuração global.

Se você redefinir o valor da opção durante a cópia/clonagem, nenhum valor para essa opção será definido nessa imagem. Isso significa que qualquer valor dessa opção especificado para o pool pai será usado se o pool pai tiver o valor configurado. Do contrário, o padrão global será usado.



## 7 Gerenciar o NFS Ganesha



### Importante

O NFS Ganesha suporta o NFS versão 4.1 e mais recente. Ele não suporta o NFS versão 3.



### Dica: Mais informações sobre o NFS Ganesha

Para obter mais informações gerais sobre o NFS Ganesha, consulte o [Capítulo 25, NFS Ganesha](#).

Para listar todas as exportações NFS disponíveis, clique em *NFS* no menu principal.

A lista mostra o diretório de cada exportação, o nome de host do daemon, o tipo de back end de armazenamento e o tipo de acesso.

<div>+ Criar</div>						
<div><div></div><div>10</div><div>Q</div><div>X</div></div>						
	Caminho	Pseudo	Cluster	Daemons	Backend de Armazenamento	Tipo de Acesso
>	/potato/potato	/exportimus-maximus	ganesha-sesdev_nfs		CephFS	MDONLY_RO
>	/root	/exportcephfs	ganesha-sesdev_nfs		CephFS	RW
>	/root/potato	/exportpotato	ganesha-sesdev_nfs		CephFS	MDONLY
0 selecionado/3 total						

FIGURA 7.1: LISTA DE EXPORTAÇÕES DO NFS

Para ver informações mais detalhadas sobre uma exportação do NFS, clique na linha dela na tabela.

Detalhes		Clientes (0)	
Tipo de Acesso	RW		
Sistema de Arquivos CephFS	sesdev_fs		
Usuário do CephFS	admin		
Cluster	ganesha-sesdev_nfs		
Daemons			
Protocolo NFS	NFSv3, NFSv4		
Caminho	/root		
Pseudo	/exportcephfs		
Squash	no_root_squash		
Backend de Armazenamento	CephFS		
Transporte	TCP, UDP		

FIGURA 7.2: DETALHES DA EXPORTAÇÃO DO NFS

## 7.1 Criando exportações do NFS

Para adicionar uma nova exportação do NFS, clique em *Criar* na parte superior esquerda da tabela de exportações e insira as informações necessárias.

Criar exportação NFS

Cluster \*

ganesha-sesdev\_nfs

Daemons

Nenhum item selecionado.

+ Adicionar daemon

Backend de Armazenamento \*

CephFS

✓

ID de Usuário do CephFS \*

admin

✓

Nome do CephFS \*

sesdev\_fs

✓

Rótulo de Segurança

☐ Habilitar rótulo de segurança

Caminho do CephFS \*

/root

✓

Novo diretório será criado

Protocolo NFS \*

☒ NFSv3
 ☒ NFSv4

Tag NFS ?

Pseudo \* ?

/exportcephfs

✓

Tipo de Acesso \*

RW

✓

Permite todas as operações

Squash \*

no\_root\_squash

✓

Protocolo de Transporte \*

☒ UDP
 ☒ TCP

Clientes

Qualquer cliente pode acessar

+ Adicionar clientes

Criar exportação NFS

Cancelar

FIGURA 7.3: ADICIONANDO UMA NOVA EXPORTAÇÃO DO NFS

1. Selecione um ou mais daemons NFS Ganesha que executarão a exportação.
2. Selecione um back end de armazenamento.



## Importante

No momento, apenas as exportações do NFS com suporte do CephFS são permitidas.

3. Selecione um ID de usuário e outras opções relacionadas ao back end.
4. Digite o caminho do diretório para a exportação NFS. Se o diretório não existir no servidor, ele será criado.
5. Especifique outras opções relacionadas ao NFS, como versão do protocolo NFS suportada, pseudo, tipo de acesso, squash ou protocolo de transporte.
6. Se você precisa limitar o acesso apenas a determinados clientes, clique em *Adicionar clientes* e adicione os endereços IP deles juntamente com o tipo de acesso e as opções de squash.
7. Clique em *Criar exportação do NFS* para confirmar.

## 7.2 Apagando exportações do NFS

Para apagar uma exportação, selecione-a e realce-a na linha da tabela. Clique na seta suspensa ao lado do botão *Editar* e selecione *Excluir*. Ative a caixa de seleção *Sim, desejo* e clique em *Excluir exportação do NFS* para confirmar.

## 7.3 Editando exportações do NFS

Para editar uma exportação existente, selecione-a e realce-a na linha da tabela e clique em *Editar* na parte superior esquerda da tabela de exportações.

Em seguida, você pode ajustar todos os detalhes da exportação NFS.

Editar exportação NFS

Cluster \*

ganesha-sesdev\_nfs

Daemons

Nenhum item selecionado.

+ Adicionar daemon

Backend de Armazenamento \*

CephFS

ID de Usuário do CephFS \*

admin

Nome do CephFS \*

sesdev\_fs

Rótulo de Segurança

☐ Habilitar rótulo de segurança

Caminho do CephFS \*

/root

Protocolo NFS \*

☒ NFSv3

☒ NFSv4

Tag NFS ?

Pseudo ?

/exportcephfs

Tipo de Acesso \*

RW

Permite todas as operações

Squash \*

no\_root\_squash

Protocolo de Transporte \*

☒ UDP

☒ TCP

Clientes

Qualquer cliente pode acessar

+ Adicionar clientes

Editar exportação NFS

Cancelar

FIGURA 7.4: EDITANDO UMA EXPORTAÇÃO DO NFS

## 8 Gerenciar o CephFS



### Dica: Para obter mais informações

Para encontrar informações detalhadas sobre o CephFS, consulte o [Capítulo 23, Sistema de arquivos em cluster](#).

### 8.1 Acessando a visão geral do CephFS

Clique em *Sistemas de arquivos* no menu principal para acessar a visão geral dos sistemas de arquivos configurados. A tabela principal mostra o nome e a data de criação de cada sistema de arquivos, e se ele está ou não habilitado.

Ao clicar na linha de um sistema de arquivos na tabela, você revela detalhes sobre sua classificação e os pools adicionados a ele.

Nome	Criado	Habilitado
sesdev_fs	30/04/2021 12:30:19	✓

Detalhes

Cientes 2

Directories

Detalhes de Desempenho

Posições

Posição	Estado	Daemon	Atividade	Dentries	Inodes
0	active	sesdev_fs.master.ingtr	Reqs: 0 /s	10	13
1 total					

Standbys

Daemons de standby
sesdev_fs.node3.jswzli

Pools

Pool	Tipo	Tamanho
cephfs.sesdev_fs. data	data	13.1 GiB
cephfs.sesdev_fs. metadata	metadata	13.1 GiB
2 total		

FIGURA 8.1: DETALHES DO CEPHFS

Na parte inferior da tela, você pode ver estatísticas que mostram o número de inodes MDS relacionados e solicitações de clientes, coletadas em tempo real.

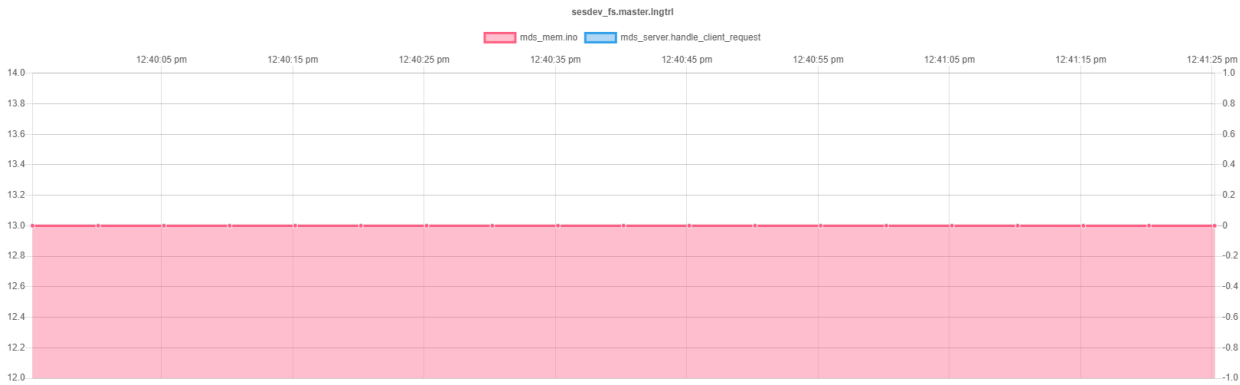


FIGURA 8.2: DETALHES DO CEPHFS

## 9 Gerenciar o Gateway de Objetos



### Importante

Antes de começar, talvez você veja a seguinte notificação ao tentar acessar o front end do Gateway de Objetos no Ceph Dashboard:

#### Information

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

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

Isso ocorre porque o Gateway de Objetos não foi configurado automaticamente pelo cephadm para o Ceph Dashboard. Se essa notificação aparecer, siga as instruções na [Seção 10.4, “Habilitando o front end de gerenciamento do Gateway de Objetos”](#) para habilitar manualmente o front end do Gateway de Objetos para o Ceph Dashboard.



### Dica: Mais informações sobre o Gateway de Objetos

Para obter mais informações gerais sobre o Gateway de Objetos, consulte o [Capítulo 21, Gateway de Objetos do Ceph](#).

## 9.1 Vendo Gateways de Objetos

Para ver uma lista de Gateways de Objetos configurados, clique em *Gateway de Objetos > Daemons*. A lista inclui o ID do gateway, o nome de host do nó do cluster onde o daemon do gateway está sendo executado e o número da versão do gateway.

Clique na seta suspensa ao lado do nome do gateway para ver informações detalhadas sobre ele. A guia *Contadores de Desempenho* mostra detalhes sobre as operações de leitura/gravação e as estatísticas do cache.



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

FIGURA 9.1: DETALHES DO GATEWAY

## 9.2 Gerenciando usuários do Gateway de Objetos

Clique em *Gateway de Objetos* > *Usuários* para ver uma lista dos usuários existentes do Gateway de Objetos.

Clique na seta suspensa ao lado do nome do usuário para ver detalhes sobre a conta do usuário, como informações de status ou os detalhes de cota do usuário e do compartimento de memória.

Detalhes	Chaves
Nome de usuário	rgw-admin
Nome completo	admin
Suspenso	Não
Sistema	Sim
Máximo de compartimentos de memória	1000
Cota do usuário	
Habilitado	Não
Tamanho máximo	-
Máximo de objetos	-
Cota de compartimento de memória	
Habilitado	Não
Tamanho máximo	-
Máximo de objetos	-

FIGURA 9.2: USUÁRIOS DO GATEWAY

### 9.2.1 Adicionando um novo usuário do gateway

Para adicionar um novo usuário do gateway, clique em *criar* na parte superior esquerda do cabeçalho da tabela. Preencha as credenciais dele, os detalhes sobre a chave S3 e as cotas do usuário e do compartimento de memória e clique em *Criar Usuário* para confirmar.

Criar Usuário

Nome de Usuário \*

example\_rgw\_user

✓

Nome completo \*

Exemplo de Usuário

✓

Endereço de e-mail

example@user.com

✓

Máx. com-  
partimentos  
de memória

Personalizado

✓ ↕

1000

☐ Suspenso

Chave S3

☒ Gerar chave automaticamente

Cota do usuário

☐ Habilitado

Cota de compartimento de memória

☒ Habilitado

☒ Tamanho ilimitado

☒ Objetos ilimitados

Criar Usuário

Cancelar

FIGURA 9.3: ADICIONANDO UM NOVO USUÁRIO DO GATEWAY

### 9.2.2 Apagando usuários do gateway

Para apagar um usuário do gateway, selecione-o e realce-o. Clique no botão suspenso ao lado de *Editar* e selecione *Excluir* na lista para apagar a conta do usuário. Ative a caixa de seleção *Sim, desejo* e clique em *Excluir usuário* para confirmar.

### 9.2.3 Editando detalhes do usuário do gateway

Para mudar os detalhes do usuário do gateway, selecione-o e realce-o. Clique em *Editar* na parte superior esquerda do cabeçalho da tabela.

Modifique as informações básicas ou adicionais do usuário, como informações de recursos, chaves, subusuários e cotas. Clique em *Editar Usuário* para confirmar.

A guia *Chaves* inclui uma lista apenas leitura de usuários do gateway e suas chaves de acesso e secretas. Para ver as chaves, clique em um nome de usuário na lista e selecione *Mostrar* na parte superior esquerda do cabeçalho da tabela. Na caixa de diálogo *Chave S3*, clique no ícone de “olho” para exibir as chaves ou clique no ícone da área de transferência para copiar a chave relacionada para a área de transferência.

## 9.3 Gerenciando compartimentos de memória do Gateway de Objetos

Os compartimentos de memória do Gateway de Objetos (OGW) implementam a funcionalidade dos containers OpenStack Swift. Os compartimentos de memória do Gateway de Objetos servem como containers para armazenar objetos de dados.

Clique em *Gateway de Objetos > Compartimentos* para ver uma lista de compartimentos de memória do Gateway de Objetos.

### 9.3.1 Adicionando um novo compartimento de memória

Para adicionar um novo compartimento de memória do Gateway de Objetos, clique em *Criar* na parte superior esquerda do cabeçalho da tabela. Insira o nome do compartimento de memória, selecione o proprietário e defina o destino de posicionamento. Clique em *Criar Compartimento* para confirmar.



## Nota

Nesta fase, você também pode habilitar o bloqueio selecionando *Habilitado*; no entanto, esse recurso poderá ser configurado após a criação. Consulte a [Seção 9.3.3, “Editando o compartimento de memória”](#) para obter mais informações.

### 9.3.2 Visualizando detalhes do compartimento de memória

Para ver informações detalhadas sobre um compartimento de memória do Gateway de Objetos, clique na seta suspensa ao lado do nome do compartimento de memória.

Detalhes	
Nome	export
ID	2a664005-94ad-432a-b873-d563fed68496.14523.1
Proprietário	rgw-admin
Tipo de índice	Normal
Regra de posicionamento	default-placement
Marcador	2a664005-94ad-432a-b873-d563fed68496.14523.1
Marcador máximo	0#1#,2#,3#,4#,5#,6#,7#,8#,9#,10#
Versão	0#1,1#1,2#1,3#1,4#1,5#1,6#1,7#1,8#1,9#1,10#1
Versão master	0#0,1#0,2#0,3#0,4#0,5#0,6#0,7#0,8#0,9#0,10#0
Tempo de modificação	24/08/20 13:24:34
Grupo de zonas	cc4ec3c6-c611-4bfd-a155-e3e05552d5cd
Controle de versão	Suspense
Apagar MFA	Desabilitado
Cota de compartimento de memória	
Habilitado	Não
Tamanho máximo	Ilimitado
Máximo de objetos	Ilimitado
Bloqueio	
Habilitado	Não

FIGURA 9.4: DETALHES DO COMPARTIMENTO DE MEMÓRIA DO GATEWAY



## Dica: Cota de compartimento de memória

Abaixo da tabela *Detalhes*, você encontra detalhes sobre as configurações de cota e bloqueio de compartimento de memória.

### 9.3.3 Editando o compartimento de memória

Selecione e realce um compartimento de memória e clique em *Editar* na parte superior esquerda do cabeçalho da tabela.

Você pode atualizar o proprietário do compartimento de memória ou habilitar o controle de versão, a autenticação multifator ou o bloqueio. Clique em *Editar Compartimento* para confirmar qualquer mudança.

Editar Compartimento de Memória

Id

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

Nome

root

Proprietário \*

asettle

✓

Destino de posicionamento

default-placement

Controle de versão

☐ Habilitado ?

Autenticação Multifator

☐ Apagar habilitado ?

Bloqueio

☐ Habilitado ?

Editar Compartimento de Memória

Cancelar

FIGURA 9.5: EDITANDO OS DETALHES DO COMPARTIMENTO DE MEMÓRIA

### 9.3.4 Apagando um compartimento de memória

Selecione e realce um compartimento de memória para apagá-lo do Gateway de Objetos. Clique no botão suspenso ao lado de *Editar* e selecione *Excluir* na lista para apagar o compartimento de memória. Ative a caixa de seleção *Sim, desejo* e clique em *Excluir compartimento* para confirmar.

## 10 Configuração manual

Esta seção apresenta informações avançadas para usuários que preferem definir as configurações do painel de controle manualmente na linha de comando.

### 10.1 Configurando o suporte a TLS/SSL

Por padrão, todas as conexões HTTP com o painel de controle são protegidas com TLS/SSL. Uma conexão segura requer um certificado SSL. Você pode usar um certificado autoassinado ou gerar um certificado assinado por uma autoridade de certificação (CA, certificate authority) reconhecida.



#### Dica: Desabilitação de SSL

Talvez você tenha algum motivo específico para desabilitar o suporte a SSL. Por exemplo, se o painel de controle for executado por meio de um proxy que não suporta SSL.

Tenha cuidado ao desabilitar o SSL, já que os **nomes de usuário e as senhas** serão enviados ao painel de controle **não criptografados**.

Para desabilitar o SSL, execute:

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



#### Dica: Reiniciando os processos do Ceph Manager

Você precisará reiniciar os processos do Ceph Manager manualmente após mudar o certificado SSL e a chave. Para fazer isso, é possível executar

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

ou desabilitar e habilitar novamente o módulo do painel de controle, o que também aciona o gerenciador para se reativar:

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



## 10.1.1 Criando certificados autoassinados

A criação de um certificado autoassinado para uma comunicação segura é simples. Dessa forma, você pode fazer com que o painel de controle seja executado rapidamente.



### Nota: Reclamação dos browsers da Web

A maioria dos browsers da Web reclamará de um certificado autoassinado e exigirá a confirmação explícita antes de estabelecer uma conexão segura com o painel de controle.

Para gerar e instalar um certificado autoassinado, use o seguinte comando incorporado:

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

## 10.1.2 Usando certificados assinados por CA

Para proteger apropriadamente a conexão com o painel de controle e eliminar as reclamações do browser da Web por causa de um certificado autoassinado, recomendamos o uso de um certificado assinado por uma CA.

Você pode gerar um par de chaves do certificado com um comando semelhante ao seguinte:

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

O comando acima resulta nos arquivos `dashboard.key` e `dashboard.crt`. Após receber o arquivo `dashboard.crt` assinado por uma CA, habilite-o para todas as instâncias do Ceph Manager executando os seguintes comandos:

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



### Dica: Certificados diferentes para cada instância do gerenciador

Se você precisar de certificados diferentes para cada instância do Ceph Manager, modifique os comandos e inclua o nome da instância da seguinte maneira. Substitua *NAME* pelo nome da instância do Ceph Manager (normalmente, o nome de host relacionado):

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

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

## 10.2 Mudando nome de host e número de porta

O Ceph Dashboard está vinculado a um endereço TCP/IP e uma porta TCP específicos. Por padrão, o Ceph Manager ativo no momento e que hospeda o painel de controle está vinculado à porta TCP 8443 (ou 8080 quando o SSL está desabilitado).



### Nota

Se um firewall estiver habilitado nos hosts que executam o Ceph Manager (e, portanto, o Ceph Dashboard), talvez seja necessário mudar a configuração para habilitar o acesso a essas portas. Para obter mais informações sobre as configurações de firewall do Ceph, consulte o Livro *“Troubleshooting Guide”, Capítulo 13 “Hints and tips”, Seção 13.7 “Firewall settings for Ceph”*.

Por padrão, o Ceph Dashboard está vinculado a “::”, que corresponde a todos os endereços IPv4 e IPv6 disponíveis. Você pode mudar o endereço IP e o número da porta do aplicativo Web para que eles se apliquem a todas as instâncias do Ceph Manager usando os seguintes comandos:

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



### Dica: Configurando as instâncias do Ceph Manager separadamente

Como cada daemon `ceph-mgr` hospeda sua própria instância do painel de controle, talvez seja necessário configurá-los separadamente. Mude o endereço IP e o número da porta para uma instância específica do gerenciador usando os seguintes comandos (substitua `NAME` pelo ID da instância do `ceph-mgr`):

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



## Dica: Listando endpoints configurados

O comando **ceph mgr services** exibe todos os endpoints que estão configurados. Procure a chave **dashboard** para obter o URL de acesso ao painel de controle.

## 10.3 Ajustando nomes de usuário e senhas

Se você não deseja usar a conta de administrador padrão, crie uma conta de usuário diferente e associe-a a pelo menos uma função. Oferecemos um conjunto de funções de sistema predefinidas que você pode usar. Para obter mais detalhes, consulte a [Capítulo 11, Gerenciar usuários e funções na linha de comando](#).

Para criar um usuário com privilégios de administrador, use o seguinte comando:

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

## 10.4 Habilitando o front end de gerenciamento do Gateway de Objetos

Para usar a funcionalidade de gerenciamento do Gateway de Objetos do painel de controle, você precisa fornecer as credenciais de login de um usuário com o flag **system** habilitado:

1. Se você não tem um usuário com o flag **system**, crie-o:

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

Anote as chaves **access\_key** e **secret\_key** na saída do comando.

2. Você também pode obter as credenciais de um usuário existente usando o comando **radosgw-admin**:

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

3. Insira as credenciais recebidas no painel de controle em arquivos separados:

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



## Nota

Por padrão, o firewall está habilitado no SUSE Linux Enterprise Server 15 SP3. Para obter informações sobre configuração de firewall, consulte o Livro *“Troubleshooting Guide”, Capítulo 13 “Hints and tips”, Seção 13.7 “Firewall settings for Ceph”*.

Há vários pontos a serem considerados:

- O nome de host e o número da porta do Gateway de Objetos são determinados automaticamente.
- Se várias zonas forem usadas, ele determinará automaticamente o host no grupo de zonas master e na zona master. Isso é o suficiente para a maioria das configurações. No entanto, em algumas circunstâncias, talvez você queira definir o nome de host e a porta manualmente:

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

- Veja a seguir outras configurações que você pode precisar:

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

- Se você usa um certificado autoassinado ([Seção 10.1, “Configurando o suporte a TLS/SSL”](#)) na configuração do Gateway de Objetos, desabilite a verificação de certificado no painel de controle para evitar conexões recusadas por causa de certificados assinados por uma CA desconhecida ou que não correspondem ao nome de host:

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

- Se o Gateway de Objetos levar muito tempo para processar as solicitações, e se a execução do painel de controle apresentar tempo de espera, o valor do tempo de espera poderá ser ajustado (o padrão é 45 segundos):

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

## 10.5 Habilitando o gerenciamento de iSCSI

O Ceph Dashboard gerencia destinos iSCSI usando a API REST fornecida pelo serviço `rbdtarget-api` do gateway Ceph iSCSI. Verifique se ela está instalada e habilitada nos gateways iSCSI.



### Nota

A funcionalidade de gerenciamento de iSCSI do Ceph Dashboard depende da versão 3 mais recente do projeto `ceph-iscsi`. Verifique se o seu sistema operacional tem a versão correta; do contrário, o Ceph Dashboard não habilitará os recursos de gerenciamento.

Se a API REST do `ceph-iscsi` estiver configurada no modo HTTPS e usar um certificado autoassinado, configure o painel de controle para evitar a verificação do certificado SSL ao acessar a API do `ceph-iscsi`.

Desabilite a verificação do SSL da API:

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

Defina os gateways iSCSI disponíveis:

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

## 10.6 Habilitando o login único

O *Login Único* (SSO, Single Sign-On) é um método de controle de acesso que permite aos usuários efetuar login com ID e senha exclusivos em vários aplicativos simultaneamente.

O Ceph Dashboard suporta autenticação externa de usuários por meio do protocolo SAML 2.0. Como a *autorização* ainda é realizada pelo painel de controle, você precisa primeiro criar as contas de usuário e associá-las às funções desejadas. No entanto, o processo de *autenticação* pode ser efetuado por um *Provedor de Identidade* (IdP, Identity Provider) existente.

Para configurar o Login Único, use o seguinte comando:

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

```
IDP_ENTITY_ID SP_X_509_CERT \
SP_PRIVATE_KEY
```

Parâmetros:

CEPH\_DASHBOARD\_BASE\_URL

URL de base para acessar o Ceph Dashboard (por exemplo, “https://cephdashboard.local”).

IDP\_METADATA

URL, caminho de arquivo ou conteúdo do XML de metadados do IdP (por exemplo, “https://myidp/metadata”).

IDP\_USERNAME\_ATTRIBUTE

Opcional. Atributo que será usado para obter o nome de usuário da resposta de autenticação. O padrão “uid” é usado.

IDP\_ENTITY\_ID

Opcional. Use quando há mais de um ID de entidade nos metadados do IdP.

SP\_X\_509\_CERT/SP\_PRIVATE\_KEY

Opcional. Caminho de arquivo ou conteúdo do certificado que será usado pelo Ceph Dashboard (Provedor de Serviços) para assinatura e criptografia. Esses caminhos de arquivos precisam estar acessíveis da instância ativa do Ceph Manager.



## Nota: Solicitações SAML

O valor do emissor das solicitações SAML seguirá este padrão:

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

Para exibir a configuração atual do SAML 2.0, execute:

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

Para desabilitar o Login Único, execute:

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

Para verificar se o SSO está habilitado, execute:

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

Para habilitar o SSO, execute:

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

# 11 Gerenciar usuários e funções na linha de comando

Esta seção descreve como gerenciar contas de usuário usadas pelo Ceph Dashboard. Ela ajuda você a criar ou modificar as contas de usuário e definir as funções e permissões de usuário apropriadas.

## 11.1 Gerenciando a política de senha

Por padrão, o recurso de política de senha está habilitado, incluindo as seguintes verificações:

- A senha tem mais de  $N$  caracteres?
- As senhas antiga e nova são iguais?

O recurso de política de senha pode ser ativado ou desativado completamente:

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

Cada uma destas verificações pode ser ativada ou desativada:

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

As seguintes opções também estão disponíveis para configurar o comportamento da política de senha.

- O tamanho mínimo da senha (padrão 8):

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

- A complexidade mínima da senha (padrão 10):

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



A complexidade da senha é calculada pela classificação de cada caractere na senha.

- Uma lista de palavras separadas por vírgula que não podem ser usadas em uma senha:

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

## 11.2 Gerenciando contas dos usuários

O Ceph Dashboard suporta o gerenciamento de várias contas de usuário. Cada conta de usuário consiste em nome de usuário, senha (armazenada em formato criptografado com bcrypt), além de nome e endereço de e-mail opcionais.

As contas dos usuários são armazenadas no banco de dados de configuração do Ceph Monitor e compartilhadas globalmente com todas as instâncias do Ceph Manager.

Use os seguintes comandos para gerenciar as contas de usuário:

**Mostrar usuários existentes:**

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

**Criar um novo usuário:**

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

**Apagar um usuário:**

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

**Mudar a senha de um usuário:**

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

**Modificar nome e e-mail de um usuário:**

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

**Desabilitar usuário**

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

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

### 11.3 Funções e permissões de usuário

Esta seção descreve quais escopos de segurança você pode atribuir a uma função de usuário, como gerenciar as funções de usuário e atribuí-las a contas de usuário.

#### 11.3.1 Definindo escopos de segurança

As contas de usuário são associadas a um conjunto de funções que definem quais partes do painel de controle podem ser acessadas pelo usuário. As partes do painel de controle são agrupadas dentro de um escopo de *segurança*. Os escopos de segurança são predefinidos e estáticos. Atualmente, os seguintes escopos de segurança estão disponíveis:

##### hosts

Inclui todos os recursos relacionados à entrada do menu *Hosts*.

##### config-opt

Inclui todos os recursos relacionados ao gerenciamento das opções de configuração do Ceph.

##### pool

Inclui todos os recursos relacionados ao gerenciamento de pools.

##### osd

Inclui todos os recursos relacionados ao gerenciamento do Ceph OSD.

##### monitor

Inclui todos os recursos relacionados ao gerenciamento do Ceph Monitor.

##### rbd-image

Inclui todos os recursos relacionados ao gerenciamento de imagens do Dispositivo de Blocos RADOS.

##### rbd-mirroring

Inclui todos os recursos relacionados ao gerenciamento de espelhamento do Dispositivo de Blocos RADOS.

#### iscsi

Inclui todos os recursos relacionados ao gerenciamento do iSCSI.

#### rgw

Inclui todos os recursos relacionados ao gerenciamento do Gateway de Objetos.

#### cephfs

Inclui todos os recursos relacionados ao gerenciamento do CephFS.

#### manager

Inclui todos os recursos relacionados ao gerenciamento do Ceph Manager.

#### registro

Inclui todos os recursos relacionados ao gerenciamento de registros do Ceph.

#### grafana

Inclui todos os recursos relacionados ao proxy do Grafana.

#### prometheus

Inclua todos os recursos relacionados ao gerenciamento de alertas do Prometheus.

#### dashboard-settings

Permite mudar as configurações do painel de controle.

### 11.3.2 Especificando funções de usuário

Uma *função* especifica um conjunto de mapeamentos entre um *escopo de segurança* e um conjunto de *permissões*. Há quatro tipos de permissões: “read”, “create”, “update” e “delete”.

O exemplo a seguir especifica uma função em que um usuário tem permissões de “leitura” e “criação” para recursos relacionados ao gerenciamento de pools e tem permissões completas para recursos relacionados ao gerenciamento de imagens RBD:

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

O painel de controle já dispõe de um conjunto de funções predefinidas que chamamos de *funções de sistema*. Você pode usá-las logo após uma instalação recente do Ceph Dashboard:

**administrator**

Concede permissões completas para todos os escopos de segurança.

**read-only**

Concede permissão de leitura para todos os escopos de segurança, exceto as configurações do painel de controle.

**block-manager**

Concede permissões completas para os escopos “rbd-image”, “rbd-mirroring” e “iscsi”.

**rgw-manager**

Concede permissões completas para o escopo “rgw”.

**cluster-manager**

Concede permissões completas para os escopos “hosts”, “osd”, “monitor”, “manager” e “config-opt”.

**pool-manager**

Concede permissões completas para o escopo “pool”.

**cephfs-manager**

Concede permissões completas para o escopo “cephfs”.

### 11.3.2.1 Gerenciando funções personalizadas

Você pode criar novas funções de usuário executando os seguintes comandos:

**Criar uma nova função:**

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

**Apagar uma função:**

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

**Adicionar permissões de escopo a uma função:**

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

Apagar permissões de escopo a uma função:

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

### 11.3.2.2 Atribuindo funções às contas dos usuários

Use os seguintes comandos para atribuir funções a usuários:

Definir funções de usuário:

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

Adicionar mais funções a um usuário:

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

Apagar funções de um usuário:

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



### Dica: Purgando funções personalizadas

Se você cria funções de usuário personalizadas e posteriormente pretende remover o cluster do Ceph com o executor **ceph.purge**, precisa purgar primeiro as funções personalizadas. Encontre mais detalhes na [Seção 13.9, “Removendo um cluster inteiro do Ceph”](#).

### 11.3.2.3 Exemplo: Criando um usuário e uma função personalizada

Esta seção ilustra um procedimento para criar uma conta de usuário capaz de gerenciar imagens RBD, ver e criar pools do Ceph e obter acesso apenas leitura a qualquer outro escopo.

1. Criar um novo usuário chamado tux:

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

2. Criar uma função e especificar permissões de escopo:

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

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

### 3. Associar as funções ao usuário tux:

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

## 11.4 Configuração de proxy

Para estabelecer um URL fixo para acessar o Ceph Dashboard, ou se você não deseja permitir conexões diretas com os nós do gerenciador, pode configurar um proxy que encaminhe automaticamente as solicitações recebidas para a instância ativa do ceph-mgr no momento.

### 11.4.1 Acessando o painel de controle com proxies reversos

Se você acessa o painel de controle por meio de uma configuração de proxy reverso, talvez seja necessário usar um prefixo de URL para acessá-lo. Para fazer com que o painel de controle use hiperlinks que incluam seu prefixo, você pode definir a configuração url\_prefix:

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

Em seguida, você pode acessar o painel de controle em [http://HOST\\_NAME:PORT\\_NUMBER/URL\\_PREFIX/](http://HOST_NAME:PORT_NUMBER/URL_PREFIX/).

### 11.4.2 Desabilitando redirecionamentos

Se o Ceph Dashboard estiver protegido por um proxy de equilíbrio de carga, como HAProxy, desabilite o comportamento de redirecionamento para evitar situações em que os URLs internos (não resolvidos) são publicados no cliente de front end. Use o seguinte comando para fazer com que o painel de controle responda com um erro HTTP (padrão 500), em vez de redirecionar para o painel de controle ativo:

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

Para redefinir a configuração para o comportamento de redirecionamento padrão, use o seguinte comando:

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

### 11.4.3 Configurando códigos de status de erro

Se o comportamento de redirecionamento estiver desabilitado, você deverá personalizar o código de status HTTP dos painéis de controle de standby. Para fazer isso, execute o seguinte comando:

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

### 11.4.4 Exemplo de configuração do HAProxy

O exemplo de configuração a seguir refere-se à transferência de TLS/SSL por meio do HAProxy.



#### Nota

A configuração funciona nas seguintes condições: Se o painel de controle falhar, o cliente de front end poderá receber uma resposta de redirecionamento HTTP (303) e será redirecionado para um host não resolvido.

Isso acontece quando o failover ocorre durante duas verificações de saúde do HAProxy. Nessa situação, o nó do painel de controle que antes estava ativo agora responderá com um código 303 que aponta para o novo nó ativo. Para evitar essa situação, você deve desabilitar o comportamento de redirecionamento nos nós de standby.

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

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

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

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

## 11.5 Fazendo auditoria das solicitações de API

A API REST do Ceph Dashboard pode gravar solicitações PUT, POST e DELETE no registro de auditoria do Ceph. Por padrão, o registro está desabilitado, mas você pode habilitá-lo com o seguinte comando:

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

Se habilitado, os seguintes parâmetros serão registrados por cada solicitação:

**from**

A origem da solicitação, por exemplo “https://[::1]:44410”.

**path**

O caminho da API REST, por exemplo /api/auth.

**method**

“PUT”, “POST” ou “DELETE”.

**user**

O nome do usuário (ou “Nenhum”).

Veja a seguir um exemplo de entrada de registro:

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





## Dica: Desabilitar o registro de payload de solicitação

Por padrão, o registro de payload de solicitação (a lista de argumentos e seus valores) está habilitado. Você pode desabilitá-lo da seguinte forma:

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

## 11.6 Configurando o NFS Ganesha no Ceph Dashboard

O Ceph Dashboard pode gerenciar exportações do NFS Ganesha que usam o CephFS ou o Gateway de Objetos como backstore. O painel de controle gerencia os arquivos de configuração do NFS Ganesha armazenados em objetos RADOS no cluster do CephFS. O NFS Ganesha deve armazenar parte da sua configuração no cluster do Ceph.

Execute o seguinte comando para configurar o local do objeto de configuração do NFS Ganesha:

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

Agora você pode gerenciar as exportações do NFS Ganesha usando o Ceph Dashboard.

### 11.6.1 Configurando vários clusters do NFS Ganesha

O Ceph Dashboard suporta o gerenciamento de exportações do NFS Ganesha pertencentes a diferentes clusters do NFS Ganesha. Recomendamos que cada cluster do NFS Ganesha armazene seus objetos de configuração em um pool/namespace RADOS diferente para isolar umas configurações das outras.

Use o seguinte comando para especificar os locais da configuração de cada cluster do NFS Ganesha:

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

O cluster\_id é uma string arbitrária que identifica exclusivamente o cluster do NFS Ganesha. Ao configurar o Ceph Dashboard com vários clusters do NFS Ganesha, a IU da Web permite escolher automaticamente a qual cluster uma exportação pertence.

## 11.7 Plug-ins de depuração

Os plug-ins do Ceph Dashboard estendem a funcionalidade do painel de controle. O plug-in de depuração permite a personalização do comportamento do painel de controle de acordo com o modo de depuração. Ele pode ser habilitado, desabilitado ou verificado com o seguinte comando:

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

Por padrão, esse modo está desabilitado. Essa é a configuração recomendada para implantações de produção. Se necessário, o modo de depuração pode ser habilitado sem necessidade de reinicialização.

## II Operação do cluster

- 12 Determinar o estado do cluster **90**
- 13 Tarefas operacionais **119**
- 14 Operação de serviços do Ceph **142**
- 15 Backup e restauração **147**
- 16 Monitoramento e alerta **150**

## 12 Determinar o estado do cluster

Quando você tem um cluster em execução, pode usar a ferramenta **ceph** para monitorá-lo. Normalmente, determinar o estado do cluster envolve verificar o status dos Ceph OSDs, Ceph Monitors, grupos de posicionamento e Servidores de Metadados.



### Dica: Modo interativo

Para executar a ferramenta **ceph** no modo interativo, digite **ceph** na linha de comando sem argumentos. O modo interativo é o mais prático quando você pretende digitar mais comandos **ceph** em uma linha. Por exemplo:

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

### 12.1 Verificando o status de um cluster

Você pode detectar o estado imediato do cluster usando **ceph status** ou **ceph -s**:

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

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

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

data:
```

```
pools: 7 pools, 169 pgs
objects: 250 objects, 10 KiB
usage: 3.1 GiB used, 27 GiB / 30 GiB avail
pgs: 169 active+clean
```

A saída apresenta as seguintes informações:

- ID do cluster
- Status de saúde do cluster
- A época do mapa do monitor e o status do quorum do monitor
- A época do mapa OSD e o status dos OSDs
- O status dos Ceph Managers
- O status dos Gateways de Objetos
- A versão do mapa do grupo de posicionamento
- O número de grupos de posicionamento e pools
- A quantidade *estimada* de dados armazenados e o número de objetos armazenados
- A quantidade total de dados armazenados.



### Dica: Como o Ceph calcula o uso de dados

O valor usado reflete o valor real do armazenamento bruto utilizado. O valor xxx GB/xxx GB indica o valor disponível (o menor número) da capacidade de armazenamento geral do cluster. O número estimado reflete o tamanho dos dados armazenados antes de serem replicados, clonados ou capturados como instantâneos. Portanto, a quantidade de dados realmente armazenados costuma exceder o valor estimado armazenado, pois o Ceph cria réplicas dos dados e também pode usar a capacidade de armazenamento para fazer clonagem e criar instantâneos.

Outros comandos que exibem informações de status imediatas são:

- ceph pg stat
- ceph osd pool stats

- `ceph df`
- `ceph df detail`

Para obter as informações atualizadas em tempo real, especifique qualquer um desses comandos (incluindo o `ceph -s`) como um argumento do comando `watch`:

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

Pressione `Ctrl-C` quando estiver cansado de observar.

## 12.2 Verificando a saúde do cluster

Após iniciar o cluster e antes de começar a leitura e/ou gravação de dados, verifique a saúde dele:

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



### Dica

Se você especificou locais diferentes do padrão em sua configuração ou no chaveiro, deve especificar estes locais:

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

O cluster do Ceph retorna um dos seguintes códigos de saúde:

#### OSD\_DOWN

Um ou mais OSDs estão marcados como inativos. O daemon OSD pode ter sido parado ou os OSDs peers talvez não conseguem acessar o OSD pela rede. As causas comuns incluem um daemon parado ou com falha, um host inativo ou uma interrupção da rede.

Verifique se o host está saudável, se o daemon foi iniciado e se a rede está funcionando. Se o daemon falhou, o arquivo de registro do daemon (`/var/log/ceph/ceph-osd.*`) pode incluir informações de depuração.

#### OSD\_tipo de crush\_DOWN. Por exemplo, OSD\_HOST\_DOWN

Todos os OSDs em uma subárvore específica do CRUSH estão marcados como inativos. Por exemplo, todos os OSDs em um host.

## OSD\_ORPHAN

Um OSD é referenciado na hierarquia do mapa CRUSH, mas não existe. O OSD pode ser removido da hierarquia do CRUSH com:

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

## OSD\_OUT\_OF\_ORDER\_FULL

Os limites de uso para *backfillfull* (padrão definido como 0,90), *nearfull* (padrão definido como 0,85), *full* (padrão definido como 0,95) e/ou *failsafe\_full* não são ascendentes. Especificamente, esperamos *backfillfull* < *nearfull*, *nearfull* < *full* e *full* < *failsafe\_full*.

Para ler os valores atuais, execute:

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

É possível ajustar os limites com os seguintes comandos:

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

## OSD\_FULL

Um ou mais OSDs excederam o limite de *full* e impedem o cluster de executar gravações. É possível verificar o uso por pool com:

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

É possível ver a cota *full* definida no momento com:

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

Uma solução alternativa de curto prazo para resolver a disponibilidade de gravação é aumentar um pouco o valor do limite de *full*:

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

Adicione o novo armazenamento ao cluster implantando mais OSDs ou apague os dados existentes para liberar espaço.

## OSD\_BACKFILLFULL

Um ou mais OSDs excederam o limite de *backfillfull*, o que impede a redistribuição dos dados no dispositivo. Trata-se de um aviso antecipado de que a redistribuição talvez não possa ser concluída e de que o cluster está quase cheio. É possível verificar o uso por pool com:

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

## OSD\_NEARFULL

Um ou mais OSDs excederam o limite de *nearfull*. Trata-se de um aviso antecipado de que o cluster está quase cheio. É possível verificar o uso por pool com:

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

## OSDMAP\_FLAGS

Um ou mais flags de interesse do cluster foram definidos. Com exceção de *full*, é possível definir ou limpar esses flags com:

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

Esses flags incluem:

### full

O cluster foi sinalizado como cheio e não pode realizar gravações.

### pauserd, pausewr

Leituras ou gravações pausadas.

### noup

Os OSDs não têm permissão para serem iniciados.

### nodown

Os relatórios de falha do OSD estão sendo ignorados, portanto, os monitores não marcarão os OSDs como *inativos*.

### noin

Os OSDs já marcados como *out* não serão remarcados como *in* quando forem iniciados.

### noout

Os OSDs *inativos* não serão automaticamente marcados como *out* após o intervalo configurado.



nobackfill, norecover, norebalance

A recuperação ou a redistribuição de dados está suspensa.

noscrub, nodeep\_scrub

A depuração (consulte a [Seção 17.6, “Depurando grupos de posicionamento”](#)) está desabilitada.

notieragent

A atividade de camadas de cache foi suspensa.

## OSD\_FLAGS

Um ou mais OSDs têm um flag de interesse definido por OSD. Esses flags incluem:

noup

O OSD não tem permissão para ser iniciado.

nodown

Os relatórios de falha para este OSD serão ignorados.

noin

Se este OSD já foi marcado como *out* automaticamente após uma falha, ele não será marcado como *in* quando for iniciado.

noout

Se este OSD estiver inativo, ele não será automaticamente marcado como *out* após o intervalo configurado.

É possível definir e limpar os flags por OSD com:

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

## OLD\_CRUSH\_TUNABLES

O Mapa CRUSH usa configurações muito antigas e deve ser atualizado. Os tunables mais antigos que podem ser usados (ou seja, a versão de cliente mais antiga que pode se conectar ao cluster) sem acionar este aviso de saúde são determinados pela opção de configuração mon\_crush\_min\_required\_version.

## OLD\_CRUSH\_STRAW\_CALC\_VERSION

O Mapa CRUSH usa um método mais antigo que não é o ideal para calcular os valores de peso intermediários para compartimentos de memória straw. O Mapa CRUSH deve ser atualizado para usar o método mais recente (straw\_calc\_version=1).

## CACHE\_POOL\_NO\_HIT\_SET

Um ou mais pools de cache não estão configurados com um conjunto de acertos para monitorar o uso, o que impede o agente de camadas de identificar objetos frios para descarregar e eliminar do cache. É possível configurar conjuntos de acertos no pool de cache com:

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

## OSD\_NO\_SORTBITWISE

Não há OSDs anteriores ao Luminous v12 em execução, mas o flag `sortbitwise` não foi definido. Você precisa definir o flag `sortbitwise` para que os OSDs do Luminous v12 ou mais recentes possam ser iniciados:

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

## POOL\_FULL

Um ou mais pools atingiram a cota e não permitem mais gravações. Você pode definir cotas e uso de pool com:

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

Você pode aumentar a cota do pool com

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

ou apagar alguns dados existentes para reduzir o uso.

## PG\_AVAILABILITY

A disponibilidade de dados está reduzida, o que significa que o cluster não pode atender a possíveis solicitações de leitura ou gravação para alguns dados no cluster. Especificamente, o estado de um ou mais PGs não permite que as solicitações de E/S sejam atendidas. Os estados dos PGs problemáticos incluem *emparelhamento*, *obsoleto*, *incompleto* e a ausência de *ativo* (se essas condições não forem resolvidas rapidamente). As informações detalhadas sobre quais PGs são afetados estão disponíveis em:

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

Na maioria dos casos, a causa raiz é que um ou mais OSDs estão inativos no momento. É possível consultar o estado dos PGs problemáticos específicos com:

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

## PG\_DEGRADED

A redundância de dados está reduzida para alguns dados, o que significa que o cluster não tem o número de réplicas desejado para todos os dados (em pools replicados) ou para fragmentos de código de eliminação (em pools codificados para eliminação). Especificamente, um ou mais PGs têm o flag *degraded* ou *undersized* definido (não há instâncias suficientes desse grupo de posicionamento no cluster) ou não tinham o flag *clean* definido durante determinado período. As informações detalhadas sobre quais PGs são afetados estão disponíveis em:

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

Na maioria dos casos, a causa raiz é que um ou mais OSDs estão inativos no momento. É possível consultar o estado dos PGs problemáticos específicos com:

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

## PG\_DEGRADED\_FULL

A redundância de dados pode estar reduzida ou em risco para alguns dados devido à ausência de espaço livre no cluster. Especificamente, um ou mais PGs têm o flag *backfill\_toofull* ou *recovery\_toofull* definido, o que significa que o cluster não pode migrar ou recuperar dados porque um ou mais OSDs estão acima do limite de *backfillfull*.

## PG\_DAMAGED

A depuração de dados (consulte a [Seção 17.6, “Depurando grupos de posicionamento”](#)) descobriu alguns problemas com a consistência de dados no cluster. Especificamente, um ou mais PGs têm o flag *inconsistent* ou *snaptrim\_error* definido, indicando que uma operação de depuração anterior detectou um problema, ou o flag *repair* definido, o que significa que um reparo para esse tipo de inconsistência está agora em andamento.

## OSD\_SCRUB\_ERRORS

Depurações recentes de OSD revelaram inconsistências.

## CACHE\_POOL\_NEAR\_FULL

Um pool de camada de cache está quase cheio. Neste contexto, “full” é determinado pelas propriedades *target\_max\_bytes* e *target\_max\_objects* no pool de cache. Quando o pool atinge o limite de destino, as solicitações de gravação para o pool podem ser bloqueadas enquanto os dados são descarregados e eliminados do cache, um estado que normalmente gera latências muito altas e baixo desempenho. É possível ajustar o tamanho do destino do pool de cache com:

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

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

As atividades normais de descarregamento e eliminação de cache também podem ficar restritas por redução na disponibilidade ou no desempenho da camada de base ou do carregamento geral do cluster.

#### TOO\_FEW\_PGS

O número de PGs em uso está abaixo do limite configurável de `mon_pg_warn_min_per_osd` PGs por OSD. Isso pode levar à distribuição e ao equilíbrio de dados abaixo do ideal em todos os OSDs no cluster, reduzindo o desempenho geral.

#### TOO\_MANY\_PGS

O número de PGs em uso está acima do limite configurável de `mon_pg_warn_max_per_osd` PGs por OSD. Isso pode levar a um aumento do uso de memória dos daemons OSD, a uma redução do emparelhamento após mudanças no estado do cluster (por exemplo, reinicializações, adições ou remoções de OSD) e a um aumento da carga nos Ceph Managers e Ceph Monitors.

Não é possível reduzir o valor `pg_num` dos pools existentes, mas é possível reduzir o valor `pgp_num`. Efetivamente, isso coloca alguns PGs nos mesmos conjuntos de OSDs, atenuando alguns dos impactos negativos descritos acima. É possível ajustar o valor `pgp_num` com:

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

#### SMALLER\_PGP\_NUM

Um ou mais pools têm um valor `pgp_num` menor do que `pg_num`. Normalmente, isso indica que a contagem de PGs foi aumentada sem aumentar também o comportamento de posicionamento. Isso costuma ser resolvido definindo `pgp_num` para corresponder a `pg_num`, o que aciona a migração de dados, com:

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

#### MANY\_OBJECTS\_PER\_PG

Um ou mais pools têm um número médio de objetos por PG que é significativamente maior do que a média geral do cluster. O limite específico é controlado pelo valor da configuração `mon_pg_warn_max_object_skew`. Isso costuma ser uma indicação de que o(s) pool(s) com a maioria dos dados no cluster está(ão) com um número muito baixo de PGs, e/ou que outros pools sem tantos dados têm PGs em excesso. É possível aumentar o limite para silenciar o aviso de saúde ajustando a opção de configuração `mon_pg_warn_max_object_skew` nos monitores.

## POOL\_APP\_NOT\_ENABLED

Existe um pool que contém um ou mais objetos, mas que não foi marcado para uso por determinado aplicativo. Resolva esse aviso identificando o pool para uso por um aplicativo. Por exemplo, se o pool é usado pelo RBD:

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

Se o pool é usado por um aplicativo personalizado “foo”, você também pode identificá-lo usando o comando de nível inferior:

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

## POOL\_FULL

Um ou mais pools atingiram (ou estão muito próximos de atingir) a cota. O limite para acionar essa condição de erro é controlado pela opção de configuração mon\_pool\_quota\_crit\_threshold. É possível aumentar ou reduzir (ou remover) as cotas de pool com:

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

A definição do valor como 0 desabilitará a cota.

## POOL\_NEAR\_FULL

Um ou mais pools estão quase atingindo a cota. O limite para acionar essa condição de aviso é controlado pela opção de configuração mon\_pool\_quota\_warn\_threshold. É possível aumentar ou reduzir (ou remover) as cotas de pool com:

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

A definição do valor como 0 desabilitará a cota.

## OBJECT\_MISPLACED

Um ou mais objetos no cluster não são armazenados no nó em que o cluster deseja que eles sejam. Isso é uma indicação de que a migração de dados causada por alguma mudança recente no cluster ainda não foi concluída. Os dados incorretamente armazenados não representam uma condição de risco por si só. A consistência de dados nunca está em risco, e as cópias antigas de objetos serão removidas apenas quando houver o número desejado de novas cópias (nos locais esperados).

## OBJECT\_UNFOUND

Um ou mais objetos no cluster não foram encontrados. Especificamente, os OSDs sabem que uma cópia nova ou atualizada de um objeto deve existir, mas uma cópia dessa versão do objeto não foi encontrada nos OSDs que estão ativos no momento. As solicitações de leitura ou gravação para os objetos “não encontrados” serão bloqueadas. No cenário ideal, o OSD inativo com a cópia mais recente do objeto não encontrado pode ser reativado. É possível identificar os OSDs candidatos com base no estado do emparelhamento referente ao(s) PG(s) responsável(is) pelo objeto não encontrado:

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

## REQUEST\_SLOW

O processamento de uma ou mais solicitações OSD está levando muito tempo. Isso pode ser uma indicação de carga extrema, um dispositivo de armazenamento lento ou um bug de software. É possível consultar a fila de solicitações no(s) OSD(s) em questão com o seguinte comando executado do host OSD:

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

Você pode ver um resumo das solicitações recentes mais lentas:

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

Você pode encontrar o local de um OSD com:

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

## REQUEST\_STUCK

Uma ou mais solicitações OSD foram bloqueadas por um período relativamente longo. Por exemplo, 4096 segundos. Isso é uma indicação de que o cluster não esteve saudável por um longo período (por exemplo, não há OSDs suficientes em execução ou PGs inativos) ou de que existe algum problema interno com o OSD.

## PG\_NOT\_SCRUBBED

Um ou mais PGs não foram depurados (consulte a [Seção 17.6, “Depurando grupos de posicionamento”](#)) recentemente. Normalmente, os PGs são depurados a cada `mon_scrub_interval` segundos, e esse aviso será acionado após decorrer `mon_warn_not_scrubbed` segundos sem uma depuração. Os PGs não serão depurados

se não forem sinalizados como limpos, o que poderá ocorrer se forem armazenados incorretamente ou estiverem degradados (consulte `PG_AVAILABILITY` e `PG_DEGRADED` acima). Você pode iniciar manualmente uma depuração de um PG limpo com:

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

### PG\_NOT\_DEEP\_SCRUBBED

Um ou mais PGs não foram depurados em detalhes (consulte a [Seção 17.6, “Depurando grupos de posicionamento”](#)) recentemente. Normalmente, os PGs são depurados a cada `osd_deep_mon_scrub_interval` segundos, e esse aviso é acionado quando `mon_warn_not_deep_scrubbed` segundos decorreram sem uma depuração. Os PGs não serão depurados (em detalhes) se não forem sinalizados como limpos, o que poderá ocorrer se forem armazenados incorretamente ou estiverem degradados (consulte `PG_AVAILABILITY` e `PG_DEGRADED` acima). Você pode iniciar manualmente uma depuração de um PG limpo com:

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



### Dica

Se você especificou locais diferentes do padrão em sua configuração ou no chaveiro, deve especificar estes locais:

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

## 12.3 Verificando as estatísticas de uso de um cluster

Para verificar o uso e a distribuição dos dados de um cluster entre pools, use o comando `ceph df`. Para obter mais detalhes, use `ceph df detail`.

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

--- POOLS ---
POOL                                ID  STORED  OBJECTS  USED    %USED  MAX AVAIL
device_health_metrics              1    0 B      0        0 B     0      8.5 GiB
```

cephfs.my_cephfs.meta	2	1.0 MiB	22	4.5 MiB	0.02	8.5 GiB
cephfs.my_cephfs.data	3	0 B	0	0 B	0	8.5 GiB
.rgw.root	4	1.9 KiB	13	2.2 MiB	0	8.5 GiB
myzone.rgw.log	5	3.4 KiB	207	6 MiB	0.02	8.5 GiB
myzone.rgw.control	6	0 B	8	0 B	0	8.5 GiB
myzone.rgw.meta	7	0 B	0	0 B	0	8.5 GiB

A seção RAW STORAGE da saída apresenta uma visão geral da quantidade de armazenamento que seu cluster usa para os dados.

- CLASS: A classe de armazenamento do dispositivo. Consulte a [Seção 17.1.1, “Classes de dispositivo”](#) para obter mais detalhes sobre classes de dispositivo.
- SIZE: A capacidade de armazenamento geral do cluster.
- AVAIL: A quantidade de espaço livre disponível no cluster.
- USED: O espaço (acumulado de todos os OSDs) alocado exclusivamente para objetos de dados mantidos em dispositivo de blocos.
- RAW USED: A soma do espaço “USED” e do espaço alocado/reservado no dispositivo de blocos para finalidade do Ceph. Por exemplo, parte do BlueFS para o BlueStore.
- % RAW USED: A porcentagem usada de armazenamento bruto. Use esse número em conjunto com full ratio e near full ratio para garantir que você não atinja a capacidade do cluster. Consulte a [Seção 12.8, “Capacidade de armazenamento”](#) para obter mais detalhes.



### Nota: Nível de preenchimento do cluster

Quando o nível de preenchimento de um armazenamento bruto está próximo de 100%, você precisa adicionar um novo armazenamento ao cluster. O uso mais alto pode resultar em OSDs únicos cheios e problemas de saúde do cluster.

Use o comando `ceph osd df tree` para listar o nível de preenchimento de todos os OSDs.

A seção P00LS da saída apresenta uma lista dos pools e o uso estimado de cada pool. A saída dessa seção *não* reflete réplicas, clones ou instantâneos. Por exemplo, se você armazenar um objeto com 1 MB de dados, o uso estimado será de 1 MB, mas o uso real poderá ser de 2 MB ou mais, dependendo do número de réplicas, clones e instantâneos.

- P00L: O nome do pool.
- ID: O ID do pool.



- STORED: A quantidade de dados armazenados pelo usuário.
- OBJECTS: O número de objetos armazenados por pool.
- USED: A quantidade de espaço alocado exclusivamente para dados por todos os nós OSD em kB.
- %USED: A porcentagem estimada de armazenamento usado por pool.
- MAX AVAIL: O espaço máximo disponível no pool especificado.



## Nota

Os números na seção POOLS são estimativas. Eles não incluem o número de réplicas, instantâneos ou clones. Como resultado, a soma dos valores USED e %USED não incluirá os valores RAW USED e %RAW USED na seção RAW STORAGE da saída.

## 12.4 Verificando o status do OSD

Você pode verificar os OSDs para garantir que estejam ativos e em execução:

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

ou

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

Você também pode ver os OSDs de acordo com a posição deles no mapa CRUSH.

O **ceph osd tree** imprimirá uma árvore CRUSH com um host, os OSDs, se eles estão ativos e o peso:

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

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

## 12.5 Verificando se há OSDs cheios

O Ceph impede você de gravar em um OSD cheio para evitar a perda de dados. Em um cluster operacional, você deve receber um aviso quando o cluster está próximo cota máxima. O padrão do valor mon osd full ratio é de 0,95, ou 95% da capacidade antes de impedir que os clientes gravem dados. O padrão do valor mon osd nearfull ratio é de 0,85, ou 85% da capacidade, quando ele gera um aviso de saúde.

O ceph health relata os nós OSD cheios:

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

ou

```
cephuser@adm > ceph health
HEALTH_ERR 1 nearfull osds, 1 full osds
osd.2 is near full at 85%
osd.3 is full at 97%
```

A melhor maneira de resolver um cluster cheio é adicionar novos hosts/discos OSD, o que permite ao cluster redistribuir os dados para o armazenamento recém-disponibilizado.



### Dica: Evitando OSDs cheios

Depois que um OSD fica cheio (usa 100% do espaço em disco), ele costuma falhar rapidamente sem aviso. Veja a seguir algumas dicas para se lembrar na hora de administrar nós OSD.

- O espaço em disco de cada OSD (normalmente montado em /var/lib/ceph/osd/osd-{1,2..}) precisa ser colocado em um disco ou partição subjacente dedicado.
- Verifique os arquivos de configuração do Ceph e certifique-se de que o Ceph não armazene o arquivo de registro em discos/partições dedicados para uso por OSDs.
- Confirme se nenhum outro processo faz gravações nos discos/partições dedicados para uso por OSDs.

## 12.6 Verificando o status do monitor

Após iniciar o cluster e antes da primeira leitura e/ou gravação de dados, verifique o status do quorum dos Ceph Monitors. Quando o cluster já estiver processando solicitações, verifique o status dos Ceph Monitors periodicamente para garantir que estejam em execução.

Para exibir o mapa do monitor, execute o seguinte:

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

ou

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

Para verificar o status do quorum para o cluster do monitor, execute o seguinte:

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

O Ceph retornará o status do quorum. Por exemplo, um cluster do Ceph com três monitores pode retornar o seguinte:

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

## 12.7 Verificando estados de grupos de posicionamento

Os grupos de posicionamento mapeiam objetos para OSDs. Ao monitorar seus grupos de posicionamento, você deseja que eles estejam ativos e limpos. Para ver uma discussão detalhada, consulte a [Seção 12.9, “Monitorando OSDs e grupos de posicionamento”](#).

## 12.8 Capacidade de armazenamento

Quando um cluster de armazenamento do Ceph está próximo da sua capacidade máxima, o Ceph impede você de gravar ou ler os Ceph OSDs como medida de segurança para evitar perda de dados. Portanto, permitir que um cluster de produção se aproxime da sua cota máxima não é uma boa prática, porque coloca em risco a alta disponibilidade. A cota máxima padrão está definida como 0,95, que significa 95% da capacidade. Essa é uma configuração muito agressiva para um cluster de teste com um número pequeno de OSDs.



### Dica: Aumentar a Capacidade de Armazenamento

Ao monitorar o cluster, fique atento aos avisos relacionados à cota `nearfull`. Eles indicam que a falha de alguns OSDs pode resultar em interrupção temporária de serviço. Considere adicionar mais OSDs para aumentar a capacidade de armazenamento.

Um cenário comum para clusters de teste envolve um administrador do sistema que remove um Ceph OSD do cluster de armazenamento do Ceph para observar a redistribuição do cluster. Em seguida, ele remove outro Ceph OSD, e assim por diante, até o cluster atingir a cota máxima e ser bloqueado. Recomendamos um pouco de planejamento de capacidade, mesmo com um cluster de teste. O planejamento permite estimar a quantidade de capacidade sobressalente que você precisará para manter a alta disponibilidade. O ideal é planejar uma série de falhas de Ceph OSDs em que o cluster possa se recuperar ao estado ativo + limpo sem substituir os Ceph OSDs imediatamente. Você pode executar um cluster no estado ativo + degradado, mas isso não é ideal em condições normais de operação.

O diagrama a seguir retrata um cluster de armazenamento do Ceph simples contendo 33 nós do Ceph com um Ceph OSD por host, cada um deles lendo e gravando em uma unidade de 3 TB. Esse cluster de exemplo tem uma capacidade máxima real de 99 TB. A opção `mon osd`

full ratio está definida como 0,95. Se o cluster atingir 5 TB da capacidade restante, ele não permitirá que os clientes leiam e gravem dados. Portanto, a capacidade operacional do cluster de armazenamento é de 95 TB, não de 99 TB.

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

FIGURA 12.1: CLUSTER DO CEPH

Nesse tipo de cluster, é normal haver falha de um ou dois OSDs. Um cenário menos frequente, mas considerável, envolve uma falha no roteador do rack ou no fornecimento de energia, o que desligaria vários OSDs ao mesmo tempo (por exemplo, OSDs 7-12). Nesse cenário, você ainda deve recorrer a um cluster capaz de se manter operante e atingir o estado ativo + limpo, mesmo que isso signifique adicionar alguns hosts com outros OSDs a curto prazo. Se o uso da capacidade for muito alto, talvez você não perca os dados. Porém, você ainda poderá colocar em risco a disponibilidade dos dados ao resolver uma interrupção em um domínio de falha, se o uso da capacidade do cluster exceder a cota máxima. Por esse motivo, recomendamos pelo menos algum planejamento de capacidade estimada.

Identifique dois números para o cluster:

1. O número de OSDs.
2. A capacidade total do cluster.

Se você dividir a capacidade total pelo número de OSDs do cluster, encontrará a capacidade média de um OSD no cluster. Considere multiplicar esse número pela quantidade de OSDs que podem falhar simultaneamente durante as operações normais (um número relativamente pequeno). Por fim, multiplique a capacidade do cluster pela cota máxima para chegar a uma capacidade operacional máxima. Em seguida, subtraia o número da quantidade de dados dos OSDs que podem falhar para chegar a uma cota máxima razoável. Repita o processo anterior com um número maior de falhas de OSD (um rack de OSDs) para chegar a um número razoável para uma cota quase máxima.

As seguintes configurações aplicam-se apenas à criação de cluster e são armazenadas no mapa OSD:

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



### Dica

Essas configurações aplicam-se apenas durante a criação do cluster. Depois disso, elas precisarão ser mudadas no Mapa OSD usando os comandos `ceph osd set-nearfull-ratio` e `ceph osd set-full-ratio`.

#### `mon osd full ratio`

A porcentagem de espaço em disco usado antes que um OSD seja considerado cheio. O padrão é 0,95

#### `mon osd backfillfull ratio`

A porcentagem de espaço em disco usado antes que um OSD seja considerado muito cheio para provisionamento. O padrão é 0,90

#### `mon osd nearfull ratio`

A porcentagem de espaço em disco usado antes que um OSD seja considerado quase cheio. O padrão é 0,85



### Dica: Verificar o peso do OSD

Se alguns OSDs estiverem quase cheios, mas outros tiverem bastante capacidade, talvez você tenha problemas com o peso do CRUSH para os OSDs quase cheios.

## 12.9 Monitorando OSDs e grupos de posicionamento

As altas disponibilidade e confiabilidade exigem uma abordagem tolerante a falhas para gerenciar problemas de hardware e software. O Ceph não tem um ponto único de falha e pode processar solicitações de dados no modo "degradado". O posicionamento de dados do Ceph

introduz uma camada de indireção para garantir que os dados não sejam diretamente vinculados a determinados endereços de OSD. Isso significa que o monitoramento de falhas no sistema requer encontrar o grupo de posicionamento e os OSDs subjacentes na raiz do problema.



### Dica: Acesso em caso de falha

Uma falha em uma parte do cluster pode impedir você de acessar um determinado objeto. Isso não significa que você não pode acessar outros objetos. Em caso de falha, siga as etapas para monitorar os OSDs e grupos de posicionamento. Em seguida, comece a solução de problemas.

Em geral, o Ceph tem a funcionalidade de conserto automático. No entanto, quando os problemas persistem, o monitoramento de OSDs e grupos de posicionamento ajuda você a identificá-los.

## 12.9.1 Monitorando OSDs

O status de um OSD é *no cluster* (“in”) ou *fora do cluster* (“out”). Ao mesmo tempo, ele é *em execução* (“up”) ou *inativo e não operante* (“down”). Se um OSD é “up”, ele pode estar no cluster (você pode ler e gravar dados) ou fora do cluster. Se ele estava no cluster e foi recentemente removido do cluster, o Ceph migra os grupos de posicionamento para outros OSDs. Se um OSD estiver fora do cluster, o CRUSH não atribuirá grupos de posicionamento a ele. Se um OSD é “down”, ele também deve ser “out”.



### Nota: Estado não saudável

Se um OSD é “down” e “in”, há um problema, e o estado do cluster não é saudável.

Se você executar um comando, como `ceph health`, `ceph -s` ou `ceph -w`, poderá perceber que o cluster nem sempre retorna `HEALTH OK`. Em relação aos OSDs, você deve esperar que o cluster *não* retorne `HEALTH OK` nas seguintes circunstâncias:

- Você ainda não iniciou o cluster (ele não responderá).
- Você iniciou ou reiniciou o cluster e ele ainda não está pronto, porque os grupos de posicionamento estão sendo criados e os OSDs estão no processo de emparelhamento.
- Você adicionou ou removeu um OSD.
- Você modificou o mapa do cluster.

Um aspecto importante do monitoramento de OSDs é garantir que, quando o cluster estiver em execução, todos os OSDs no cluster também estejam funcionando. Para ver se todos os OSDs estão funcionando, execute:

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

O resultado deve indicar o número total de OSDs (x), quantos estão “up” (y), quantos estão “in” (z) e a época do mapa (eNNNN). Se o número de OSDs que estão “in” no cluster for maior do que o número de OSDs que estão “up”, execute o seguinte comando para identificar os daemons `ceph-osd` que não estão funcionando:

```
# ceph osd tree
#ID CLASS WEIGHT  TYPE NAME                STATUS REWEIGHT PRI-AFF
-1          2.00000 pool openstack
-3          2.00000 rack dell-2950-rack-A
-2          2.00000 host dell-2950-A1
0  ssd 1.00000    osd.0                up  1.00000 1.00000
1  ssd 1.00000    osd.1                down 1.00000 1.00000
```

Por exemplo, se um OSD com ID 1 estiver desativado, inicie-o:

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

Consulte a *Livro “Troubleshooting Guide”, Capítulo 4 “Troubleshooting OSDs”, Seção 4.3 “OSDs not running”* para ver os problemas associados aos OSDs que estão parados ou que não serão reiniciados.

## 12.9.2 Atribuindo conjuntos de grupos de posicionamento

Quando o CRUSH atribui grupos de posicionamento a OSDs, ele analisa o número de réplicas para o pool e atribui o grupo de posicionamento aos OSDs de modo que cada réplica do grupo de posicionamento seja atribuída a um OSD diferente. Por exemplo, se o pool exigir três réplicas de um grupo de posicionamento, o CRUSH poderá atribuí-las a `osd.1`, `osd.2` e `osd.3`, respectivamente. Na verdade, o CRUSH busca um posicionamento pseudo-aleatório que considera os domínios de falha definidos no seu Mapa CRUSH. Sendo assim, você raramente verá grupos de posicionamento atribuídos aos OSDs vizinhos mais próximos em um cluster de grande porte. Nossa referência é ao conjunto de OSDs que deve incluir as réplicas de um determinado



grupo de posicionamento como o *conjunto de atuação*. Em alguns casos, um OSD no conjunto de atuação está inativo ou, de alguma outra forma, não pode processar as solicitações para objetos no grupo de posicionamento. Nesses tipos de situação, um dos seguintes cenários pode ocorrer:

- Você adicionou ou removeu um OSD. Em seguida, o CRUSH reatribuiu o grupo de posicionamento a outros OSDs e, portanto, mudou a composição do *conjunto de atuação*, provocando a migração dos dados com um processo de “provisionamento”.
- Um OSD estava “down”, foi reiniciado e agora está se recuperando.
- Um OSD no *conjunto de atuação* está “down” ou não pode processar as solicitações, e outro OSD assumiu temporariamente as tarefas dele.

O Ceph processa uma solicitação de cliente usando o *conjunto ativo*, que é o conjunto de OSDs que processará de fato as solicitações. Na maioria dos casos, o *conjunto ativo* e o *conjunto de atuação* são praticamente idênticos. Quando não são, isso pode indicar que o Ceph está migrando dados, que um OSD está se recuperando ou que há um problema. Por exemplo, o Ceph costuma retornar o estado HEALTH\_WARN com a mensagem “stuck stale” em cenários assim.

Para recuperar uma lista de grupos de posicionamento, execute:

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

Para ver quais OSDs estão no *conjunto de atuação* ou no *conjunto ativo* para um determinado grupo de posicionamento, execute:

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

O resultado deve indicar a época do osdmap (eNNN), o número do grupo de posicionamento (PG\_NUM), os OSDs no *conjunto ativo* (“up”) e os OSDs no *conjunto de atuação* (“acting”):



### Dica: Indicador de problema no cluster

Se o *conjunto ativo* e o *conjunto de atuação* não forem os mesmos, isso pode ser um indicador de que o cluster está realizando a própria redistribuição ou de um possível problema com o cluster.

### 12.9.3 Emparelhamento

Antes que você possa gravar dados em um grupo de posicionamento, ele deve estar no estado ativo e, de preferência, limpo. Para o Ceph determinar o estado atual de um grupo de posicionamento, o OSD principal do grupo de posicionamento (o primeiro OSD no *conjunto de atuação*), é emparelhado com os OSDs secundários e terciários para estabelecer um acordo em relação ao estado atual do grupo de posicionamento (considerando um pool com três réplicas do PG).

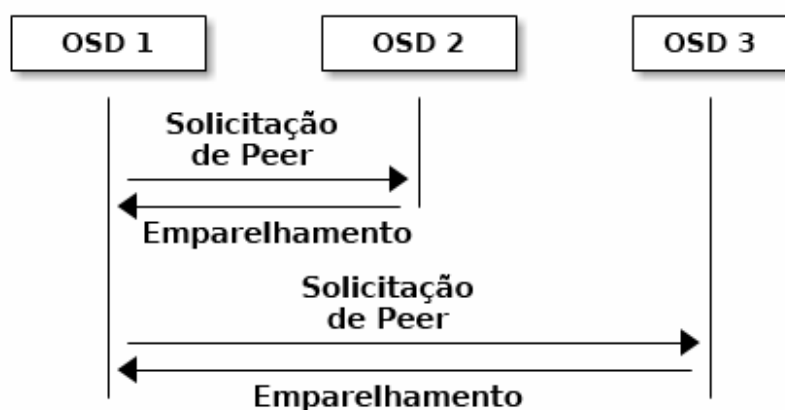


FIGURA 12.2: ESQUEMA DE EMPARELHAMENTO

### 12.9.4 Monitorando estados de grupos de posicionamento

Se você executar um comando, como `ceph health`, `ceph -s` ou `ceph -w`, poderá perceber que o cluster nem sempre retorna a mensagem HEALTH OK. Após verificar se os OSDs estão em execução, verifique também os estados do grupo de posicionamento.

É esperado que o cluster **não** retorne HEALTH OK em várias circunstâncias relacionadas ao emparelhamento de grupos de posicionamento:

- Você criou um pool, e os grupos de posicionamento ainda não foram emparelhados.
- Os grupos de posicionamento estão se recuperando.
- Você adicionou ou removeu um OSD do cluster.
- Você modificou seu Mapa CRUSH, e seus grupos de posicionamento estão sendo migrados.

- Há dados inconsistentes em diferentes réplicas de um grupo de posicionamento.
- O Ceph está depurando as réplicas de um grupo de posicionamento.
- O Ceph não tem capacidade de armazenamento suficiente para concluir as operações de provisionamento.

Se uma das circunstâncias mencionadas acima fizer com que o Ceph retorne `HEALTH_WARN`, não se preocupe. Em muitos casos, o cluster se recuperará por conta própria. Em alguns casos, pode ser necessário tomar medidas. Um aspecto importante do monitoramento dos grupos de posicionamento é garantir que, quando o cluster estiver em funcionamento, todos os grupos de posicionamento estejam "ativos" e, de preferência, no "estado limpo". Para ver o status de todos os grupos de posicionamento, execute:

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

O resultado deve indicar o número total de grupos de posicionamento (x), quantos grupos de posicionamento estão em determinado estado, como “ativo + limpo” (y) e a quantidade de dados armazenados (z).

Além dos estados do grupo de posicionamento, o Ceph também retornará a quantidade de capacidade de armazenamento utilizada (aa), a quantidade de capacidade de armazenamento restante (bb) e a capacidade total de armazenamento do grupo de posicionamento. Esses números podem ser importantes em alguns casos:

- Você está atingindo a cota quase máxima ou cota máxima.
- Seus dados não estão sendo distribuídos por todo o cluster por causa de um erro na configuração do CRUSH.



## Dica: IDs dos grupos de posicionamento

Os IDs dos grupos de posicionamento consistem no número do pool (não no nome do pool) seguido de um ponto (.) e no ID do grupo de posicionamento: um número hexadecimal. Você pode ver os números de pool e os respectivos nomes na saída do comando `ceph osd lspools`. Por exemplo, o pool padrão `rb` corresponde ao número do pool 0. Um ID de grupo de posicionamento totalmente qualificado tem o seguinte formato:

```
POOL_NUM.PG_ID
```

Normalmente, ele tem a seguinte aparência:

```
0.1f
```

Para recuperar uma lista de grupos de posicionamento, execute o seguinte:

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

Você também pode formatar a saída em JSON e gravá-la em um arquivo:

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

Para consultar um determinado grupo de posicionamento, execute o seguinte:

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

A lista a seguir descreve os estados comuns do grupo de posicionamento em detalhes.

## CREATING

Quando você cria um pool, ele cria o número de grupos de posicionamento que você especificou. O Ceph retornará “creating” (criando) durante a criação de um ou mais grupos de posicionamento. Quando são criados, os OSDs que fazem parte do *conjunto de atuação* do grupo de posicionamento são emparelhados. Quando o emparelhamento é concluído, o status do grupo de posicionamento deve ser “ativo + limpo”, o que significa que um cliente Ceph pode começar a gravar no grupo de posicionamento.

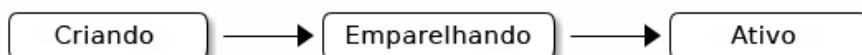


FIGURA 12.3: STATUS DOS GRUPOS DE POSICIONAMENTO

## PEERING

Quando o Ceph está emparelhando um grupo de posicionamento, ele leva os OSDs que armazenam as réplicas do grupo de posicionamento a um acordo em relação ao estado dos objetos e metadados nesse grupo. Quando o Ceph conclui o emparelhamento, isso significa que os OSDs que armazenam o grupo de posicionamento concordam sobre o estado atual desse grupo. No entanto, a conclusão do processo de emparelhamento **não** significa que cada réplica tem o conteúdo mais recente.



## Nota: Histórico de autorização

O Ceph **não** confirmará uma operação de gravação em um cliente até que todos os OSDs do *conjunto de atuação* continuem a operação de gravação. Esta prática garante que pelo menos um membro do *conjunto de atuação* tenha um registro de cada operação de gravação confirmada desde a última operação de emparelhamento bem-sucedida.

Com um registro preciso de cada operação de gravação confirmada, o Ceph pode construir e ampliar um novo histórico de autorização do grupo de posicionamento: um conjunto completo e totalmente organizado de operações que, se executadas, mantêm uma cópia do OSD de um grupo de posicionamento atualizada.

### ACTIVE

Quando o Ceph conclui o processo de emparelhamento, um grupo de posicionamento pode se tornar ativo. O estado ativo significa que os dados no grupo de posicionamento geralmente estão disponíveis no grupo de posicionamento principal e nas réplicas para as operações de leitura e gravação.

### CLEAN

Quando um grupo de posicionamento está no estado limpo, o OSD principal e os OSDs de réplica foram emparelhados com êxito e não há réplicas perdidas para o grupo de posicionamento. O Ceph replicou o número correto de vezes todos os objetos no grupo de posicionamento.

### DEGRADED

Quando um cliente grava um objeto no OSD principal, esse OSD é responsável por gravar as réplicas nos OSDs de réplica. Depois que o OSD principal grava o objeto no armazenamento, o grupo de posicionamento permanecerá no estado "degradado" até que o OSD principal receba uma confirmação dos OSDs de réplica de que o Ceph criou os objetos de réplica com êxito.

O motivo pelo qual um grupo de posicionamento pode ser “ativo + degradado” é que um OSD pode ser “ativo” mesmo que ainda não armazene todos os objetos. Se um OSD ficar inativo, o Ceph marcará cada grupo de posicionamento atribuído ao OSD como “degradado”. Os OSDs deverão ser emparelhados novamente quando o OSD voltar a ficar ativo. No entanto, um cliente ainda poderá gravar um novo objeto em um grupo de posicionamento degradado se ele for “ativo”.

Se um OSD for “inativo” e a condição “degradado” permanecer, o Ceph poderá marcar o OSD inativo como “fora” do cluster e remapear os dados do OSD “inativo” para outro OSD. O tempo entre ser marcado como “inativo” e como “fora” é controlado pela opção `mon osd down out interval`, que por padrão está definida como 600 segundos.

Um grupo de posicionamento também pode ser “degradado” porque o Ceph não encontra um ou mais objetos que deveriam estar no grupo de posicionamento. Embora você não possa ler ou gravar em objetos não encontrados, ainda pode acessar todos os outros objetos no grupo de posicionamento “degradado”.

## RECOVERING

O Ceph foi projetado para tolerância a falhas em escala, quando os problemas de hardware e software são constantes. Quando um OSD fica “inativo”, o conteúdo dele pode não acompanhar o estado atual das outras réplicas nos grupos de posicionamento. Quando o OSD volta a ficar “ativo”, o conteúdo dos grupos de posicionamento deve ser atualizado para refletir o estado atual. Durante esse período, o OSD pode refletir um estado de “recuperação”.

A recuperação nem sempre é comum, porque uma falha de hardware pode causar uma falha em cascata de vários OSDs. Por exemplo, em uma possível falha do switch de rede para um rack ou gabinete, os OSDs de várias máquinas host podem não acompanhar o estado atual do cluster. Cada um dos OSDs deverá se recuperar quando a falha for resolvida. O Ceph oferece uma série de configurações para equilibrar a contenção de recursos entre as novas solicitações de serviço e a necessidade de recuperar os objetos de dados e restaurar os grupos de posicionamento ao estado atual. A configuração `osd recovery delay start` permite que um OSD reinicie, emparelhe novamente e até processe algumas solicitações de reprodução antes do início do processo de recuperação. A `osd recovery thread timeout` define um tempo de espera do thread porque vários OSDs podem falhar, ser reiniciados e novamente emparelhados em fases. A configuração `osd recovery max active` limita o número de solicitações de recuperação que um OSD processará simultaneamente para impedir falha no processamento do OSD. A configuração `osd recovery max chunk` limita o tamanho dos blocos de dados recuperados para evitar congestionamento de rede.

## BACK FILLING

Quando um novo OSD ingressa no cluster, o CRUSH reatribui os grupos de posicionamento dos OSDs no cluster para o OSD recém-adicionado. Forçar o novo OSD a aceitar os grupos de posicionamento reatribuídos imediatamente pode sobrecarregar o novo OSD.

O provisionamento do OSD com os grupos de posicionamento permite que este processo seja iniciado em segundo plano. Quando o provisionamento for concluído, o novo OSD começará a processar solicitações quando estiver pronto.

Durante as operações de provisionamento, você pode ver um dos vários estados: “backfill\_wait” indica que uma operação de provisionamento está pendente, mas ainda não está em andamento; “backfill” indica que uma operação de provisionamento está em andamento; “backfill\_too\_full” indica que uma operação de provisionamento foi solicitada, mas não pôde ser concluída devido à capacidade de armazenamento insuficiente. Quando não é possível provisionar um grupo de posicionamento, ele pode ser considerado “incompleto”.

O Ceph dispõe de várias configurações para gerenciar a carga associada à reatribuição de grupos de posicionamento para um OSD (especialmente um novo OSD). Por padrão, `osd max backfills` define o número máximo de provisionamentos simultâneos de/para um OSD como 10. A cota máxima de provisionamento permite que um OSD recuse uma solicitação de provisionamento se ele estiver próximo da sua cota máxima (90%, por padrão) e faça a modificação com o comando **`ceph osd set-backfillfull-ratio`**. Se um OSD recusar uma solicitação de provisionamento, o `osd backfill retry interval` permitirá que um OSD repita a solicitação (após 10 segundos, por padrão). Os OSDs também podem definir `osd backfill scan min` e `osd backfill scan max` para gerenciar intervalos de exploração (64 e 512, por padrão).

## REMAPPED

Quando o *conjunto de atuação* que processa um grupo de posicionamento é modificado, os dados são migrados do *conjunto de atuação* antigo para o *conjunto de atuação* novo. Pode levar algum tempo para um novo OSD principal processar as solicitações. Por isso, talvez seja solicitado para o conjunto principal antigo continuar processando as solicitações até que a migração do grupo de posicionamento seja concluída. Quando a migração dos dados for concluída, o mapeamento usará o OSD principal do novo *conjunto de atuação*.

## STALE

Enquanto o Ceph usa os heartbeats para garantir que os hosts e daemons estejam em execução, os daemons `ceph-osd` também podem entrar em um estado “travado” quando não estiverem relatando estatísticas em tempo hábil (por exemplo, uma falha temporária na rede). Por padrão, os daemons OSD relatam suas estatísticas de grupo de posicionamento, inicialização e falha a cada meio segundo (0,5), que é mais frequente do que os limites de heartbeat. Se o OSD principal do *conjunto de atuação* de um grupo

de posicionamento não puder relatar para o monitor ou se outros OSDs relataram o OSD principal como “inativo”, os monitores marcarão o grupo de posicionamento como “obsoleto”.

Quando você inicia o cluster, é comum ver o estado "obsoleto" até o processo de emparelhamento ser concluído. Depois que o cluster estiver funcionando por um tempo, ver grupos de posicionamento no estado "obsoleto" indicará que o OSD principal para esses grupos está inativo ou não está relatando as estatísticas de grupo de posicionamento para o monitor.

## 12.9.5 Encontrando o local de um objeto

Para armazenar dados de objetos no Armazenamento de Objetos do Ceph, um cliente Ceph precisa definir um nome de objeto e especificar um pool relacionado. O cliente Ceph recupera o mapa do cluster mais recente e o algoritmo CRUSH calcula como mapear o objeto para um grupo de posicionamento e, em seguida, calcula como atribuir o grupo de posicionamento a um OSD dinamicamente. Para encontrar o local do objeto, tudo o que você precisa é do nome do objeto e do nome do pool. Por exemplo:

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

### EXEMPLO 12.1: LOCALIZANDO UM OBJETO

Como exemplo, vamos criar um objeto. Especifique o nome do objeto “test-object-1”, um caminho para o arquivo de exemplo “testfile.txt” com alguns dados do objeto e o nome do pool “data” usando o comando **rados put** na linha de comando:

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

Para verificar se o Armazenamento de Objetos do Ceph armazenou o objeto, execute o seguinte:

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

Agora, identifique o local do objeto. O Ceph retornará o local do objeto:

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

Para remover o objeto de exemplo, basta apagá-lo usando o comando **rados rm**:

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



## 13 Tarefas operacionais

### 13.1 Modificando a configuração do cluster

Para modificar a configuração de um cluster do Ceph existente, siga estas etapas:

1. Exporte a configuração atual do cluster para um arquivo:

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

2. Edite o arquivo com a configuração e atualize as linhas relevantes. Encontre exemplos de especificações em *Livro “Guia de Implantação”, Capítulo 8 “Implantando os serviços principais restantes com o cephadm”* e *Seção 13.4.3, “Adicionando OSDs por meio da especificação DriveGroups”*.

3. Aplique a nova configuração:

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

### 13.2 Adicionando nós

Para adicionar um novo nó a um cluster do Ceph, siga estas etapas:

1. Instale o SUSE Linux Enterprise Server e o SUSE Enterprise Storage no novo host. Consulte o *Livro “Guia de Implantação”, Capítulo 5 “Instalando e configurando o SUSE Linux Enterprise Server”* para obter mais informações.
2. Configure o host como um Minion Salt de um Master Salt existente. Consulte o *Livro “Guia de Implantação”, Capítulo 6 “Implantando o Salt”* para obter mais informações.
3. Adicione o novo host ao `ceph-salt` e informe ao cephadm, por exemplo:

```
root@master # ceph-salt config /ceph_cluster/minions add ses-min5.example.com
root@master # ceph-salt config /ceph_cluster/roles/cephadm add ses-min5.example.com
```

Consulte o *Livro “Guia de Implantação”, Capítulo 7 “Implantando o cluster de boot usando ceph-salt”, Seção 7.2.2 “Adicionando minions Salt”* para obter mais informações.

4. Verifique se o nó foi adicionado ao `ceph-salt`:

```
root@master # ceph-salt config /ceph_cluster/minions ls
```

```
o- minions ..... [Minions: 5]
[...]
o- ses-min5.example.com ..... [no roles]
```

5. Aplique a configuração ao novo host de cluster:

```
root@master # ceph-salt apply ses-min5.example.com
```

6. Verifique se o host recém-adicionado agora pertence ao ambiente do cephadm:

```
cephuser@adm > ceph orch host ls
HOST                ADDR                LABELS      STATUS
[...]
ses-min5.example.com  ses-min5.example.com
```

## 13.3 Removendo nós



### Dica: Remover OSDs

Se o nó que você vai remover executa OSDs, remova-os primeiro e verifique se nenhum OSD está sendo executado nesse nó. Consulte [Seção 13.4.4, “Removendo OSDs”](#) para obter mais detalhes sobre como remover OSDs.

Para remover um nó de um cluster, faça o seguinte:

1. Para todos os tipos de serviço do Ceph, exceto `node-exporter` e `crash`, remova o nome de host do nó do arquivo de especificação de posicionamento do cluster (por exemplo, `cluster.yml`). Consulte a *Livro “Guia de Implantação”, Capítulo 8 “Implantando os serviços principais restantes com o cephadm”, Seção 8.2 “Especificação de serviço e posicionamento”* para obter mais detalhes. Por exemplo, se você remover o host chamado `ses-min2`, remova todas as ocorrências de `- ses-min2` de todas as seções `placement`:

#### Atualização

```
service_type: rgw
service_id: EXAMPLE_NFS
placement:
  hosts:
    - ses-min2
    - ses-min3
```

para

```
service_type: rgw
service_id: EXAMPLE_NFS
placement:
  hosts:
    - ses-min3
```

Aplique suas mudanças ao arquivo de configuração:

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

2. Remova o nó do ambiente do cephadm:

```
cephuser@adm > ceph orch host rm ses-min2
```

3. Se o nó estiver executando os serviços `crash.osd.1` e `crash.osd.2`, remova-os executando o seguinte comando no host:

```
root@minion > cephadm rm-daemon --fsid CLUSTER_ID --name SERVICE_NAME
```

Por exemplo:

```
root@minion > cephadm rm-daemon --fsid b4b30c6e... --name crash.osd.1
root@minion > cephadm rm-daemon --fsid b4b30c6e... --name crash.osd.2
```

4. Remova todas as funções do minion que deseja apagar:

```
cephuser@adm > ceph-salt config /ceph_cluster/roles/tuned/throughput remove ses-min2
cephuser@adm > ceph-salt config /ceph_cluster/roles/tuned/latency remove ses-min2
cephuser@adm > ceph-salt config /ceph_cluster/roles/cephadm remove ses-min2
cephuser@adm > ceph-salt config /ceph_cluster/roles/admin remove ses-min2
```

Se o minion que você deseja remover for de boot, você também precisará remover a função de boot:

```
cephuser@adm > ceph-salt config /ceph_cluster/roles/bootstrap reset
```

5. Após remover todos os OSDs de um único host, remova o host do mapa CRUSH:

```
cephuser@adm > ceph osd crush remove bucket-name
```



## Nota

O nome do compartimento de memória deve ser igual ao nome de host.

6. Agora você pode remover o minion do cluster:

```
cephuser@adm > ceph-salt config /ceph_cluster/minions remove ses-min2
```



## Importante

Em caso de falha e se o minion que você está tentando remover estiver em um estado permanentemente desligado, você precisará remover o nó do Master Salt:

```
root@master # salt-key -d minion_id
```

Em seguida, remova o nó de *pillar\_root/ceph-salt.sls*. Normalmente, ele está localizado em */srv/pillar/ceph-salt.sls*.

## 13.4 Gerenciamento de OSD

Esta seção descreve como adicionar, apagar ou remover OSDs de um cluster do Ceph.

### 13.4.1 Listando dispositivos de disco

Para identificar os dispositivos de disco usados e não usados em todos os nós do cluster, liste-os executando o seguinte comando:

```
cephuser@adm > ceph orch device ls
```

HOST	PATH	TYPE	SIZE	DEVICE	AVAIL	REJECT	REASONS
ses-master	/dev/vda	hdd	42.0G		False	locked	
ses-min1	/dev/vda	hdd	42.0G		False	locked	
ses-min1	/dev/vdb	hdd	8192M	387836	False	locked, LVM detected, Insufficient space	(<5GB) on vgs
ses-min2	/dev/vdc	hdd	8192M	450575	True		

## 13.4.2 Apagando dispositivos de disco

Para reutilizar um dispositivo de disco, você precisa apagá-lo (ou *zap*) primeiro:

```
ceph orch device zap HOST_NAME DISK_DEVICE
```

Por exemplo:

```
cephuser@adm > ceph orch device zap ses-min2 /dev/vdc
```



### Nota

Se você já implantou os OSDs por meio de DriveGroups ou da opção `--all-available-devices` sem o flag `unmanaged` definido, o `cephadm` implantará esses OSDs automaticamente depois que você os apagar.

## 13.4.3 Adicionando OSDs por meio da especificação DriveGroups

Os *DriveGroups* especificam os layouts dos OSDs no cluster do Ceph. Eles são definidos em um único arquivo YAML. Nesta seção, usaremos `drive_groups.yml` como exemplo.

Um administrador deve especificar manualmente um grupo de OSDs que estão interligados (OSDs híbridos implantados em uma combinação de HDDs e SDDs) ou compartilhar opções idênticas de implantação (por exemplo, mesmo armazenamento de objetos, mesma opção de criptografia, OSDs independentes). Para evitar a listagem explícita de dispositivos, os DriveGroups usam uma lista de itens de filtro que correspondem a poucos campos selecionados dos relatórios de inventário do **ceph-volume**. O `cephadm` fornecerá o código que converte esses DriveGroups em listas reais de dispositivos para inspeção pelo usuário.

O comando para aplicar a especificação de OSD ao cluster é:

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

Para obter uma visualização das ações e testar seu aplicativo, você pode usar a opção `--dry-run` junto com o comando **ceph orch apply osd**. Por exemplo:

```
cephuser@adm > ceph orch apply osd -i drive_groups.yml --dry-run
...
+-----+-----+-----+-----+-----+
|SERVICE|NAME  |HOST  |DATA      |DB  |WAL  |
+-----+-----+-----+-----+-----+
|osd     |test  |mgr0  |/dev/sda  |-   |-   |
```

```
|osd      |test  |mgr0   |/dev/sdb  |-   |-   |
+-----+-----+-----+-----+-----+-----+
```

Se a saída `--dry-run` atender às suas expectativas, basta executar novamente o comando sem a opção `--dry-run`.

#### 13.4.3.1 OSDs não gerenciados

Todos os dispositivos de disco limpos disponíveis que correspondem à especificação DriveGroups serão usados como OSDs automaticamente depois que você os adicionar ao cluster. Esse comportamento é chamado de modo *gerenciado*.

Para desabilitar o modo *gerenciado*, adicione a linha `unmanaged: true` às especificações relevantes, por exemplo:

```
service_type: osd
service_id: example_drvgrp_name
placement:
  hosts:
    - ses-min2
    - ses-min3
encrypted: true
unmanaged: true
```



#### Dica

Para mudar os OSDs já implantados do modo *gerenciado* para *não gerenciado*, adicione as linhas `unmanaged: true` aos locais aplicáveis durante o procedimento descrito na [Seção 13.1, “Modificando a configuração do cluster”](#).

#### 13.4.3.2 Especificação DriveGroups

Veja a seguir um exemplo do arquivo de especificação DriveGroups:

```
service_type: osd
service_id: example_drvgrp_name
placement:
  host_pattern: '*'
data_devices:
  drive_spec: DEVICE_SPECIFICATION
db_devices:
  drive_spec: DEVICE_SPECIFICATION
```

```
wal_devices:
  drive_spec: DEVICE_SPECIFICATION
block_wal_size: '5G' # (optional, unit suffixes permitted)
block_db_size: '5G' # (optional, unit suffixes permitted)
encrypted: true      # 'True' or 'False' (defaults to 'False')
```



## Nota

A opção antes chamada de "criptografia" no DeepSea foi renomeada para "criptografada". Ao usar o DriveGroups no SUSE Enterprise Storage 7, adote essa nova terminologia na especificação do serviço; do contrário, haverá falha na operação **ceph orch apply**.

### 13.4.3.3 Correspondendo dispositivos de disco

Você pode descrever a especificação usando os seguintes filtros:

- Por um modelo de disco:

```
model: DISK_MODEL_STRING
```

- Por um fornecedor de disco:

```
vendor: DISK_VENDOR_STRING
```



## Dica

Insira *DISK\_VENDOR\_STRING* sempre em letras minúsculas.

Para obter detalhes sobre o modelo e o fornecedor do disco, observe a saída do seguinte comando:

```
cephuser@adm > ceph orch device ls
HOST    PATH    TYPE  SIZE DEVICE_ID                MODEL          VENDOR
ses-min1 /dev/sdb ssd   29.8G SATA_SSD_AF34075704240015  SATA SSD      ATA
ses-min2 /dev/sda ssd   223G Micron_5200_MTFDDAK240TDN  Micron_5200_MTFD ATA
[...]
```

- Se um disco é ou não rotacional. Os SSDs e as unidades NVMe não são rotacionais.

```
rotational: 0
```

- Implante um nó usando *todas* as unidades disponíveis para OSDs:

```
data_devices:  
  all: true
```

- Limite também o número de discos correspondentes:

```
limit: 10
```

#### 13.4.3.4 Filtrando dispositivos por tamanho

Você pode filtrar dispositivos de disco por tamanho, seja por um tamanho exato ou por uma faixa de tamanhos. O parâmetro `size`: aceita argumentos no seguinte formato:

- “10G”: Inclui discos de um tamanho exato.
- “10G:40G”: Inclui discos cujo tamanho está dentro da faixa.
- “:10G”: Inclui discos menores do que ou iguais a 10 GB.
- “40G:”: Inclui discos iguais ou maiores do que 40 GB.

##### EXEMPLO 13.1: CORRESPONDENDO POR TAMANHO DO DISCO

```
service_type: osd  
service_id: example_drvgrp_name  
placement:  
  host_pattern: '*'  
data_devices:  
  size: '40TB:'  
db_devices:  
  size: ':2TB'
```



#### Nota: Aspas obrigatórias

Ao usar o delimitador “:”, você precisa colocar o tamanho entre aspas; do contrário, o sinal “:” será interpretado como um novo hash de configuração.



#### Dica: Atalhos de unidade

Em vez de Gigabytes (G), você pode especificar os tamanhos em Megabytes (M) ou Terabytes (T).



### 13.4.3.5 Exemplos de DriveGroups

Esta seção inclui exemplos de configurações de OSD diferentes.

#### EXEMPLO 13.2: CONFIGURAÇÃO SIMPLES

Este exemplo descreve dois nós com a mesma configuração:

- 20 HDDs
  - Fornecedor: Intel
  - Modelo: SSD-123-foo
  - Tamanho: 4 TB
- 2 SSDs
  - Fornecedor: Micron
  - Modelo: MC-55-44-ZX
  - Tamanho: 512 GB

O arquivo `drive_groups.yml` correspondente será da seguinte maneira:

```
service_type: osd
service_id: example_drvgrp_name
placement:
  host_pattern: '*'
data_devices:
  model: SSD-123-foo
db_devices:
  model: MC-55-44-XZ
```

Essa configuração é simples e válida. O problema é que um administrador pode adicionar discos de diferentes fornecedores no futuro, e eles não serão incluídos. Você pode melhorá-la reduzindo os filtros nas propriedades de núcleo das unidades:

```
service_type: osd
service_id: example_drvgrp_name
placement:
  host_pattern: '*'
data_devices:
  rotational: 1
db_devices:
```

```
rotational: 0
```

No exemplo anterior, impomos todos os dispositivos rotacionais que serão declarados como "dispositivos de dados", e todos os dispositivos não rotacionais serão usados como "dispositivos compartilhados" (wal, db).

Se você sabe que as unidades com mais de 2 TB sempre serão os dispositivos de dados mais lentos, pode filtrar por tamanho:

```
service_type: osd
service_id: example_drvgrp_name
placement:
  host_pattern: '*'
data_devices:
  size: '2TB:'
db_devices:
  size: ':2TB'
```

#### EXEMPLO 13.3: CONFIGURAÇÃO AVANÇADA

Este exemplo descreve duas configurações distintas: 20 HDDs devem compartilhar 2 SSDs, enquanto 10 SSDs devem compartilhar 2 NVMeS.

- 20 HDDs
  - Fornecedor: Intel
  - Modelo: SSD-123-foo
  - Tamanho: 4 TB
- 12 SSDs
  - Fornecedor: Micron
  - Modelo: MC-55-44-ZX
  - Tamanho: 512 GB
- 2 NVMeS
  - Fornecedor: Samsung
  - Modelo: NVME-YYYY-987
  - Tamanho: 256 GB

Essa configuração pode ser definida com dois layouts da seguinte maneira:

```
service_type: osd
service_id: example_drvgrp_name
placement:
  host_pattern: '*'
data_devices:
  rotational: 0
db_devices:
  model: MC-55-44-XZ
```

```
service_type: osd
service_id: example_drvgrp_name2
placement:
  host_pattern: '*'
data_devices:
  model: MC-55-44-XZ
db_devices:
  vendor: samsung
  size: 256GB
```

#### EXEMPLO 13.4: CONFIGURAÇÃO AVANÇADA COM NÓS NÃO UNIFORMES

Os exemplos anteriores consideraram que todos os nós tinham as mesmas unidades. No entanto, esse nem sempre é o caso:

Nós 1-5:

- 20 HDDs
  - Fornecedor: Intel
  - Modelo: SSD-123-foo
  - Tamanho: 4 TB
- 2 SSDs
  - Fornecedor: Micron
  - Modelo: MC-55-44-ZX
  - Tamanho: 512 GB

Nós 6-10:

- 5 NVMe

- Fornecedor: Intel
- Modelo: SSD-123-foo
- Tamanho: 4 TB
- 20 SSDs
  - Fornecedor: Micron
  - Modelo: MC-55-44-ZX
  - Tamanho: 512 GB

Você pode usar a chave de “destino” no layout para direcionar nós específicos. A notação de destino do Salt ajuda a simplificar as coisas:

```
service_type: osd
service_id: example_drvgrp_one2five
placement:
  host_pattern: 'node[1-5]'
data_devices:
  rotational: 1
db_devices:
  rotational: 0
```

seguido de

```
service_type: osd
service_id: example_drvgrp_rest
placement:
  host_pattern: 'node[6-10]'
data_devices:
  model: MC-55-44-XZ
db_devices:
  model: SSD-123-foo
```

#### EXEMPLO 13.5: CONFIGURAÇÃO TÉCNICA

Todos os casos anteriores consideraram que os WALs e BDs usavam o mesmo dispositivo. No entanto, também é possível implantar o WAL em um dispositivo dedicado:

- 20 HDDs

- Fornecedor: Intel
- Modelo: SSD-123-foo
- Tamanho: 4 TB
- 2 SSDs
  - Fornecedor: Micron
  - Modelo: MC-55-44-ZX
  - Tamanho: 512 GB
- 2 NVMeS
  - Fornecedor: Samsung
  - Modelo: NVME-QQQQ-987
  - Tamanho: 256 GB

```
service_type: osd
service_id: example_drvgrp_name
placement:
  host_pattern: '*'
data_devices:
  model: MC-55-44-XZ
db_devices:
  model: SSD-123-foo
wal_devices:
  model: NVME-QQQQ-987
```

#### EXEMPLO 13.6: CONFIGURAÇÃO COMPLEXA (E IMPROVÁVEL)

Na configuração a seguir, tentamos definir:

- 20 HDDs com 1 NVMe
- 2 HDDs com 1 SSD(db) e 1 NVMe(wal)
- 8 SSDs com 1 NVMe
- 2 SSDs independentes (criptografados)
- 1 HDD é sobressalente e não deve ser implantado

Veja a seguir o resumo das unidades usadas:

- 23 HDDs
  - Fornecedor: Intel
  - Modelo: SSD-123-foo
  - Tamanho: 4 TB
- 10 SSDs
  - Fornecedor: Micron
  - Modelo: MC-55-44-ZX
  - Tamanho: 512 GB
- 1 NVMe
  - Fornecedor: Samsung
  - Modelo: NVME-QQQQ-987
  - Tamanho: 256 GB

A definição dos DriveGroups será a seguinte:

```
service_type: osd
service_id: example_drvgrp_hdd_nvme
placement:
  host_pattern: '*'
data_devices:
  rotational: 0
db_devices:
  model: NVME-QQQQ-987
```

```
service_type: osd
service_id: example_drvgrp_hdd_ssd_nvme
placement:
  host_pattern: '*'
data_devices:
  rotational: 0
db_devices:
  model: MC-55-44-XZ
wal_devices:
```

```
model: NVME-QQQQ-987
```

```
service_type: osd
service_id: example_drvgrp_ssd_nvme
placement:
  host_pattern: '*'
data_devices:
  model: SSD-123-foo
db_devices:
  model: NVME-QQQQ-987
```

```
service_type: osd
service_id: example_drvgrp_standalone_encrypted
placement:
  host_pattern: '*'
data_devices:
  model: SSD-123-foo
encrypted: True
```

Um HDD permanecerá enquanto o arquivo está sendo analisado de cima para baixo.

### 13.4.4 Removendo OSDs

Antes de remover um nó OSD do cluster, verifique se o cluster tem mais espaço livre em disco do que o disco OSD que será removido. Saiba que a remoção de um OSD provoca a redistribuição do cluster inteiro.

1. Identifique o OSD que será removido obtendo seu ID:

```
cephuser@adm > ceph orch ps --daemon_type osd
NAME    HOST          STATUS          REFRESHED  AGE  VERSION
osd.0   target-ses-090 running (3h)    7m ago     3h   15.2.7.689 ...
osd.1   target-ses-090 running (3h)    7m ago     3h   15.2.7.689 ...
osd.2   target-ses-090 running (3h)    7m ago     3h   15.2.7.689 ...
osd.3   target-ses-090 running (3h)    7m ago     3h   15.2.7.689 ...
```

2. Remova um ou mais OSDs do cluster:

```
cephuser@adm > ceph orch osd rm OSD1_ID OSD2_ID ...
```

Por exemplo:

```
cephuser@adm > ceph orch osd rm 1 2
```

### 3. Você pode consultar o estado da operação de remoção:

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

OSD_ID	HOST	STATE	PG_COUNT	REPLACE	FORCE	STARTED_AT
2	cephadm-dev	done, waiting for purge	0	True	False	2020-07-17 13:01:43.147684
3	cephadm-dev	draining	17	False	True	2020-07-17 13:01:45.162158
4	cephadm-dev	started	42	False	True	2020-07-17 13:01:45.162158

#### 13.4.4.1 Interrompendo a remoção do OSD

Após programar uma remoção do OSD, você poderá interrompê-la, se necessário. O comando a seguir redefinirá o estado inicial do OSD e o removerá da fila:

```
cephuser@adm > ceph orch osd rm stop OSD_SERVICE_ID
```

#### 13.4.5 Substituindo OSDs

Há várias razões pelas quais você pode precisar substituir um disco OSD. Por exemplo:

- Houve falha no disco OSD ou ele está prestes a falhar com base nas informações do SMART e não poderá mais ser usado para armazenar dados com segurança.
- Por exemplo, você precisa fazer upgrade do disco OSD para aumentar o tamanho dele.
- Você precisa mudar o layout do disco OSD.
- Você planeja mudar de um layout não LVM para um layout baseado em LVM.

Para substituir um OSD mantendo seu ID, execute:

```
cephuser@adm > ceph orch osd rm OSD_SERVICE_ID --replace
```

Por exemplo:

```
cephuser@adm > ceph orch osd rm 4 --replace
```

A substituição de um OSD é idêntica à remoção de um OSD (consulte a [Seção 13.4.4, “Removendo OSDs”](#) para obter mais detalhes), exceto que o OSD não é permanentemente removido da hierarquia CRUSH e recebe um flag destroyed.



O flag `destroyed` é usado para determinados IDs de OSD que serão reutilizados durante a próxima implantação de OSD. Os discos recém-adicionados que corresponderem à especificação `DriveGroups` (consulte a [Seção 13.4.3, “Adicionando OSDs por meio da especificação `DriveGroups`”](#) para obter mais detalhes) receberão os IDs de OSD da contraparte substituída.



### Dica

Anexar a opção `--dry-run` não executará a substituição real, mas apresentará uma visualização das etapas que normalmente ocorrem.



### Nota

No caso de substituir um OSD após uma falha, é altamente recomendável acionar uma remoção profunda dos grupos de posicionamento. Visite a [Seção 17.6, “Depurando grupos de posicionamento”](#) para obter mais detalhes.

Execute o seguinte comando para iniciar uma remoção profunda:

```
cephuser@adm > ceph osd deep-scrub osd.OSD_NUMBER
```



### Importante: Falha no dispositivo compartilhado

Em caso de falha em um dispositivo compartilhado para BD/WAL, você precisará executar o procedimento de substituição para todos os OSDs que compartilham o dispositivo com falha.

## 13.5 Movendo o Master Salt para um novo nó

Se você precisar substituir o host Master Salt por um novo, siga estas etapas:

1. Exporte a configuração do cluster e faça backup do arquivo JSON exportado. Encontre mais detalhes na *Livro “Guia de Implantação”, Capítulo 7 “Implantando o cluster de boot usando ceph-salt”, Seção 7.2.14 “Exportando as configurações do cluster”*.
2. Se o Master Salt antigo também for o único nó de administração no cluster, mova manualmente `/etc/ceph/ceph.client.admin.keyring` e `/etc/ceph/ceph.conf` para o novo Master Salt.

3. Pare e desabilite o serviço `systemd` do Master Salt no nó do Master Salt antigo:

```
root@master # systemctl stop salt-master.service
root@master # systemctl disable salt-master.service
```

4. Se o nó do Master Salt antigo não estiver mais no cluster, também pare e desabilite o serviço `systemd` do Minion Salt:

```
root@master # systemctl stop salt-minion.service
root@master # systemctl disable salt-minion.service
```



### Atenção

Não pare nem desabilite o `salt-minion.service` se o nó do Master Salt antigo tiver daemons do Ceph (MON, MGR, OSD, MDS, gateway, monitoramento) em execução.

5. Instale o SUSE Linux Enterprise Server 15 SP3 no novo Master Salt seguindo o procedimento descrito no Livro *“Guia de Implantação”, Capítulo 5 “Instalando e configurando o SUSE Linux Enterprise Server”*.



### Dica: Transição de Minion Salt

Para simplificar a transição dos Minions Salt para o novo Master Salt, remova a chave pública do Master Salt original de cada um deles:

```
root@minion > rm /etc/salt/pki/minion/minion_master.pub
root@minion > systemctl restart salt-minion.service
```

6. Instale o pacote `salt-master` e, se aplicável, o pacote `salt-minion` no novo Master Salt.
7. Instale o `ceph-salt` no novo nó do Master Salt:

```
root@master # zypper install ceph-salt
root@master # systemctl restart salt-master.service
root@master # salt '*' saltutil.sync_all
```



### Importante

Execute todos os três comandos antes de continuar. Os comandos são idempotentes; não faz diferença se eles são repetidos.

8. Inclua o novo Master Salt no cluster, conforme descrito no Livro *“Guia de Implantação”, Capítulo 7 “Implantando o cluster de boot usando ceph-salt”, Seção 7.1 “Instalando ceph-salt”, Livro “Guia de Implantação”, Capítulo 7 “Implantando o cluster de boot usando ceph-salt”, Seção 7.2.2 “Adicionando minions Salt” e Livro “Guia de Implantação”, Capítulo 7 “Implantando o cluster de boot usando ceph-salt”, Seção 7.2.4 “Especificando o nó de admin”.*
9. Importe a configuração do cluster de backup e aplique-a:

```
root@master # ceph-salt import CLUSTER_CONFIG.json
root@master # ceph-salt apply
```



### Importante

Renomeie o `minion id` do Master Salt no arquivo `CLUSTER_CONFIG.json` exportado antes de importá-lo.

## 13.6 Atualizando os nós do cluster

Mantenha os nós do cluster do Ceph atualizados aplicando as atualizações sequenciais regularmente.

### 13.6.1 Repositórios do software

Antes de corrigir o cluster com os pacotes de software mais recentes, verifique se todos os nós do cluster têm acesso aos repositórios relevantes. Consulte o Livro *“Guia de Implantação”, Capítulo 10 “Upgrade do SUSE Enterprise Storage 6 para 7.1”, Seção 10.1.5.1 “Repositórios do software”* para obter uma lista completa dos repositórios necessários.

### 13.6.2 Propagação em fases do repositório

Se você usa uma ferramenta de propagação em fases; por exemplo, SUSE Manager ou SMT (Subscription Management Tool), que processa os repositórios do software nos nós do cluster, verifique se as fases dos dois repositórios "Updates" para o SUSE Linux Enterprise Server e o SUSE Enterprise Storage foram criadas no mesmo momento.

É altamente recomendável usar uma ferramenta de propagação em fases para aplicar os patches que têm níveis congelados ou em fases. Isso garante que os novos nós que ingressarem no cluster tenham o mesmo nível de patch que os nós que já estão em execução no cluster. Dessa forma, você não precisa aplicar os patches mais recentes a todos os nós do cluster antes que os novos nós possam ingressar no cluster.

### 13.6.3 Tempo de espera dos serviços do Ceph

Dependendo da configuração, os nós do cluster podem ser reinicializados durante a atualização. Se houver um ponto único de falha para os serviços, como Gateway de Objetos, Samba Gateway, NFS Ganesha ou iSCSI, as máquinas cliente poderão ser temporariamente desconectadas dos serviços cujos nós estão sendo reinicializados.

### 13.6.4 Executando a atualização

Para atualizar os pacotes de software em todos os nós do cluster para a versão mais recente, execute o seguinte comando:

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

## 13.7 Atualizando o Ceph

Você pode instruir o cephadm a atualizar o Ceph de uma versão de correção de bug para outra. A atualização automatizada dos serviços do Ceph respeita a ordem recomendada: ela começa com os Ceph Managers, os Ceph Monitors e continua com os outros serviços, como Ceph OSDs, Servidores de Metadados e Gateways de Objetos. Cada daemon será reiniciado apenas depois que o Ceph indicar que o cluster permanecerá disponível.



#### Nota

O procedimento de atualização a seguir usa o comando **`ceph orch upgrade`**. Lembre-se de que as instruções a seguir detalham como atualizar o cluster do Ceph com uma versão do produto (por exemplo, uma atualização de manutenção) e *não* fornecem instruções sobre como fazer upgrade do cluster de uma versão do produto para outra.

## 13.7.1 Iniciando a atualização

Antes de iniciar a atualização, verifique se todos os nós estão online e se o cluster está saudável:

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

Para atualizar para uma versão específica do Ceph:

```
cephuser@adm > ceph orch upgrade start --image REGISTRY_URL
```

Por exemplo:

```
cephuser@adm > ceph orch upgrade start --image registry.suse.com/ses/7.1/ceph/ceph:latest
```

Fazer upgrade de pacotes nos hosts:

```
cephuser@adm > ceph-salt update
```

## 13.7.2 Monitorando a atualização

Execute o seguinte comando para determinar se uma atualização está em andamento:

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

Enquanto a atualização estiver em andamento, você verá uma barra de andamento na saída de status do Ceph:

```
cephuser@adm > ceph -s
[...]
progress:
  Upgrade to registry.suse.com/ses/7.1/ceph/ceph:latest (00h 20m 12s)
    [=====.....] (time remaining: 01h 43m 31s)
```

Você também pode observar o registro do cephadm:

```
cephuser@adm > ceph -W cephadm
```

## 13.7.3 Cancelando uma atualização

Você pode interromper o processo de atualização a qualquer momento:

```
cephuser@adm > ceph orch upgrade stop
```

## 13.8 Parando ou reiniciando o cluster

Em alguns casos, talvez seja necessário parar ou reinicializar o cluster inteiro. Recomendamos verificar com cuidado as dependências dos serviços em execução. As seguintes etapas apresentam uma descrição de como parar e iniciar o cluster:

1. Especifique para o cluster do Ceph não marcar os OSDs com o flag “out”:

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

2. Pare os daemons e os nós na seguinte ordem:

1. Clientes de armazenamento
2. Gateways. Por exemplo, NFS Ganesha ou Gateway de Objetos
3. Servidor de Metadados
4. Ceph OSD
5. Ceph Manager
6. Ceph Monitor

3. Se necessário, execute as tarefas de manutenção.

4. Inicie os nós e os servidores na ordem inversa do processo de encerramento:

1. Ceph Monitor
2. Ceph Manager
3. Ceph OSD
4. Servidor de Metadados
5. Gateways. Por exemplo, NFS Ganesha ou Gateway de Objetos
6. Clientes de armazenamento

5. Remova o flag “noout”:

```
cephuser@adm > ceph osd unset noout
```

## 13.9 Removendo um cluster inteiro do Ceph

O comando **`ceph-salt purge`** remove todo o cluster do Ceph. Se houver mais clusters do Ceph implantados, será purgado aquele que for relatado pelo **`ceph -s`**. Desta forma, você pode limpar o ambiente do cluster ao testar configurações diferentes.

Para evitar a exclusão acidental, a orquestração verifica se a segurança está desligada. Você pode desligar as medidas de segurança e remover o cluster do Ceph executando:

```
root@master # ceph-salt disengage-safety
root@master # ceph-salt purge
```

## 14 Operação de serviços do Ceph

Você pode operar os serviços do Ceph no nível do daemon, nó ou cluster. Dependendo da abordagem necessária, use o comando `cephadm` ou `systemctl`.

### 14.1 Operando serviços individuais

Se você precisa operar um serviço individual, identifique-o primeiro:

```
cephuser@adm > ceph orch ps
```

NAME	HOST	STATUS	REFRESHED	[...]
mds.my_cephfs.ses-min1.oterul	ses-min1	running (5d)	8m ago	
mgr.ses-min1.gpijpm	ses-min1	running (5d)	8m ago	
mgr.ses-min2.oopvyh	ses-min2	running (5d)	8m ago	
mon.ses-min1	ses-min1	running (5d)	8m ago	
mon.ses-min2	ses-min2	running (5d)	8m ago	
mon.ses-min4	ses-min4	running (5d)	7m ago	
osd.0	ses-min2	running (61m)	8m ago	
osd.1	ses-min3	running (61m)	7m ago	
osd.2	ses-min4	running (61m)	7m ago	
rgw.myrealm.myzone.ses-min1.kwazo	ses-min1	running (5d)	8m ago	
rgw.myrealm.myzone.ses-min2.jngabw	ses-min2	error	8m ago	

Para identificar um serviço em um nó específico, execute:

```
ceph orch ps NODE_HOST_NAME
```

Por exemplo:

```
cephuser@adm > ceph orch ps ses-min2
```

NAME	HOST	STATUS	REFRESHED
mgr.ses-min2.oopvyh	ses-min2	running (5d)	3m ago
mon.ses-min2	ses-min2	running (5d)	3m ago
osd.0	ses-min2	running (67m)	3m ago



#### Dica

O comando `ceph orch ps` suporta vários formatos de saída. Para mudá-lo, anexe a opção `--format FORMAT`, em que *FORMAT* é `json`, `json-pretty` ou `yaml`. Por exemplo:

```
cephuser@adm > ceph orch ps --format yaml
```



Depois de saber o nome do serviço, você poderá iniciá-lo, reiniciá-lo ou pará-lo:

```
ceph orch daemon COMMAND SERVICE_NAME
```

Por exemplo, para reiniciar o serviço OSD com ID 0, execute:

```
cephuser@adm > ceph orch daemon restart osd.0
```

## 14.2 Operando tipos de serviço

Se você precisar operar um tipo específico de serviço em todo o cluster do Ceph, use o seguinte comando:

```
ceph orch COMMAND SERVICE_TYPE
```

Substitua *COMMAND* por start, stop ou restart.

Por exemplo, o comando a seguir reinicia todos os MONs no cluster, sejam quais forem os nós em que eles são executados de fato:

```
cephuser@adm > ceph orch restart mon
```

## 14.3 Operando serviços em um único nó

Usando o comando **systemctl**, você pode operar os serviços e destinos do systemd relacionados ao Ceph em um único nó.

### 14.3.1 Identificando serviços e destinos

Antes de operar os serviços e destinos do systemd relacionados ao Ceph, você precisa identificar os nomes dos seus arquivos unitários. Os nomes de arquivo dos serviços têm o seguinte padrão:

```
ceph-FSID@SERVICE_TYPE.ID.service
```

Por exemplo:

```
ceph-b4b30c6e-9681-11ea-ac39-525400d7702d@mon.doc-ses-min1.service
```

```
ceph-b4b30c6e-9681-11ea-ac39-525400d7702d@rgw.myrealm.myzone.doc-ses-min1.kwwazo.service
```

#### FSID

ID exclusivo do cluster do Ceph. Você pode encontrá-lo na saída do comando **`ceph fsid`**.

#### SERVICE\_TYPE

Tipo de serviço, por exemplo `osd`, `mon` ou `rgw`.

#### ID

String de identificação do serviço. Para os OSDs, esse é o número de ID do serviço. Para outros serviços, ele pode ser um nome de host do nó ou strings adicionais relevantes ao tipo de serviço.



#### Dica

O `SERVICE_TYPE`. A parte `ID` é idêntica ao conteúdo da coluna `NAME` na saída do comando **`ceph orch ps`**.

### 14.3.2 Operando todos os serviços em um nó

Ao usar os destinos do `systemd` do Ceph, você pode ao mesmo tempo operar *todos* os serviços em um nó ou todos os serviços que *pertencem a um cluster* identificado por seu `FSID`.

Por exemplo, para parar todos os serviços do Ceph em um nó, isija qual for o cluster ao qual eles pertençam, execute:

```
root@minion > systemctl stop ceph.target
```

Para reiniciar todos os serviços que pertencem a um cluster do Ceph com ID `b4b30c6e-9681-11ea-ac39-525400d7702d`, execute:

```
root@minion > systemctl restart ceph-b4b30c6e-9681-11ea-ac39-525400d7702d.target
```

### 14.3.3 Operando um serviço individual em um nó

Depois de identificar o nome de um serviço específico, opere-o da seguinte maneira:

```
systemctl COMMAND SERVICE_NAME
```

Por exemplo, para reiniciar um único serviço OSD com ID 1 em um cluster com ID b4b30c6e-9681-11ea-ac39-525400d7702d, execute:

```
# systemctl restart ceph-b4b30c6e-9681-11ea-ac39-525400d7702d@osd.1.service
```

### 14.3.4 Consultando o status do serviço

É possível consultar o `systemd` para saber o status dos serviços. Por exemplo:

```
# systemctl status ceph-b4b30c6e-9681-11ea-ac39-525400d7702d@osd.0.service
```

## 14.4 Encerrando e reiniciando todo o cluster do Ceph

Pode ser necessário encerrar e reiniciar o cluster em caso de queda de energia planejada. Para parar todos os serviços relacionados ao Ceph e reiniciá-los sem problemas, siga as etapas abaixo.

#### PROCEDIMENTO 14.1: ENCERRANDO TODO O CLUSTER DO CEPH

1. Encerre ou desconecte todos os clientes que acessam o cluster.
2. Para impedir que o CRUSH reequilibre automaticamente o cluster, defina o cluster como noout:

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

3. Pare todos os serviços do Ceph em todos os nós do cluster:

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

4. Desligue todos os nós do cluster:

```
root@master # salt -G 'ceph-salt:member' cmd.run "shutdown -h"
```

#### PROCEDIMENTO 14.2: INICIANDO TODO O CLUSTER DO CEPH

1. Ligue o Nó de Admin.
2. Ligue os nós do Ceph Monitor.
3. Ligue os nós do Ceph OSD.

4. Cancele a definição do flag noout:

```
root@master # ceph osd unset noout
```

5. Ligue todos os gateways configurados.
6. Ligue ou conecte os clientes do cluster.

## 15 Backup e restauração

Este capítulo explica quais partes do cluster do Ceph devem ser incluídas no backup para que seja possível restaurar a funcionalidade dele.

### 15.1 Fazer backup da configuração e dos dados do cluster

#### 15.1.1 Fazer backup da configuração do ceph-salt

Exporte a configuração do cluster. Mais informações podem ser encontradas em *Livro “Guia de Implantação”, Capítulo 7 “Implantando o cluster de boot usando ceph-salt”, Seção 7.2.14 “Exportando as configurações do cluster”*.

#### 15.1.2 Fazer backup da configuração do Ceph

Faça backup do diretório `/etc/ceph`. Ele inclui a configuração essencial do cluster. Por exemplo, você precisará fazer backup do `/etc/ceph` quando for necessário substituir o Nó de Admin.

#### 15.1.3 Fazer backup da configuração do Salt

Você precisa fazer backup do diretório `/etc/salt/`. Ele contém os arquivos de configuração do Salt, por exemplo, a chave do Master Salt e as chaves aceitas do cliente.

Os arquivos do Salt não são estritamente necessários para fazer backup do Nó de Admin, mas facilitam a reimplantação do cluster do Salt. Se não houver nenhum backup desses arquivos, os minions Salt precisarão ser registrados novamente no novo Nó de Admin.



#### Nota: Segurança da chave privada master do Salt

Garanta que o backup da chave privada do Master Salt seja armazenada em um local seguro. A chave do Master Salt pode ser usada para manipular todos os nós do cluster.

### 15.1.4 Fazer backup das configurações personalizadas

- Dados e personalização do Prometheus.
- Personalização do Grafana.
- Mudanças manuais na configuração do iSCSI.
- Chaves do Ceph.
- Mapa CRUSH e regras CRUSH. Execute o comando a seguir para gravar o Mapa CRUSH descompilado incluindo as regras CRUSH no `crushmap-backup.txt`:

```
cephuser@adm > ceph osd getcrushmap | crushtool -d - -o crushmap-backup.txt
```

- Configuração do Gateway do Samba. Se você usa um único gateway, faça backup do `/etc/samba/smb.conf`. Se você usa uma configuração de HA, faça backup também dos arquivos de configuração do CTDB e do Pacemaker. Consulte o [Capítulo 24, Exportar dados do Ceph por meio do Samba](#) para obter detalhes sobre a configuração que é usada pelos Gateways do Samba.
- Configuração do NFS Ganesha. Necessário apenas ao usar uma configuração de HA. Consulte o [Capítulo 25, NFS Ganesha](#) para obter detalhes sobre a configuração que é usada pelo NFS Ganesha.

## 15.2 Restaurando um nó do Ceph

O procedimento para recuperar um nó do backup é reinstalar o nó, substituir seus arquivos de configuração e, em seguida, reorquestrar o cluster para que o nó de substituição seja adicionado novamente.

Se você precisar reimplantar o Nó de Admin, consulte a [Seção 13.5, “Movendo o Master Salt para um novo nó”](#).

Para os minions, geralmente é mais fácil apenas reconstruir e reimplantar.

1. Reinstale o nó. Mais informações podem ser encontradas em *Livro “Guia de Implantação”, Capítulo 5 “Instalando e configurando o SUSE Linux Enterprise Server”*
2. Instale o Salt. Encontre mais informações no *Livro “Guia de Implantação”, Capítulo 6 “Implantando o Salt”*.

3. Após restaurar o diretório `/etc/salt` de um backup, habilite e reinicie os serviços do Salt aplicáveis, por exemplo:

```
root@master # systemctl enable salt-master
root@master # systemctl start salt-master
root@master # systemctl enable salt-minion
root@master # systemctl start salt-minion
```

4. Remova a chave master pública do nó do Master Salt antigo de todos os minions.

```
root@master # rm /etc/salt/pki/minion/minion_master.pub
root@master # systemctl restart salt-minion
```

5. Restaure para o Nó de Admin tudo o que era local.
6. Importe a configuração do cluster do arquivo JSON exportado anteriormente. Consulte Livro *“Guia de Implantação”, Capítulo 7 “Implantando o cluster de boot usando ceph-salt”, Seção 7.2.14 “Exportando as configurações do cluster”* para obter mais detalhes.
7. Aplique a configuração importada do cluster:

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

## 16 Monitoramento e alerta

No SUSE Enterprise Storage 7.1, o `cephadm` implanta uma pilha de monitoramento e alerta. Os usuários precisam definir os serviços (como Prometheus, Alertmanager e Grafana) que desejam implantar com o `cephadm` em um arquivo de configuração YAML ou eles podem usar a CLI para implantá-los. Quando vários serviços do mesmo tipo são implantados, uma configuração altamente disponível é implantada. O exportador de nós é uma exceção a essa regra.

É possível usar o `cephadm` para implantar os seguintes serviços de monitoramento:

- **Prometheus** é o kit de ferramentas de monitoramento e alerta. Ele coleta os dados fornecidos pelos exportadores do Prometheus e dispara alertas pré-configurados se os limites predefinidos forem atingidos.
- O **Alertmanager** processa os alertas enviados pelo servidor Prometheus. Ele elimina a duplicação, agrupa e roteia os alertas para o receptor correto. Por padrão, o Ceph Dashboard será configurado automaticamente como o receptor.
- **Grafana** é o software de visualização e alerta. A funcionalidade de alerta do Grafana não é usada por esta pilha de monitoramento. Para alertas, o Alertmanager é usado.
- O **exportador de nós** do Prometheus é que fornece os dados sobre o nó em que ele está instalado. É recomendável instalar o exportador de nós em todos os nós.

O Módulo do Gerenciador do Prometheus inclui um exportador do Prometheus para transmitir os contadores de desempenho do Ceph do ponto de coleta no `ceph-mgr`.

A configuração do Prometheus, incluindo os destinos de *scrape* (daemons que extraem métricas), é definida automaticamente pelo `cephadm`. O `cephadm` também implanta uma lista de alertas padrão, por exemplo, erro de saúde, 10% dos OSDs inativos ou páginas inativas.

Por padrão, o tráfego para o Grafana é criptografado com TLS. Você pode fornecer seu próprio certificado TLS ou usar um certificado autoassinado. Se nenhum certificado personalizado foi configurado antes da implantação do Grafana, um certificado autoassinado é criado e configurado automaticamente para o Grafana.

Você pode configurar certificados personalizados para o Grafana seguindo estas etapas:

### 1. Configure os arquivos de certificado:

```
cephuser@adm > ceph config-key set mgr/cephadm/grafana_key -i $PWD/key.pem
cephuser@adm > ceph config-key set mgr/cephadm/grafana.crt -i $PWD/certificate.pem
```



## 2. Reinicie o serviço Ceph Manager:

```
cephuser@adm > ceph orch restart mgr
```

## 3. Reconfigure o serviço Grafana para refletir os caminhos dos novos certificados e defina o URL correto para o Ceph Dashboard:

```
cephuser@adm > ceph orch reconfig grafana
```

O Alertmanager processa os alertas enviados pelo servidor Prometheus. Ele cuida da eliminação de duplicação, do agrupamento e do processamento deles para o receptor correto. É possível usar o Alertmanager para silenciar os alertas, mas as configurações de silenciamento também podem ser gerenciadas no Ceph Dashboard.

Recomendamos que o `Node exporter` seja implantado em todos os nós. Isso pode ser feito usando o arquivo `monitoring.yaml` com o tipo de serviço `node-exporter`. Consulte o *Livro "Guia de Implantação", Capítulo 8 "Implantando os serviços principais restantes com o cephadm", Seção 8.3.8 "Implantando a pilha de monitoramento"* para obter mais informações sobre como implantar serviços.

## 16.1 Configurando imagens personalizadas ou locais



### Dica

Esta seção descreve como mudar a configuração das imagens de container usadas quando os serviços são implantados ou atualizados. Ela não inclui os comandos necessários para implantar ou reimplantar os serviços.

O método recomendado para implantar a pilha de monitoramento é aplicando a respectiva especificação, conforme descrito no *Livro "Guia de Implantação", Capítulo 8 "Implantando os serviços principais restantes com o cephadm", Seção 8.3.8 "Implantando a pilha de monitoramento"*.

Para implantar imagens de container personalizadas ou locais, elas precisam ser definidas no cephadm. Para fazer isso, você precisa executar o seguinte comando:

```
cephuser@adm > ceph config set mgr mgr/cephadm/OPTION_NAME VALUE
```

Em que *OPTION\_NAME* é qualquer um dos seguintes nomes:

- container\_image\_prometheus
- container\_image\_node\_exporter
- container\_image\_alertmanager
- container\_image\_grafana

Se nenhuma opção for definida ou se a configuração for removida, as seguintes imagens serão usadas como *VALUE*:

- registry.suse.com/ses/7.1/ceph/prometheus-server:2.32.1
- registry.suse.com/ses/7.1/ceph/prometheus-node-exporter:1.1.2
- registry.suse.com/ses/7.1/ceph/prometheus-alertmanager:0.21.0
- registry.suse.com/ses/7.1/ceph/grafana:7.5.12

Por exemplo:

```
cephuser@adm > ceph config set mgr mgr/cephadm/container_image_prometheus prom/  
prometheus:v1.4.1
```



## Nota

Ao definir uma imagem personalizada, o valor padrão será anulado (mas não sobregravado). O valor padrão muda quando as atualizações ficam disponíveis. Ao definir uma imagem personalizada, você não poderá atualizar o componente para o qual definiu a imagem personalizada automaticamente. Você precisará atualizar manualmente a configuração (nome e tag da imagem) para poder instalar as atualizações.

Em vez disso, se você seguir as recomendações, poderá redefinir a imagem personalizada. Após esse procedimento, o valor padrão será usado novamente. Use **ceph config rm** para redefinir a opção de configuração:

```
cephuser@adm > ceph config rm mgr mgr/cephadm/OPTION_NAME
```

Por exemplo:

```
cephuser@adm > ceph config rm mgr mgr/cephadm/container_image_prometheus
```

## 16.2 Atualizando os serviços de monitoramento

Conforme mencionado na [Seção 16.1, “Configurando imagens personalizadas ou locais”](#), o `cephadm` é fornecido com os URLs das imagens de container recomendadas e testadas, e eles são usados por padrão.

Quando há atualizações dos pacotes do Ceph, novas versões desses URLs podem ser fornecidas. Isso apenas atualiza o local de onde as imagens de container são extraídas, mas não atualiza os serviços.

Depois que os URLs para as novas imagens de container forem atualizados, seja manualmente (conforme descrito na [Seção 16.1, “Configurando imagens personalizadas ou locais”](#)) ou automaticamente por meio de uma atualização do pacote do Ceph, os serviços de monitoramento poderão ser atualizados.

Para fazer isso, use `ceph orch reconfig` da seguinte maneira:

```
cephuser@adm > ceph orch reconfig node-exporter
cephuser@adm > ceph orch reconfig prometheus
cephuser@adm > ceph orch reconfig alertmanager
cephuser@adm > ceph orch reconfig grafana
```

Atualmente, não existe um único comando para atualizar todos os serviços de monitoramento. A ordem em que esses serviços são atualizados não é importante.



### Nota

Se você usar imagens de container personalizadas, os URLs especificados para os serviços de monitoramento não serão automaticamente modificados se os pacotes do Ceph forem atualizados. Se você especificou imagens de container personalizadas, precisa especificar os URLs das novas imagens de container manualmente. Esse poderá ser o caso se você usar um registro de container local.

Você encontra os URLs das imagens de container recomendadas para uso na [seção Seção 16.1, “Configurando imagens personalizadas ou locais”](#).

## 16.3 Desabilitando o monitoramento

Para desabilitar a pilha de monitoramento, execute os seguintes comandos:

```
cephuser@adm > ceph orch rm grafana
```

```
cephuser@adm > ceph orch rm prometheus --force # this will delete metrics data
collected so far
cephuser@adm > ceph orch rm node-exporter
cephuser@adm > ceph orch rm alertmanager
cephuser@adm > ceph mgr module disable prometheus
```

## 16.4 Configurando o Grafana

O back end do Ceph Dashboard requer que o URL do Grafana possa verificar a existência de Grafana Dashboards antes mesmo de serem carregados pelo front end. Devido à natureza da implementação do Grafana no Ceph Dashboard, isso significa que duas conexões de trabalho são necessárias para poder ver os gráficos do Grafana no Ceph Dashboard:

- O back end (módulo MGR do Ceph) precisa verificar a existência do gráfico solicitado. Se essa solicitação for bem-sucedida, ela informará ao front end que ele pode acessar o Grafana com segurança.
- Em seguida, o front end solicita os gráficos do Grafana diretamente do browser do usuário usando um iframe. A instância do Grafana é acessada diretamente sem qualquer desvio pelo Ceph Dashboard.

Agora, talvez seja o caso de o seu ambiente dificultar o acesso direto do browser do usuário ao URL configurado no Ceph Dashboard. Para resolver esse problema, é possível configurar um URL separado que será usado exclusivamente para informar ao front end (o browser do usuário) qual URL ele deve usar para acessar o Grafana.

Para mudar o URL que é retornado ao front end, execute o seguinte comando:

```
cephuser@adm > ceph dashboard set-grafana-frontend-api-url GRAFANA-SERVER-URL
```

Se nenhum valor for definido para essa opção, ela simplesmente retornará para o valor da opção `GRAFANA_API_URL`, que é definida automaticamente e atualizada com frequência pelo `cephadm`. Se definida, ela instruirá o browser a usar esse URL para acessar o Grafana.

## 16.5 Configurando o módulo do gerenciador do Prometheus

O Módulo do Gerenciador do Prometheus é um módulo do Ceph que estende a funcionalidade do Ceph. O módulo lê os (meta)dados do Ceph sobre seu estado e saúde, fornecendo os dados (extraídos) em um formato consumível pelo Prometheus.



### Nota

O Módulo do Gerenciador do Prometheus precisa ser reiniciado para que as mudanças de configuração sejam aplicadas.

### 16.5.1 Configurando a interface de rede

Por padrão, o Módulo do Gerenciador do Prometheus aceita solicitações HTTP na porta 9283 em todos os endereços IPv4 e IPv6 no host. A porta e o endereço de escuta podem ser configurados usando o `ceph config-key set`, com as chaves `mgr/prometheus/server_addr` e `mgr/prometheus/server_port`. Essa porta está registrada no registro do Prometheus.

Para atualizar o `server_addr`, execute o seguinte comando:

```
cephuser@adm > ceph config set mgr mgr/prometheus/server_addr 0.0.0.0
```

Para atualizar o `server_port`, execute o seguinte comando:

```
cephuser@adm > ceph config set mgr mgr/prometheus/server_port 9283
```

### 16.5.2 Configurando o `scrape_interval`

Por padrão, o Módulo do Gerenciador do Prometheus está configurado com um intervalo de extração de 15 segundos. Não é recomendável usar um intervalo de extração abaixo de 10 segundos. Para definir um intervalo de extração diferente no módulo do Prometheus, defina `scrape_interval` com o valor desejado:



### Importante

Para funcionar corretamente e não causar problemas, o `scrape_interval` desse módulo deve ser definido sempre com o mesmo valor do intervalo de extração do Prometheus.

```
cephuser@adm > ceph config set mgr mgr/prometheus/scrape_interval 15
```

### 16.5.3 Configurando o cache

Em clusters grandes (mais de 1.000 OSDs), o tempo para buscar as métricas pode aumentar significativamente. Sem o cache, o Módulo do Gerenciador do Prometheus pode sobrecarregar o gerenciador e fazer com que as instâncias do Ceph Manager não respondam ou travem. Como resultado, o cache é habilitado por padrão e não pode ser desabilitado, mas isso significa que o cache pode se tornar obsoleto. O cache é considerado obsoleto quando o tempo para buscar as métricas do Ceph excede o `scrape_interval` configurado.

Se esse for o caso, um aviso será registrado e o módulo:

- Responderá com um código de status HTTP 503 (serviço não disponível).
- Retornará o conteúdo do cache, mesmo que ele seja obsoleto.

Esse comportamento pode ser configurado usando os comandos `ceph config set`.

Para instruir o módulo a responder com dados possivelmente obsoletos, defina-o como `return`:

```
cephuser@adm > ceph config set mgr mgr/prometheus/stale_cache_strategy return
```

Para instruir o módulo a responder com `serviço não disponível`, defina-o como `fail`:

```
cephuser@adm > ceph config set mgr mgr/prometheus/stale_cache_strategy fail
```

### 16.5.4 Habilitando o monitoramento de imagens RBD

O Módulo do Gerenciador do Prometheus pode coletar estatísticas de E/S por imagem RBD habilitando contadores de desempenho OSD dinâmicos. As estatísticas são coletadas de todas as imagens nos pools especificados no parâmetro de configuração `mgr/prometheus/rbd_stats_pools`.

O parâmetro é uma lista separada por vírgulas ou espaços de entradas `pool[/namespace]`. Se o namespace não for especificado, as estatísticas serão coletadas para todos os namespaces no pool.

Por exemplo:

```
cephuser@adm > ceph config set mgr mgr/prometheus/rbd_stats_pools "pool1,pool2,poolN"
```

O módulo explora os pools e namespaces especificados, cria uma lista de todas as imagens disponíveis e a atualiza periodicamente. É possível configurar o intervalo usando o parâmetro `mgr/prometheus/rbd_stats_pools_refresh_interval` (em segundos). O padrão é 300 segundos (cinco minutos).

Por exemplo, se você mudou o intervalo de sincronização para 10 minutos:

```
cephuser@adm > ceph config set mgr mgr/prometheus/rbd_stats_pools_refresh_interval 600
```

## 16.6 Modelo de segurança do Prometheus

O modelo de segurança do Prometheus presume que os usuários não confiáveis têm acesso ao endpoint HTTP e aos registros do Prometheus. Os usuários não confiáveis têm acesso a todos os (meta)dados que o Prometheus coleta que estão contidos no banco de dados, além de uma variedade de informações operacionais e de depuração.

No entanto, a API HTTP do Prometheus é limitada a operações apenas leitura. As configurações não podem ser mudadas usando a API, e os segredos não são expostos. Além disso, o Prometheus tem algumas medidas incorporadas para mitigar o impacto dos ataques de negação de serviço.

## 16.7 Gateway SNMP do Alertmanager do Prometheus

Para ser notificado sobre alertas do Prometheus por meio de detecções de SNMP, você pode instalar o gateway SNMP do Alertmanager do Prometheus por meio do `cephadm` ou do Ceph Dashboard. Para fazer isso com o SNMPv2c, por exemplo, você precisa criar um arquivo de especificação de serviço e posicionamento com o seguinte conteúdo:



### Nota

Para obter mais informações sobre arquivos de serviço e posicionamento, consulte o *Livro "Guia de Implantação", Capítulo 8 "Implantando os serviços principais restantes com o cephadm", Seção 8.2 "Especificação de serviço e posicionamento"*.

```
service_type: snmp-gateway
service_name: snmp-gateway
```

```
placement:
  ADD_PLACEMENT_HERE
spec:
  credentials:
    snmp_community: ADD_COMMUNITY_STRING_HERE
    snmp_destination: ADD_FQDN_HERE:ADD_PORT_HERE
    snmp_version: V2c
```

Se preferir, use o Ceph Dashboard para implantar o serviço de gateway SNMP para SNMPv2c e SNMPv3. Para obter mais detalhes, visite [Seção 4.4, “Exibindo serviços”](#).



### III Armazenando dados em um cluster

- 17 Gerenciamento de dados armazenados **160**
- 18 Gerenciar pools de armazenamento **192**
- 19 Pools codificados para eliminação **213**
- 20 Dispositivo de blocos RADOS **220**

## 17 Gerenciamento de dados armazenados

O algoritmo CRUSH determina como armazenar e recuperar dados calculando os locais de armazenamento de dados. O CRUSH permite que os clientes do Ceph se comuniquem diretamente com os OSDs, sem a necessidade de um servidor centralizado ou um controlador. Com um método de armazenamento e recuperação de dados determinado por algoritmo, o Ceph evita um ponto único de falha, gargalo no desempenho e limite físico à escalabilidade.

O CRUSH requer um mapa do cluster e usa o Mapa CRUSH para armazenar e recuperar dados de forma pseudo-aleatória nos OSDs com uma distribuição uniforme dos dados pelo cluster.

Os mapas CRUSH contêm uma lista de OSDs, uma lista de “compartimentos de memória” para agregar os dispositivos em locais físicos e uma lista de regras que orientam como o CRUSH deve replicar os dados nos pools de um cluster do Ceph. Ao refletir a organização física adjacente da instalação, o CRUSH pode moldar (e, portanto, resolver) fontes potenciais de falhas de dispositivos correlacionados. As fontes comuns incluem proximidade física, fonte de energia compartilhada e rede compartilhada. Ao codificar essas informações no mapa do cluster, as políticas de posicionamento do CRUSH podem separar réplicas de objetos em diferentes domínios de falha enquanto ainda mantêm a distribuição desejada. Por exemplo, para evitar a possibilidade de falhas simultâneas, convém usar diferentes prateleiras, racks, fontes de alimentação, controladoras e/ou locais físicos para os dispositivos nos quais as réplicas de dados são armazenadas.

Depois que você implantar um cluster do Ceph, um Mapa CRUSH padrão será gerado. Isso é bom para o seu ambiente de área de segurança do Ceph. No entanto, ao implantar um cluster de dados em grande escala, você deve considerar significativamente o desenvolvimento de um Mapa CRUSH personalizado, pois ele o ajudará a gerenciar o cluster do Ceph, melhorar o desempenho e garantir a segurança dos dados.

Por exemplo, se um OSD ficar inativo, um Mapa CRUSH poderá ajudá-lo a localizar o data center físico, a sala, a fileira e o rack do host com o OSD que falhou, caso seja necessário usar o suporte no local ou substituir o hardware.

Da mesma forma, o CRUSH pode ajudá-lo a identificar falhas mais rapidamente. Por exemplo, se todos os OSDs em determinado rack ficarem inativos ao mesmo tempo, a falha poderá estar associada ao comutador de rede ou à energia que abastece o rack, e não aos próprios OSDs.

Um Mapa CRUSH personalizado também pode ajudar você a identificar os locais físicos onde o Ceph armazena as cópias redundantes dos dados quando o(s) grupo(s) de posicionamento (consulte a [Seção 17.4, “Grupos de posicionamento”](#)) associado(s) ao host com falha está(ão) prejudicado(s).

Há três seções principais para um Mapa CRUSH.

- *Dispositivos OSD* representam qualquer dispositivo de armazenamento de objetos correspondente a um daemon `ceph-osd`.
- *Compartimentos de memória* representam uma agregação hierárquica de locais de armazenamento (por exemplo, fileiras, racks, hosts, etc.) e os pesos atribuídos.
- *Conjuntos de regras* representam o modo de seleção dos compartimentos de memória.

## 17.1 Dispositivos OSD

Para mapear os grupos de posicionamento para OSDs, o Mapa CRUSH requer uma lista de dispositivos OSD (o nome do daemon OSD). A lista de dispositivos aparece primeiro no Mapa CRUSH.

```
#devices
device NUM osd.OSD_NAME class CLASS_NAME
```

Por exemplo:

```
#devices
device 0 osd.0 class hdd
device 1 osd.1 class ssd
device 2 osd.2 class nvme
device 3 osd.3 class ssd
```

Como regra geral, um daemon OSD é mapeado para um único disco.

### 17.1.1 Classes de dispositivo

A flexibilidade do Mapa CRUSH para controlar o posicionamento de dados é um dos pontos fortes do Ceph. É também uma das partes mais difíceis de gerenciamento do cluster. As *classes de dispositivo* automatizam as mudanças mais comuns nos Mapas CRUSH que o administrador antes precisava fazer manualmente.

### 17.1.1.1 Problema de gerenciamento do CRUSH

Os clusters do Ceph costumam ser criados com vários tipos de dispositivos de armazenamento: HDD, SSD, NVMe ou até classes combinadas dos elementos acima. Denominamos esses diferentes tipos de dispositivos de armazenamento como *classes de dispositivo* para evitar confusão entre a propriedade *tipo* dos compartimentos de memória do CRUSH (por exemplo, host, rack, linha. Consulte a [Seção 17.2, “Compartimentos de memória”](#) para obter mais detalhes). Os Ceph OSDs com SSDs são muito mais rápidos do que os com discos giratórios, o que os torna mais adequados a determinadas cargas de trabalho. O Ceph facilita a criação de pools RADOS para diferentes conjuntos de dados ou cargas de trabalho e a atribuição de regras CRUSH diferentes para controlar o posicionamento de dados para esses pools.

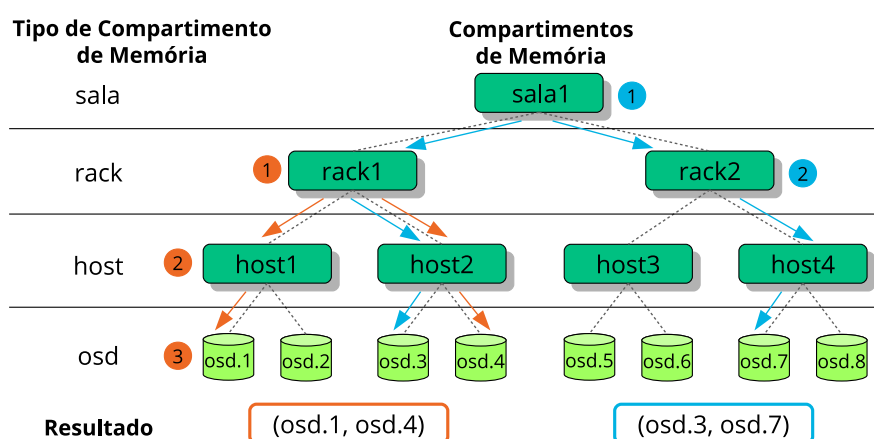


FIGURA 17.1: OSDS COM CLASSES DE DISPOSITIVO COMBINADAS

No entanto, a configuração de regras CRUSH para posicionar dados apenas em uma determinada classe de dispositivo é desgastante. As regras funcionam nos termos da hierarquia CRUSH, mas se os dispositivos forem combinados com os mesmos hosts ou racks (como na hierarquia da amostra acima), eles (por padrão) serão combinados e aparecerão nas mesmas subárvores da hierarquia. A separação manual deles em árvores separadas envolvia a criação de várias versões de cada nó intermediário para cada classe de dispositivo nas versões anteriores do SUSE Enterprise Storage.

### 17.1.1.2 Classes de dispositivo

Uma solução refinada que o Ceph oferece é adicionar uma propriedade denominada *classe de dispositivo* a cada OSD. Por padrão, os OSDs definirão automaticamente suas classes de dispositivo como “hdd”, “ssd” ou “nvme” com base nas propriedades de hardware expostas pelo kernel do Linux. Essas classes de dispositivo são relatadas em uma nova coluna da saída do comando **ceph osd tree**:

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

ID	CLASS	WEIGHT	TYPE	NAME	STATUS	REWEIGHT	PRI-AFF
-1		83.17899	root	default			
-4		23.86200	host	cpach			
2	hdd	1.81898		osd.2	up	1.00000	1.00000
3	hdd	1.81898		osd.3	up	1.00000	1.00000
4	hdd	1.81898		osd.4	up	1.00000	1.00000
5	hdd	1.81898		osd.5	up	1.00000	1.00000
6	hdd	1.81898		osd.6	up	1.00000	1.00000
7	hdd	1.81898		osd.7	up	1.00000	1.00000
8	hdd	1.81898		osd.8	up	1.00000	1.00000
15	hdd	1.81898		osd.15	up	1.00000	1.00000
10	nvme	0.93100		osd.10	up	1.00000	1.00000
0	ssd	0.93100		osd.0	up	1.00000	1.00000
9	ssd	0.93100		osd.9	up	1.00000	1.00000

Se houver falha na detecção automática da classe de dispositivo; por exemplo, porque o driver do dispositivo não expõe apropriadamente as informações sobre o dispositivo usando o `/sys/block`, você poderá ajustar as classes de dispositivo pela linha de comando:

```
cephuser@adm > ceph osd crush rm-device-class osd.2 osd.3
done removing class of osd(s): 2,3
cephuser@adm > ceph osd crush set-device-class ssd osd.2 osd.3
set osd(s) 2,3 to class 'ssd'
```

### 17.1.1.3 Definindo regras de posicionamento CRUSH

As regras CRUSH podem restringir o posicionamento a uma classe de dispositivo específica. Por exemplo, é possível criar um pool **replicado** “rápido” que distribui os dados apenas entre os discos SSD executando o seguinte comando:

```
cephuser@adm > ceph osd crush rule create-
replicated RULE_NAME ROOT FAILURE_DOMAIN_TYPE DEVICE_CLASS
```

Por exemplo:

```
cephuser@adm > ceph osd crush rule create-replicated fast default host ssd
```

Crie um pool denominado “fast\_pool” e atribua-o à regra “fast” (rápido):

```
cephuser@adm > ceph osd pool create fast_pool 128 128 replicated fast
```

O processo para criar as regras de **código de eliminação** é um pouco diferente. Primeiramente, você cria um perfil de código de eliminação que inclui uma propriedade para sua classe de dispositivo desejada. Em seguida, usa esse perfil ao criar o pool codificado para eliminação:

```
cephuser@adm > ceph osd erasure-code-profile set myprofile \
k=4 m=2 crush-device-class=ssd crush-failure-domain=host
cephuser@adm > ceph osd pool create mypool 64 erasure myprofile
```

Caso você precise editar manualmente o Mapa CRUSH para personalizar sua regra, a sintaxe foi estendida para permitir que a classe de dispositivo seja especificada. Por exemplo, a regra CRUSH gerada pelos comandos acima tem o seguinte formato:

```
rule ecpool {
  id 2
  type erasure
  min_size 3
  max_size 6
  step set_chooseleaf_tries 5
  step set_choose_tries 100
  step take default class ssd
  step chooseleaf indep 0 type host
  step emit
}
```

A diferença importante aqui é que o comando “take” inclui o sufixo “NOME\_CLASSE da classe” adicional.

#### 17.1.1.4 Comandos adicionais

Para listar as classes de dispositivo usadas em um mapa CRUSH, execute:

```
cephuser@adm > ceph osd crush class ls
[
  "hdd",
  "ssd"
]
```

Para listar as regras CRUSH existentes, execute:

```
cephuser@adm > ceph osd crush rule ls
replicated_rule
fast
```

Para ver os detalhes da regra CRUSH denominada “fast”, execute:

```
cephuser@adm > ceph osd crush rule dump fast
{
  "rule_id": 1,
  "rule_name": "fast",
  "ruleset": 1,
  "type": 1,
  "min_size": 1,
  "max_size": 10,
  "steps": [
    {
      "op": "take",
      "item": -21,
      "item_name": "default~ssd"
    },
    {
      "op": "chooseleaf_firstn",
      "num": 0,
      "type": "host"
    },
    {
      "op": "emit"
    }
  ]
}
```

Para listar os OSDs que pertencem a uma classe “ssd”, execute:

```
cephuser@adm > ceph osd crush class ls-osd ssd
0
1
```

#### 17.1.1.5 Migrando de uma regra SSD legada para classes de dispositivo

No SUSE Enterprise Storage anterior à versão 5, você precisava editar manualmente o Mapa CRUSH e manter uma hierarquia paralela para cada tipo de dispositivo especializado (como SSD) a fim de gravar regras que se aplicassem a esses dispositivos. A partir do SUSE Enterprise Storage 5, o recurso de classe de dispositivo permitiu que isso fosse feito de maneira transparente.

É possível transformar uma regra e hierarquia legadas nas novas regras com base em classe usando o comando **crushtool**. Há vários tipos de transformação possíveis:

#### **crushtool --reclassify-root** *ROOT\_NAME* *DEVICE\_CLASS*

Esse comando considera tudo o que está na hierarquia abaixo de *ROOT\_NAME* e ajusta quaisquer regras que fazem referência à raiz por meio de

```
take ROOT_NAME
```

para

```
take ROOT_NAME class DEVICE_CLASS
```

Ele enumera novamente os compartimentos de memória para que os IDs antigos sejam usados na “árvore de sombra” da classe especificada. Como consequência, nenhum movimento de dados ocorre.

#### EXEMPLO 17.1: **crushtool --reclassify-root**

Considere a seguinte regra existente:

```
rule replicated_ruleset {
  id 0
  type replicated
  min_size 1
  max_size 10
  step take default
  step chooseleaf firstn 0 type rack
  step emit
}
```

Se você reclassificar a raiz “default” como a classe “hdd”, a regra se tornará

```
rule replicated_ruleset {
  id 0
  type replicated
  min_size 1
  max_size 10
  step take default class hdd
  step chooseleaf firstn 0 type rack
  step emit
}
```

#### **crushtool --set-subtree-class** *BUCKET\_NAME* *DEVICE\_CLASS*

Esse método marca cada dispositivo na subárvore com raiz em *BUCKET\_NAME* com a classe de dispositivo especificada.



Normalmente, a `--set-subtree-class` é usada em conjunto com a opção `--reclassify-root` para garantir que todos os dispositivos nessa raiz sejam rotulados com a classe correta. No entanto, alguns desses dispositivos podem intencionalmente ter uma classe diferente e, portanto, você não deseja rotulá-los outra vez. Nesses casos, exclua a opção `--set-subtree-class`. Saiba que esse tipo de remapeamento não será perfeito, porque a regra anterior é distribuída pelos dispositivos das várias classes, mas as regras ajustadas serão mapeadas apenas para os dispositivos da classe especificada.

#### **`crushtool --reclassify-bucket MATCH_PATTERN DEVICE_CLASS DEFAULT_PATTERN`**

Esse método permite a fusão de uma hierarquia específica do tipo paralelo com a hierarquia normal. Por exemplo, muitos usuários têm Mapas CRUSH semelhantes aos seguintes:

##### EXEMPLO 17.2: `crushtool --reclassify-bucket`

```
host node1 {
    id -2          # do not change unnecessarily
    # weight 109.152
    alg straw
    hash 0 # rjenkins1
    item osd.0 weight 9.096
    item osd.1 weight 9.096
    item osd.2 weight 9.096
    item osd.3 weight 9.096
    item osd.4 weight 9.096
    item osd.5 weight 9.096
    [...]
}

host node1-ssd {
    id -10         # do not change unnecessarily
    # weight 2.000
    alg straw
    hash 0 # rjenkins1
    item osd.80 weight 2.000
    [...]
}

root default {
    id -1          # do not change unnecessarily
    alg straw
    hash 0 # rjenkins1
    item node1 weight 110.967
    [...]
}
```

```

root ssd {
    id -18          # do not change unnecessarily
    # weight 16.000
    alg straw
    hash 0 # rjenkins1
    item node1-ssd weight 2.000
    [...]
}

```

Essa função reclassifica cada compartimento de memória que corresponde a um determinado padrão. O padrão pode ter um formato parecido com %suffix ou prefix %. No exemplo acima, você usará o padrão %-ssd. Para cada compartimento de memória combinado, a parte restante do nome que corresponde ao curinga “%” especifica o compartimento de memória de base. Todos os dispositivos no compartimento de memória combinado são rotulados com a classe de dispositivo especificada e, em seguida, movidos para o compartimento de memória de base. Se o compartimento de memória de base não existe (por exemplo, se “node12-ssd” existe, mas “node12” não), ele é criado e vinculado abaixo do compartimento de memória pai padrão especificado. Os IDs de compartimentos de memória antigos são preservados nos novos compartimentos de memória de sombra para evitar a movimentação de dados. As regras com as etapas take que fazem referência a compartimentos de memória antigos são ajustadas.

**crushtool --reclassify-bucket** *BUCKET\_NAME* *DEVICE\_CLASS* *BASE\_BUCKET*

É possível usar a opção --reclassify-bucket sem um curinga para mapear um único compartimento de memória. Como no exemplo anterior, em que desejamos que o compartimento de memória “ssd” seja mapeado para o compartimento de memória padrão. O comando final para converter o mapa composto dos fragmentos acima é o seguinte:

```

cephuser@adm > ceph osd getcrushmap -o original
cephuser@adm > crushtool -i original --reclassify \
    --set-subtree-class default hdd \
    --reclassify-root default hdd \
    --reclassify-bucket %-ssd ssd default \
    --reclassify-bucket ssd ssd default \
    -o adjusted

```

Para verificar se a conversão está correta, há uma opção --compare que testa uma grande amostra de entradas no Mapa CRUSH e compara se o mesmo resultado é retornado. Essas entradas são controladas pelas mesmas opções aplicadas a --test. Para o exemplo acima, o comando é o seguinte:

```

cephuser@adm > crushtool -i original --compare adjusted

```

```
rule 0 had 0/10240 mismatched mappings (0)
rule 1 had 0/10240 mismatched mappings (0)
maps appear equivalent
```



## Dica

Se houvesse diferenças, você veria a proporção das entradas que seriam mapeadas novamente entre parênteses.

Se você estiver satisfeito com o Mapa CRUSH ajustado, poderá aplicá-lo ao cluster:

```
cephuser@adm > ceph osd setcrushmap -i adjusted
```

### 17.1.1.6 Para obter mais informações

Encontre mais detalhes sobre os Mapas CRUSH na [Seção 17.5, “Manipulação de mapa CRUSH”](#).

Encontre mais detalhes em geral sobre os pools do Ceph no [Capítulo 18, Gerenciar pools de armazenamento](#).

Encontre mais detalhes sobre os pools codificados para eliminação no [Capítulo 19, Pools codificados para eliminação](#).

## 17.2 Compartimentos de memória

Os mapas CRUSH contêm uma lista de OSDs, que podem ser organizados em uma estrutura de árvore de compartimentos de memória para agregar os dispositivos em locais físicos. Cada OSD abrange as folhas da árvore.

0	osd	Um dispositivo ou OSD específico ( <code>osd.1</code> , <code>osd.2</code> , etc).
1	host	O nome de um host que contém um ou mais OSDs.
2	Chassi	Identificador do chassi que contém o <code>host</code> no rack.
3	rack	Um rack de computador. O padrão é <code>unknownrack</code> .
4	row	Uma fileira em uma série de racks.

5	pdu	Abreviação de “Power Distribution Unit” (Unidade de Distribuição de Energia).
6	pod	Abreviação de “Point of Delivery” (Ponto de Entrega): neste contexto, um grupo de PDUs ou um grupo de fileiras de racks.
7	room	Uma sala com fileiras de racks.
8	centro de dados	Um centro de dados físico com uma ou mais salas.
9	region	Região geográfica global (por exemplo, NAM, LAM, EMEA, APAC etc.)
10	usuário	O nó raiz da árvore de compartimentos de memória OSD (normalmente definido como <u>default</u> ).



## Dica

Você pode modificar os tipos existentes e criar seus próprios tipos de compartimento de memória.

As ferramentas de implantação do Ceph geram um Mapa CRUSH que contém um compartimento de memória para cada host e uma raiz denominada “default”, que é útil para o pool rbd padrão. Os tipos de compartimento de memória restantes oferecem um meio de armazenar informações sobre o local físico dos nós/compartimentos de memória, o que facilita bastante a administração do cluster em caso de mal funcionamento dos OSDs, dos hosts ou do hardware de rede e quando o administrador precisa acessar o hardware físico.

Um compartimento de memória tem um tipo, um nome exclusivo (string), um ID único indicado por um número inteiro negativo, um peso relativo à capacidade total do(s) item(ns), o algoritmo do compartimento de memória (por padrão, straw2) e o hash (por padrão, 0, refletindo o Hash CRUSH rjenkins1). Um compartimento de memória pode ter um ou mais itens. Os itens podem ser constituídos de outros compartimentos de memória ou OSDs. Os itens podem ter um peso que reflete o peso relativo do item.

```
[bucket-type] [bucket-name] {
  id [a unique negative numeric ID]
  weight [the relative capacity/capability of the item(s)]
```

```
alg [the bucket type: uniform | list | tree | straw2 | straw ]
hash [the hash type: 0 by default]
item [item-name] weight [weight]
}
```

O exemplo a seguir ilustra como você pode usar compartimentos de memória para agregar um pool e locais físicos, como data center, sala, rack e fileira.

```
host ceph-osd-server-1 {
    id -17
    alg straw2
    hash 0
    item osd.0 weight 0.546
    item osd.1 weight 0.546
}

row rack-1-row-1 {
    id -16
    alg straw2
    hash 0
    item ceph-osd-server-1 weight 2.00
}

rack rack-3 {
    id -15
    alg straw2
    hash 0
    item rack-3-row-1 weight 2.00
    item rack-3-row-2 weight 2.00
    item rack-3-row-3 weight 2.00
    item rack-3-row-4 weight 2.00
    item rack-3-row-5 weight 2.00
}

rack rack-2 {
    id -14
    alg straw2
    hash 0
    item rack-2-row-1 weight 2.00
    item rack-2-row-2 weight 2.00
    item rack-2-row-3 weight 2.00
    item rack-2-row-4 weight 2.00
    item rack-2-row-5 weight 2.00
}

rack rack-1 {
    id -13
```

```

    alg straw2
    hash 0
    item rack-1-row-1 weight 2.00
    item rack-1-row-2 weight 2.00
    item rack-1-row-3 weight 2.00
    item rack-1-row-4 weight 2.00
    item rack-1-row-5 weight 2.00
}

room server-room-1 {
    id -12
    alg straw2
    hash 0
    item rack-1 weight 10.00
    item rack-2 weight 10.00
    item rack-3 weight 10.00
}

datacenter dc-1 {
    id -11
    alg straw2
    hash 0
    item server-room-1 weight 30.00
    item server-room-2 weight 30.00
}

root data {
    id -10
    alg straw2
    hash 0
    item dc-1 weight 60.00
    item dc-2 weight 60.00
}

```

## 17.3 Conjuntos de regras

Os mapas CRUSH suportam a noção de “regras CRUSH”, que determinam o posicionamento dos dados em um pool. Para clusters grandes, convém criar muitos pools, em que cada um pode ter seu próprio conjunto de regras CRUSH e suas próprias regras. O Mapa CRUSH padrão tem uma regra para a raiz padrão. Se você deseja mais raízes e regras, precisa criá-las no futuro, ou elas serão criadas automaticamente quando novos pools forem criados.



## Nota

Na maioria dos casos, você não precisará modificar as regras padrão. Quando você cria um novo pool, o conjunto de regras padrão dele é 0.

Uma regra apresenta o seguinte formato:

```
rule rulename {  
  
    ruleset ruleset  
    type type  
    min_size min-size  
    max_size max-size  
    step step  
  
}
```

### ruleset

Um número inteiro. Classifica uma regra como pertencente a um conjunto de regras. Ativado quando o conjunto de regras é definido em um pool. Essa opção é obrigatória. O padrão é 0.

### type

Uma string. Descreve uma regra para um pool com codificação "replicado" ou "de eliminação". Essa opção é obrigatória. O padrão é replicado.

### min\_size

Um número inteiro. Se um grupo de pools gerar menos réplicas do que esse número, o CRUSH NÃO selecionará essa regra. Essa opção é obrigatória. O padrão é 2.

### max\_size

Um número inteiro. Se um grupo de pools gerar mais réplicas do que esse número, o CRUSH NÃO selecionará essa regra. Essa opção é obrigatória. O padrão é 10.

### step take bucket

Usa um compartimento de memória especificado por um nome e inicia a iteração descendente na árvore. Essa opção é obrigatória. Para obter uma explicação sobre iteração na árvore, consulte a [Seção 17.3.1, "Iterando a árvore de nós"](#).

`step targetmodenum type bucket-type`

`target` pode ser `choose` ou `chooseleaf`. Quando definido como `choose`, um número de compartimentos de memória é selecionado. `chooseleaf` seleciona diretamente os OSDs (nós folha) da subárvore de cada compartimento de memória no conjunto de compartimentos de memória.

`mode` pode ser `firstn` ou `indep`. Consulte a [Seção 17.3.2, “firstn e indep”](#).

Seleciona o número de compartimentos de memória de determinado tipo. Em que N é o número de opções disponíveis, se `num > 0` && `< N`, escolha essa mesma quantidade de compartimentos de memória; se `num < 0`, isso significa `N - num` e, se `num == 0`, escolha N compartimentos de memória (todos disponíveis). Segue `step take` ou `step choose`.

`step emit`

Gera o valor atual e esvazia a pilha. Normalmente usado no fim de uma regra, mas também pode ser usado para estruturar árvores diferentes na mesma regra. Segue `step choose`.

### 17.3.1 Iterando a árvore de nós

É possível ver a estrutura definida com os compartimentos de memória como uma árvore de nós. Os compartimentos de memória são os nós, e os OSDs são as folhas da árvore.

As regras no Mapa CRUSH definem como os OSDs são selecionados dessa árvore. Uma regra começa com um nó e, em seguida, faz a iteração descendente pela árvore para retornar um conjunto de OSDs. Não é possível definir qual ramificação precisa ser selecionada. Em vez disso, o algoritmo CRUSH garante que o conjunto de OSDs atende aos requisitos de replicação e distribui os dados igualmente.

Com `step take bucket`, a iteração pela árvore de nós começa no compartimento de memória especificado (sem tipo de compartimento de memória). Se os OSDs de todas as ramificações na árvore tiverem que ser retornados, o compartimento de memória deverá ser a raiz. Do contrário, as etapas a seguir apenas fará a iteração na subárvore.

Após `step take`, uma ou mais entradas `step choose` vêm a seguir na definição da regra. Cada `step choose` escolhe um número definido de nós (ou ramificações) do nó superior selecionado anteriormente.

No fim, os OSDs selecionados são retornados com `step emit`.

`step chooseleaf` é uma prática função que seleciona os OSDs diretamente das ramificações do compartimento de memória especificado.



A *Figura 17.2, “Exemplo de árvore”* mostra um exemplo de como o `step` é usado para iterar em uma árvore. As setas e os números laranjas correspondem a `example1a` e `example1b`, e os azuis correspondem a `example2` nas definições de regra a seguir.

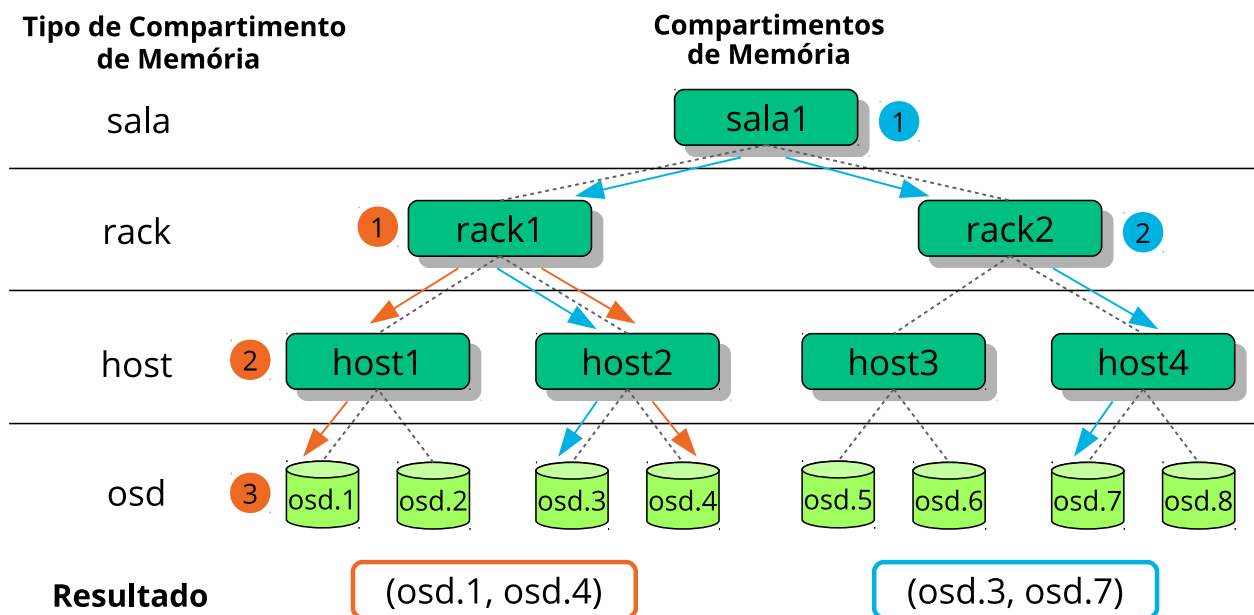


FIGURA 17.2: EXEMPLO DE ÁRVORE

```
# orange arrows
rule example1a {
    ruleset 0
    type replicated
    min_size 2
    max_size 10
    # orange (1)
    step take rack1
    # orange (2)
    step choose firstn 0 host
    # orange (3)
    step choose firstn 1 osd
    step emit
}

rule example1b {
    ruleset 0
    type replicated
    min_size 2
    max_size 10
    # orange (1)
    step take rack1
```

```

# orange (2) + (3)
step chooseleaf firstn 0 host
step emit
}

# blue arrows
rule example2 {
    ruleset 0
    type replicated
    min_size 2
    max_size 10
    # blue (1)
    step take room1
    # blue (2)
    step chooseleaf firstn 0 rack
    step emit
}

```

### 17.3.2 firstn e indep

Uma regra CRUSH define substituições para nós ou OSDs com falha (consulte a [Seção 17.3, “Conjuntos de regras”](#)). A palavra-chave `step` requer `firstn` ou `indep` como parâmetro. A [Figura 17.3, “Métodos de substituição de nó”](#) apresenta um exemplo.

O `firstn` adiciona nós de substituição ao fim da lista de nós ativos. No caso de um nó com falha, os seguintes nós saudáveis são deslocados para a esquerda para preencher a lacuna do nó com falha. Esse é o método padrão desejado para *pools replicados*, porque um nó secundário já tem todos os dados e, portanto, pode assumir as tarefas do nó principal imediatamente.

O `indep` seleciona nós de substituição fixos para cada nó ativo. A substituição de um nó com falha não muda a ordem dos nós restantes. Esse é o método desejado para *pools codificados para eliminação*. Nos pools codificados para eliminação, os dados armazenados em um nó dependem da posição dele na seleção do nó. Quando a ordem dos nós muda, todos os dados nos nós afetados precisam ser realocados.

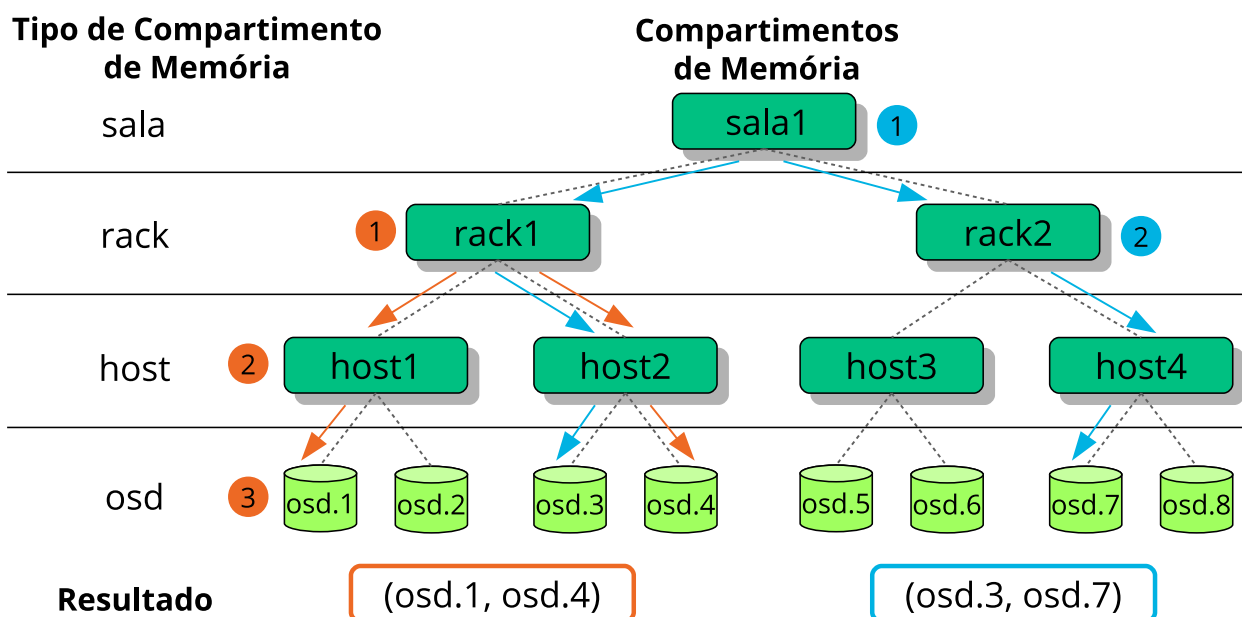


FIGURA 17.3: MÉTODOS DE SUBSTITUIÇÃO DE NÓ

## 17.4 Grupos de posicionamento

O Ceph mapeia objetos para grupos de posicionamento (PGs, placement groups). Os grupos de posicionamento são fragmentos de um pool de objetos lógicos que armazenam os objetos como um grupo em OSDs. Os grupos de posicionamento reduzem a quantidade de metadados por objeto quando o Ceph armazena os dados em OSDs. Um número maior de grupos de posicionamento, por exemplo, 100 por OSD, proporciona um melhor equilíbrio.

### 17.4.1 Usando grupos de posicionamento

Um grupo de posicionamento (PG) agrega objetos em um pool. O principal motivo é que o monitoramento do posicionamento de objetos e dos metadados por objeto é caro em termos de capacidade de computação. Por exemplo, um sistema com milhões de objetos não pode monitorar o posicionamento de cada um dos seus objetos diretamente.

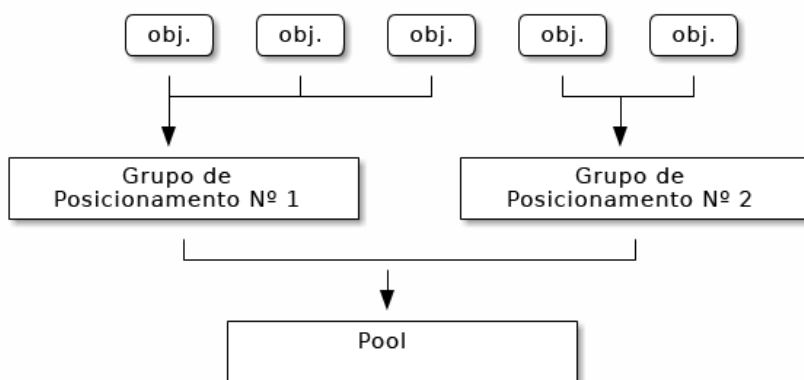


FIGURA 17.4: GRUPOS DE POSICIONAMENTO EM UM POOL

O cliente Ceph calculará a que grupo de posicionamento um objeto pertencerá. Ele faz isso por meio de hashing do ID do objeto e da aplicação de uma operação com base no número de PGs no pool definido e no ID do pool.

O conteúdo do objeto em um grupo de posicionamento é armazenado em um conjunto de OSDs. Por exemplo, em um pool replicado de tamanho dois, cada grupo de posicionamento armazenará objetos em dois OSDs:

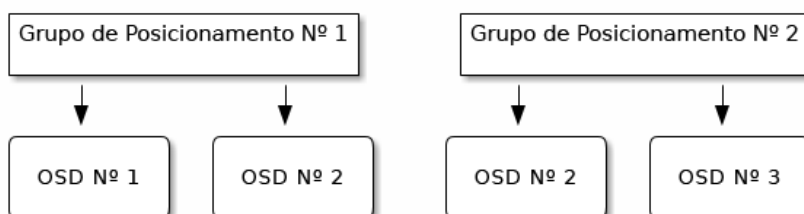


FIGURA 17.5: GRUPOS DE POSICIONAMENTO E OSDs

Se o OSD nº 2 falhar, outro OSD será atribuído ao grupo de posicionamento nº 1 e será preenchido com cópias de todos os objetos no OSD nº 1. Se o tamanho do pool for mudado de dois para três, um OSD adicional será atribuído ao grupo de posicionamento e receberá cópias de todos os objetos no grupo de posicionamento.

Os grupos de posicionamento não têm a propriedade do OSD, eles o compartilham com outros grupos de posicionamento do mesmo pool ou até de outros pools. Se o OSD nº 2 falhar, o grupo de posicionamento nº 2 também precisará restaurar cópias dos objetos usando o OSD nº 3.

Quando o número de grupos de posicionamento aumentar, os novos grupos de posicionamento receberão os OSDs. O resultado da função CRUSH também mudará, e alguns objetos dos antigos grupos de posicionamento serão copiados para os novos grupos de posicionamento e removidos dos antigos.

## 17.4.2 Determinando o valor de *PG\_NUM*



### Nota

A partir do Ceph Nautilus (v14.x), você pode usar o módulo `pg_autoscaler` do Ceph Manager para dimensionar automaticamente os PGs, conforme necessário. Para habilitar esse recurso, consulte o Livro *“Deploying and Administering SUSE Enterprise Storage with Rook”*, Capítulo 8 “Configuration”, Seção 8.1.1.1 “Default PG and PGP counts”.

Ao criar um novo pool, você ainda pode escolher o valor de *PG\_NUM* manualmente:

```
# ceph osd pool create POOL_NAME PG_NUM
```

Não é possível calcular o *PG\_NUM* automaticamente. Veja a seguir alguns valores mais usados, dependendo do número de OSDs no cluster:

**Menos do que 5 OSDs:**

Defina *PG\_NUM* como 128.

**Entre 5 e 10 OSDs:**

Defina *PG\_NUM* como 512.

**Entre 10 e 50 OSDs:**

Defina *PG\_NUM* como 1024.

Conforme o número de OSDs aumenta, a escolha do valor certo para *PG\_NUM* torna-se mais importante. O *PG\_NUM* afeta altamente o comportamento do cluster e também a durabilidade dos dados em caso de falha no OSD.

### 17.4.2.1 Calculando os grupos de posicionamento para mais do que 50 OSDs

Se você tem menos do que 50 OSDs, use a pré-seleção descrita na [Seção 17.4.2, “Determinando o valor de PG\\_NUM”](#). Se você tem mais do que 50 OSDs, recomendamos aproximadamente de 50 a 100 grupos de posicionamento por OSD para equilibrar o uso de recursos, a durabilidade e a distribuição dos dados. Para um único pool de objetos, você pode usar a seguinte fórmula para obter uma linha de base:

```
total PGs = (OSDs * 100) / POOL_SIZE
```

Em que `POOL_SIZE` é o número de réplicas para os pools replicados ou a soma de “k” + “m” para os pools codificados para eliminação, conforme retornado pelo comando `ceph osd erasure-code-profile get`. Você deve arredondar o resultado para cima até a potência mais próxima de 2. O arredondamento é recomendado para que o algoritmo CRUSH equilibre igualmente o número de objetos entre os grupos de posicionamento.

Como exemplo, para um cluster com 200 OSDs e tamanho de pool de 3 réplicas, você estima o número de PGs da seguinte maneira:

```
(200 * 100) / 3 = 6667
```

A potência mais próxima de 2 é **8192**.

Ao usar vários pools de dados para armazenar objetos, você precisa garantir o equilíbrio entre o número de grupos de posicionamento por pool e o número de grupos de posicionamento por OSD. Você precisa atingir um número total razoável de grupos de posicionamento que ofereça variação suficientemente baixa por OSD sem sobrecarregar os recursos do sistema nem tornar o processo de emparelhamento muito lento.

Por exemplo, um cluster de 10 pools, cada um com 512 grupos de posicionamento em 10 OSDs, representa um total de 5.120 grupos de posicionamento distribuídos por 10 OSDs, ou seja, 512 grupos de posicionamento por OSD. Esse tipo de configuração não usa muitos recursos. No entanto, se 1.000 grupos foram criados com 512 grupos de posicionamento cada, os OSDs processam aproximadamente 50.000 grupos de posicionamento cada, e isso requer muito mais recursos e tempo de emparelhamento.

### 17.4.3 Definindo o número de grupos de posicionamento



#### Nota

A partir do Ceph Nautilus (v14.x), você pode usar o módulo `pg_autoscaler` do Ceph Manager para dimensionar automaticamente os PGs, conforme necessário. Para habilitar esse recurso, consulte *Livro “Deploying and Administering SUSE Enterprise Storage with Rook”, Capítulo 8 “Configuration”, Seção 8.1.1.1 “Default PG and PGP counts”*.

Se você ainda precisa especificar o número de grupos de posicionamento em um pool manualmente, é necessário especificá-los no momento da criação do pool (consulte a [Seção 18.1, “Criando um pool”](#)). Após definir os grupos de posicionamento para um pool, você poderá aumentar o número de grupos de posicionamento executando o seguinte comando:

```
# ceph osd pool set POOL_NAME pg_num PG_NUM
```

Após aumentar o número de grupos de posicionamento, você também precisará aumentar esse número para o posicionamento (`PGP_NUM`) antes que o cluster seja redistribuído. O `PGP_NUM` será o número de grupos de posicionamento que serão considerados para o posicionamento pelo algoritmo CRUSH. O aumento do `PG_NUM` divide os grupos de posicionamento, mas os dados não serão migrados para os grupos de posicionamento mais recentes até que `PGP_NUM` seja aumentado. O `PGP_NUM` deve ser igual ao `PG_NUM`. Para aumentar o número de grupos para o posicionamento, execute o seguinte:

```
# ceph osd pool set POOL_NAME pgp_num PGP_NUM
```

### 17.4.4 Encontrando o número de grupos de posicionamento

Para encontrar o número de grupos de posicionamento em um pool, execute o seguinte comando **get**:

```
# ceph osd pool get POOL_NAME pg_num
```

## 17.4.5 Encontrando as estatísticas de PG de um cluster

Para encontrar as estatísticas dos grupos de posicionamento em seu cluster, execute o seguinte comando:

```
# ceph pg dump [--format FORMAT]
```

Os formatos válidos são “plain” (padrão) e “json”.

## 17.4.6 Encontrando as estatísticas de PGs travados

Para encontrar as estatísticas de todos os grupos de posicionamento travados em um determinado estado, execute o seguinte:

```
# ceph pg dump_stuck STATE \  
  [--format FORMAT] [--threshold THRESHOLD]
```

O *STATE* é um dos seguintes: “inactive” (PGs não podem processar leituras ou gravações porque estão aguardando por um OSD com os dados mais atualizados), “unclean” (PGs contêm objetos que não foram replicados o número desejado de vezes), “stale” (PGs em estado desconhecido: os OSDs que os hospedam não foram relatados ao cluster do monitor no intervalo especificado pela opção `mon_osd_report_timeout`), “undersized” ou “degraded”.

Os formatos válidos são “plain” (padrão) e “json”.

O limite define o número mínimo de segundos que o grupo de posicionamento permanece travado antes de incluí-lo nas estatísticas retornadas (por padrão, 300 segundos).

## 17.4.7 Pesquisando o mapa de um grupo de posicionamento

Para pesquisar o mapa de um determinado grupo de posicionamento, execute o seguinte:

```
# ceph pg map PG_ID
```

O Ceph retornará o mapa do grupo de posicionamento, o grupo de posicionamento e o status do OSD:

```
# ceph pg map 1.6c  
osdmap e13 pg 1.6c (1.6c) -> up [1,0] acting [1,0]
```



## 17.4.8 Recuperando as estatísticas de grupos de posicionamento

Para recuperar as estatísticas de um determinado grupo de posicionamento, execute o seguinte:

```
# ceph pg PG_ID query
```

## 17.4.9 Depurando um grupo de posicionamento

Para depurar (*Seção 17.6, “Depurando grupos de posicionamento”*) um grupo de posicionamento, execute o seguinte:

```
# ceph pg scrub PG_ID
```

O Ceph verifica os nós primários e de réplica, gera um catálogo de todos os objetos no grupo de posicionamento e os compara para garantir que nenhum objeto esteja ausente ou seja incompatível e que seu conteúdo seja consistente. Supondo que todas as réplicas sejam correspondentes, uma varredura semântica final garante que todos os metadados de objetos relacionados a instantâneo sejam consistentes. Os erros são relatados por meio de registros.

## 17.4.10 Priorizando o provisionamento e a recuperação de grupos de posicionamento

Você pode enfrentar uma situação em que vários grupos de posicionamento exigem recuperação e/ou provisionamento, enquanto alguns grupos armazenam dados mais importantes do que outros. Por exemplo, alguns PGs podem armazenar dados para imagens usadas por máquinas em execução, e outros PGs podem ser usados por máquinas inativas ou dados menos relevantes. Nesse caso, você pode priorizar a recuperação desses grupos para que o desempenho e a disponibilidade dos dados armazenados neles sejam restaurados com mais antecedência. Para marcar grupos de posicionamento específicos como priorizados durante o provisionamento ou a recuperação, execute o seguinte:

```
# ceph pg force-recovery PG_ID1 [PG_ID2 ... ]  
# ceph pg force-backfill PG_ID1 [PG_ID2 ... ]
```

Isso fará com que o Ceph realize a recuperação ou o provisionamento dos grupos de posicionamento especificados primeiro, antes dos outros grupos de posicionamento. Esse procedimento não interrompe os provisionamentos ou a recuperação que está em andamento, mas faz com que os PGs especificados sejam processados o mais rápido possível. Se você mudar de ideia ou priorizar grupos errados, cancele a priorização:

```
# ceph pg cancel-force-recovery PG_ID1 [PG_ID2 ... ]  
# ceph pg cancel-force-backfill PG_ID1 [PG_ID2 ... ]
```

Os comandos **cancel-\*** removem o flag “force” dos PGs para que sejam processados na ordem padrão. Mais uma vez, isso não afeta os grupos de posicionamento que estão sendo processados, apenas aqueles que ainda estão na fila. O flag “force” será limpo automaticamente após a recuperação ou o provisionamento do grupo.

### 17.4.11 Revertendo objetos perdidos

Se o cluster perdeu um ou mais objetos, e você decidiu parar de procurar os dados perdidos, precisará marcar os objetos não encontrados como “perdidos”.

Se os objetos ainda continuarem perdidos depois de ter consultado todos os locais possíveis, você talvez tenha de desistir deles. Isso é possível considerando as combinações incomuns de falhas que permitem que o cluster reconheça as gravações que foram realizadas antes de serem recuperadas.

Atualmente, a única opção suportada é “revert”, que voltará para uma versão anterior do objeto ou o esquecerá completamente, no caso de um novo objeto. Para marcar os objetos “não encontrados” como “perdidos”, execute o seguinte:

```
cephuser@adm > ceph pg PG_ID mark_unfound_lost revert|delete
```

### 17.4.12 Habilitando o dimensionador automático de PG

Os grupos de posicionamento (PGs, Placement Groups) são um detalhe interno de implementação de como o Ceph distribui os dados. Ao habilitar o pg-autoscaling, você pode permitir que o cluster crie ou ajuste os PGs automaticamente com base no modo como o cluster é usado.

Cada pool no sistema tem uma propriedade `pg_autoscale_mode` que pode ser definida como `off`, `on` ou `warn`:

O dimensionador automático é configurado por pool e pode ser executado em três modos:

#### off

Desabilite o dimensionamento automático para este pool. O administrador é quem escolhe um número apropriado de PGs para cada pool.

#### on

Habilite os ajustes automatizados da contagem de PGs para o pool especificado.

#### aviso

Emita alertas de saúde quando a contagem de PGs precisar ser ajustada.

Para definir o modo de dimensionamento automático para os pools existentes:

```
cephuser@adm > ceph osd pool set POOL_NAME pg_autoscale_mode mode
```

Você também pode configurar o `pg_autoscale_mode` padrão que será aplicado a qualquer pool criado no futuro com:

```
cephuser@adm > ceph config set global osd_pool_default_pg_autoscale_mode MODE
```

Você pode ver cada pool, sua utilização relativa e quaisquer mudanças sugeridas na contagem de PGs com este comando:

```
cephuser@adm > ceph osd pool autoscale-status
```

## 17.5 Manipulação de mapa CRUSH

Esta seção apresenta os modos de manipulação do Mapa CRUSH básico. Por exemplo, editar um Mapa CRUSH, mudar parâmetros do Mapa CRUSH e adicionar/mover/remover um OSD.

### 17.5.1 Editando um mapa CRUSH

Para editar um mapa CRUSH existente, faça o seguinte:

1. Obtenha um Mapa CRUSH. Para obter o Mapa CRUSH para seu cluster, execute o seguinte:

```
cephuser@adm > ceph osd getcrushmap -o compiled-crushmap-filename
```

O Ceph gerará (-o) um Mapa CRUSH compilado com o nome de arquivo que você especificou. Como o Mapa CRUSH está em um formato compilado, você deve descompilá-lo antes que você possa editá-lo.

2. Descompile um Mapa CRUSH. Para descompilar um Mapa CRUSH, execute o seguinte:

```
cephuser@adm > crushtool -d compiled-crushmap-filename \  
-o decompiled-crushmap-filename
```

O Ceph descompilará (-d) o Mapa CRUSH compilado e o gerará (-o) com o nome de arquivo que você especificou.

3. Edite pelo menos um dos parâmetros de Dispositivos, Compartimentos de Memória e Regras.
4. Compile um Mapa CRUSH. Para compilar um Mapa CRUSH, execute o seguinte:

```
cephuser@adm > crushtool -c decompiled-crush-map-filename \  
-o compiled-crush-map-filename
```

O Ceph armazenará um Mapa CRUSH compilado com o nome de arquivo que você especificou.

5. Defina um Mapa CRUSH. Para definir o Mapa CRUSH para o cluster, execute o seguinte:

```
cephuser@adm > ceph osd setcrushmap -i compiled-crushmap-filename
```

O Ceph inserirá o Mapa CRUSH compilado do nome de arquivo que você especificou como o Mapa CRUSH para o cluster.



### Dica: Usar o sistema de controle de versão

Use um sistema de controle de versão, como git ou svn, para os arquivos de Mapa CRUSH exportados e modificados. Isso simplifica um possível rollback.



### Dica: Testar o novo mapa CRUSH

Teste o novo Mapa CRUSH ajustado usando o comando **`crushtool --test`** e compare com o estado antes da aplicação do novo Mapa CRUSH. Você pode achar útil os seguintes switches de comando: `--show-statistics`, `--show-mappings`, `--show-bad-mappings`, `--show-utilization`, `--show-utilization-all`, `--show-choose-tries`

## 17.5.2 Adicionando ou movendo um OSD

Para adicionar ou mover um OSD no Mapa CRUSH de um cluster em execução, faça o seguinte:

```
cephuser@adm > ceph osd crush set id_or_name weight root=pool-name  
bucket-type=bucket-name ...
```

### id

Um número inteiro. O ID numérico do OSD. Essa opção é obrigatória.

### name

Uma string. O nome completo do OSD. Essa opção é obrigatória.

### weight

Um duplo. O peso do CRUSH para o OSD. Essa opção é obrigatória.

### usuário

Um par de chave/valor. Por padrão, a hierarquia do CRUSH contém o pool padrão como raiz. Essa opção é obrigatória.

### bucket-type

Pares de chave/valor. Você pode especificar o local do OSD na hierarquia do CRUSH.

O exemplo a seguir adiciona `osd.0` à hierarquia ou move o OSD de um local anterior.

```
cephuser@adm > ceph osd crush set osd.0 1.0 root=data datacenter=dc1 room=room1 \  
row=foo rack=bar host=foo-bar-1
```

## 17.5.3 Diferença entre **ceph osd reweight** e **ceph osd crush reweight**

Há dois comandos similares que mudam o “peso” de um Ceph OSD. O contexto do uso deles é diferente e pode causar confusão.

### 17.5.3.1 **ceph osd reweight**

Uso:

```
cephuser@adm > ceph osd reweight OSD_NAME NEW_WEIGHT
```

O **`ceph osd reweight`** define um peso de substituição no Ceph OSD. Esse valor está na faixa de 0 a 1 e força o CRUSH a reposicionar os dados que, de outra forma, residiriam nesta unidade. Ele **não** muda os pesos atribuídos aos compartimentos de memória acima do OSD e é uma medida corretiva em caso de não funcionamento ou mau funcionamento da distribuição normal do CRUSH. Por exemplo, se um dos OSDs estiver em 90% e os outros estiverem em 40%, você poderá reduzir esse peso para tentar compensá-lo.



### Nota: O peso do OSD é temporário

Observe que a configuração **`ceph osd reweight`** não é persistente. Quando um OSD é marcado para ser removido, seu peso é definido como 0. Quando ele é marcado para ser incluído novamente, o peso é modificado para 1.

#### 17.5.3.2 **`ceph osd crush reweight`**

Uso:

```
cephuser@adm > ceph osd crush reweight OSD_NAME NEW_WEIGHT
```

**`ceph osd crush reweight`** define o peso de **CRUSH** do OSD. Esse peso é um valor arbitrário; normalmente, o tamanho do disco em TB, e controla a quantidade de dados que o sistema tenta alocar para o OSD.

#### 17.5.4 **Removendo um OSD**

Para remover um OSD do Mapa CRUSH de um cluster em execução, faça o seguinte:

```
cephuser@adm > ceph osd crush remove OSD_NAME
```

#### 17.5.5 **Adicionando um compartimento de memória**

Para adicionar um compartimento de memória ao Mapa CRUSH de um cluster em execução, use o comando **`ceph osd crush add-bucket`**:

```
cephuser@adm > ceph osd crush add-bucket BUCKET_NAME BUCKET_TYPE
```

## 17.5.6 Movendo um compartimento de memória

Para mover um compartimento de memória para outro local ou posição na hierarquia do Mapa CRUSH, execute o seguinte:

```
cephuser@adm > ceph osd crush move BUCKET_NAME BUCKET_TYPE=BUCKET_NAME [...]
```

Por exemplo:

```
cephuser@adm > ceph osd crush move bucket1 datacenter=dc1 room=room1 row=foo rack=bar  
host=foo-bar-1
```

## 17.5.7 Removendo um compartimento de memória

Para remover um compartimento de memória da hierarquia do Mapa CRUSH, execute o seguinte:

```
cephuser@adm > ceph osd crush remove BUCKET_NAME
```



### Nota: Apenas compartimento de memória vazio

Um compartimento de memória deve estar vazio antes de removê-lo da hierarquia do CRUSH.

## 17.6 Depurando grupos de posicionamento

Além de fazer várias cópias dos objetos, o Ceph garante a integridade dos dados *depurando* os grupos de posicionamento (há mais informações sobre grupos de posicionamento no *Livro “Guia de Implantação”, Capítulo 1 “SES e Ceph”, Seção 1.3.2 “Grupos de posicionamento”*). A depuração do Ceph equivale à execução do **fsck** na camada de armazenamento de objetos. Para cada grupo de posicionamento, o Ceph gera um catálogo de todos os objetos e compara cada objeto principal e suas réplicas para garantir que nenhum objeto esteja ausente ou seja incompatível. A depuração diária simples verifica o tamanho e os atributos dos objetos, enquanto a depuração semanal profunda lê os dados e usa checksums para garantir a integridade dos dados.

A depuração é importante para manter a integridade dos dados, mas ela pode reduzir o desempenho. Você pode ajustar as seguintes configurações para aumentar ou diminuir as operações de depuração:

#### osd max scrubs

O número máximo de operações de depuração simultâneas para um Ceph OSD. O padrão é 1.

#### osd scrub begin hour, osd scrub end hour

As horas do dia (0 a 24) que definem o intervalo para a execução da depuração. Por padrão, esse valor começa em 0 e termina em 24.



### Importante

Se o intervalo de depuração do grupo de posicionamento exceder a configuração osd scrub max interval, a depuração será executada independentemente do intervalo definido para ela.

#### osd scrub during recovery

Permite depurações durante a recuperação. Ao defini-la como “false”, a programação de novas depurações é desabilitada durante uma recuperação ativa. As depurações que já estão em execução continuam. Essa opção é útil para reduzir a carga em clusters ocupados. O padrão é “true”.

#### osd scrub thread timeout

O tempo máximo em segundos antes que um thread de depuração esgote o tempo de espera. O padrão é 60.

#### osd scrub finalize thread timeout

O tempo máximo em segundos antes que um thread de finalização da depuração esgote o tempo de espera. O padrão é 60\*10.

#### osd scrub load threshold

A carga máxima normalizada. O Ceph não efetuará a depuração quando a carga do sistema (conforme definido pela proporção de getloadavg()/número de cpus online) for superior a esse número. O padrão é 0.5.

#### osd scrub min interval

O intervalo mínimo em segundos para depuração do Ceph OSD quando a carga do cluster do Ceph está baixa. O padrão é 60\*60\*24 (uma vez por dia).



#### osd scrub max interval

O intervalo máximo em segundos para depuração do Ceph OSD, independentemente da carga do cluster. O padrão é 7\*60\*60\*24 (uma vez por semana).

#### osd scrub chunk min

O número mínimo de pacotes de armazenamento de objetos para depurar durante uma única operação. O Ceph bloqueia as gravações em um único pacote durante uma depuração. O padrão é 5.

#### osd scrub chunk max

O número máximo de pacotes de armazenamento de objetos para depurar durante uma única operação. O padrão é 25.

#### osd scrub sleep

O tempo no modo adormecido antes da depuração do próximo grupo de pacotes. O aumento desse valor desacelera toda a operação de depuração, enquanto as operações de cliente são menos afetadas. O padrão é 0.

#### osd deep scrub interval

O intervalo da depuração “profunda” (com leitura completa de todos os dados). A opção osd scrub load threshold não afeta essa configuração. O padrão é 60\*60\*24\*7 (uma vez por semana).

#### osd scrub interval randomize ratio

Adicione um atraso aleatório ao valor osd scrub min interval ao programar a próxima tarefa de depuração para um grupo de posicionamento. O atraso é um valor aleatório menor do que o resultado de osd scrub min interval \* osd scrub interval randomized ratio. Portanto, a configuração padrão distribui as depurações quase aleatoriamente dentro do período permitido de  $[1, 1,5] * \text{osd scrub min interval}$ . O padrão é 0.5.

#### osd deep scrub stride

Tamanho da leitura ao efetuar uma depuração profunda. O padrão é 524288 (512 KB).

## 18 Gerenciar pools de armazenamento

O Ceph armazena dados em pools. Pools são grupos lógicos para armazenamento de objetos. Quando você implanta um cluster pela primeira vez sem criar um pool, o Ceph usa os pools padrão para armazenar os dados. Os destaques importantes a seguir são relacionados aos pools do Ceph:

- *Resiliência:* Os pools do Ceph oferecem resiliência replicando ou codificando os dados contidos neles. É possível definir cada pool como replicado ou codificação de eliminação. Para pools replicados, você ainda define o número de réplicas, ou cópias, que cada objeto de dados terá no pool. O número de cópias (OSDs, compartimentos de memória/folhas CRUSH) que podem ser perdidas é um a menos que o número de réplicas. Com a codificação de eliminação, você define os valores de  $k$  e  $m$ , em que  $k$  é o número de pacotes de dados e  $m$  é o número de pacotes de codificação. Para os pools codificados para eliminação, esse é o número de pacotes de codificação que determina quantos OSDs (compartimentos de memória/folhas CRUSH) podem ser perdidos sem perda de dados.
- *Grupos de Posicionamento:* Você pode definir o número de grupos de posicionamento para o pool. Uma configuração típica usa aproximadamente 100 grupos de posicionamento por OSD para possibilitar o equilíbrio ideal sem usar muitos recursos de computação. Ao configurar vários pools, tenha cuidado para garantir que você defina um número adequado de grupos de posicionamento para o pool e o cluster como um todo.
- *Regras CRUSH:* Quando você armazena dados em um pool, os objetos e suas réplicas (ou blocos, no caso de pools codificados para eliminação) são posicionados de acordo com o conjunto de regras CRUSH mapeado para o pool. Você pode criar uma regra CRUSH personalizada para o pool.
- *Instantâneos:* Ao criar instantâneos com `ceph osd pool mksnap`, você efetivamente captura um instantâneo de determinado pool.

Para organizar dados em pools, você pode listar, criar e remover pools. Você também pode ver as estatísticas de uso para cada pool.

## 18.1 Criando um pool

Um pool pode ser criado como replicated para recuperar OSDs perdidos mantendo várias cópias dos objetos, ou como erasure para ter um recurso RAID5 ou 6 generalizado. Os pools replicados exigem mais armazenamento bruto, enquanto os pools codificados para eliminação exigem menos armazenamento bruto. A configuração padrão é replicated. Para obter mais informações sobre pools codificados para eliminação, consulte o [Capítulo 19, Pools codificados para eliminação](#).

Para criar um pool replicado, execute:

```
cephuser@adm > ceph osd pool create POOL_NAME
```



### Nota

O dimensionador automático cuidará dos argumentos opcionais restantes. Para obter mais informações, consulte a [Seção 17.4.12, “Habilitando o dimensionador automático de PG”](#).

Para criar um pool codificado para eliminação, execute:

```
cephuser@adm > ceph osd pool create POOL_NAME erasure CRUSH_RULESET_NAME \
EXPECTED_NUM_OBJECTS
```

O comando **ceph osd pool create** poderá falhar se você exceder o limite de grupos de posicionamento por OSD. O limite é definido com a opção mon\_max\_pg\_per\_osd.

#### POOL\_NAME

O nome do pool. Ele deve ser exclusivo. Essa opção é obrigatória.

#### POOL\_TYPE

O tipo de pool, que pode ser replicated para recuperação de OSDs perdidos mantendo várias cópias dos objetos, ou erasure para aplicar um tipo de recurso RAID5 generalizado. Os pools replicados exigem mais armazenamento bruto, porém implementam todas as operações do Ceph. Os pools de eliminação exigem menos armazenamento bruto, porém implementam apenas um subconjunto das operações disponíveis. O padrão de POOL\_TYPE é replicated.

#### CRUSH\_RULESET\_NAME

O nome do conjunto de regras CRUSH para este pool. Se o conjunto de regras especificado não existir, haverá falha na criação dos pools replicados com -ENOENT. Para pools replicados, trata-se do conjunto de regras especificado pela variável de configuração osd

`pool default CRUSH replicated ruleset`. Esse conjunto de regras deve existir. Para pools de eliminação, trata-se do “erasure-code”, se o perfil de código de eliminação for usado ou, do contrário, `POOL_NAME`. Esse conjunto de regras será criado implicitamente se ainda não existir.

`erasure_code_profile=profile`

Apenas para pools codificados para eliminação. Use o perfil de código de eliminação. Ele deve ser um perfil existente, conforme definido por `osd erasure-code-profile set`.



## Nota

Por qualquer motivo, se o dimensionador automático tiver sido desabilitado (`pg_autoscale_mode` definido como desativado) em um pool, você poderá calcular e definir os números de PGs manualmente. Consulte a [Seção 17.4, “Grupos de posicionamento”](#) para obter detalhes sobre como calcular um número apropriado de grupos de posicionamento para seu pool.

`EXPECTED_NUM_OBJECTS`

O número esperado de objetos para este pool. Ao definir esse valor (juntamente com um limite de fusão de armazenamento de arquivos negativo), a divisão da pasta PG é feita no momento da criação do pool. Isso evita o impacto da latência com uma divisão de pasta em tempo de execução.

## 18.2 Listando os pools

Para listar os pools do cluster, execute:

```
cephuser@adm > ceph osd pool ls
```

## 18.3 Renomeando um pool

Para renomear um pool, execute:

```
cephuser@adm > ceph osd pool rename CURRENT_POOL_NAME NEW_POOL_NAME
```

Se você renomear um pool e tiver recursos por pool para um usuário autenticado, deverá atualizar os recursos do usuário com o novo nome do pool.

## 18.4 Apagando um pool



### Atenção: A exclusão do pool não é reversível

Os pools podem conter dados importantes. Apagar um pool faz com que todos os dados nele desapareçam, e não é possível recuperá-los.

Como a exclusão acidental do pool é um perigo real, o Ceph implementa dois mecanismos que impedem que os pools sejam apagados. Os dois mecanismos devem ser desabilitados antes que um pool possa ser apagado.

O primeiro mecanismo é o flag `NODELETE`. Cada pool tem esse flag, e seu valor padrão é “false”. Para saber o valor desse flag em um pool, execute o seguinte comando:

```
cephuser@adm > ceph osd pool get pool_name nodelete
```

Se a saída for `nodelete: true`, não será possível apagar o pool até você mudar o flag usando o seguinte comando:

```
cephuser@adm > ceph osd pool set pool_name nodelete false
```

O segundo mecanismo é o parâmetro de configuração de todo o cluster `mon allow pool delete`, que assume como padrão “false”. Por padrão, isso significa que não é possível apagar um pool. A mensagem de erro exibida é:

```
Error EPERM: pool deletion is disabled; you must first set the
mon_allow_pool_delete config option to true before you can destroy a pool
```

Para apagar o pool mesmo com essa configuração de segurança, você pode definir `mon allow pool delete` temporariamente como “true”, apagar o pool e, em seguida, reverter o parâmetro para “false”:

```
cephuser@adm > ceph tell mon.* injectargs --mon-allow-pool-delete=true
cephuser@adm > ceph osd pool delete pool_name pool_name --yes-i-really-really-mean-it
cephuser@adm > ceph tell mon.* injectargs --mon-allow-pool-delete=false
```

O comando `injectargs` exibe a seguinte mensagem:

```
injectargs:mon_allow_pool_delete = 'true' (not observed, change may require restart)
```

Trata-se apenas de uma confirmação de que o comando foi executado com êxito. Isso não é um erro.

Se você criou seus próprios conjuntos de regras e suas próprias regras para um pool, convém removê-los quando ele não for mais necessário.

## 18.5 Outras operações

### 18.5.1 Associando pools a um aplicativo

Antes de usar os pools, você precisa associá-los a um aplicativo. Os pools que serão usados com o CephFS ou os pools criados automaticamente pelo Gateway de Objetos são associados de forma automática.

Nos outros casos, você pode associar manualmente um nome de aplicativo de formato livre a um pool:

```
cephuser@adm > ceph osd pool application enable POOL_NAME APPLICATION_NAME
```



#### Dica: Nomes de aplicativos padrão

O CephFS usa o nome do aplicativo cephfs, o Dispositivo de Blocos RADOS usa o rbd e o Gateway de Objetos usa o rgw.

É possível associar um pool a vários aplicativos, e cada aplicativo tem seus próprios metadados. Para listar o(s) aplicativo(s) associado(s) a um pool, emita o seguinte comando:

```
cephuser@adm > ceph osd pool application get pool_name
```

### 18.5.2 Definindo cotas de pool

Você pode definir cotas do pool para o número máximo de bytes e/ou para o número máximo de objetos por pool.

```
cephuser@adm > ceph osd pool set-quota POOL_NAME MAX_OBJECTS OBJ_COUNT MAX_BYTES BYTES
```

Por exemplo:

```
cephuser@adm > ceph osd pool set-quota data max_objects 10000
```

Para remover uma cota, defina o valor como 0.

## 18.5.3 Mostrando as estatísticas do pool

Para mostrar as estatísticas de uso de um pool, execute:

```
cephuser@adm > rados df
```

POOL_NAME		USED		OBJECTS	CLONES	COPIES	MISSING_ON_PRIMARY		UNFOUND
DEGRADED	RD_OPS	RD	WR_OPS	WR	USED	COMPR	UNDER	COMPR	
.rgw.root			768 KiB	4	0	12			0 0
0	44 44 KiB	4	4 KiB	0 B		0 B			
cephfs_data			960 KiB	5	0	15			0 0
0	5502 2.1 MiB	14	11 KiB	0 B		0 B			
cephfs_metadata			1.5 MiB	22	0	66			0 0
0	26 78 KiB	176	147 KiB	0 B		0 B			
default.rgw.buckets.index			0 B	1	0	3			0 0
0	4 4 KiB	1	0 B	0 B		0 B			
default.rgw.control			0 B	8	0	24			0 0
0	0 0 B	0	0 B	0 B		0 B			
default.rgw.log			0 B	207	0	621			0 0
0	5372132 5.1 GiB	3579618	0 B	0 B		0 B			
default.rgw.meta			961 KiB	6	0	18			0 0
0	155 140 KiB	14	7 KiB	0 B		0 B			
example_rbd_pool			2.1 MiB	18	0	54			0 0
0	3350841 2.7 GiB	118	98 KiB	0 B		0 B			
iscsi-images			769 KiB	8	0	24			0 0
0	1559261 1.3 GiB	61	42 KiB	0 B		0 B			
mirrored-pool			1.1 MiB	10	0	30			0 0
0	475724 395 MiB	54	48 KiB	0 B		0 B			
pool2			0 B	0	0	0			0 0
0	0 0 B	0	0 B	0 B		0 B			
pool3			333 MiB	37	0	111			0 0
0	3169308 2.5 GiB	14847	118 MiB	0 B		0 B			
pool4			1.1 MiB	13	0	39			0 0
0	1379568 1.1 GiB	16840	16 MiB	0 B		0 B			

Veja a seguir uma descrição de cada coluna:

### USED

Número de bytes usados pelo pool.

### OBJECTS

Número de objetos armazenados no pool.

### CLONES

Número de clones armazenados no pool. Quando um instantâneo é criado e faz gravações em um objeto, em vez de modificar o objeto original, o clone dele é criado para que o conteúdo do objeto original do qual foi feito o instantâneo não seja modificado.

#### COPIES

Número de réplicas do objeto. Por exemplo, se um pool replicado com o fator de replicação 3 tiver “x” objetos, ele normalmente terá  $3 * x$  cópias.

#### MISSING\_ON\_PRIMARY

Número de objetos no estado degradado (nem todas as cópias existem) enquanto a cópia está ausente no OSD principal.

#### UNFOUND

Número de objetos não encontrados.

#### DEGRADED

Número de objetos degradados.

#### RD\_OPS

Número total de operações de leitura solicitadas para este pool.

#### RD

Número total de bytes lidos deste pool.

#### WR\_OPS

Número total de operações de gravação solicitadas para este pool.

#### WR

Número total de bytes gravados no pool. Observe que isso não é igual ao uso do pool porque você pode gravar no mesmo objeto várias vezes. O resultado é que o uso do pool permanecerá o mesmo, mas o número de bytes gravados no pool aumentará.

#### USED COMPR

Número de bytes alocados para dados comprimidos.

#### UNDER COMPR

Número de bytes que os dados comprimidos ocupam quando não são comprimidos.

## 18.5.4 Obtendo valores do pool

Para obter um valor de um pool, execute o seguinte comando **get**:

```
cephuser@adm > ceph osd pool get POOL_NAME KEY
```



Você pode obter valores para as chaves listadas na [Seção 18.5.5, “Definindo valores de um pool”](#) e as chaves a seguir:

#### PG\_NUM

O número de grupos de posicionamento para o pool.

#### PGP\_NUM

O número efetivo de grupos de posicionamento a ser usado ao calcular o posicionamento dos dados. A faixa válida é igual a ou menor do que PG\_NUM.



### Dica: Todos os valores de um pool

Para listar todos os valores relacionados a um pool específico, execute:

```
cephuser@adm > ceph osd pool get POOL_NAME all
```

## 18.5.5 Definindo valores de um pool

Para definir um valor para um pool, execute:

```
cephuser@adm > ceph osd pool set POOL_NAME KEY VALUE
```

Veja a seguir uma lista de valores de pool classificados por tipo de pool:

#### VALORES COMUNS DE POOL

##### crash\_replay\_interval

O número de segundos para permitir que os clientes reproduzam solicitações confirmadas, mas não comprometidas.

##### pg\_num

O número de grupos de posicionamento para o pool. Se você adicionar novos OSDs ao cluster, verifique o valor para os grupos de posicionamento em todos os pools direcionados para os novos OSDs.

##### pgp\_num

O número efetivo de grupos de posicionamento a ser usado ao calcular o posicionamento dos dados.

### **crush\_ruleset**

O conjunto de regras a ser usado para mapear o posicionamento de objetos no cluster.

### **hashpspool**

Defina (1) ou não defina (0) o flag HASHPSPOOL em um pool específico. A habilitação desse flag muda o algoritmo para distribuir melhor os PGs pelos OSDs. Após a habilitação desse flag em um pool com o flag HASHPSPOOL definido como o padrão 0, o cluster iniciará o provisionamento para reposicionar todos os PGs corretamente. Saiba que isso pode gerar uma carga de E/S muito significativa em um cluster, portanto, não habilite o flag de 0 a 1 em clusters de produção altamente carregados.

### **nodelete**

Impede que o pool seja removido.

### **nopgchange**

Impede que pg\_num e pgp\_num do pool sejam modificados.

### **noscrub,nodeep-scrub**

Desabilita a depuração (profunda) dos dados para o pool específico a fim de resolver uma alta carga de E/S temporária.

### **write\_fadvise\_dontneed**

Defina ou cancele a definição do flag WRITE\_FADVISE\_DONTNEED nas solicitações de leitura/gravação de um determinado pool para ignorar a colocação dos dados em cache. O padrão é false. Aplica-se aos pools tanto replicados quanto EC.

### **scrub\_min\_interval**

O intervalo mínimo em segundos para depuração do pool quando a carga do cluster está baixa. O padrão 0 significa que o valor osd\_scrub\_min\_interval do arquivo de configuração do Ceph foi usado.

### **scrub\_max\_interval**

O intervalo máximo em segundos para depuração do pool, independentemente da carga do cluster. O padrão 0 significa que o valor osd\_scrub\_max\_interval do arquivo de configuração do Ceph foi usado.

### **deep\_scrub\_interval**

O intervalo em segundos para depuração do pool *em detalhes*. O padrão 0 significa que o valor osd\_deep\_scrub do arquivo de configuração do Ceph foi usado.

### size

Define o número de réplicas para os objetos no pool. Consulte a [Seção 18.5.6, “Definindo o número de réplicas do objeto”](#) para obter mais detalhes. Apenas pools replicados.

### min\_size

Define o número mínimo de réplicas necessárias para E/S. Consulte a [Seção 18.5.6, “Definindo o número de réplicas do objeto”](#) para obter mais detalhes. Apenas pools replicados.

### nosizechange

Impede que o tamanho do pool seja modificado. Quando um pool é criado, o valor padrão é obtido do valor do parâmetro `osd_pool_default_flag_nosizechange`, que é `false` por padrão. Aplica-se apenas a pools replicados porque não é possível mudar o tamanho dos pools EC.

### hit\_set\_type

Habilita o monitoramento de conjunto de acertos para pools de cache. Consulte [Filtro de Bloom](#) ([http://en.wikipedia.org/wiki/Bloom\\_filter](http://en.wikipedia.org/wiki/Bloom_filter)) [↗](#) para obter informações adicionais. Essa opção pode ter os seguintes valores: `bloom`, `explicit_hash`, `explicit_object`. O padrão é `bloom`, os outros valores são apenas para teste.

### hit\_set\_count

O número de conjuntos de acertos para armazenar nos pools de cache. Quanto maior o número, mais RAM é consumida pelo daemon `ceph-osd`. O padrão é `0`.

### hit\_set\_period

A duração em segundos de um período do conjunto de acertos para os pools de cache. Quanto maior o número, mais RAM é consumida pelo daemon `ceph-osd`. Quando um pool é criado, o valor padrão é obtido do valor do parâmetro `osd_tier_default_cache_hit_set_period`, que é `1200` por padrão. Aplica-se apenas a pools replicados porque os pools EC não podem ser usados como camada de cache.

### hit\_set\_fpp

A probabilidade de falsos positivos para o tipo de conjunto de acertos bloom. Consulte [Filtro de Bloom](#) ([http://en.wikipedia.org/wiki/Bloom\\_filter](http://en.wikipedia.org/wiki/Bloom_filter)) [↗](#) para obter informações adicionais. A faixa válida é de `0,0` a `1,0`. O padrão é `0,05`

#### `use_gmt_hitset`

Force os OSDs a usar marcações de horário em GMT (Horário de Greenwich) ao criar um conjunto de acertos para camadas de cache. Isso garante que os nós em fusos horários diferentes retornem o mesmo resultado. O padrão é 1. Esse valor não deve ser mudado.

#### `cache_target_dirty_ratio`

A porcentagem do pool de cache que contém os objetos modificados antes que o agente de camadas de cache os descarregue para o pool de armazenamento de suporte. O padrão é 0.4.

#### `cache_target_dirty_high_ratio`

A porcentagem do pool de cache que contém os objetos modificados antes que o agente de camadas de cache os descarregue para o pool de armazenamento de suporte com uma velocidade maior. O padrão é 0.6.

#### `cache_target_full_ratio`

A porcentagem do pool de cache que contém os objetos não modificados (limpos) antes que o agente de camadas de cache os elimine do pool de cache. O padrão é 0.8.

#### `target_max_bytes`

O Ceph iniciará o descarregamento ou a eliminação de objetos quando o limite max\_bytes for acionado.

#### `target_max_objects`

O Ceph iniciará o descarregamento ou a eliminação de objetos quando o limite max\_objects for acionado.

#### `hit_set_grade_decay_rate`

Taxa de redução de temperatura entre dois hit\_sets sucessivos. O padrão é 20.

#### `hit_set_search_last_n`

Considera no máximo N aparições nos hit\_sets para o cálculo da temperatura. O padrão é 1.

#### `cache_min_flush_age`

O tempo (em segundos) antes que o agente de camadas de cache descarregue um objeto do pool de cache para o pool de armazenamento.

#### `cache_min_evict_age`

O tempo (em segundos) antes que o agente de camadas de cache elimine um objeto do pool de cache.

**fast\_read**

Se esse flag estiver habilitado nos pools com codificação de eliminação, a solicitação de leitura emitirá subleituradas para todos os fragmentos e aguardará até receber fragmentos suficientes para decodificar e atender ao cliente. No caso dos plug-ins de eliminação *jerasure* e *isa*, quando as primeiras  $K$  respostas são retornadas, a solicitação do cliente é atendida imediatamente, usando os dados decodificados dessas respostas. Essa abordagem gera mais carga de CPU e menos carga de disco/rede. No momento, esse flag é suportado apenas para pools com codificação de eliminação. O padrão é `0`.

## 18.5.6 Definindo o número de réplicas do objeto

Para definir o número de réplicas do objeto em um pool replicado, execute o seguinte:

```
cephuser@adm > ceph osd pool set poolname size num-replicas
```

O `num-replicas` inclui o próprio objeto. Por exemplo, se você deseja o objeto e duas cópias dele para um total de três instâncias do objeto, especifique 3.



### Atenção: Não defina menos do que 3 réplicas

Se você definir `num-replicas` como 2, haverá apenas *uma* cópia dos dados. Se você perder uma instância do objeto, precisará confiar que a outra cópia não foi corrompida desde a última depuração durante a recuperação, por exemplo (consulte a [Seção 17.6, “Depurando grupos de posicionamento”](#) para obter detalhes).

A definição de um pool para uma réplica significa que existe exatamente *uma* instância do objeto de dados no pool. Se houver falha no OSD, você perderá os dados. Um uso possível para um pool com uma réplica é armazenar dados temporários por um curto período.



### Dica: Definindo mais do que 3 réplicas

A definição de 4 réplicas para um pool aumenta a confiabilidade em 25%.

No caso de dois data centers, você precisa definir pelo menos 4 réplicas para que um pool tenha duas cópias em cada data center. Desse modo, se um data center for perdido, ainda haverá duas cópias, e você poderá perder um disco sem perder os dados.



## Nota

Um objeto pode aceitar E/S no modo degradado com menos do que `pool size` réplicas. Para definir um número mínimo de réplicas necessárias para E/S, você deve usar a configuração `min_size`. Por exemplo:

```
cephuser@adm > ceph osd pool set data min_size 2
```

Isso garante que nenhum objeto no pool de dados receba E/S com menos do que `min_size` réplicas.



## Dica: Obter o número de réplicas do objeto

Para obter o número de réplicas do objeto, execute o seguinte:

```
cephuser@adm > ceph osd dump | grep 'replicated size'
```

O Ceph listará os pools, com o atributo `replicated size` realçado. Por padrão, o Ceph cria duas réplicas de um objeto (um total de três cópias, ou um tamanho de 3).

## 18.6 Migração de pool

Ao criar um pool (consulte a [Seção 18.1, “Criando um pool”](#)), você precisa especificar os parâmetros iniciais, como o tipo de pool ou o número de grupos de posicionamento. Posteriormente, se você decidir mudar qualquer um desses parâmetros, por exemplo, ao converter um pool replicado em um codificado para eliminação ou reduzir o número de grupos de posicionamento, será necessário migrar os dados do pool para outro cujos parâmetros sejam mais adequados à sua implantação.

Esta seção descreve dois métodos de migração: *camada de cache* para migração geral de dados do pool, e o método que usa os subcomandos `rbd migrate` para migrar imagens RBD para um novo pool. Cada método tem suas especificações e limitações.

## 18.6.1 Limitações

- Você pode usar o método de *camada de cache* para migrar de um pool replicado para um pool EC ou para outro replicado. A migração de um pool EC não é suportada.
- Não é possível migrar imagens RBD e exportações do CephFS de um pool replicado para um pool com EC. O motivo é que os pools EC não suportam `omap`, e o RBD e o CephFS usam o `omap` para armazenar seus metadados. Por exemplo, haverá falha ao descarregar o objeto de cabeçalho do RBD. Porém, você pode migrar os dados para o pool EC, deixando os metadados no pool replicado.
- O método **rbd migration** permite migrar imagens com tempo de espera mínimo do cliente. Você apenas precisa parar o cliente antes da etapa de preparação e iniciá-lo depois. Observe que apenas um cliente `librbd` com suporte a esse recurso (Ceph Nautilus ou mais recente) poderá abrir a imagem logo após a etapa de preparação. Os clientes `librbd` mais antigos ou os clientes `krbd` não poderão abrir a imagem antes da execução da etapa de confirmação.

## 18.6.2 Migração usando a camada de cache

O princípio é simples: incluir o pool que você precisa migrar para a camada de cache na ordem inversa. O exemplo a seguir migra um pool replicado chamado “testpool” para um pool codificado para eliminação:

### PROCEDIMENTO 18.1: MIGRANDO UM POOL REPLICADO PARA UM CODIFICADO PARA ELIMINAÇÃO

1. Crie um novo pool codificado para eliminação chamado “newpool”. Consulte a [Seção 18.1, “Criando um pool”](#) para obter uma explicação detalhada dos parâmetros de criação de pool.

```
cephuser@adm > ceph osd pool create newpool erasure default
```

Verifique se o chaveiro do cliente usado oferece pelo menos os mesmos recursos do “testpool” para o “newpool”.

Agora você tem dois pools: o “testpool” replicado original preenchido com dados e o novo “newpool” codificado para eliminação vazio:

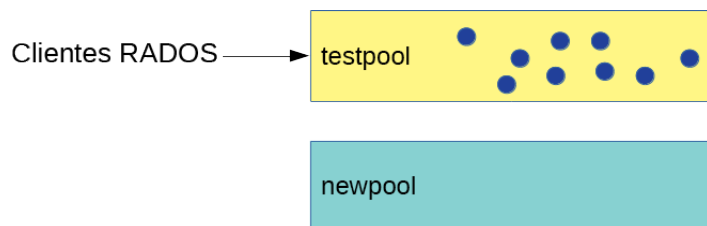


FIGURA 18.1: POOLS ANTES DA MIGRAÇÃO

2. Configure a camada de cache e defina o pool replicado “testpool” como o pool de cache. A opção `--force-nonempty` permite adicionar uma camada de cache mesmo que o pool já tenha dados:

```
cephuser@adm > ceph tell mon.* injectargs \
'--mon_debug_unsafe_allow_tier_with_nonempty_snaps=1'
cephuser@adm > ceph osd tier add newpool testpool --force-nonempty
cephuser@adm > ceph osd tier cache-mode testpool proxy
```

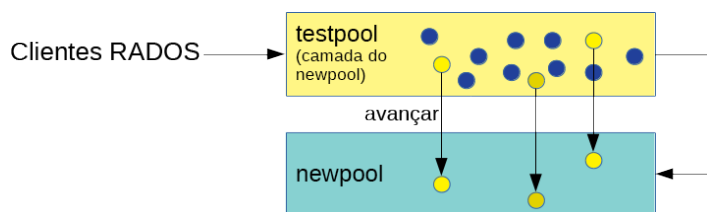


FIGURA 18.2: CONFIGURAÇÃO DA CAMADA DE CACHE

3. Force o pool de cache a mover todos os objetos para o novo pool:

```
cephuser@adm > rados -p testpool cache-flush-evict-all
```

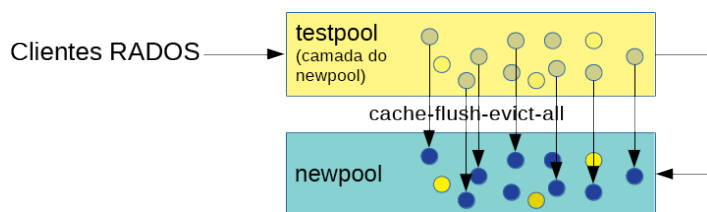


FIGURA 18.3: DESCARREGANDO DADOS



- Até todos os dados serem descarregados para o novo pool codificado para eliminação, você precisa especificar uma sobreposição para que esses objetos sejam pesquisados no pool antigo:

```
cephuser@adm > ceph osd tier set-overlay newpool testpool
```

Com a sobreposição, todas as operações são encaminhadas para o “testpool” replicado antigo:

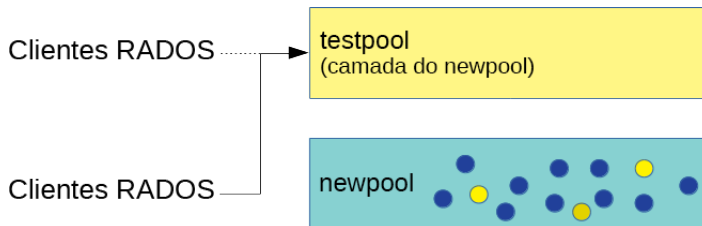


FIGURA 18.4: DEFININDO A SOBREPOSIÇÃO

Agora você pode alternar todos os clientes para acessar objetos no novo pool.

- Após a migração de todos os dados para o “newpool” codificado para eliminação, remova a sobreposição e o pool de cache antigo “testpool”:

```
cephuser@adm > ceph osd tier remove-overlay newpool  
cephuser@adm > ceph osd tier remove newpool testpool
```

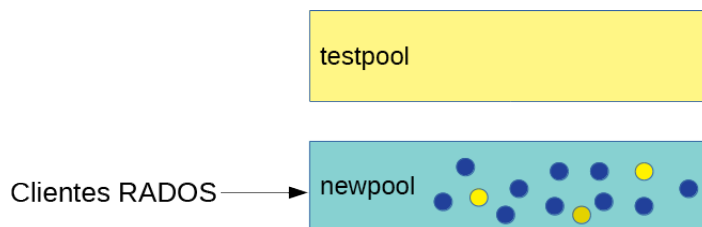


FIGURA 18.5: MIGRAÇÃO CONCLUÍDA

- Execute:

```
cephuser@adm > ceph tell mon.* injectargs \  
'--mon_debug_unsafe_allow_tier_with_nonempty_snaps=0'
```

### 18.6.3 Migrando imagens RBD

Veja a seguir a maneira recomendada de migrar imagens RBD entre dois pools replicados.

1. Impeça os clientes (como uma máquina virtual) de acessar a imagem RBD.
2. Crie uma nova imagem no pool de destino, com o pai definido como a imagem de origem:

```
cephuser@adm > rbd migration prepare SRC_POOL/IMAGE TARGET_POOL/IMAGE
```



### Dica: Migrar apenas dados para um pool codificado para eliminação

Se você precisa migrar apenas os dados da imagem para um novo pool EC e deixar os metadados no pool replicado original, execute o seguinte comando:

```
cephuser@adm > rbd migration prepare SRC_POOL/IMAGE \  
--data-pool TARGET_POOL/IMAGE
```

3. Permita que os clientes acessem a imagem no pool de destino.
4. Migre os dados para o pool de destino:

```
cephuser@adm > rbd migration execute SRC_POOL/IMAGE
```

5. Remova a imagem antiga:

```
cephuser@adm > rbd migration commit SRC_POOL/IMAGE
```

## 18.7 Instantâneos de pool

Os instantâneos de pool são capturados com base no estado do pool inteiro do Ceph. Com os instantâneos de pool, você pode manter o histórico de estado do pool. A criação de instantâneos de pool consome espaço de armazenamento proporcional ao tamanho do pool. Confira sempre se há espaço em disco suficiente no armazenamento relacionado antes de criar um instantâneo de um pool.

### 18.7.1 Criando um instantâneo de um pool

Para criar um instantâneo de um pool, execute:

```
cephuser@adm > ceph osd pool mksnap POOL-NAME SNAP-NAME
```

Por exemplo:

```
cephuser@adm > ceph osd pool mksnap pool1 snap1
created pool pool1 snap snap1
```

## 18.7.2 Listando instantâneos de um pool

Para listar os instantâneos existentes de um pool, execute:

```
cephuser@adm > rados lssnap -p POOL_NAME
```

Por exemplo:

```
cephuser@adm > rados lssnap -p pool1
1 snap1 2018.12.13 09:36:20
2 snap2 2018.12.13 09:46:03
2 snaps
```

## 18.7.3 Removendo um instantâneo de um pool

Para remover um instantâneo de um pool, execute:

```
cephuser@adm > ceph osd pool rmsnap POOL-NAME SNAP-NAME
```

## 18.8 Compactação de dados

O BlueStore (encontre mais detalhes no *Livro “Guia de Implantação”, Capítulo 1 “SES e Ceph”, Seção 1.4 “BlueStore”*) oferece compactação de dados sob demanda para economizar espaço no disco. A taxa de compactação depende dos dados armazenados no sistema. Observe que a compactação/descompactação requer mais capacidade da CPU.

Você pode configurar a compactação de dados globalmente (consulte a [Seção 18.8.3, “Opções globais de compactação”](#)) e, em seguida, anular as configurações de compactação específicas para cada pool individual.

Você pode habilitar ou desabilitar a compactação de dados do pool ou mudar o algoritmo e o modo de compactação a qualquer momento, para um pool tanto com dados quanto sem dados. Nenhuma compactação será aplicada aos dados existentes após habilitar a compactação do pool. Após desabilitar a compactação de um pool, todos os dados dele serão descompactados.

## 18.8.1 Habilitando a compactação

Para habilitar a compactação de dados para um pool denominado *POOL\_NAME*, execute o seguinte comando:

```
cephuser@adm > ceph osd pool set POOL_NAME compression_algorithm COMPRESSION_ALGORITHM  
cephuser@adm > ceph osd pool set POOL_NAME compression_mode COMPRESSION_MODE
```



### Dica: Desabilitando a compactação do pool

Para desabilitar a compactação de dados para um pool, use “none” (nenhum) como o algoritmo de compactação:

```
cephuser@adm > ceph osd pool set POOL_NAME compression_algorithm none
```

## 18.8.2 Opções de compactação do pool

Uma lista completa de configurações de compactação:

### compression\_algorithm

Os valores possíveis são none, zstd e snappy. O padrão é snappy.

O algoritmo de compactação a ser usado depende do caso de uso específico. Veja a seguir várias recomendações:

- Use o padrão snappy se não tiver um bom motivo para mudá-lo.
- O zstd oferece uma boa taxa de compactação, mas provoca alto overhead da CPU ao compactar pequenas quantidades de dados.
- Realize um benchmark desses algoritmos em uma amostra dos dados reais e observe o uso de CPU e memória do cluster.

### compression\_mode

Os valores possíveis são none, aggressive, passive e force. O padrão é none.

- none: nunca comprimir
- passive: comprimir se houver a dica COMPRESSIBLE

- aggressive: comprimir, exceto se houver a dica INCOMPRESSIBLE
- force: sempre comprimir

#### **compression\_required\_ratio**

Valor: Duplo, Taxa =  $\text{SIZE\_COMPRESSED}/\text{SIZE\_ORIGINAL}$ . O padrão é 0,875, o que significa que, se a compactação não reduzir o espaço ocupado em pelo menos 12,5%, o objeto não será comprimido.

Os objetos acima dessa taxa não serão comprimidos por causa do baixo ganho líquido.

#### **compression\_max\_blob\_size**

Valor: Número Inteiro Não Assinado, tamanho em bytes. Padrão: 0

Tamanho máximo dos objetos que serão comprimidos.

#### **compression\_min\_blob\_size**

Valor: Número Inteiro Não Assinado, tamanho em bytes. Padrão: 0

Tamanho mínimo dos objetos que serão comprimidos.

### 18.8.3 Opções globais de compactação

As seguintes opções de configuração podem ser definidas na configuração do Ceph e aplicam-se a todos os OSDs, não apenas a um único pool. A configuração específica do pool listada na [Seção 18.8.2, “Opções de compactação do pool”](#) tem prioridade.

#### **bluestore\_compression\_algorithm**

Consulte a [compression\\_algorithm](#)

#### **bluestore\_compression\_mode**

Consulte a [compression\\_mode](#)

#### **bluestore\_compression\_required\_ratio**

Consulte a [compression\\_required\\_ratio](#)

#### **bluestore\_compression\_min\_blob\_size**

Valor: Número Inteiro Não Assinado, tamanho em bytes. Padrão: 0

Tamanho mínimo dos objetos que serão comprimidos. Por padrão, a configuração é ignorada a favor de bluestore\_compression\_min\_blob\_size\_hdd e bluestore\_compression\_min\_blob\_size\_ssd. Ela tem prioridade quando definida como um valor diferente de zero.

#### **bluestore\_compression\_max\_blob\_size**

Valor: Número Inteiro Não Assinado, tamanho em bytes. Padrão: 0

Tamanho máximo dos objetos que são comprimidos antes de serem divididos em blocos menores. Por padrão, a configuração é ignorada a favor de bluestore\_compression\_max\_blob\_size\_hdd e bluestore\_compression\_max\_blob\_size\_ssd. Ela tem prioridade quando definida como um valor diferente de zero.

#### **bluestore\_compression\_min\_blob\_size\_ssd**

Valor: Número Inteiro Não Assinado, tamanho em bytes. Padrão: 8K

Tamanho mínimo dos objetos que serão comprimidos e armazenados na unidade de estado sólido.

#### **bluestore\_compression\_max\_blob\_size\_ssd**

Valor: Número Inteiro Não Assinado, tamanho em bytes. Padrão: 64K

Tamanho máximo dos objetos que são comprimidos e armazenados em unidade de estado sólido antes de serem divididos em blocos menores.

#### **bluestore\_compression\_min\_blob\_size\_hdd**

Valor: Número Inteiro Não Assinado, tamanho em bytes. Padrão: 128K

Tamanho mínimo dos objetos que serão comprimidos e armazenados em discos rígidos.

#### **bluestore\_compression\_max\_blob\_size\_hdd**

Valor: Número Inteiro Não Assinado, tamanho em bytes. Padrão: 512K

Tamanho máximo dos objetos que são comprimidos e armazenados em discos rígidos antes de serem divididos em blocos menores.

## 19 Pools codificados para eliminação

O Ceph oferece uma alternativa à replicação normal de dados em pools conhecidos como de *eliminação* ou *codificados para eliminação*. Os pools de eliminação não oferecem todas as funcionalidades que os pools *replicados* (por exemplo, eles não podem armazenar metadados para pools RBD), mas exigem menos armazenamento bruto. Um pool de eliminação padrão capaz de armazenar 1 TB de dados requer 1,5 TB de armazenamento bruto, o que permite a falha de um único disco. Isso equivale a um pool replicado que precisa de 2 TB de armazenamento bruto para a mesma finalidade.

Para obter informações sobre o Código de Eliminação, visite [https://en.wikipedia.org/wiki/Erasure\\_code](https://en.wikipedia.org/wiki/Erasure_code).

Para obter uma lista de valores de pool relacionados a pools EC, consulte [Valores de pool codificado para eliminação](#).

### 19.1 Pré-requisito para pools codificados para eliminação

Para usar a codificação de eliminação, você precisa:

- Definir uma regra de eliminação no Mapa CRUSH.
- Definir um perfil de código de eliminação que especifica o algoritmo de codificação a ser usado.
- Criar um pool usando a regra e o perfil mencionados anteriormente.

Lembre-se de que a modificação do perfil e dos detalhes no perfil não será possível depois que o pool for criado e tiver dados.

Verifique se as regras CRUSH para os *pools de eliminação* usam `indep` para `step`. Para saber os detalhes, consulte a [Seção 17.3.2, “firstn e indep”](#).

## 19.2 Criando um pool codificado para eliminação de exemplo

O pool codificado para eliminação mais simples é equivalente ao RAID5 e requer pelo menos três hosts. Este procedimento descreve como criar um pool para fins de teste.

1. O comando **ceph osd pool create** é usado para criar um pool do tipo de *eliminação*. 12 representa o número de grupos de posicionamento. Com os parâmetros padrão, o pool é capaz de resolver a falha de um OSD.

```
cephuser@adm > ceph osd pool create ecpool 12 12 erasure  
pool 'ecpool' created
```

2. A string ABCDEFGHI é gravada em um objeto denominado NYAN.

```
cephuser@adm > echo ABCDEFGHI | rados --pool ecpool put NYAN -
```

3. Para fins de teste, os OSDs agora podem ser desabilitados. Por exemplo, desconecte-os da rede.
4. Para testar se o pool é capaz de resolver a falha de dispositivos, o conteúdo do arquivo pode ser acessado com o comando **rados**.

```
cephuser@adm > rados --pool ecpool get NYAN -  
ABCDEFGHI
```

## 19.3 Perfis de código de eliminação

Quando o comando **ceph osd pool create** é invocado para criar um *pool de eliminação*, o perfil padrão é usado, a menos que outro perfil seja especificado. Os perfis definem a redundância dos dados. Para fazer isso, defina dois parâmetros denominados aleatoriamente k e m. k e m definem em quantos pacotes os dados são divididos e quantos pacotes de codificação são criados. Em seguida, os pacotes redundantes são armazenados em OSDs diferentes.

Definições necessárias para perfis de pool de eliminação:

### chunk

quando a função de codificação é chamada, ela retorna pacotes do mesmo tamanho: pacotes de dados que podem ser concatenados para reconstruir o objeto original e pacotes de codificação que podem ser usados para reconstruir um pacote perdido.



**k**

o número de pacotes de dados, que é o número de pacotes em que objeto original é dividido. Por exemplo, se  $k = 2$ , um objeto de 10 kB será dividido em  $k$  objetos de 5 kB cada um. O `min_size` padrão nos pools codificados para eliminação é  $k + 1$ . No entanto, recomendamos que o `min_size` seja no mínimo  $k + 2$  para evitar perda de gravações e dados.

**m**

o número de pacotes de codificação, que é o número de pacotes adicionais calculado pelas funções de codificação. Se houver 2 pacotes de codificação, isso significa que 2 OSDs poderão ser eliminados sem perda de dados.

#### **crush-failure-domain**

define para quais dispositivos os pacotes são distribuídos. Um tipo de compartimento de memória precisa ser definido como valor. Para todos os tipos de compartimento de memória, consulte a [Seção 17.2, “Compartimentos de memória”](#). Se o domínio de falha for `rack`, os pacotes serão armazenados em racks diferentes para aumentar a resiliência em caso de falhas no rack. Observe que isso exige  $k + m$  racks.

Com o perfil de código de eliminação padrão usado na [Seção 19.2, “Criando um pool codificado para eliminação de exemplo”](#), você não perderá os dados do cluster se houver falha em um único OSD ou host. Dessa forma, para armazenar 1 TB de dados, ele precisa de mais 0,5 TB de armazenamento bruto. Isso significa que 1,5 TB de armazenamento bruto é necessário para 1 TB de dados (porque  $k = 2$ ,  $m = 1$ ). Isso equivale a uma configuração RAID 5 comum. Para comparação, um pool replicado precisa de 2 TB de armazenamento bruto para armazenar 1 TB de dados.

As configurações do perfil padrão podem ser exibidas com:

```
cephuser@adm > ceph osd erasure-code-profile get default
directory=.libs
k=2
m=1
plugin=jerasure
crush-failure-domain=host
technique=reed_sol_van
```

A escolha do perfil correto é importante, porque ele não poderá ser modificado após a criação do pool. Um novo pool com um perfil diferente precisa ser criado, e todos os objetos do pool anterior precisam ser movidos para o novo (consulte a [Seção 18.6, “Migração de pool”](#)).

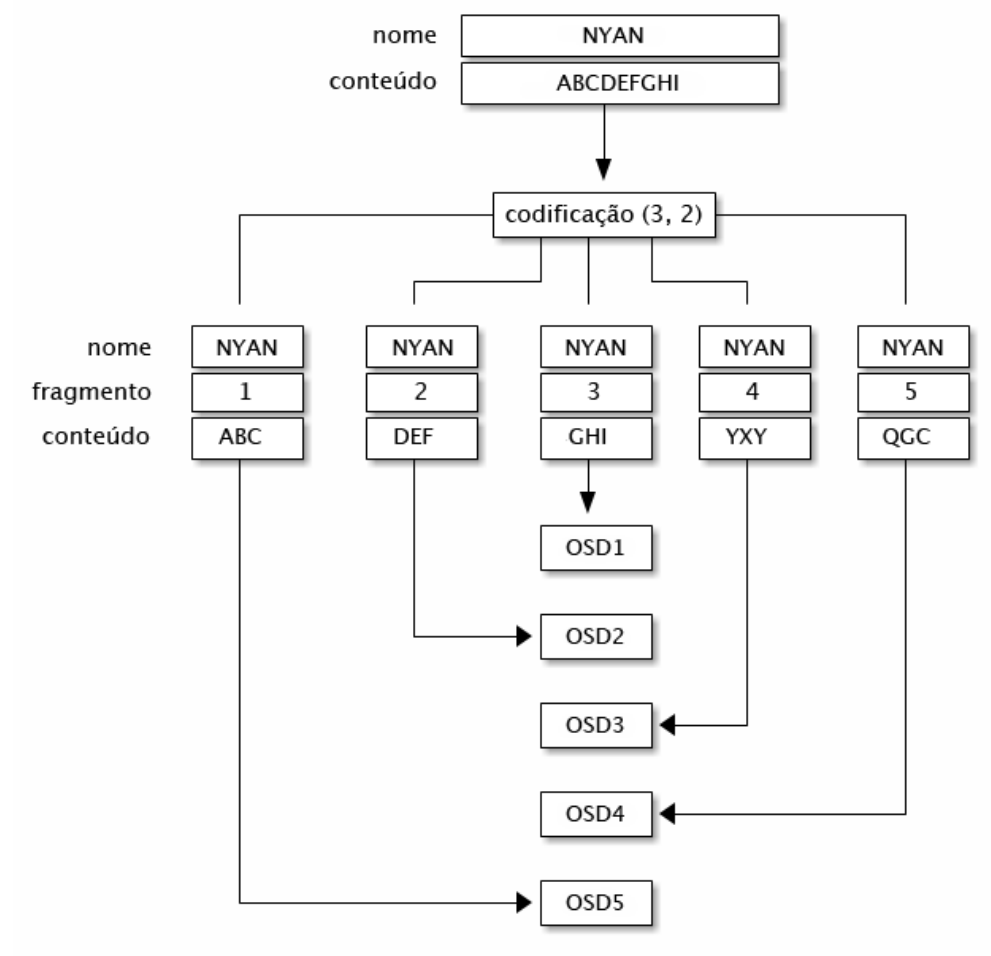
Os parâmetros mais importantes do perfil são k, m e crush-failure-domain porque definem o overhead de armazenamento e a durabilidade dos dados. Por exemplo, se a arquitetura desejada tiver que sustentar a perda de dois racks com um overhead de armazenamento de 66%, o seguinte perfil poderá ser definido. Observe que isso é válido apenas com um Mapa CRUSH que tenha compartimentos de memória do tipo “rack”:

```
cephuser@adm > ceph osd erasure-code-profile set myprofile \  
k=3 \  
m=2 \  
crush-failure-domain=rack
```

O exemplo na [Seção 19.2, “Criando um pool codificado para eliminação de exemplo”](#) pode ser repetido com este novo perfil:

```
cephuser@adm > ceph osd pool create ecpool 12 12 erasure myprofile  
cephuser@adm > echo ABCDEFGHI | rados --pool ecpool put NYAN -  
cephuser@adm > rados --pool ecpool get NYAN -  
ABCDEFGHI
```

O objeto NYAN será dividido em três (k=3), e dois pacotes adicionais serão criados (m=2). O valor de m define quantos OSDs podem ser perdidos simultaneamente sem nenhuma perda de dados. O crush-failure-domain=rack criará um conjunto de regras CRUSH para garantir que dois pacotes não sejam armazenados no mesmo rack.



### 19.3.1 Criando um novo perfil de código de eliminação

O comando a seguir cria um novo perfil de código de eliminação:

```
# ceph osd erasure-code-profile set NAME \
  directory=DIRECTORY \
  plugin=PLUGIN \
  stripe_unit=STRIPE_UNIT \
  KEY=VALUE ... \
  --force
```

#### DIRECTORY

Opcional. Defina o nome do diretório do qual o plug-in de código de eliminação é carregado. O padrão é /usr/lib/ceph/erasure-code.

## PLUGIN

Opcional. Use o plug-in de código de eliminação para calcular blocos de codificação e recuperar blocos ausentes. Os plug-ins disponíveis são “jerasure”, “isa”, “lrc” e “shes”. O padrão é “jerasure”.

## STRIPE\_UNIT

Opcional. A quantidade de dados em um pacote, por distribuição. Por exemplo, um perfil com 2 pacotes de dados e `stripe_unit = 4K` coloca a faixa 0-4K no pacote 0, 4K-8K no pacote 1 e 8K-12K no pacote 0 novamente. Esse valor deve ser um múltiplo de 4K para obter o melhor desempenho. O valor padrão é extraído da opção de configuração do monitor `osd_pool_erasure_code_stripe_unit` quando um pool é criado. O "stripe\_width" de um pool que usa este perfil será o número de pacotes de dados multiplicado por esta "stripe\_unit".

## KEY=VALUE

Os pares de chave/valor das opções específicas ao plug-in de código de eliminação selecionado.

## --force

Opcional. Substitua um perfil existente pelo mesmo nome e permita definir uma `stripe_unit` sem alinhamento de 4K.

## 19.3.2 Removendo um perfil de código de eliminação

O comando a seguir remove um perfil de código de eliminação conforme identificado por seu NAME:

```
# ceph osd erasure-code-profile rm NAME
```



### Importante

Se o perfil for referenciado por um pool, haverá falha na exclusão.

### 19.3.3 Exibindo detalhes do perfil de código de eliminação

O comando a seguir exibe os detalhes de um perfil de código de eliminação conforme identificado por seu NAME:

```
# ceph osd erasure-code-profile get NAME
```

### 19.3.4 Listando perfis de código de eliminação

O comando a seguir lista os nomes de todos os perfis de código de eliminação:

```
# ceph osd erasure-code-profile ls
```

## 19.4 Marcando pools codificados para eliminação com dispositivo de blocos RADOS

Para marcar um pool EC como pool RBD, sinalize-o de acordo:

```
cephuser@adm > ceph osd pool application enable rbd ec_pool_name
```

O RBD pode armazenar *dados* da imagem em pools EC. No entanto, o cabeçalho e os metadados da imagem ainda precisam ser armazenados em um pool replicado. Considerando que você tem um pool chamado “rbd” para esta finalidade:

```
cephuser@adm > rbd create rbd/image_name --size 1T --data-pool ec_pool_name
```

Você pode usar a imagem normalmente, como qualquer outra, com exceção de que todos os dados serão armazenados no pool ec\_pool\_name em vez do “rbd”.

## 20 Dispositivo de blocos RADOS

Um bloco é uma sequência de bytes. Por exemplo, um bloco de dados de 4 MB. As interfaces de armazenamento com base em blocos são a maneira mais comum para armazenar dados com mídia rotativa, como discos rígidos, CDs e disquetes. A onipresença de interfaces de dispositivo de blocos faz do dispositivo de blocos virtual o candidato ideal para interagir com um sistema de armazenamento de dados em massa, como o Ceph.

Os dispositivos de blocos do Ceph permitem o compartilhamento de recursos físicos e são redimensionáveis. Eles armazenam dados distribuídos por vários OSDs em um cluster do Ceph. Os dispositivos de blocos do Ceph aproveitam os recursos do RADOS, como criação de instantâneos, replicação e consistência. Os Dispositivo de Blocos RADOS (RBD) do Ceph interagem com os OSDs usando os módulos do kernel ou a biblioteca `librbd`.

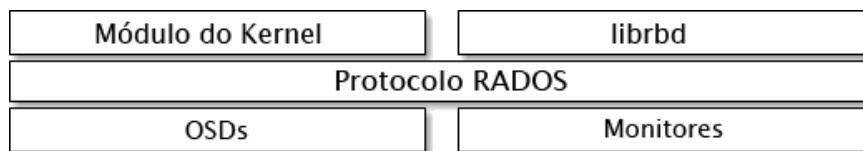


FIGURA 20.1: PROTOCOLO RADOS

Os dispositivos de blocos do Ceph oferecem alto desempenho com escalabilidade infinita aos módulos do kernel. Eles suportam soluções de virtualização, como QEMU, ou sistemas de computação com base em nuvem, como OpenStack, que utilizam a `libvirt`. Você pode usar o mesmo cluster para operar o Gateway de Objetos, o CephFS e os Dispositivos de Blocos RADOS simultaneamente.

### 20.1 Comandos do dispositivo de blocos

O comando `rbd` permite criar, listar, avaliar e remover imagens de dispositivo de blocos. Você também pode usá-lo, por exemplo, para clonar imagens, criar instantâneos, voltar uma imagem para um instantâneo ou ver um instantâneo.

## 20.1.1 Criando uma imagem de dispositivo de blocos em um pool replicado

Antes que você possa adicionar um dispositivo de blocos a um cliente, precisa criar uma imagem relacionada em um pool existente (consulte o [Capítulo 18, Gerenciar pools de armazenamento](#)):

```
cephuser@adm > rbd create --size MEGABYTES POOL-NAME/IMAGE-NAME
```

Por exemplo, para criar uma imagem de 1 GB denominada “myimage” que armazena informações em um pool chamado “mypool”, execute o seguinte:

```
cephuser@adm > rbd create --size 1024 mypool/myimage
```



### Dica: Unidades de tamanho de imagem

Se você omitir um atalho de unidade de tamanho (“G” ou “T”), o tamanho da imagem será em megabytes. Use “G” ou “T” após o número do tamanho para especificar gigabytes ou terabytes.

## 20.1.2 Criando uma imagem de dispositivo de blocos em um pool codificado para eliminação

É possível armazenar dados de uma imagem de dispositivo de blocos diretamente em pools codificados para eliminação (EC, Erasure Coded). A imagem de um Dispositivo de Blocos RADOS consiste nas partes de *dados* e *metadados*. Você pode armazenar apenas a parte de dados de uma imagem de Dispositivo de Blocos RADOS em um pool EC. O pool precisa ter o flag overwrite definido como *true*, e isso apenas será possível se todos os OSDs em que o pool está armazenado usarem o BlueStore.

Você não pode armazenar a parte de metadados da imagem em um pool EC. Você pode especificar o pool replicado para armazenar os metadados da imagem com a opção --pool= do comando **rbd create** ou especificar pool/ como prefixo para o nome da imagem.

Crie um pool EC:

```
cephuser@adm > ceph osd pool create EC_POOL 12 12 erasure
cephuser@adm > ceph osd pool set EC_POOL allow_ec_overwrites true
```

Especifique o pool replicado para armazenar os metadados:

```
cephuser@adm > rbd create IMAGE_NAME --size=1G --data-pool EC_POOL --pool=POOL
```

Ou:

```
cephuser@adm > rbd create POOL/IMAGE_NAME --size=1G --data-pool EC_POOL
```

### 20.1.3 Listando imagens de dispositivo de blocos

Para listar os dispositivos de blocos em um pool chamado “mypool”, execute o seguinte:

```
cephuser@adm > rbd ls mypool
```

### 20.1.4 Recuperando informações da imagem

Para recuperar as informações de uma imagem “myimage” em um pool chamado “mypool”, execute o seguinte:

```
cephuser@adm > rbd info mypool/myimage
```

### 20.1.5 Redimensionando uma imagem de dispositivo de blocos

As imagens de Dispositivo de Blocos RADOS são aprovisionadas dinamicamente, elas não usam nenhum armazenamento físico até você começar a gravar dados nelas. No entanto, elas têm uma capacidade máxima que você define com a opção `--size`. Para aumentar (ou diminuir) o tamanho máximo da imagem, execute o seguinte:

```
cephuser@adm > rbd resize --size 2048 POOL_NAME/IMAGE_NAME # to increase  
cephuser@adm > rbd resize --size 2048 POOL_NAME/IMAGE_NAME --allow-shrink # to decrease
```

### 20.1.6 Removendo uma imagem de dispositivo de blocos

Para remover um dispositivo de blocos correspondente a uma imagem “myimage” no pool chamado “mypool”, execute o seguinte:

```
cephuser@adm > rbd rm mypool/myimage
```



## 20.2 Montando e desmontando

Após criar um Dispositivo de Blocos RADOS, você poderá usá-lo como qualquer outro dispositivo de disco: formatá-lo, montá-lo para poder trocar arquivos e desmontá-lo depois de concluído.

O comando **rbd** usa como padrão o acesso ao cluster por meio da conta do usuário admin do Ceph. Essa conta tem acesso administrativo completo ao cluster. Esse comportamento cria o risco de danos acidentais, similar ao login em uma estação de trabalho Linux como root. Portanto, é preferível criar contas dos usuários com menos privilégios e usá-las para acesso normal de leitura/gravação ao Dispositivo de Blocos RADOS.

### 20.2.1 Criando uma conta do usuário do Ceph

Para criar uma nova conta do usuário com os recursos Ceph Manager, Ceph Monitor e Ceph OSD, use o comando **ceph** com o subcomando **auth get-or-create**:

```
cephuser@adm > ceph auth get-or-create client.ID mon 'profile rbd' osd 'profile profile
name \
  [pool=pool-name] [, profile ...]' mgr 'profile rbd [pool=pool-name]'
```

Por exemplo, para criar um usuário chamado qemu com acesso de leitura/gravação ao pool vms e acesso apenas leitura ao pool images, execute o seguinte:

```
ceph auth get-or-create client.qemu mon 'profile rbd' osd 'profile rbd pool=vms, profile
rbd-read-only pool=images' \
  mgr 'profile rbd pool=images'
```

A saída do comando **ceph auth get-or-create** será o chaveiro do usuário especificado, que poderá gravar em /etc/ceph/ceph.client.ID.keyring.



#### Nota

Ao usar o comando **rbd**, você pode especificar o ID de usuário inserindo o argumento **--id ID** opcional.

Para obter mais detalhes sobre o gerenciamento de contas dos usuários do Ceph, consulte o [Capítulo 30, Autenticação com cephx](#).

## 20.2.2 Autenticação de usuário

Para especificar um nome de usuário, utilize `--id user-name`. Se você usa a autenticação do `cephx`, também precisa especificar um segredo. Ele pode vir de um chaveiro ou de um arquivo que contém o segredo:

```
cephuser@adm > rbd device map --pool rbd myimage --id admin --keyring /path/to/keyring
```

ou

```
cephuser@adm > rbd device map --pool rbd myimage --id admin --keyfile /path/to/file
```

## 20.2.3 Preparando um Dispositivo de Blocos RADOS para uso

1. Verifique se o cluster do Ceph inclui um pool com a imagem do disco que você deseja mapear. Considere o pool chamado `mypool` e a imagem `myimage`.

```
cephuser@adm > rbd list mypool
```

2. Mapeie a imagem para um novo dispositivo de blocos:

```
cephuser@adm > rbd device map --pool mypool myimage
```

3. Liste todos os dispositivos mapeados:

```
cephuser@adm > rbd device list
id pool  image  snap device
0  mypool myimage -   /dev/rbd0
```

O dispositivo no qual desejamos trabalhar é `/dev/rbd0`.



### Dica: Caminho do dispositivo RBD

Em vez do `/dev/rbdDEVICE_NUMBER`, você pode usar `/dev/rbd/POOL_NAME/IMAGE_NAME` como o caminho de um dispositivo persistente. Por exemplo:

```
/dev/rbd/mypool/myimage
```

4. Crie um sistema de arquivos XFS no dispositivo `/dev/rbd0`:

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

```
log stripe unit (4194304 bytes) is too large (maximum is 256KiB)
log stripe unit adjusted to 32KiB
meta-data=/dev/rbd0          isize=256    agcount=9, agsize=261120 blks
=                               sectsz=512   attr=2, projid32bit=1
=                               crc=0        finobt=0
data      =                               bsize=4096   blocks=2097152, imaxpct=25
=                               sunit=1024   swidth=1024 blks
naming    =version 2               bsize=4096   ascii-ci=0 ftype=0
log       =internal log           bsize=4096   blocks=2560, version=2
=                               sectsz=512   sunit=8 blks, lazy-count=1
realtime  =none                   extsz=4096   blocks=0, rtextents=0
```

5. Substitua `/mnt` por seu ponto de montagem, monte o dispositivo e verifique se ele foi montado corretamente:

```
# mount /dev/rbd0 /mnt
# mount | grep rbd0
/dev/rbd0 on /mnt type xfs (rw,relatime,attr2,inode64,sunit=8192,...
```

Agora, você pode mover os dados de e para o dispositivo como se ele fosse um diretório local.



### Dica: Aumentando o tamanho do dispositivo RBD

Se você acha que o tamanho do dispositivo RBD não é mais suficiente, pode aumentá-lo com facilidade.

1. Aumente o tamanho da imagem RBD. Por exemplo, até 10 GB.

```
cephuser@adm > rbd resize --size 10000 mypool/myimage
Resizing image: 100% complete...done.
```

2. Expanda o sistema de arquivos para preencher o novo tamanho do dispositivo:

```
# xfs_growfs /mnt
[...]
data blocks changed from 2097152 to 2560000
```

6. Após terminar de acessar o dispositivo, você poderá anular o mapeamento e desmontá-lo.

```
cephuser@adm > rbd device unmap /dev/rbd0
# umount /mnt
```



## Dica: Montagem e desmontagem manuais

Um script **rbdmap** e a unidade **systemd** são fornecidos para facilitar o processo de mapeamento e montagem de RBDs após a inicialização e de desmontagem deles antes do encerramento. Consulte a [Seção 20.2.4, “rbdmap Mapear dispositivos RBD no momento da inicialização”](#).

### 20.2.4 **rbdmap** Mapear dispositivos RBD no momento da inicialização

**rbdmap** é um script de shell que automatiza as operações **rbd map** e **rbd device unmap** em uma ou mais imagens RBD. Embora você possa executar o script manualmente a qualquer momento, a principal vantagem é mapear e montar automaticamente as imagens RBD no momento da inicialização (e desmontar e anular o mapeamento no encerramento), conforme acionado pelo sistema Init. Um arquivo da unidade **systemd**, **rbdmap.service**, está incluído no pacote **ceph-common** para essa finalidade.

O script aplica um único argumento, que pode ser **map** ou **unmap**. Em qualquer um dos casos, o script analisa um arquivo de configuração. Ele assume como padrão **/etc/ceph/rbdmap**, mas pode ser anulado por meio de uma variável de ambiente **RBDMAPIFILE**. Cada linha do arquivo de configuração corresponde a uma imagem RBD que será mapeada ou que terá o mapeamento anulado.

O arquivo de configuração tem o seguinte formato:

```
image_specification rbd_options
```

#### image\_specification

Caminho para uma imagem em um pool. Especifique como nome\_do\_pool/nome\_da\_imagem.

#### rbd\_options

Uma lista opcional de parâmetros a serem passados para o comando **rbd device map** de base. Esses parâmetros e seus valores devem ser especificados como uma string separada por vírgula. Por exemplo:

```
PARAM1=VAL1,PARAM2=VAL2,...
```

O exemplo faz com que o script **rbdmmap** execute o seguinte comando:

```
cephuser@adm > rbd device map POOL_NAME/IMAGE_NAME --PARAM1 VAL1 --PARAM2 VAL2
```

No exemplo a seguir, você pode ver como especificar um nome de usuário e um chaveiro com um segredo correspondente:

```
cephuser@adm > rbdmap device map mypool/myimage id=rbd_user,keyring=/etc/ceph/ceph.client.rbd.keyring
```

Quando executado como **rbdmmap map**, o script analisa o arquivo de configuração e, para cada imagem RBD especificada, ele tenta primeiro mapear a imagem (usando o comando **rbd device map**) e, na sequência, montar a imagem.

Quando executado como **rbdmmap unmap**, as imagens listadas no arquivo de configuração serão desmontadas e o mapeamento delas será anulado.

**rbdmmap unmap-all** tenta desmontar e, na sequência, anular o mapeamento de todas as imagens RBD mapeadas, independentemente de estarem listadas no arquivo de configuração.

Se bem-sucedida, a operação **rbd device map** mapeia a imagem para um dispositivo `/dev/rbdX` e, nesse ponto, uma regra udev é acionada para criar um link simbólico do nome do dispositivo amigável `/dev/rbd/nome_do_pool/nome_da_imagem` apontando para o dispositivo real mapeado.

Para que a montagem e a desmontagem sejam bem-sucedidas, o nome “amigável” do dispositivo precisa ter uma entrada correspondente em `/etc/fstab`. Ao gravar entradas `/etc/fstab` em imagens RBD, especifique a opção de montagem “noauto” (ou “nofail”). Isso impede que o sistema Init tente montar o dispositivo com muita antecedência, antes mesmo de ele existir, pois `rbdmmap.service` é normalmente acionado mais adiante na sequência de boot.

Para obter uma lista de opções **rbd**, consulte a página de manual do **rbd** ([man 8 rbd](#)).

Para ver exemplos de uso do **rbdmmap**, consulte a página de manual do **rbdmmap** ([man 8 rbdmap](#)).

## 20.2.5 Aumentando o tamanho dos dispositivos RBD

Se você acha que o tamanho do dispositivo RBD não é mais suficiente, pode aumentá-lo com facilidade.

1. Aumente o tamanho da imagem RBD. Por exemplo, até 10 GB.

```
cephuser@adm > rbd resize --size 10000 mypool/myimage
```

```
Resizing image: 100% complete...done.
```

2. Expanda o sistema de arquivos para preencher o novo tamanho do dispositivo.

```
# xfs_growfs /mnt
[...]  
data blocks changed from 2097152 to 2560000
```

## 20.3 Instantâneos

Um instantâneo RBD é aquele de uma imagem do Dispositivo de Blocos RADOS. Com os instantâneos, você pode manter o histórico de estado da imagem. O Ceph também suporta camadas de instantâneo, o que permite clonar imagens de VM de forma rápida e fácil. O Ceph suporta instantâneos de dispositivo de blocos usando o comando **rbd** e muitas interfaces de nível mais alto, incluindo QEMU, libvirt, OpenStack e CloudStack.



### Nota

Pare as operações de entrada e saída e descarregue todas as gravações pendentes antes de criar o instantâneo de uma imagem. Se a imagem tiver um sistema de arquivos, o estado dele deverá ser consistente no momento da criação do instantâneo.

### 20.3.1 Habilitando e configurando o cephx

Quando o cephx está habilitado, você deve especificar um nome de usuário ou ID e um caminho para o chaveiro que contém a chave correspondente para o usuário. Consulte o [Capítulo 30, Autenticação com cephx](#) para obter mais detalhes. É possível também adicionar a variável de ambiente CEPH\_ARGS para evitar uma nova entrada dos parâmetros a seguir.

```
cephuser@adm > rbd --id user-ID --keyring=/path/to/secret commands  
cephuser@adm > rbd --name username --keyring=/path/to/secret commands
```

Por exemplo:

```
cephuser@adm > rbd --id admin --keyring=/etc/ceph/ceph.keyring commands  
cephuser@adm > rbd --name client.admin --keyring=/etc/ceph/ceph.keyring commands
```



## Dica

Adicione o usuário e o segredo à variável de ambiente `CEPH_ARGS` para que você não precise digitá-los toda vez.

## 20.3.2 Aspectos básicos do instantâneo

Os procedimentos a seguir demonstram como criar, listar e remover instantâneos usando o comando `rbd` na linha de comando.

### 20.3.2.1 Criando instantâneos

Para criar um instantâneo com `rbd`, especifique a opção `snap create`, o nome do pool e o nome da imagem.

```
cephuser@adm > rbd --pool pool-name snap create --snap snap-name image-name
cephuser@adm > rbd snap create pool-name/image-name@snap-name
```

Por exemplo:

```
cephuser@adm > rbd --pool rbd snap create --snap snapshot1 image1
cephuser@adm > rbd snap create rbd/image1@snapshot1
```

### 20.3.2.2 Listando instantâneos

Para listar os instantâneos de uma imagem, especifique o nome do pool e o nome da imagem.

```
cephuser@adm > rbd --pool pool-name snap ls image-name
cephuser@adm > rbd snap ls pool-name/image-name
```

Por exemplo:

```
cephuser@adm > rbd --pool rbd snap ls image1
cephuser@adm > rbd snap ls rbd/image1
```

### 20.3.2.3 Revertendo instantâneos

Para voltar a um instantâneo com `rbd`, especifique a opção `snap rollback`, o nome do pool, o nome da imagem e o nome do instantâneo.

```
cephuser@adm > rbd --pool pool-name snap rollback --snap snap-name image-name
cephuser@adm > rbd snap rollback pool-name/image-name@snap-name
```

Por exemplo:

```
cephuser@adm > rbd --pool pool1 snap rollback --snap snapshot1 image1
cephuser@adm > rbd snap rollback pool1/image1@snapshot1
```



## Nota

Voltar uma imagem para um instantâneo significa sobregravar a versão atual da imagem com os dados de um instantâneo. O tempo necessário para executar um rollback aumenta de acordo com o tamanho da imagem. É *mais rápido clonar* de um instantâneo *do que voltar* uma imagem para um instantâneo, e é o método preferencial para reverter a um estado preexistente.

### 20.3.2.4 Apagando um instantâneo

Para apagar um instantâneo com **rbd**, especifique a opção `snap rm`, o nome do pool, o nome da imagem e o nome de usuário.

```
cephuser@adm > rbd --pool pool-name snap rm --snap snap-name image-name
cephuser@adm > rbd snap rm pool-name/image-name@snap-name
```

Por exemplo:

```
cephuser@adm > rbd --pool pool1 snap rm --snap snapshot1 image1
cephuser@adm > rbd snap rm pool1/image1@snapshot1
```



## Nota

Os Ceph OSDs apagam dados de forma assíncrona, portanto, apagar um instantâneo não libera o espaço em disco imediatamente.

### 20.3.2.5 Purgando instantâneos

Para apagar todos os instantâneos de uma imagem com **rbd**, especifique a opção `snap purge` e o nome da imagem.



```
cephuser@adm > rbd --pool pool-name snap purge image-name
cephuser@adm > rbd snap purge pool-name/image-name
```

Por exemplo:

```
cephuser@adm > rbd --pool pool1 snap purge image1
cephuser@adm > rbd snap purge pool1/image1
```

### 20.3.3 Camadas de instantâneo

O Ceph permite criar vários clones de cópia em gravação (COW, Copy-On-Write) de um instantâneo de dispositivo de blocos. As camadas de instantâneo permitem que os clientes de dispositivo de blocos do Ceph criem imagens muito rapidamente. Por exemplo, você pode criar uma imagem de dispositivo de blocos com uma VM Linux gravada nela e, em seguida, capturar um instantâneo da imagem, proteger o instantâneo e criar quantos clones de cópia em gravação desejar. Um instantâneo é apenas leitura, portanto, sua clonagem simplifica a semântica, possibilitando criar clones rapidamente.



#### Nota

Os termos “pai” e “filho” mencionados nos exemplos de linha de comando a seguir indicam um instantâneo de dispositivo de blocos do Ceph (pai) e a imagem correspondente clonada do instantâneo (filho).

Cada imagem clonada (filho) armazena uma referência à imagem pai, o que permite que a imagem clonada abra o instantâneo pai e o leia.

Um clone COW de um instantâneo funciona exatamente como qualquer outra imagem de dispositivo de blocos do Ceph. Você pode ler, gravar, clonar e redimensionar imagens clonadas. Não há nenhuma restrição especial em relação às imagens clonadas. No entanto, o clone de cópia em gravação de um instantâneo refere-se ao instantâneo. Sendo assim, você *deve* proteger o instantâneo antes de cloná-lo.



#### Nota: `--image-format 1` não suportado

Você não pode criar instantâneos de imagens criadas com a opção `rbd create --image-format 1` descontinuada. O Ceph suporta apenas a clonagem de imagens no *formato 2* padrão.

### 20.3.3.1 Introdução às camadas

As camadas de dispositivo de blocos do Ceph são um processo simples. Você deve ter uma imagem. Você deve criar um instantâneo da imagem. Você deve proteger o instantâneo. Após executar essas etapas, você poderá iniciar a clonagem do instantâneo.

A imagem clonada tem uma referência ao instantâneo pai e inclui os IDs do pool, da imagem e do instantâneo. A inclusão do ID do pool significa que você pode clonar instantâneos de um pool para imagens em outro pool.

- *Gabarito de Imagem*: Um caso de uso comum para camadas de dispositivo de blocos é criar uma imagem master e um instantâneo que serve como gabarito para os clones. Por exemplo, um usuário pode criar uma imagem para uma distribuição Linux (por exemplo, SUSE Linux Enterprise Server) e criar um instantâneo para ela. Periodicamente, o usuário pode atualizar a imagem e criar um novo instantâneo (por exemplo, `zypper ref && zypper patch` seguido por `rbd snap create`). Durante a maturação da imagem, o usuário pode clonar qualquer um dos instantâneos.
- *Gabarito Estendido*: Um caso de uso mais avançado inclui estender a imagem de um gabarito que fornece mais informações do que uma imagem de base. Por exemplo, um usuário pode clonar uma imagem (um gabarito de VM) e instalar outro software (por exemplo, um banco de dados, um sistema de gerenciamento de conteúdo ou um sistema de análise) e, em seguida, capturar um instantâneo da imagem estendida que, por si só, pode ser atualizada da mesma forma que a imagem de base.
- *Pool de Gabarito*: Uma maneira de usar as camadas de dispositivo de blocos é criar um pool que contém imagens master que atuam como gabaritos e instantâneos desses gabaritos. Em seguida, você pode estender os privilégios apenas leitura aos usuários para que eles possam clonar os instantâneos sem a capacidade de gravação ou execução no pool.
- *Migração/Recuperação de Imagens*: Uma maneira de usar as camadas de dispositivo de blocos é migrar ou recuperar os dados de um pool para outro.

### 20.3.3.2 Protegendo um instantâneo

Os clones acessam os instantâneos pai. Todos os clones serão destruídos se um usuário apagar o instantâneo pai por engano. Para evitar a perda de dados, você precisa proteger o instantâneo antes de cloná-lo.

```
cephuser@adm > rbd --pool pool-name snap protect \
```

```
--image image-name --snap snapshot-name  
cephuser@adm > rbd snap protect pool-name/image-name@snapshot-name
```

Por exemplo:

```
cephuser@adm > rbd --pool pool1 snap protect --image image1 --snap snapshot1  
cephuser@adm > rbd snap protect pool1/image1@snapshot1
```



## Nota

Você não pode apagar um instantâneo protegido.

### 20.3.3.3 Clonando um instantâneo

Para clonar um instantâneo, você precisa especificar o pool pai, a imagem, o instantâneo, o pool filho e o nome da imagem. Você precisa proteger o instantâneo antes de cloná-lo.

```
cephuser@adm > rbd clone --pool pool-name --image parent-image \  
--snap snap-name --dest-pool pool-name \  
--dest child-image  
cephuser@adm > rbd clone pool-name/parent-image@snap-name \  
pool-name/child-image-name
```

Por exemplo:

```
cephuser@adm > rbd clone pool1/image1@snapshot1 pool1/image2
```



## Nota

Você pode clonar um instantâneo de um pool para uma imagem em outro pool. Por exemplo, você pode manter as imagens e os instantâneos apenas leitura como gabaritos em um pool e os clones graváveis em outro pool.

### 20.3.3.4 Anulando a proteção de um instantâneo

Antes de apagar um instantâneo, você deve anular a proteção dele. Além disso, você *não* pode apagar instantâneos com referências de clones. Você precisa nivelar cada clone de um instantâneo antes de apagar o instantâneo.

```
cephuser@adm > rbd --pool pool-name snap unprotect --image image-name \  

```

```
--snap snapshot-name  
cephuser@adm > rbd snap unprotect pool-name/image-name@snapshot-name
```

Por exemplo:

```
cephuser@adm > rbd --pool pool1 snap unprotect --image image1 --snap snapshot1  
cephuser@adm > rbd snap unprotect pool1/image1@snapshot1
```

### 20.3.3.5 Listando os filhos de um instantâneo

Para listar os filhos de um instantâneo, execute o seguinte:

```
cephuser@adm > rbd --pool pool-name children --image image-name --snap snap-name  
cephuser@adm > rbd children pool-name/image-name@snapshot-name
```

Por exemplo:

```
cephuser@adm > rbd --pool pool1 children --image image1 --snap snapshot1  
cephuser@adm > rbd children pool1/image1@snapshot1
```

### 20.3.3.6 Nivelando uma imagem clonada

As imagens clonadas mantêm uma referência ao instantâneo pai. Ao remover a referência do clone filho para o instantâneo pai, você “nivela” com eficiência a imagem copiando as informações do instantâneo para o clone. O tempo necessário para nivelar um clone aumenta de acordo com o tamanho do instantâneo. Para apagar um instantâneo, você deve primeiro nivelar as imagens filho.

```
cephuser@adm > rbd --pool pool-name flatten --image image-name  
cephuser@adm > rbd flatten pool-name/image-name
```

Por exemplo:

```
cephuser@adm > rbd --pool pool1 flatten --image image1  
cephuser@adm > rbd flatten pool1/image1
```



#### Nota

Como uma imagem nivelada contém todas as informações do instantâneo, ela ocupa mais espaço de armazenamento do que um clone em camadas.

## 20.4 Espelhos de imagens RBD

É possível espelhar as imagens RBD de forma assíncrona entre dois clusters do Ceph. Esse recurso está disponível em dois modos:

### Com base em diário

Esse modo usa o recurso de registro de imagens RBD em diário para garantir a replicação consistente com o ponto no tempo e a falha entre os clusters. Cada gravação na imagem RBD é registrada primeiro no diário associado antes de modificar a imagem real. O cluster remoto fará a leitura do diário e reproduzirá as atualizações em sua cópia local da imagem. Como cada gravação na imagem RBD resultará em duas gravações no cluster do Ceph, é esperado que as latências de gravação quase dobrem ao usar o recurso de registro de imagens RBD em diário.

### Com base em instantâneo

Esse modo usa instantâneos de espelho de imagens RBD programados periodicamente ou criados manualmente para replicar imagens RBD consistentes com a falha entre os clusters. O cluster remoto determinará quaisquer atualizações de dados ou metadados entre dois instantâneos de espelho e copiará os deltas para sua cópia local da imagem. Com a ajuda do recurso de imagem RBD fast-diff, os blocos de dados atualizados podem ser rapidamente calculados sem a necessidade de explorar a imagem RBD completa. Como esse modo não é consistente com o ponto no tempo, o delta de instantâneo completo precisará ser sincronizado antes do uso durante um cenário de failover. Quaisquer deltas de instantâneo parcialmente aplicados voltarão ao último instantâneo totalmente sincronizado antes do uso.

O espelhamento é configurado por pool nos clusters de peer. Ele pode ser configurado em um subconjunto específico de imagens no pool ou configurado para espelhar automaticamente todas as imagens em um pool ao usar apenas o espelhamento com base em diário. O espelhamento é configurado usando o comando **rbd**. O daemon rbd-mirror é responsável por capturar as atualizações da imagem do cluster de peer remoto e aplicá-las à imagem no cluster local.

Dependendo das necessidades de replicação desejadas, o espelhamento de RBD pode ser configurado para replicação unidirecional ou bidirecional:

### Replicação unidirecional

Quando os dados são espelhados apenas de um cluster principal para um cluster secundário, o daemon rbd-mirror é executado somente no cluster secundário.

## Replicação bidirecional

Quando os dados são espelhados de imagens principais em um cluster para imagens não principais em outro cluster (e vice-versa), o daemon `rbd-mirror` é executado nos dois clusters.

### Importante

Cada instância do daemon `rbd-mirror` precisa ser capaz de se conectar aos clusters do Ceph `local` e `remoto` simultaneamente. Por exemplo, todos os hosts de monitor e OSD. Além disso, a rede precisa ter largura de banda suficiente entre os dois data centers para processar a carga de trabalho de espelhamento.

## 20.4.1 Configuração do pool

Os procedimentos a seguir demonstram como executar as tarefas administrativas básicas para configurar o espelhamento usando o comando `rbd`. O espelhamento é configurado por pool nos clusters do Ceph.

Você precisa executar as etapas de configuração do pool em ambos os clusters peer. Estes procedimentos consideram a existência de dois clusters, chamados de `local` e `remoto`, acessíveis de um único host, por motivos de clareza.

Consulte a página de manual do `rbd` ([man 8 rbd](#)) para obter mais detalhes sobre como se conectar a diferentes clusters do Ceph.

### Dica: Vários clusters

O nome do cluster nos exemplos a seguir corresponde a um arquivo de configuração do Ceph com o mesmo nome `/etc/ceph/remote.conf` e ao arquivo de chaveiro do Ceph com o mesmo nome `/etc/ceph/remote.client.admin.keyring`.

### 20.4.1.1 Permitir o espelhamento em um pool

Para habilitar o espelhamento em um pool, especifique o subcomando `mirror pool enable`, o nome do pool e o modo de espelhamento. O modo de espelhamento pode ser `pool` ou `image`:

#### `pool`

Todas as imagens no pool com o recurso de registro em diário habilitado são espelhadas.

## image

O espelhamento precisa ser habilitado explicitamente em cada imagem. Consulte a [Seção 20.4.2.1, “Habilitando o espelhamento de imagens”](#) para obter mais informações.

Por exemplo:

```
cephuser@adm > rbd --cluster local mirror pool enable POOL_NAME pool
cephuser@adm > rbd --cluster remote mirror pool enable POOL_NAME pool
```

### 20.4.1.2 Desabilitar o espelhamento

Para desabilitar o espelhamento em um pool, especifique o subcomando **mirror pool disable** e o nome do pool. Quando o espelhamento é desabilitado dessa maneira em um pool, ele também é desabilitado em todas as imagens (no pool) para as quais ele foi explicitamente habilitado.

```
cephuser@adm > rbd --cluster local mirror pool disable POOL_NAME
cephuser@adm > rbd --cluster remote mirror pool disable POOL_NAME
```

### 20.4.1.3 Inicializando peers

Para que o daemon **rbd-mirror** descubra seu cluster de peer, o peer precisa ser registrado no pool, e uma conta do usuário precisa ser criada. Esse processo pode ser automatizado com o **rbd** e os comandos **mirror pool peer bootstrap create** e **mirror pool peer bootstrap import**. Para criar manualmente um novo token de boot com o **rbd**, especifique o comando **mirror pool peer bootstrap create**, o nome de um pool, junto com o nome amigável de um site, para descrever o cluster **local**:

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

A saída do **mirror pool peer bootstrap create** será um token que deve ser inserido no comando **mirror pool peer bootstrap import**. Por exemplo, no cluster **local**:

```
cephuser@local > rbd --cluster local mirror pool peer bootstrap create --site-name local
image-pool
eyJmc2lkIjoiaWY1MjgyZGI0Yjg5ODU0NTk2LTgwOTgtMzIwYzFmYzY5MjYyY2xpZW50X2lkIjoicmJkLWlpcnJvcilwZWVYIiw
\
joiQVFBUnczOWQwdkhvQmhBQVlMM1I4RmR5dHNJQU50bkFTZ0l0TVE9PSIsIm1vbl9ob3N0IjoiaW3Yy0jE5Mi4xNjguMS4z0jY4MjAs
```

Para importar manualmente o token de boot criado por outro cluster com o comando `rbd`, use a seguinte sintaxe:

```
rbd mirror pool peer bootstrap import \  
  [--site-name LOCAL_SITE_NAME] \  
  [--direction DIRECTION \  
  POOL_NAME TOKEN_PATH
```

Onde:

LOCAL\_SITE\_NAME

O nome amigável opcional de um site para descrever o cluster local.

DIRECTION

Uma direção de espelhamento. Assume `rx-tx` como padrão para espelhamento bidirecional, mas também pode ser definido como `rx-only` para espelhamento unidirecional.

POOL\_NAME

O nome do pool.

TOKEN\_PATH

O caminho de arquivo para o token criado (ou `-` para lê-lo da entrada padrão).

Por exemplo, no cluster remoto:

```
cephuser@remote > cat <<EOF > token  
eyJmc2lkIjo1OWY1MjgyZGI0NTk2LTgwOTgtMzIwYzFmYzYzM5NmYzIiwiaWY2xpZW50X2lkIjoicmJkLWlpcnJvcilwZWVyIiw  
EOF
```

```
cephuser@adm > rbd --cluster remote mirror pool peer bootstrap import \  
  --site-name remote image-pool token
```

#### 20.4.1.4 Adicionando um peer do cluster manualmente

Como alternativa à inicialização de peers, conforme descrito na [Seção 20.4.1.3, “Inicializando peers”](#), você pode especificar os peers manualmente. O daemon remoto `rbd-mirror` precisará de acesso ao cluster local para executar o espelhamento. Crie um novo usuário do Ceph local que o daemon `rbd-mirror` remoto usará, por exemplo `rbd-mirror-peer`:

```
cephuser@adm > ceph auth get-or-create client.rbd-mirror-peer \  

```



```
mon 'profile rbd' osd 'profile rbd'
```

Use a seguinte sintaxe para adicionar um cluster de peer de espelhamento do Ceph com o comando **rbd**:

```
rbd mirror pool peer add POOL_NAME CLIENT_NAME@CLUSTER_NAME
```

Por exemplo:

```
cephuser@adm > rbd --cluster site-a mirror pool peer add image-pool client.rbd-mirror-  
peer@site-b  
cephuser@adm > rbd --cluster site-b mirror pool peer add image-pool client.rbd-mirror-  
peer@site-a
```

Por padrão, o daemon **rbd-mirror** precisa ter acesso ao arquivo de configuração do Ceph localizado em `/etc/ceph/.CLUSTER_NAME.conf`. Ele inclui os endereços IP dos MONs do cluster de peer e um chaveiro para um cliente denominado *CLIENT\_NAME*, localizado nos caminhos padrão ou personalizado de pesquisa de chaveiro, por exemplo `/etc/ceph/CLUSTER_NAME.CLIENT_NAME.keyring`.

Como alternativa, é possível gravar o MON do cluster de peer e/ou a chave do cliente com segurança no armazenamento de chave de configuração local do Ceph. Para especificar os atributos de conexão do cluster de peer ao adicionar um peer de espelhamento, use as opções `--remote-mon-host` e `--remote-key-file`. Por exemplo:

```
cephuser@adm > rbd --cluster site-a mirror pool peer add image-pool \  
client.rbd-mirror-peer@site-b --remote-mon-host 192.168.1.1,192.168.1.2 \  
--remote-key-file /PATH/TO/KEY_FILE  
cephuser@adm > rbd --cluster site-a mirror pool info image-pool --all  
Mode: pool  
Peers:  
  UUID          NAME    CLIENT                      MON_HOST                      KEY  
  587b08db... site-b client.rbd-mirror-peer 192.168.1.1,192.168.1.2 AQAeuZdb...
```

### 20.4.1.5 Remover o peer de cluster

Para remover um cluster peer de espelhamento, especifique o subcomando **mirror pool peer remove**, o nome do pool e o UUID do peer (disponível no comando **rbd mirror pool info**):

```
cephuser@adm > rbd --cluster local mirror pool peer remove POOL_NAME \  
55672766-c02b-4729-8567-f13a66893445  
cephuser@adm > rbd --cluster remote mirror pool peer remove POOL_NAME \  
55672766-c02b-4729-8567-f13a66893445
```

#### 20.4.1.6 Pools de dados

Ao criar imagens no cluster de destino, o `rbd-mirror` seleciona um pool de dados da seguinte maneira:

- Se o cluster de destino tiver um pool de dados padrão configurado (com a opção de configuração `rbd_default_data_pool`), ele será usado.
- Do contrário, se a imagem de origem usar um pool de dados separado, e existir um pool com o mesmo nome no cluster de destino, esse pool será usado.
- Se nenhuma das opções acima for verdadeira, nenhum pool de dados será definido.

### 20.4.2 Configuração de imagens RBD

Diferentemente da configuração do pool, a configuração da imagem precisa apenas ser executada em um único cluster peer de espelhamento do Ceph.

As imagens RBD espelhadas são designadas como *principais* ou *não principais*. Essa é uma propriedade da imagem, e não do pool. Não é possível modificar as imagens designadas como não principais.

As imagens são automaticamente promovidas a principais quando o espelhamento é habilitado primeiro em uma imagem (seja implicitamente, se o modo de espelhamento do pool for “pool” e a imagem tiver o recurso de registro de imagens em diário habilitado, ou explicitamente (consulte a [Seção 20.4.2.1, “Habilitando o espelhamento de imagens”](#)) pelo comando `rbd`).

#### 20.4.2.1 Habilitando o espelhamento de imagens

Se o espelhamento for configurado no modo `image`, será necessário habilitar explicitamente o espelhamento para cada imagem no pool. Para habilitar o espelhamento de uma imagem específica com o `rbd`, especifique o subcomando `mirror image enable` junto com o nome do pool e da imagem:

```
cephuser@adm > rbd --cluster local mirror image enable \  
POOL_NAME/IMAGE_NAME
```

O modo de espelhamento da imagem pode ser journal ou snapshot:

#### journal (padrão)

Quando configurado no modo journal, o espelhamento usará o recurso de registro de imagens RBD em diário para replicar o conteúdo da imagem. Se o recurso de registro de imagens RBD em diário ainda não estiver habilitado na imagem, ele será habilitado automaticamente.

#### snapshot

Quando configurado no modo snapshot, o espelhamento usará o instantâneo de espelho de imagens RBD para replicar o conteúdo da imagem. Uma vez habilitado, um instantâneo de espelho inicial será criado automaticamente. É possível criar mais instantâneos de espelho de imagens RBD com o comando rbd.

Por exemplo:

```
cephuser@adm > rbd --cluster local mirror image enable image-pool/image-1 snapshot
cephuser@adm > rbd --cluster local mirror image enable image-pool/image-2 journal
```

### 20.4.2.2 Habilitando o recurso de registro de imagens em diário

O espelhamento de RBD usa o recurso de registro de RBD em diário para garantir que a imagem replicada permaneça sempre consistente com a falha. Ao usar o modo de espelhamento de imagem, o recurso de registro em diário será habilitado automaticamente se o espelhamento estiver habilitado na imagem. Ao usar o modo de espelhamento de pool, antes que uma imagem possa ser espelhada para um cluster de peer, o recurso de registro de imagens RBD em diário deve ser habilitado. O recurso pode ser habilitado no momento da criação da imagem inserindo a opção --image-feature exclusive-lock, journaling no comando rbd.

Se preferir, o recurso de registro em diário pode ser habilitado dinamicamente nas imagens RBD preexistentes. Para habilitar o registro em diário, especifique o subcomando feature enable, o nome do pool e da imagem e o nome do recurso:

```
cephuser@adm > rbd --cluster local feature enable POOL_NAME/IMAGE_NAME exclusive-lock
cephuser@adm > rbd --cluster local feature enable POOL_NAME/IMAGE_NAME journaling
```



## Nota: Dependência da opção

O recurso de registro em diário depende do recurso de bloqueio exclusivo. Se o recurso de bloqueio exclusivo ainda não estiver habilitado, você precisará habilitá-lo antes de habilitar o recurso de registro em diário.



## Dica

É possível habilitar o registro em diário em todas as imagens novas por padrão, adicionando rbd default features = layering,exclusive-lock,object-map,deep-flatten,journaling ao arquivo de configuração do Ceph.

### 20.4.2.3 Criando instantâneos de espelho de imagem

Ao usar o espelhamento com base em instantâneo, os instantâneos de espelho precisarão ser criados sempre que você desejar espelhar o conteúdo modificado da imagem RBD. Para criar um instantâneo de espelho manualmente com o **rbd**, especifique o comando **mirror image snapshot**, junto com o nome do pool e da imagem:

```
cephuser@adm > rbd mirror image snapshot POOL_NAME/IMAGE_NAME
```

Por exemplo:

```
cephuser@adm > rbd --cluster local mirror image snapshot image-pool/image-1
```

Por padrão, apenas três instantâneos de espelho serão criados por imagem. O instantâneo de espelho mais recente será removido automaticamente se o limite for atingido. O limite poderá ser anulado por meio da opção de configuração rbd\_mirroring\_max\_mirroring\_snapshots, se necessário. Além disso, os instantâneos de espelho são automaticamente apagados quando a imagem é removida ou quando o espelhamento é desabilitado.

Os instantâneos de espelho também poderão ser criados automaticamente de maneira periódica, se as programações de instantâneos de espelho forem definidas. O instantâneo de espelho pode ser programado nos níveis global, por pool ou por imagem. Várias programações de instantâneos de espelho podem ser definidas em qualquer nível, mas apenas as programações de instantâneo mais específicas correspondentes a uma imagem espelhada individual serão executadas.

Para criar uma programação de instantâneo de espelho com o **rbd**, especifique o comando **mirror snapshot schedule add**, junto com o nome de um pool ou imagem opcional, o intervalo e o horário de início opcional.

O intervalo pode ser especificado em dias, horas ou minutos usando os sufixos `d`, `h` ou `m`, respectivamente. O horário de início opcional pode ser especificado usando o formato de horário ISO 8601. Por exemplo:

```
cephuser@adm > rbd --cluster local mirror snapshot schedule add --pool image-pool 24h
14:00:00-05:00
cephuser@adm > rbd --cluster local mirror snapshot schedule add --pool image-pool --image
image1 6h
```

Para remover uma programação de instantâneo de espelho com o `rbd`, especifique o comando **`mirror snapshot schedule remove`** com as opções correspondentes ao comando de adição da programação.

Para listar todas as programações de instantâneo para um nível específico (global, pool ou imagem) com o `rbd`, especifique o comando **`mirror snapshot schedule ls`**, junto com o nome de um pool ou imagem opcional. Além disso, a opção `--recursive` pode ser usada para listar todas as programações no nível especificado e abaixo dele. Por exemplo:

```
cephuser@adm > rbd --cluster local mirror schedule ls --pool image-pool --recursive
POOL      NAMESPACE IMAGE  SCHEDULE
image-pool -          -      every 1d starting at 14:00:00-05:00
image-pool          image1 every 6h
```

Para saber quando os próximos instantâneos serão criados para imagens RBD de espelhamento com base em instantâneo com o `rbd`, especifique o comando **`mirror snapshot schedule status`**, junto com o nome de um pool ou imagem opcional. Por exemplo:

```
cephuser@adm > rbd --cluster local mirror schedule status
SCHEDULE TIME      IMAGE
2020-02-26 18:00:00 image-pool/image1
```

#### 20.4.2.4 Desabilitando o espelhamento de imagens

Para desabilitar o espelhamento em determinada imagem, especifique o subcomando **`mirror image disable`** juntamente com o nome do pool e da imagem:

```
cephuser@adm > rbd --cluster local mirror image disable POOL_NAME/IMAGE_NAME
```

### 20.4.2.5 Promovendo e retrocedendo imagens

Em um cenário de failover em que a designação principal precisa ser movida para a imagem no cluster peer, você precisa interromper o acesso à imagem principal, rebaixar a imagem principal atual, promover a nova imagem principal e continuar o acesso à imagem no cluster alternativo.



#### Nota: Promoção forçada

A promoção pode ser forçada usando a opção `--force`. A promoção forçada é necessária quando o rebaixamento não pode ser propagado para o cluster peer (por exemplo, em caso de falha do cluster ou interrupção da comunicação). Isso resultará em um cenário de split-brain (dupla personalidade) entre os dois peers, e a imagem não será mais sincronizada até que um subcomando `resync` seja emitido.

Para rebaixar determinada imagem para não principal, especifique o subcomando `mirror image demote` juntamente com o nome do pool e da imagem:

```
cephuser@adm > rbd --cluster local mirror image demote POOL_NAME/IMAGE_NAME
```

Para rebaixar todas as imagens principais em um pool para não principais, especifique o subcomando `mirror pool demote` juntamente com o nome do pool:

```
cephuser@adm > rbd --cluster local mirror pool demote POOL_NAME
```

Para promover determinada imagem para principal, especifique o subcomando `mirror image promote` juntamente com o nome do pool e da imagem:

```
cephuser@adm > rbd --cluster remote mirror image promote POOL_NAME/IMAGE_NAME
```

Para promover todas as imagens não principais em um pool para principais, especifique o subcomando `mirror pool promote` juntamente com o nome do pool:

```
cephuser@adm > rbd --cluster local mirror pool promote POOL_NAME
```



#### Dica: Dividir a carga de E/S

Como o status principal ou não principal refere-se a cada imagem, é possível ter dois clusters que dividem a carga de E/S e o failover ou fallback por fase.

### 20.4.2.6 Forçando a ressincronização da imagem

Se um evento de split-brain for detectado pelo daemon `rbd-mirror`, ele não tentará espelhar a imagem afetada até o problema ser corrigido. Para continuar o espelhamento de uma imagem, primeiro rebaixe a imagem que foi identificada como desatualizada e, em seguida, solicite uma ressincronização com a imagem principal. Para solicitar a ressincronização de uma imagem, especifique o subcomando **`mirror image resync`** juntamente com o nome do pool e da imagem:

```
cephuser@adm > rbd mirror image resync POOL_NAME/IMAGE_NAME
```

### 20.4.3 Verificando o status do espelho

O status de replicação do cluster peer é armazenado para cada imagem principal espelhada. Esse status pode ser recuperado usando os subcomandos **`mirror image status`** e **`mirror pool status`**:

Para solicitar o status da imagem de espelho, especifique o subcomando **`mirror image status`** juntamente com o nome do pool e da imagem:

```
cephuser@adm > rbd mirror image status POOL_NAME/IMAGE_NAME
```

Para solicitar o status do resumo do pool de espelhos, especifique o subcomando **`mirror pool status`** juntamente com o nome do pool:

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



#### Dica:

A adição da opção `--verbose` ao subcomando **`mirror pool status`** resultará também nos detalhes de status para cada imagem de espelhamento no pool.

## 20.5 Configurações de cache

A implementação do espaço de usuário do dispositivo de blocos do Ceph (`librbd`) não pode se beneficiar do cache de página do Linux. Portanto, ela inclui seu próprio cache na memória. O cache RBD tem comportamento semelhante ao cache de disco rígido. Quando o OS envia uma solicitação de barreira ou descarga, todos os dados "modificados" são gravados nos OSDs. Isso significa que o uso do cache de write-back é tão seguro quanto o uso de um disco rígido físico de

bom funcionamento com uma VM que envia descarregamentos apropriadamente. O cache usa um algoritmo *menos utilizado recentemente* (LRU, Least Recently Used) e, no modo write-back, ele pode fundir solicitações adjacentes para um melhor throughput.

O Ceph suporta cache de write-back para RBD. Para habilitá-lo, execute

```
cephuser@adm > ceph config set client rbd_cache true
```

Por padrão, o librbd não executa armazenamento em cache. As gravações e leituras seguem diretamente para o cluster de armazenamento, e as gravações são retornadas apenas quando os dados estão no disco em todas as réplicas. Com o cache habilitado, as gravações são retornadas imediatamente, a menos que haja mais bytes descarregados do que o que foi definido na opção rbd cache max dirty. Nesse caso, a gravação aciona write-back e blocos até que sejam descarregados bytes suficientes.

O Ceph suporta cache de write-through para RBD. Você pode definir o tamanho do cache, destinos e limites para alternar do cache de write-back para o cache de write-through. Para habilitar o modo write-through, execute

```
cephuser@adm > ceph config set client rbd_cache_max_dirty 0
```

Isso significa que as gravações são retornadas apenas quando os dados estão no disco em todas as réplicas, mas as leituras podem vir do cache. O cache está na memória do cliente, e cada imagem RBD tem seu próprio cache. Como o cache é local ao cliente, não fará sentido se houver outras pessoas acessando a imagem. A execução do GFS ou OCFS no RBD não funcionará com o cache habilitado.

Os parâmetros a seguir afetam o comportamento dos Dispositivos de Blocos RADOS. Para defini-los, use a categoria client:

```
cephuser@adm > ceph config set client PARAMETER VALUE
```

#### rbd cache

Habilitar o cache para Dispositivo de blocos RADOS (RBD, RADOS Block Device). O padrão é “true”.

#### rbd cache size

O tamanho do cache RBD em bytes. O padrão é 32 MB.

#### rbd cache max dirty

O limite de "modificação" em bytes em que o cache aciona o write-back. rbd cache max dirty precisa ser inferior a rbd cache size. Se definido como 0, usa o cache de write-through. O padrão é 24 MB.



#### rbd cache target dirty

O “destino de modificação” antes de o cache começar a gravar dados no armazenamento de dados. Não bloqueia as gravações para o cache. O padrão é 16 MB.

#### rbd cache max dirty age

Por quantos segundos os dados modificados permanecem no cache antes que o write-back seja iniciado. O padrão é 1.

#### rbd cache writethrough until flush

Comece no modo write-through e alterne para write-back depois de receber a primeira solicitação de descarregamento. A habilitação dessa configuração é conservadora, porém segura, caso as máquinas virtuais executadas no rbd sejam muito antigas para enviar descarregamentos (por exemplo, o driver virtio no Linux antes do kernel 2.6.32). O padrão é “true”.

## 20.6 Configurações de QoS

Em geral, a Qualidade do Serviço (QoS) refere-se aos métodos de priorização de tráfego e reserva de recursos. Ela é importante principalmente para o transporte de tráfego com requisitos especiais.



### Importante: Não suportadas pelo iSCSI

As seguintes configurações de QoS são usadas apenas pela implementação RBD librbd do espaço de usuário, e *não* são usadas pela implementação krbd. Como o iSCSI usa o krbd, ele não usa as configurações de QoS. No entanto, para o iSCSI, você pode configurar a QoS na camada do dispositivo de blocos do kernel usando os recursos padrão do kernel.

#### rbd qos iops limit

O limite desejado das operações de E/S por segundo. O padrão é 0 (sem limite).

#### rbd qos bps limit

O limite desejado de bytes de E/S por segundo. O padrão é 0 (sem limite).

#### rbd qos read iops limit

O limite desejado das operações de leitura por segundo. O padrão é 0 (sem limite).

#### rbd qos write iops limit

O limite desejado das operações de gravação por segundo. O padrão é 0 (sem limite).

rbd qos read bps limit

O limite desejado de bytes de leitura por segundo. O padrão é 0 (sem limite).

rbd qos write bps limit

O limite desejado de bytes de gravação por segundo. O padrão é 0 (sem limite).

rbd qos iops burst

O limite de burst desejado das operações de E/S. O padrão é 0 (sem limite).

rbd qos bps burst

O limite de burst desejado de bytes de E/S. O padrão é 0 (sem limite).

rbd qos read iops burst

O limite de burst desejado das operações de leitura. O padrão é 0 (sem limite).

rbd qos write iops burst

O limite de burst desejado das operações de gravação. O padrão é 0 (sem limite).

rbd qos read bps burst

O limite de burst desejado de bytes de leitura. O padrão é 0 (sem limite).

rbd qos write bps burst

O limite de burst desejado de bytes de gravação. O padrão é 0 (sem limite).

rbd qos schedule tick min

O tique de horário mínimo (em milissegundos) para QoS. O padrão é 50.

## 20.7 Configurações de leitura com ajuda

O Dispositivo de Blocos RADOS suporta leitura com ajuda/pré-busca para otimizar pequenas leituras sequenciais. Isso costuma ser processado pelo OS convidado no caso de uma máquina virtual, mas os carregadores de boot talvez não emitam leituras suficientes. A leitura com ajuda será automaticamente desabilitada se o cache for desabilitado.



## Importante: Não suportadas pelo iSCSI

As seguintes configurações de leitura com ajuda são usadas apenas pela implementação RBD `librbd` do espaço de usuário, e *não* são usadas pela implementação `kRBD`. Como o iSCSI usa o `kRBD`, ele não usa as configurações de leitura com ajuda. No entanto, para o iSCSI, você pode configurar a leitura com ajuda na camada do dispositivo de blocos do kernel usando os recursos padrão do kernel.

### `rbt readahead trigger requests`

Número de solicitações de leitura sequenciais necessárias para acionar a leitura com ajuda. O padrão é 10.

### `rbt readahead max bytes`

Tamanho máximo de uma solicitação de leitura com ajuda. Se definido como 0, a leitura com ajuda será desabilitada. O padrão é 512 KB.

### `rbt readahead disable after bytes`

Após a leitura dessa quantidade de bytes de uma imagem RBD, a leitura com ajuda será desabilitada para essa imagem até ser fechada. Isso permite que o OS convidado controle a leitura com ajuda quando é inicializado. Se definido como 0, a leitura com ajuda permanecerá habilitada. O padrão é 50 MB.

## 20.8 Recursos avançados

O Dispositivo de Blocos RADOS suporta recursos avançados que melhoram a funcionalidade das imagens RBD. Você pode especificar os recursos na linha de comando ao criar uma imagem RBD ou no arquivo de configuração do Ceph usando a opção `rbt_default_features`.

Você pode especificar os valores da opção `rbd_default_features` de duas maneiras:

- Como a soma dos valores internos dos recursos. Cada recurso tem seu próprio valor interno. Por exemplo, “layering” tem 1 e “fast-diff” tem 16. Portanto, para ativar esses dois recursos por padrão, inclua o seguinte:

```
rbd_default_features = 17
```

- Como uma lista de recursos separada por vírgula. O exemplo anterior terá a seguinte aparência:

```
rbd_default_features = layering,fast-diff
```



### Nota: Recursos não suportados pelo iSCSI

As imagens RBD com os seguintes recursos não serão suportadas pelo iSCSI: deep-flatten, object-map, journaling, fast-diff, striping

Veja a seguir uma lista de recursos RBD avançados:

#### layering

As camadas (layering) permitem usar a clonagem.

O valor interno é 1, o padrão é “yes”.

#### striping

A distribuição difunde os dados por vários objetos e ajuda com paralelismo para cargas de trabalho de leitura/gravação sequenciais. Ele evita gargalos de nó único para Dispositivo de Blocos RADOS grandes ou ocupados.

O valor interno é 2, o padrão é “yes”.

#### exclusive-lock

Quando habilitado, ele requer que um cliente crie um bloqueio em um objeto antes de fazer uma gravação. Habilite o bloqueio exclusivo apenas quando um único cliente está acessando uma imagem ao mesmo tempo. O valor interno é 4. O padrão é “yes”.

#### object-map

O suporte ao mapa de objetos depende do suporte ao bloqueio exclusivo. Os dispositivos de blocos são aprovisionados dinamicamente, o que significa que eles apenas armazenam dados que realmente existem. O suporte ao mapa de objetos ajuda a monitorar os objetos

que realmente existem (com dados armazenados em uma unidade). A habilitação do suporte ao mapa de objetos acelera as operações de E/S para clonagem, importação e exportação de uma imagem pouco preenchida, além da exclusão.

O valor interno é 8, o padrão é “yes”.

#### fast-diff

O suporte à comparação rápida (fast-diff) depende do suporte ao mapa de objetos e do suporte ao bloqueio exclusivo. Ele adiciona outra propriedade ao mapa de objetos, o que o torna muito mais rápido para gerar comparações entre instantâneos de uma imagem e o uso real dos dados de um instantâneo.

O valor interno é 16, o padrão é “yes”.

#### deep-flatten

O nivelamento profundo (deep-flatten) faz com que o **rbd flatten** (consulte a [Seção 20.3.3.6, “Nivelando uma imagem clonada”](#)) funcione em todos os instantâneos de uma imagem, além da própria imagem. Sem ele, os instantâneos de uma imagem ainda dependerão do pai, portanto, você não poderá apagar a imagem pai até os instantâneos serem apagados. O deep-flatten torna um pai independente de seus clones, mesmo que eles tenham instantâneos.

O valor interno é 32, o padrão é “yes”.

#### journaling

O suporte ao registro em diário depende do suporte ao bloqueio exclusivo. O diário registra todas as modificações em uma imagem na ordem em que elas ocorrem. O espelhamento RBD (consulte a [Seção 20.4, “Espelhos de imagens RBD”](#)) usa o diário para replicar uma imagem consistente de falha para um cluster remoto.

O valor interno é 64, o padrão é “no”.

## 20.9 Mapeando o RBD por meio de clientes antigos do kernel

Clientes antigos (por exemplo, SLE11 SP4) podem não conseguir mapear imagens RBD porque um cluster implantado com o SUSE Enterprise Storage 7.1 força alguns recursos (no nível tanto da imagem RBD quanto do RADOS) que esses clientes antigos não suportam. Quando isso acontece, os registros do OSD mostram mensagens semelhantes às seguintes:

```
2019-05-17 16:11:33.739133 7fcb83a2e700 0 -- 192.168.122.221:0/1006830 >> \
```

```
192.168.122.152:6789/0 pipe(0x65d4e0 sd=3 :57323 s=1 pgs=0 cs=0 l=1 c=0x65d770).connect \
protocol feature mismatch, my 2fffffffffff < peer 4010ff8ffacffff missing 4010000000000000
```



## Atenção: Mudança de tipos de compartimento de memória do Mapa CRUSH provoca rebalanceamento massivo

Se você pretende alternar os tipos de compartimento de memória do Mapa CRUSH entre “straw” e “straw2”, faça um planejamento para isso. Espere um impacto significativo na carga do cluster, porque a mudança do tipo de compartimento de memória provoca uma redistribuição massiva do cluster.

1. Desabilite quaisquer recursos da imagem RBD que não sejam suportados. Por exemplo:

```
cephuser@adm > rbd feature disable pool1/image1 object-map
cephuser@adm > rbd feature disable pool1/image1 exclusive-lock
```

2. Mude os tipos de compartimento de memória do Mapa CRUSH de “straw2” para “straw”:

- a. Grave o Mapa CRUSH:

```
cephuser@adm > ceph osd getcrushmap -o crushmap.original
```

- b. Descompile o Mapa CRUSH:

```
cephuser@adm > crushtool -d crushmap.original -o crushmap.txt
```

- c. Edite o Mapa CRUSH e substitua “straw2” por “straw”.

- d. Recompile o Mapa CRUSH:

```
cephuser@adm > crushtool -c crushmap.txt -o crushmap.new
```

- e. Defina o novo Mapa CRUSH:

```
cephuser@adm > ceph osd setcrushmap -i crushmap.new
```

## 20.10 Habilitando dispositivos de blocos e Kubernetes

Você pode usar o Ceph RBD com o Kubernetes 1.13 e versões mais recentes por meio do driver `ceph-csi`. Esse driver provisiona dinamicamente as imagens RBD para suportar os volumes do Kubernetes e mapeia essas imagens RBD como dispositivos de blocos (opcionalmente, montando um sistema de arquivos contido na imagem) em nós de trabalho que executam pods que fazem referência a um volume com suporte do RBD.

Para usar dispositivos de blocos do Ceph com o Kubernetes, você deve instalar e configurar o `ceph-csi` no seu ambiente do Kubernetes.



### Importante

O `ceph-csi` usa os módulos do kernel RBD por padrão, que podem não suportar todos os tunables do Ceph CRUSH ou os recursos de imagem RBD.

1. Por padrão, os dispositivos de blocos do Ceph usam o pool RBD. Crie um pool para armazenamento de volume do Kubernetes. Verifique se o cluster do Ceph está em execução e crie o pool:

```
cephuser@adm > ceph osd pool create kubernetes
```

2. Use a ferramenta RBD para inicializar o pool:

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

3. Crie um novo usuário para o Kubernetes e o `ceph-csi`. Execute o seguinte e registre a chave gerada:

```
cephuser@adm > ceph auth get-or-create client.kubernetes mon 'profile rbd' osd  
'profile rbd pool=kubernetes' mgr 'profile rbd pool=kubernetes'  
[client.kubernetes]  
key = AQD9o0Fd6hQRChAAt7fMaSZXduT3NWEqylNpmg==
```

4. O `ceph-csi` requer um objeto ConfigMap armazenado no Kubernetes para definir os endereços do monitor do Ceph para o cluster do Ceph. Colete o FSID exclusivo do cluster do Ceph e os endereços do monitor:

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

```
fsid b9127830-b0cc-4e34-aa47-9d1a2e9949a8
<...>
0: [v2:192.168.1.1:3300/0,v1:192.168.1.1:6789/0] mon.a
1: [v2:192.168.1.2:3300/0,v1:192.168.1.2:6789/0] mon.b
2: [v2:192.168.1.3:3300/0,v1:192.168.1.3:6789/0] mon.c
```

5. Gere um arquivo `csi-config-map.yaml` semelhante ao exemplo abaixo, substituindo o FSID por `clusterID` e os endereços do monitor por `monitors`:

```
kubectl@adm > cat <<EOF > csi-config-map.yaml
---
apiVersion: v1
kind: ConfigMap
data:
  config.json: |-
    [
      {
        "clusterID": "b9127830-b0cc-4e34-aa47-9d1a2e9949a8",
        "monitors": [
          "192.168.1.1:6789",
          "192.168.1.2:6789",
          "192.168.1.3:6789"
        ]
      }
    ]
metadata:
  name: ceph-csi-config
EOF
```

6. Quando gerado, armazene o novo objeto ConfigMap no Kubernetes:

```
kubectl@adm > kubectl apply -f csi-config-map.yaml
```

7. O `ceph-csi` requer as credenciais do `cephx` para se comunicar com o cluster do Ceph. Gere um arquivo `csi-rbd-secret.yaml` semelhante ao exemplo abaixo, usando o ID de usuário do Kubernetes recém-criado e a chave do `cephx`:

```
kubectl@adm > cat <<EOF > csi-rbd-secret.yaml
---
apiVersion: v1
kind: Secret
metadata:
  name: csi-rbd-secret
  namespace: default
stringData:
  userID: kubernetes
```



```
userKey: AQD9o0Fd6hQRChAAt7fMaSZXduT3NWEqylNpmg==
EOF
```

8. Quando gerado, armazene o novo objeto secreto no Kubernetes:

```
kubect@adm > kubectl apply -f csi-rbd-secret.yaml
```

9. Crie os objetos do Kubernetes ServiceAccount e RBAC ClusterRole/ClusterRoleBinding necessários. Esses objetos não precisam ser necessariamente personalizados para seu ambiente do Kubernetes e, portanto, podem ser usados diretamente dos arquivos YAML de implantação do `ceph-csi`:

```
kubect@adm > kubectl apply -f https://raw.githubusercontent.com/ceph/ceph-csi/master/deploy/rbd/kubernetes/csi-provisioner-rbac.yaml
kubect@adm > kubectl apply -f https://raw.githubusercontent.com/ceph/ceph-csi/master/deploy/rbd/kubernetes/csi-nodeplugin-rbac.yaml
```

10. Crie o provisionador `ceph-csi` e os plug-ins de nó:

```
kubect@adm > wget https://raw.githubusercontent.com/ceph/ceph-csi/master/deploy/rbd/kubernetes/csi-rbdplugin-provisioner.yaml
kubect@adm > kubectl apply -f csi-rbdplugin-provisioner.yaml
kubect@adm > wget https://raw.githubusercontent.com/ceph/ceph-csi/master/deploy/rbd/kubernetes/csi-rbdplugin.yaml
kubect@adm > kubectl apply -f csi-rbdplugin.yaml
```



### Importante

Por padrão, os arquivos YAML do provisionador e do plug-in de nó extrairão a versão de desenvolvimento do container `ceph-csi`. Os arquivos YAML devem ser atualizados para usar uma versão de lançamento.

## 20.10.1 Usando dispositivos de blocos do Ceph no Kubernetes

A StorageClass do Kubernetes define uma classe de armazenamento. Vários objetos StorageClass podem ser criados para mapear para diferentes níveis e recursos de qualidade de serviço. Por exemplo, pools com base em NVMe em relação aos pools com base em HDD.

Para criar uma `StorageClass` do `ceph-csi` que mapeie para o pool do Kubernetes criado acima, o seguinte arquivo YAML pode ser usado, depois de garantir que a propriedade `clusterID` corresponda ao FSID do cluster do Ceph:

```
kubectl@adm > cat <<EOF > csi-rbd-sc.yaml
---
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: csi-rbd-sc
provisioner: rbd.csi.ceph.com
parameters:
  clusterID: b9127830-b0cc-4e34-aa47-9d1a2e9949a8
  pool: kubernetes
  csi.storage.k8s.io/provisioner-secret-name: csi-rbd-secret
  csi.storage.k8s.io/provisioner-secret-namespace: default
  csi.storage.k8s.io/node-stage-secret-name: csi-rbd-secret
  csi.storage.k8s.io/node-stage-secret-namespace: default
reclaimPolicy: Delete
mountOptions:
  - discard
EOF
kubectl@adm > kubectl apply -f csi-rbd-sc.yaml
```

Um `PersistentVolumeClaim` é uma solicitação de recursos de armazenamento abstratos feita por um usuário. O `PersistentVolumeClaim` é associado a um recurso de pod para provisionar um `PersistentVolume`, que recebe suporte de uma imagem de bloco do Ceph. É possível incluir um `volumeMode` opcional para selecionar entre um sistema de arquivos montado (padrão) ou um volume bruto com base em dispositivo de blocos.

Usando o `ceph-csi`, a especificação de `Filesystem` para `volumeMode` pode suportar ambos os requerimentos `ReadWriteOnce` e `ReadOnlyMany` `accessMode`, e a especificação de `Block` para `volumeMode` pode suportar os requerimentos `ReadWriteOnce`, `ReadWriteMany` e `ReadOnlyMany` `accessMode`.

Por exemplo, para criar um `PersistentVolumeClaim` baseado em blocos que usa o `ceph-csi-based` `StorageClass` criado acima, o seguinte arquivo YAML pode ser usado para solicitar o armazenamento de blocos bruto de `csi-rbd-sc` `StorageClass`:

```
kubectl@adm > cat <<EOF > raw-block-pvc.yaml
---
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: raw-block-pvc
```

```
spec:
  accessModes:
    - ReadWriteOnce
  volumeMode: Block
  resources:
    requests:
      storage: 1Gi
  storageClassName: csi-rbd-sc
EOF
kubectl@adm > kubectl apply -f raw-block-pvc.yaml
```

Veja a seguir uma demonstração de como vincular o PersistentVolumeClaim acima a um recurso de pod como um dispositivo de blocos bruto:

```
kubectl@adm > cat <<EOF > raw-block-pod.yaml
---
apiVersion: v1
kind: Pod
metadata:
  name: pod-with-raw-block-volume
spec:
  containers:
    - name: fc-container
      image: fedora:26
      command: ["/bin/sh", "-c"]
      args: ["tail -f /dev/null"]
      volumeDevices:
        - name: data
          devicePath: /dev/xvda
  volumes:
    - name: data
      persistentVolumeClaim:
        claimName: raw-block-pvc
EOF
kubectl@adm > kubectl apply -f raw-block-pod.yaml
```

Para criar um PersistentVolumeClaim baseado em sistema de arquivos que usa o ceph-csi-based StorageClass criado acima, o seguinte arquivo YAML pode ser usado para solicitar um sistema de arquivos montado (com suporte de uma imagem RBD) de csi-rbd-sc StorageClass:

```
kubectl@adm > cat <<EOF > pvc.yaml
---
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
```

```
name: rbd-pvc
spec:
  accessModes:
    - ReadWriteOnce
  volumeMode: Filesystem
  resources:
    requests:
      storage: 1Gi
  storageClassName: csi-rbd-sc
EOF
kubectl@adm > kubectl apply -f pvc.yaml
```

Veja a seguir uma demonstração de como vincular o PersistentVolumeClaim acima a um recurso de pod como um sistema de arquivos montado:

```
kubectl@adm > cat <<EOF > pod.yaml
---
apiVersion: v1
kind: Pod
metadata:
  name: csi-rbd-demo-pod
spec:
  containers:
    - name: web-server
      image: nginx
      volumeMounts:
        - name: mypvc
          mountPath: /var/lib/www/html
  volumes:
    - name: mypvc
      persistentVolumeClaim:
        claimName: rbd-pvc
        readOnly: false
EOF
kubectl@adm > kubectl apply -f pod.yaml
```

## IV Acessando os dados do cluster

- 21 Gateway de Objetos do Ceph **260**
- 22 Ceph iSCSI Gateway **316**
- 23 Sistema de arquivos em cluster **334**
- 24 Exportar dados do Ceph por meio do Samba **345**
- 25 NFS Ganesha **363**

## 21 Gateway de Objetos do Ceph

Este capítulo apresenta detalhes sobre as tarefas de administração relacionadas ao Gateway de Objetos, como verificação de status do serviço, gerenciamento de contas, gateways multissite ou autenticação LDAP.

### 21.1 Restrições e limitações de nomeação do Gateway de Objetos

Veja a seguir uma lista dos limites importantes do Gateway de Objetos:

#### 21.1.1 Limitações de compartimento de memória

Ao usar o Gateway de Objetos por meio da API do S3, há um limite para os nomes de compartimento de memória que devem ser compatíveis com DNS e podem ter um traço “-”. Ao usar o Gateway de Objetos por meio da API do Swift, você pode aplicar qualquer combinação de caracteres UTF-8 permitidos, exceto a barra “/”. O tamanho máximo do nome de um compartimento de memória é de 255 caracteres. Os nomes de compartimento de memória devem ser exclusivos.



**Dica:** Usar nomes de compartimento de memória compatíveis com DNS

Embora seja possível usar qualquer nome de compartimento de memória baseado em UTF-8 por meio da API do Swift, é recomendável nomear os compartimentos de memória de acordo com as limitações de nomeação do S3 para evitar problemas ao acessar o mesmo compartimento de memória pela API do S3.

#### 21.1.2 Limitações de objetos armazenados

**Número máximo de objetos por usuário**

Por padrão, nenhuma restrição (limitado por  $\sim 2^{63}$ ).

**Número máximo de objetos por compartimento de memória**

Por padrão, nenhuma restrição (limitado por  $\sim 2^{63}$ ).

## Tamanho máximo de um objeto para upload/armazenamento

Cada upload está restrito a 5 GB. Use várias partes para tamanhos de objetos maiores. O número máximo de pacotes de várias partes é 10.000.

### 21.1.3 Limitações de cabeçalho HTTP

A limitação de cabeçalho HTTP e de solicitação depende do front end da Web usado. O Beast padrão restringe o tamanho do cabeçalho HTTP a 16 kB.

## 21.2 Implantando o Gateway de Objetos

A implantação do Gateway de Objetos do Ceph segue o mesmo procedimento da implantação de outros serviços do Ceph: por meio do `cephadm`. Para obter mais detalhes, consulte o *Livro “Guia de Implantação”, Capítulo 8 “Implantando os serviços principais restantes com o cephadm”, Seção 8.2 “Especificação de serviço e posicionamento”, especificamente o Livro “Guia de Implantação”, Capítulo 8 “Implantando os serviços principais restantes com o cephadm”, Seção 8.3.4 “Implantando Gateways de Objetos”.*

### 21.3 Operando o serviço Gateway de Objetos

Você pode operar os Gateways de Objetos da mesma forma que os outros serviços do Ceph, identificando primeiro o nome do serviço com o comando `ceph orch ps` e executando o seguinte comando para os serviços operacionais, por exemplo:

```
ceph orch daemon restart OGW_SERVICE_NAME
```

Consulte o [Capítulo 14, Operação de serviços do Ceph](#) para obter informações completas sobre como operar os serviços do Ceph.

### 21.4 Opções de configuração

Consulte a [Seção 28.5, “Gateway de Objetos do Ceph”](#) para ver uma lista de opções de configuração do Gateway de Objetos.

## 21.5 Gerenciando o acesso ao Gateway de Objetos

Você pode se comunicar com o Gateway de Objetos usando qualquer interface compatível com S3 ou Swift. A interface do S3 é compatível com um grande subconjunto da API RESTful do Amazon S3. A interface do Swift é compatível com um grande subconjunto da API do OpenStack Swift.

As duas interfaces exigem que você crie um usuário específico e instale o software cliente relevante para comunicação com o gateway usando a chave secreta do usuário.

### 21.5.1 Acessando o Gateway de Objetos

#### 21.5.1.1 Acesso à interface do S3

Para acessar a interface do S3, você precisa de um cliente REST. **S3cmd** é um cliente S3 de linha de comando. Você pode encontrá-lo em [OpenSUSE Build Service \(https://build.opensuse.org/package/show/Cloud:Tools/s3cmd\)](https://build.opensuse.org/package/show/Cloud:Tools/s3cmd). O repositório contém as versões para ambas as distribuições baseadas no SUSE Linux Enterprise e no openSUSE.

Para testar o acesso à interface do S3, você também pode gravar um pequeno script do Python. O script se conectará ao Gateway de Objetos, criará um novo compartimento de memória e listará todos os compartimentos de memória. Os valores para `aws_access_key_id` e `aws_secret_access_key` são extraídos dos valores de `access_key` e `secret_key` retornados pelo comando `radosgw_admin` da [Seção 21.5.2.1, “Adicionando usuários do S3 e do Swift”](#).

1. Instale o pacote `python-boto`:

```
# zypper in python-boto
```

2. Crie um novo script do Python denominado `s3test.py` com o seguinte conteúdo:

```
import boto
import boto.s3.connection
access_key = '11BS02LGFB6AL6H1ADMW'
secret_key = 'vzCEkuryfn060dfec4fgQPqFrncKEIkh3Zcd0ANY'
conn = boto.connect_s3(
    aws_access_key_id = access_key,
    aws_secret_access_key = secret_key,
    host = 'HOSTNAME',
    is_secure=False,
```



```
calling_format = boto.s3.connection.OrdinaryCallingFormat(),
)
bucket = conn.create_bucket('my-new-bucket')
for bucket in conn.get_all_buckets():
    print "NAME\tCREATED".format(
        name = bucket.name,
        created = bucket.creation_date,
    )
```

Substitua HOSTNAME pelo nome de host no qual você configurou o serviço do Gateway de Objetos. Por exemplo, gateway\_host.

### 3. Execute o script:

```
python s3test.py
```

A saída do script é parecida com o seguinte:

```
my-new-bucket 2015-07-22T15:37:42.000Z
```

#### 21.5.1.2 Acesso à interface do Swift

Para acessar o Gateway de Objetos pela interface do Swift, você precisa do cliente de linha de comando **swift**. A página de manual dele [man 1 swift](#) apresenta mais detalhes sobre as opções de linha de comando.

O pacote está incluído no módulo “Public Cloud” para o SUSE Linux Enterprise 12 a partir do SP3 e o SUSE Linux Enterprise 15. Antes de instalar o pacote, você precisa ativar o módulo e atualizar o repositório de software:

```
# SUSEConnect -p sle-module-public-cloud/12/SYSTEM-ARCH
sudo zypper refresh
```

Ou

```
# SUSEConnect -p sle-module-public-cloud/15/SYSTEM-ARCH
# zypper refresh
```

Para instalar o comando **swift**, execute o seguinte:

```
# zypper in python-swiftclient
```

O acesso ao swift usa a seguinte sintaxe:

```
> swift -A http://IP_ADDRESS/auth/1.0 \
```

```
-U example_user:swift -K 'SWIFT_SECRET_KEY' list
```

Substitua `IP_ADDRESS` pelo endereço IP do servidor gateway, e `_SECRET_KEY` pelo respectivo valor da saída do comando **radosgw-admin key create** executado para o usuário `swiftswift` na [Seção 21.5.2.1, “Adicionando usuários do S3 e do Swift”](#).

Por exemplo:

```
> swift -A http://gateway.example.com/auth/1.0 -U example_user:swift \
-K 'r5wWIXj0CeE07DixD1FjTLmNYIViaC6JVhi3013h' list
```

A saída é:

```
my-new-bucket
```

## 21.5.2 Gerenciar contas do S3 e do Swift

### 21.5.2.1 Adicionando usuários do S3 e do Swift

É necessário criar um usuário, uma chave de acesso e um segredo para permitir que os usuários finais interajam com o gateway. Há dois tipos de usuário: *usuário* e *subusuário*. Os *usuários* são usados para interagir com a interface do S3, os *subusuários* são usuários da interface do Swift. Cada subusuário está associado a um usuário.

Para criar um usuário do Swift, siga as etapas:

1. Para criar um usuário do Swift, que é um *subusuário* em nossa terminologia, você precisa criar primeiro o *usuário* associado.

```
cephuser@adm > radosgw-admin user create --uid=USERNAME \
--display-name="DISPLAY-NAME" --email=EMAIL
```

Por exemplo:

```
cephuser@adm > radosgw-admin user create \
--uid=example_user \
--display-name="Example User" \
--email=penguin@example.com
```

2. Para criar um subusuário (interface do Swift) para o usuário, você deve especificar o ID de usuário (`--uid=USERNAME`), um ID de subusuário e o nível de acesso para o subusuário.

```
cephuser@adm > radosgw-admin subuser create --uid=UID \  
--subuser=UID \  
--access=[ read | write | readwrite | full ]
```

Por exemplo:

```
cephuser@adm > radosgw-admin subuser create --uid=example_user \  
--subuser=example_user:swift --access=full
```

### 3. Gere uma chave secreta para o usuário.

```
cephuser@adm > radosgw-admin key create \  
--gen-secret \  
--subuser=example_user:swift \  
--key-type=swift
```

### 4. Os dois comandos resultarão em dados formatados em JSON que mostram o estado do usuário. Observe as linhas a seguir e lembre-se do valor secret\_key:

```
"swift_keys": [  
  { "user": "example_user:swift",  
    "secret_key": "r5wWIXj0CeE07DixD1FjTLmNYIViaC6JVhi3013h"}],
```

Ao acessar o Gateway de Objetos por meio da interface do S3, você precisa criar um usuário do S3 executando:

```
cephuser@adm > radosgw-admin user create --uid=USERNAME \  
--display-name="DISPLAY-NAME" --email=EMAIL
```

Por exemplo:

```
cephuser@adm > radosgw-admin user create \  
--uid=example_user \  
--display-name="Example User" \  
--email=penguin@example.com
```

O comando também cria o acesso do usuário e a chave secreta. Verifique a saída para as palavras-chave access\_key e secret\_key e seus valores:

```
[...]  
"keys": [  
  { "user": "example_user",  
    "access_key": "11BS02LGFB6AL6H1ADMW",  
    "secret_key": "vzCEkuryfn060dfce4fgQPqFrncKEIkh3Zcd0ANY"}],  
[...]
```

### 21.5.2.2 Removendo usuários do S3 e do Swift

O procedimento para apagar usuários é semelhante para os usuários do S3 e do Swift. No caso dos usuários do Swift, porém, você pode precisar apagar o usuário com os subusuários incluídos. Para remover um usuário do S3 ou do Swift (incluindo todos os subusuários), especifique `user rm` e o ID de usuário no seguinte comando:

```
cephuser@adm > radosgw-admin user rm --uid=example_user
```

Para remover um subusuário, especifique `subuser rm` e o ID de subusuário.

```
cephuser@adm > radosgw-admin subuser rm --uid=example_user:swift
```

Você pode usar as seguintes opções:

#### `--purge-data`

Purga todos os dados associados ao ID de usuário.

#### `--purge-keys`

Purga todas as chaves associadas ao ID de usuário.



#### Dica: Removendo um subusuário

Ao remover um subusuário, você remove o acesso à interface do Swift. O usuário permanecerá no sistema.

### 21.5.2.3 Mudando o acesso e as chaves secretas do usuário do S3 e do Swift

Os parâmetros `access_key` e `secret_key` identificam o usuário do Gateway de Objetos ao acessar o gateway. A mudança das chaves existentes de usuário é o mesmo que criar novas chaves, pois as chaves antigas são sobregravadas.

Para usuários do S3, execute o seguinte:

```
cephuser@adm > radosgw-admin key create --uid=EXAMPLE_USER --key-type=s3 --gen-access-key --gen-secret
```

Para usuários do Swift, execute o seguinte:

```
cephuser@adm > radosgw-admin key create --subuser=EXAMPLE_USER:swift --key-type=swift --gen-secret
```

--key-type=TYPE

Especifica o tipo de chave. Pode ser swift ou s3.

--gen-access-key

Gera uma chave de acesso aleatória (por padrão, para o usuário do S3).

--gen-secret

Gera uma chave secreta aleatória.

--secret=KEY

Especifica uma chave secreta. Por exemplo, gerada manualmente.

#### 21.5.2.4 Habilitando o gerenciamento de cotas de usuários

O Gateway de Objetos do Ceph permite definir cotas para usuários e compartimentos de memória pertencentes aos usuários. As cotas incluem o número máximo de objetos em um compartimento de memória e o tamanho máximo de armazenamento em megabytes.

Antes de habilitar uma cota de usuário, você precisa definir os respectivos parâmetros:

```
cephuser@adm > radosgw-admin quota set --quota-scope=user --uid=EXAMPLE_USER \
--max-objects=1024 --max-size=1024
```

--max-objects

Especifica o número máximo de objetos. Um valor negativo desabilita a verificação.

--max-size

Especifica o número máximo de bytes. Um valor negativo desabilita a verificação.

--quota-scope

Define o escopo para a cota. As opções são bucket e user. As cotas de compartimento de memória aplicam-se aos compartimentos de memória que um usuário possui. As cotas de usuário aplicam-se a um usuário.

Após definir uma cota de usuário, você poderá habilitá-la:

```
cephuser@adm > radosgw-admin quota enable --quota-scope=user --uid=EXAMPLE_USER
```

Para desabilitar uma cota:

```
cephuser@adm > radosgw-admin quota disable --quota-scope=user --uid=EXAMPLE_USER
```

Para listar as configurações de cota:

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

Para atualizar as estatísticas de cota:

```
cephuser@adm > radosgw-admin user stats --uid=EXAMPLE_USER --sync-stats
```

## 21.6 Front ends HTTP

O Gateway de Objetos do Ceph suporta dois front ends HTTP incorporados: *Beast* e *Civetweb*.

O front end *Beast* usa a biblioteca *Boost.Beast* para análise de HTTP e a biblioteca *Boost.Asio* para E/S de rede assíncrona.

O front end *Civetweb* usa a biblioteca HTTP *Civetweb*, que é uma bifurcação do *Mongoose*.

Você pode configurá-la com a opção `rgw_frontends`. Consulte a [Seção 28.5, “Gateway de Objetos do Ceph”](#) para ver uma lista de opções de configuração.

## 21.7 Habilitar HTTPS/SSL para Gateways de Objetos

Para habilitar a comunicação segura do Gateway de Objetos por meio de SSL, você precisa ter um certificado emitido por uma CA ou criar um autoassinado.

### 21.7.1 Criando um certificado autoassinado



#### Dica

Ignore esta seção se você já tem um certificado válido assinado por uma CA.

O procedimento a seguir descreve como gerar um certificado SSL autoassinado no Master Salt.

1. Se você precisar que o Gateway de Objetos seja reconhecido por outras identidades de assunto, adicione-as à opção `subjectAltName` na seção `[v3_req]` do arquivo `/etc/ssl/openssl.cnf`:

```
[...]
[ v3_req ]
```

```
subjectAltName = DNS:server1.example.com DNS:server2.example.com
[...]
```



### Dica: Endereços IP em subjectAltName

Para usar endereços IP no lugar de nomes de domínio na opção `subjectAltName`, substitua a linha de exemplo pelo seguinte:

```
subjectAltName = IP:10.0.0.10 IP:10.0.0.11
```

2. Crie a chave e o certificado usando **openssl**. Insira todos os dados que você precisa incluir em seu certificado. É recomendável inserir o FQDN como nome comum. Antes de assinar o certificado, verifique se “X509v3 Subject Alternative Name:” está incluído nas extensões solicitadas e se o certificado resultante tem "X509v3 Subject Alternative Name:" definido.

```
root@master # openssl req -x509 -nodes -days 1095 \
-newkey rsa:4096 -keyout rgw.key
-out rgw.pem
```

3. Anexe a chave ao arquivo de certificado:

```
root@master # cat rgw.key >> rgw.pem
```

## 21.7.2 Configurando o Gateway de Objetos com SSL

Para configurar o Gateway de Objetos para usar certificados SSL, use a opção `rgw_frontends`. Por exemplo:

```
cephuser@adm > ceph config set WHO rgw_frontends \
beast ssl_port=443 ssl_certificate=config://CERT ssl_key=config://KEY
```

Se você não especificar as chaves de configuração `CERT` e `KEY`, o serviço Gateway de Objetos procurará o certificado SSL e a chave nas seguintes chaves de configuração:

```
rgw/cert/RGW_REALM/RGW_ZONE.key
rgw/cert/RGW_REALM/RGW_ZONE.crt
```

Para anular a chave SSL padrão e o local do certificado, importe-os para o banco de dados de configuração usando o seguinte comando:

```
ceph config-key set CUSTOM_CONFIG_KEY -i PATH_TO_CERT_FILE
```

Em seguida, use as chaves de configuração personalizadas com a diretiva `config://`.

## 21.8 Módulos de sincronização

O Gateway de Objetos é implantado como um serviço multissite, enquanto você pode espelhar dados e metadados entre as zonas. Os *módulos de sincronização* foram desenvolvidos com base na estrutura multissite, que permite encaminhar dados e metadados para uma camada externa diferente. Um módulo de sincronização permite a execução de um conjunto de ações sempre que há uma mudança nos dados (por exemplo, operações de metadados como criação de compartimento de memória ou de usuário). Como as mudanças de multissite do Gateway de Objetos acabam sendo consistentes em sites remotos, elas são propagadas de forma assíncrona. Isso abrange casos de uso como backup de armazenamento de objetos em um cluster de nuvem externo, solução de backup personalizada que usa unidades de fita ou indexação de metadados no ElasticSearch.

### 21.8.1 Configurando módulos de sincronização

Todos os módulos de sincronização são configurados de forma semelhante. Você precisa criar uma nova zona (consulte a [Seção 21.13, “Gateways de Objetos multissite”](#) para obter mais detalhes) e definir a opção `--tier_type` dela, por exemplo, `--tier-type=cloud` para o módulo de sincronização de nuvem:

```
cephuser@adm > radosgw-admin zone create --rgw-zonegroup=ZONE-GROUP-NAME \  
--rgw-zone=ZONE-NAME \  
--endpoints=http://endpoint1.example.com,http://endpoint2.example.com, [...] \  
--tier-type=cloud
```

Você pode configurar a camada específica usando o seguinte comando:

```
cephuser@adm > radosgw-admin zone modify --rgw-zonegroup=ZONE-GROUP-NAME \  
--rgw-zone=ZONE-NAME \  
--tier-config=KEY1=VALUE1,KEY2=VALUE2
```

A *KEY* (CHAVE) na configuração especifica a variável de configuração que você deseja atualizar, e o *VALUE* (VALOR) especifica o novo valor dela. É possível usar um ponto para acessar os valores aninhados. Por exemplo:

```
cephuser@adm > radosgw-admin zone modify --rgw-zonegroup=ZONE-GROUP-NAME \  
--tier-config=tier.config.bucket=cloud
```



```
--rgw-zone=ZONE-NAME \  
--tier-config=connection.access_key=KEY,connection.secret=SECRET
```

Você pode acessar entradas de matriz anexando colchetes “[ ]” com a entrada referenciada. Você pode adicionar uma nova entrada de matriz usando colchetes “[ ]”. O valor do índice de -1 faz referência à última entrada na matriz. Não é possível criar uma nova entrada e fazer referência a ela novamente no mesmo comando. Por exemplo, veja a seguir um comando para criar um novo perfil para compartimentos de memória que começam com *PREFIX*:

```
cephuser@adm > radosgw-admin zone modify --rgw-zonegroup=ZONE-GROUP-NAME \  
--rgw-zone=ZONE-NAME \  
--tier-config=profiles[].source_bucket=PREFIX'*'  
cephuser@adm > radosgw-admin zone modify --rgw-zonegroup=ZONE-GROUP-NAME \  
--rgw-zone=ZONE-NAME \  
--tier-config=profiles[-1].connection_id=CONNECTION_ID,profiles[-1].acls_id=ACLS_ID
```



### Dica: Adicionando e removendo entradas de configuração

Você pode adicionar uma nova entrada de configuração de camada usando o parâmetro `--tier-config-add=KEY=VALUE`.

Você pode remover uma entrada existente usando `--tier-config-rm=KEY`.

## 21.8.2 Sincronizando zonas

A configuração de um módulo de sincronização é local para uma zona. O módulo de sincronização determina se a zona exporta os dados ou apenas pode consumir os dados que foram modificados em outra zona. A partir do Luminous, os plug-ins de sincronização suportados são *ElasticSearch*, *rgw*, que é o plug-in padrão que sincroniza dados entre zonas, e *log*, que é o plug-in comum que registra a operação de metadados executada nas zonas remotas. As seções a seguir foram elaboradas com o exemplo de uma zona que usa o módulo de sincronização *ElasticSearch*. O mesmo processo pode ser aplicado para configurar qualquer outro plug-in de sincronização.



### Nota: Plug-in de sincronização padrão

*rgw* é o plug-in de sincronização padrão, e não há necessidade de configurá-lo explicitamente.

### 21.8.2.1 Requisitos e considerações

Vamos considerar uma configuração multissite simples, conforme descrito na [Seção 21.13, “Gateways de Objetos multissite”](#), com 2 zonas: us-east e us-west. Agora, adicionamos uma terceira zona us-east-es, que processará apenas os metadados de outros sites. Essa zona pode estar no mesmo ou em um cluster do Ceph diferente do us-east. Essa zona consumirá apenas os metadados de outras zonas, e os Gateways de Objetos nela não atenderão diretamente nenhuma solicitação de usuário final.

### 21.8.2.2 Configurando zonas

1. Crie a terceira zona semelhante às aquelas descritas na [Seção 21.13, “Gateways de Objetos multissite”](#). Por exemplo,

```
cephuser@adm > radosgw-admin zone create --rgw-zonegroup=us --rgw-zone=us-east-es \
--access-key=SYSTEM-KEY --secret=SECRET --endpoints=http://rgw-es:80
```

2. É possível configurar um módulo de sincronização para essa zona por meio do seguinte comando:

```
cephuser@adm > radosgw-admin zone modify --rgw-zone=ZONE-NAME --tier-type=TIER-TYPE \
--tier-config={set of key=value pairs}
```

3. Por exemplo, no módulo de sincronização ElasticSearch

```
cephuser@adm > radosgw-admin zone modify --rgw-zone=ZONE-NAME --tier-
type=elasticsearch \
--tier-config=endpoint=http://localhost:9200,num_shards=10,num_replicas=1
```

Para as várias opções de configuração de camada suportadas, consulte a [Seção 21.8.3, “Módulo de sincronização ElasticSearch”](#).

4. Por fim, atualize o período

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

5. Agora, inicie o Gateway de Objetos na zona

```
cephuser@adm > ceph orch start rgw.REALM-NAME.ZONE-NAME
```

### 21.8.3 Módulo de sincronização Elasticsearch

Esse módulo de sincronização grava os metadados de outras zonas no Elasticsearch. A partir do Luminous, é o JSON dos campos de dados que armazenamos no Elasticsearch atualmente.

```
{
  "_index" : "rgw-gold-ee5863d6",
  "_type" : "object",
  "_id" : "34137443-8592-48d9-8ca7-160255d52ade.34137.1:object1:null",
  "_score" : 1.0,
  "_source" : {
    "bucket" : "testbucket123",
    "name" : "object1",
    "instance" : "null",
    "versioned_epoch" : 0,
    "owner" : {
      "id" : "user1",
      "display_name" : "user1"
    },
    "permissions" : [
      "user1"
    ],
    "meta" : {
      "size" : 712354,
      "mtime" : "2017-05-04T12:54:16.462Z",
      "etag" : "7ac66c0f148de9519b8bd264312c4d64"
    }
  }
}
```

#### 21.8.3.1 Parâmetros de configuração de tipo de camada do Elasticsearch

##### endpoint

Especifica o endpoint do servidor Elasticsearch a ser acessado.

##### num\_shards

*(número inteiro)* O número de fragmentos com os quais o Elasticsearch será configurado na inicialização da sincronização de dados. Observe que ele não pode ser mudado após a inicialização. Qualquer mudança aqui requer a reconstrução do índice do Elasticsearch e a reinicialização do processo de sincronização de dados.

#### num\_replicas

*(número inteiro)* O número de réplicas com as quais o Elasticsearch será configurado na inicialização da sincronização de dados.

#### explicit\_custom\_meta

*(true | false)* Especifica se todos os metadados personalizados do usuário serão indexados ou se o usuário precisará configurar (no nível do compartimento de memória) quais entradas de metadados do cliente devem ser indexadas. Por padrão, isso é “false”

#### index\_buckets\_list

*(lista de strings separadas por vírgulas)* Se vazia, todos os compartimentos de memória serão indexados. Do contrário, apenas os compartimentos de memória especificados nela serão indexados. É possível inserir prefixos (por exemplo, “foo\*”) ou sufixos (por exemplo, “\*bar”) de compartimento de memória.

#### approved\_owners\_list

*(lista de strings separadas por vírgulas)* Se vazia, os compartimentos de memória de todos os proprietários serão indexados (sujeito a outras restrições); do contrário, apenas os compartimentos de memória pertencentes a determinados proprietários serão indexados. É possível também inserir prefixos e sufixos.

#### override\_index\_path

*(string)* Se não estiver vazia, essa string será usada como o caminho do índice do Elasticsearch. Do contrário, o caminho do índice será determinado e gerado na inicialização da sincronização.

#### username

Especifica um nome de usuário para o Elasticsearch se a autenticação for necessária.

#### password

Especifica uma senha para o Elasticsearch se a autenticação for necessária.

### 21.8.3.2 Consultas de metadados

Como o cluster do Elasticsearch agora armazena metadados de objetos, é importante não expor o endpoint do Elasticsearch ao público e mantê-lo acessível apenas aos administradores de cluster. A própria exposição das consultas de metadados ao usuário final representa um problema, já

que desejamos que o usuário consulte apenas os metadados dele, e não de quaisquer outros usuários. Para isso, o cluster do Elasticsearch deve autenticar os usuários de modo similar ao RGW, o que representa um problema.

A partir do Luminous, o RGW na zona master de metadados agora pode atender às solicitações de usuários finais. Isso evita a exposição do endpoint do Elasticsearch ao público e resolve também o problema de autenticação e autorização, pois o próprio RGW pode autenticar as solicitações de usuário final. Para essa finalidade, o RGW inclui uma nova consulta nas APIs de compartimento de memória que pode atender às solicitações de serviço do Elasticsearch. Todas essas solicitações devem ser enviadas para a zona master de metadados.

### Obter uma consulta do Elasticsearch

```
GET /BUCKET?query=QUERY-EXPR
```

parâmetros de solicitação:

- max-keys: número máx. de entradas a retornar
- marker: marcador de paginação

```
expression := [(]<arg> <op> <value> [)][<and|or> ...]
```

op é um dos seguintes: <, <=, ==, >=, >

Por exemplo:

```
GET /?query=name==foo
```

Retornará todas as chaves indexadas para as quais o usuário tem permissão de leitura e que são denominadas “foo”. A saída será uma lista de chaves em XML, que é semelhante à resposta de compartimentos de memória da lista do S3.

### Configurar campos personalizados de metadados

Defina quais entradas de metadados personalizados devem ser indexadas (no compartimento de memória especificado) e quais são os tipos das chaves. Se for configurada a indexação explícita de metadados personalizados, esse procedimento será necessário para o rgw indexar os valores de metadados personalizados especificados. Do contrário, ele será necessário nos casos em que as chaves dos metadados indexados são de um tipo diferente de string.

```
POST /BUCKET?mdsearch
```

```
x-amz-meta-search: <key [; type]> [, ...]
```

Vários campos de metadados devem ser separados por vírgula. É possível forçar um tipo para um campo com “;”. Os tipos permitidos atualmente são string (padrão), número inteiro e data. Por exemplo, para indexar metadados de um objeto personalizado x-amz-meta-year como número inteiro, x-amz-meta-date como o tipo data e x-amz-meta-title como string, faça o seguinte

```
POST /mybooks?mdsearch
x-amz-meta-search: x-amz-meta-year;int, x-amz-meta-release-date;date, x-amz-meta-
title;string
```

#### Apague a configuração de metadados personalizados

Apague a configuração de compartimento de memória dos metadados personalizados.

```
DELETE /BUCKET?mdsearch
```

#### Obter a configuração dos metadados personalizados

Recupere a configuração de compartimento de memória dos metadados personalizados.

```
GET /BUCKET?mdsearch
```

## 21.8.4 Módulo de sincronização de nuvem

Esta seção apresenta um módulo que sincroniza os dados da zona com um serviço de nuvem remoto. A sincronização é apenas unidirecional, os dados não são sincronizados de volta da zona remota. O principal objetivo deste módulo é habilitar a sincronização de dados com vários provedores de serviços de nuvem. Atualmente, ele suporta provedores de nuvem compatíveis com a AWS (S3).

Para sincronizar os dados com um serviço de nuvem remoto, você precisa configurar as credenciais do usuário. Como muitos serviços de nuvem apresentam limites quanto ao número de compartimentos de memória que cada usuário pode criar, é possível configurar o mapeamento de objetos e compartimentos de memória de origem, destinos diferentes para compartimentos de memória distintos e prefixos de compartimento de memória. Observe que as listas de acesso de origem (ACLs) não serão preservadas. É possível mapear permissões de usuários de origem específicos para usuários de destino específicos.

Devido às limitações da API, não existe um modo de preservar o horário de modificação do objeto original e a tag da entidade HTTP (ETag). O módulo de sincronização de nuvem armazena essas informações como atributos de metadados nos objetos de destino.

### 21.8.4.1 Configurando o módulo de sincronização de nuvem

Veja a seguir exemplos de uma configuração comum e não comum para o módulo de sincronização de nuvem. Observe que a configuração comum pode ser diferente da não comum.

#### EXEMPLO 21.1: CONFIGURAÇÃO COMUM

```
{
  "connection": {
    "access_key": ACCESS,
    "secret": SECRET,
    "endpoint": ENDPOINT,
    "host_style": path | virtual,
  },
  "acls": [ { "type": id | email | uri,
    "source_id": SOURCE_ID,
    "dest_id": DEST_ID } ... ],
  "target_path": TARGET_PATH,
}
```

#### EXEMPLO 21.2: CONFIGURAÇÃO NÃO COMUM

```
{
  "default": {
    "connection": {
      "access_key": ACCESS,
      "secret": SECRET,
      "endpoint": ENDPOINT,
      "host_style" path | virtual,
    },
    "acls": [
      {
        "type": id | email | uri, # optional, default is id
        "source_id": ID,
        "dest_id": ID
      } ... ]
    "target_path": PATH # optional
  },
  "connections": [
    {
      "connection_id": ID,
      "access_key": ACCESS,
      "secret": SECRET,
      "endpoint": ENDPOINT,
      "host_style": path | virtual, # optional
    } ... ],
  "acl_profiles": [
```

```

{
  "acls_id": ID, # acl mappings
  "acls": [ {
    "type": id | email | uri,
    "source_id": ID,
    "dest_id": ID
  } ... ]
},
],
"profiles": [
{
  "source_bucket": SOURCE,
  "connection_id": CONNECTION_ID,
  "acls_id": MAPPINGS_ID,
  "target_path": DEST,          # optional
} ... ],
}

```

Veja a seguir a explicação dos termos de configuração usados:

#### connection

Representa uma conexão com o serviço de nuvem remota. Contém "connection\_id", "access\_key", "secret", "endpoint" e "host\_style".

#### access\_key

A chave de acesso à nuvem remota que será usada para a conexão específica.

#### secret

A chave secreta para o serviço de nuvem remota.

#### endpoint

URL do endpoint do serviço de nuvem remota.

#### host\_style

Tipo de estilo do host ("path" ou "virtual") a ser usado quando acessar o endpoint da nuvem remota. O padrão é "path" (caminho).

#### acls

Matriz de mapeamentos da lista de acesso.

#### acl\_mapping

Cada estrutura "acl\_mapping" contém "type", "source\_id" e "dest\_id". Eles definirão a mutação da ACL para cada objeto. Uma mutação da ACL permite converter o ID de usuário de origem em um ID de destino.



## type

Tipo de ACL: "id" define o ID de usuário, "email" define o usuário por e-mail e "uri" define o usuário por URI (grupo).

## source\_id

ID do usuário na zona de origem.

## dest\_id

ID do usuário no destino.

## target\_path

Uma string que define como o caminho de destino é criado. O caminho de destino especifica um prefixo ao qual o nome do objeto de origem é anexado. O caminho de destino configurável pode incluir qualquer uma das seguintes variáveis:

### SID

Uma string exclusiva que representa o ID da instância de sincronização.

### ZONEGROUP

Nome do grupo de zonas.

### ZONEGROUP\_ID

ID do grupo de zonas.

### ZONE

Nome da zona.

### ZONE\_ID

ID da zona.

### BUCKET

Nome do compartimento de memória de origem.

### OWNER

ID do proprietário do compartimento de memória de origem.

Por exemplo: target\_path = rgwx-ZONE-SID/OWNER/BUCKET

## acl\_profiles

Uma matriz de perfis da lista de acesso.

## acl\_profile

Cada perfil contém: "acls\_id", que representa o perfil, e uma matriz de "acls", que armazena uma lista de "acl\_mappings".

## profiles

Uma lista de perfis. Cada perfil contém o seguinte:

### source\_bucket

Nome ou prefixo do compartimento de memória (se terminar com \*), que define o(s) compartimento(s) de memória de origem para este perfil.

### target\_path

Veja a explicação acima.

### connection\_id

ID da conexão que será usada para este perfil.

### acls\_id

ID do perfil da ACL que será usado para este perfil.

## 21.8.4.2 Elementos de configuração específicos do S3

O módulo de sincronização de nuvem apenas funcionará com back ends compatíveis com o AWS S3. Há alguns elementos de configuração que podem ser usados para ajustar o comportamento ao acessar serviços de nuvem do S3:

```
{
  "multipart_sync_threshold": OBJECT_SIZE,
  "multipart_min_part_size": PART_SIZE
}
```

### multipart\_sync\_threshold

Os objetos cujo tamanho é igual ou maior do que esse valor serão sincronizados com o serviço de nuvem por meio do upload de várias partes.

### multipart\_min\_part\_size

Tamanho mínimo das partes para usar na sincronização de objetos por meio do upload de várias partes.

## 21.8.5 Módulo de sincronização de arquivo

O *módulo de sincronização de arquivo* usa o recurso de controle de versão dos objetos do S3 no Gateway de Objetos. Você pode configurar uma *zona de arquivo*, que captura as diferentes versões dos objetos do S3 à medida que surgem nas outras zonas ao longo do tempo. O histórico de versões que a zona de arquivo mantém apenas pode ser eliminado pelos gateways associados à zona de arquivo.

Com essa arquitetura, várias zonas sem controle versão podem espelhar seus dados e metadados por meio de seus gateways de zona, oferecendo alta disponibilidade aos usuários finais, enquanto a zona de arquivo captura todas as atualizações de dados para consolidá-los como versões dos objetos do S3.

Ao incluir a zona de arquivo em uma configuração de várias zonas, você ganha a flexibilidade de um histórico de objetos do S3 em uma zona, além de economizar o espaço que as réplicas dos objetos do S3 com controle de versão consomem nas zonas restantes.

### 21.8.5.1 Configurando o módulo de sincronização de arquivo



#### Dica: Mais informações

Consulte a [Seção 21.13, “Gateways de Objetos multissite”](#) para obter detalhes sobre a configuração de gateways multissite.

Consulte a [Seção 21.8, “Módulos de sincronização”](#) para obter detalhes sobre a configuração de módulos de sincronização.

Para usar o módulo de sincronização de arquivo, você precisa criar uma nova zona com o tipo de camada definido como arquivo:

```
cephuser@adm > radosgw-admin zone create --rgw-zonegroup=ZONE_GROUP_NAME \
--rgw-zone=OGW_ZONE_NAME \
--endpoints=http://OGW_ENDPOINT1_URL[,http://OGW_ENDPOINT2_URL,...]
--tier-type=archive
```

## 21.9 Autenticação LDAP

Além da autenticação de usuário local padrão, o Gateway de Objetos pode usar os serviços do servidor LDAP para autenticar também os usuários.

## 21.9.1 Mecanismo de autenticação

O Gateway de Objetos extrai as credenciais de LDAP do usuário de um token. Um filtro de pesquisa é construído com base no nome de usuário. O Gateway de Objetos usa a conta de serviço configurada para pesquisar uma entrada correspondente no diretório. Se uma entrada for encontrada, o Gateway de Objetos tentará se vincular ao nome exclusivo encontrado com a senha do token. Se as credenciais forem válidas, o vínculo será bem-sucedido, e o Gateway de Objetos concederá o acesso.

Você pode limitar os usuários permitidos definindo a base para a pesquisa como uma unidade organizacional específica ou especificando um filtro de pesquisa personalizado. Por exemplo, exigir a participação em um grupo específico, classes de objetos ou atributos personalizados.

## 21.9.2 Requisitos

- *LDAP ou Active Directory*: Uma instância LDAP em execução acessível pelo Gateway de Objetos.
- *Conta de serviço*: Credenciais LDAP para uso do Gateway de Objetos com permissões de pesquisa.
- *Conta de usuário*: Pelo menos, uma conta do usuário no diretório LDAP.



### Importante: Não sobreponha usuários LDAP e locais

Você não deve usar os mesmos nomes para usuários locais e usuários autenticados por LDAP. O Gateway de Objetos não pode diferenciá-los e os trata como se fossem os mesmos usuários.



### Dica: Verificações de integridade

Use o utilitário **ldapsearch** para verificar a conta de serviço ou a conexão LDAP. Por exemplo:

```
> ldapsearch -x -D "uid=ceph,ou=system,dc=example,dc=com" -W \  
-H ldaps://example.com -b "ou=users,dc=example,dc=com" 'uid=*' dn
```

Use os mesmos parâmetros LDAP que o arquivo de configuração do Ceph para evitar possíveis problemas.

## 21.9.3 Configurando o Gateway de Objetos para usar a autenticação LDAP

Os parâmetros a seguir estão relacionados à autenticação LDAP:

### rgw\_s3\_auth\_use\_ldap

Defina essa opção como true para habilitar a autenticação S3 com LDAP.

### rgw\_ldap\_uri

Especifica o servidor LDAP a ser usado. Use o parâmetro ldaps://FQDN:PORT para evitar a transmissão aberta de credenciais de texto simples.

### rgw\_ldap\_binddn

O DN (Distinguished Name – Nome Exclusivo) da conta de serviço usada pelo Gateway de Objetos.

### rgw\_ldap\_secret

A senha para a conta de serviço.

### rgw\_ldap\_searchdn

Especifica a base na árvore de informações do diretório para pesquisar usuários. Ela pode ser a unidade organizacional de usuários ou alguma OU (Organizational Unit – Unidade Organizacional) mais específica.

### rgw\_ldap\_dnattr

O atributo que está sendo usado no filtro de pesquisa construído para corresponder um nome de usuário. Dependendo da DIT (Directory Information Tree – Árvore de Informações do Diretório), ele provavelmente será uid ou cn.

### rgw\_search\_filter

Se não for especificado, o Gateway de Objetos construirá automaticamente o filtro de pesquisa com a configuração rgw\_ldap\_dnattr. Use esse parâmetro para restringir a lista de usuários permitidos com muita flexibilidade. Consulte a [Seção 21.9.4, “Usando um filtro de pesquisa personalizado para limitar o acesso do usuário”](#) para obter detalhes.

## 21.9.4 Usando um filtro de pesquisa personalizado para limitar o acesso do usuário

Você pode usar o parâmetro rgw\_search\_filter de duas maneiras.

#### 21.9.4.1 Filtro parcial para limitar ainda mais o filtro de pesquisa construído

Veja a seguir um exemplo de filtro parcial:

```
"objectclass=inetorgperson"
```

O Gateway de Objetos gerará o filtro de pesquisa como de costume com o nome de usuário extraído do token e o valor de `rgw_ldap_dnattr`. Em seguida, o filtro construído será combinado ao filtro parcial com base no atributo `rgw_search_filter`. Dependendo do nome de usuário e das configurações, o filtro de pesquisa final poderá ser:

```
"(&(uid=hari)(objectclass=inetorgperson))"
```

Nesse caso, o usuário “hari” apenas receberá acesso se for encontrado no diretório LDAP, se tiver uma classe de objeto “inetorgperson” e se especificar uma senha válida.

#### 21.9.4.2 Filtro completo

Um filtro completo deve conter um token `USERNAME` que será substituído pelo nome de usuário durante a tentativa de autenticação. O parâmetro `rgw_ldap_dnattr` não é mais usado neste caso. Por exemplo, para limitar os usuários válidos a um grupo específico, use o filtro a seguir:

```
"(&(uid=USERNAME)(memberOf=cn=ceph-users,ou=groups,dc=mycompany,dc=com))"
```



#### Nota: Atributo `memberOf`

O uso do atributo `memberOf` nas pesquisas LDAP requer suporte da sua implementação de servidor LDAP específica.

### 21.9.5 Gerando um token de acesso para autenticação LDAP

O utilitário `radosgw-token` gera o token de acesso com base no nome de usuário e na senha LDAP. Ele emite uma string codificada com base64, que é o token de acesso real. Use seu cliente S3 favorito (consulte a [Seção 21.5.1, “Acessando o Gateway de Objetos”](#)), especifique o token como a chave de acesso e use uma chave secreta vazia.

```
> export RGW_ACCESS_KEY_ID="USERNAME"
> export RGW_SECRET_ACCESS_KEY="PASSWORD"
```

```
cephuser@adm > radosgw-token --encode --ttype=ldap
```



### Importante: Credenciais de texto sem criptografia

O token de acesso é uma estrutura JSON codificada com base64 que contém as credenciais LDAP como texto sem criptografia.



### Nota: Active Directory

Para o Active Directory, use o parâmetro `--ttype=ad`.

## 21.10 Fragmentação de índice do compartimento de memória

O Gateway de Objetos armazena os dados de índice do compartimento de memória em um pool de índice, que assume `.rgw.buckets.index` como padrão. Se você colocar um número excessivo (centenas de milhares) de objetos em um único compartimento de memória, e a cota para o número máximo de objetos por compartimento de memória (`rgw bucket default quota max objects`) não for definida, o desempenho do pool de índice poderá ser prejudicado. A *fragmentação de índice do compartimento de memória* impede essa redução no desempenho e permite um alto número de objetos por compartimento de memória.

### 21.10.1 Refragmentação de índice do compartimento de memória

Se um compartimento de memória ficar muito grande e sua configuração inicial não for mais suficiente, será necessário refragmentar o pool de índice dele. Você pode usar a refragmentação de índice do compartimento de memória automática online (consulte a [Seção 21.10.1.1, “Refragmentação dinâmica”](#)) ou refragmentar o índice do compartimento de memória offline manualmente (consulte a [Seção 21.10.1.2, “Refragmentação manual”](#)).

### 21.10.1.1 Refragmentação dinâmica

A partir do SUSE Enterprise Storage 5, oferecemos suporte à refragmentação do compartimento de memória online. Ela detecta se o número de objetos por compartimento de memória atinge determinado limite e aumenta automaticamente o número de fragmentos usados pelo índice do compartimento de memória. Esse processo reduz o número de entradas em cada fragmento de índice do compartimento de memória.

O processo de detecção é executado:

- Quando novos objetos são adicionados ao compartimento de memória.
- Em um processo em segundo plano que explora periodicamente todos os compartimentos de memória. Isso é necessário para resolver a questão de compartimentos de memória existentes que não são atualizados.

Um compartimento de memória que requer refragmentação é adicionado à fila `reshard_log` e será programado para ser refragmentado posteriormente. Os threads de refragmentação são executados em segundo plano e executam a refragmentação programada, uma de cada vez.

#### CONFIGURANDO A REFRAGMENTAÇÃO DINÂMICA

##### rgw\_dynamic\_resharding

Habilita ou desabilita a refragmentação dinâmica de índice do compartimento de memória.

Os valores possíveis são “true” (verdadeiro) ou “false” (falso). O padrão é “true”.

##### rgw\_reshard\_num\_logs

Número de fragmentos para o registro da refragmentação. O padrão é 16.

##### rgw\_reshard\_bucket\_lock\_duration

Duração do bloqueio do objeto do compartimento de memória durante a refragmentação.

O padrão é 120 segundos.

##### rgw\_max\_objs\_per\_shard

Número máximo de objetos por fragmento de índice do compartimento de memória. O padrão é 100.000 objetos.

##### rgw\_reshard\_thread\_interval

Tempo máximo de refragmentação entre os ciclos de processamento do . O padrão é 600 segundos.



## COMANDOS PARA ADMINISTRAR O PROCESSO DE REFRAGMENTAÇÃO

Adicionar um compartimento de memória à fila de refragmentação:

```
cephuser@adm > radosgw-admin reshard add \  
--bucket BUCKET_NAME \  
--num-shards NEW_NUMBER_OF_SHARDS
```

Listar a fila de refragmentação:

```
cephuser@adm > radosgw-admin reshard list
```

Processar/Programar a refragmentação de um compartimento de memória:

```
cephuser@adm > radosgw-admin reshard process
```

Exibir o status da refragmentação do compartimento de memória:

```
cephuser@adm > radosgw-admin reshard status --bucket BUCKET_NAME
```

Cancelar uma refragmentação pendente do compartimento de memória:

```
cephuser@adm > radosgw-admin reshard cancel --bucket BUCKET_NAME
```

### 21.10.1.2 Refragmentação manual

A refragmentação dinâmica mencionada na [Seção 21.10.1.1, “Refragmentação dinâmica”](#) é suportada apenas nas configurações simples do Gateway de Objetos. Para configurações multissite, use a refragmentação manual descrita nesta seção.

Para refragmentar o índice do compartimento de memória manualmente offline, use o seguinte comando:

```
cephuser@adm > radosgw-admin bucket reshard
```

O comando **bucket reshard** executa o seguinte:

- Cria um novo conjunto de objetos de índice do compartimento de memória para o objeto especificado.
- Distribui todas as entradas desses objetos de índice.
- Cria uma nova instância do compartimento de memória.

- Vincula a nova instância do compartimento de memória ao compartimento de memória para que todas as novas operações de índice passem pelos novos índices do compartimento de memória.
- Imprime o ID do compartimento de memória antigo e novo para a saída padrão.



## Dica

Ao escolher um número de fragmentos, observe o seguinte: especifique no máximo 100.000 entradas por fragmento. Os fragmentos de índice do compartimento de memória que são números primos costumam funcionar melhor na distribuição uniforme das entradas de índice do compartimento de memória entre os fragmentos. Por exemplo, 503 fragmentos de índice do compartimento de memória são melhores do que 500, pois o primeiro é número primo.

### PROCEDIMENTO 21.1: REFRAGMENTANDO O ÍNDICE DO COMPARTIMENTO DE MEMÓRIA

1. Verifique se todas as operações no compartimento de memória foram interrompidas.
2. Faça backup do índice original do compartimento de memória:

```
cephuser@adm > radosgw-admin bi list \  
--bucket=BUCKET_NAME \  
> BUCKET_NAME.list.backup
```

3. Refragmente o índice do compartimento de memória:

```
cephuser@adm > radosgw-admin bucket reshard \  
--bucket=BUCKET_NAME \  
--num-shards=NEW_SHARDS_NUMBER
```



## Dica: ID do compartimento de memória antigo

Como parte da saída, esse comando também imprime o ID do compartimento de memória novo e antigo.

## 21.10.2 Fragmentação de índice para novos compartimentos de memória

Há duas opções que afetam a fragmentação de índice do compartimento de memória:

- Use a opção `rgw_override_bucket_index_max_shards` para configurações simples.
- Use a opção `bucket_index_max_shards` para configurações multissite.

A definição das opções como `0` desabilita a fragmentação de índice do compartimento de memória. Um valor maior do que `0` habilita a fragmentação de índice do compartimento de memória e define o número máximo de fragmentos.

A fórmula a seguir ajuda você a calcular o número recomendado de fragmentos:

```
number_of_objects_expected_in_a_bucket / 100000
```

Esteja ciente de que o número máximo de fragmentos é 7877.

### 21.10.2.1 Configurações multissite

As configurações multissite podem ter um pool de índice diferente para gerenciar o failover. Para configurar um número consistente de fragmentos para as zonas em um grupo de zonas, defina a opção `bucket_index_max_shards` na configuração do grupo de zonas:

1. Exporte a configuração do grupo de zonas para o arquivo `zonegroup.json`:

```
cephuser@adm > radosgw-admin zonegroup get > zonegroup.json
```

2. Edite o arquivo `zonegroup.json` e defina a opção `bucket_index_max_shards` para cada zona nomeada.

3. Redefina o grupo de zonas:

```
cephuser@adm > radosgw-admin zonegroup set < zonegroup.json
```

4. Atualize o período. Consulte a [Seção 21.13.2.6, “Atualize o período”](#).

## 21.11 Integração do OpenStack Keystone

O OpenStack Keystone é um serviço de identidade que faz parte do produto OpenStack. Você pode integrar o Gateway de Objetos ao Keystone para configurar um gateway que aceita o token de autenticação do Keystone. Um usuário autorizado pelo Keystone a acessar o gateway será verificado no Gateway de Objetos do Ceph e criado automaticamente, se necessário. O Gateway de Objetos consulta o Keystone periodicamente para obter uma lista de tokens revogados.

### 21.11.1 Configurando o OpenStack

Antes de configurar o Gateway de Objetos do Ceph, você precisa configurar o OpenStack Keystone para habilitar o serviço Swift e apontá-lo para o Gateway de Objetos do Ceph:

1. *Defina o serviço Swift.* Para usar o OpenStack para validar usuários do Swift, crie primeiro o serviço Swift:

```
> openstack service create \
  --name=swift \
  --description="Swift Service" \
  object-store
```

2. *Defina os endpoints.* Após criar o serviço Swift, aponte para o Gateway de Objetos do Ceph. Substitua REGION\_NAME pelo nome do grupo de zonas ou da região do gateway.

```
> openstack endpoint create --region REGION_NAME \
  --publicurl "http://radosgw.example.com:8080/swift/v1" \
  --adminurl "http://radosgw.example.com:8080/swift/v1" \
  --internalurl "http://radosgw.example.com:8080/swift/v1" \
  swift
```

3. *Verifique as configurações.* Após criar o serviço Swift e definir os endpoints, mostre os endpoints para verificar se todas as configurações estão corretas.

```
> openstack endpoint show object-store
```

## 21.11.2 Configurando o Gateway de Objetos do Ceph

### 21.11.2.1 Configurar certificados SSL

O Gateway de Objetos do Ceph consulta o Keystone periodicamente para obter uma lista de tokens revogados. Essas solicitações são codificadas e assinadas. É possível também configurar o Keystone para fornecer tokens autoassinados, que também são codificados e assinados. Você precisa configurar o gateway para que possa decodificar e verificar essas mensagens assinadas. Portanto, os certificados OpenSSL que o Keystone usa para criar as solicitações precisam ser convertidos no formato “nss db”:

```
# mkdir /var/ceph/nss
# openssl x509 -in /etc/keystone/ssl/certs/ca.pem \
  -pubkey | certutil -d /var/ceph/nss -A -n ca -t "TCu,Cu,Tuw"
rootopenssl x509 -in /etc/keystone/ssl/certs/signing_cert.pem \
  -pubkey | certutil -A -d /var/ceph/nss -n signing_cert -t "P,P,P"
```

Para permitir que o Gateway de Objetos do Ceph interaja com o OpenStack Keystone, o OpenStack Keystone pode usar um certificado SSL autoassinado. Instale o certificado SSL do Keystone no nó que executa o Gateway de Objetos do Ceph ou, se preferir, defina o valor da opção `rgw keystone verify ssl` como “false”. A definição de `rgw keystone verify ssl` como “false” indica que o gateway não tentará verificar o certificado.

### 21.11.2.2 Configurar as opções do Gateway de Objetos

Você pode configurar a integração com o Keystone usando as seguintes opções:

`rgw keystone api version`

Versão da API do Keystone. As opções válidas são 2 ou 3. O padrão é 2.

`rgw keystone url`

O URL e o número da porta da API RESTful administrativa no servidor Keystone. Segue o padrão `URL_SERVIDOR:NÚMERO_DA_PORTA`.

`rgw keystone admin token`

O token ou segredo compartilhado configurado internamente no Keystone para solicitações administrativas.

`rgw keystone accepted roles`

As funções necessárias para atender às solicitações. O padrão é “Member, admin”.

#### rgw keystone accepted admin roles

A lista de funções que permite a um usuário obter privilégios administrativos.

#### rgw keystone token cache size

O número máximo de entradas no cache de token do Keystone.

#### rgw keystone revocation interval

O número de segundos antes de verificar se há tokens revogados. O padrão é 15 \* 60.

#### rgw keystone implicit tenants

Criar novos usuários em seus próprios locatários de mesmo nome. O padrão é “false”.

#### rgw s3 auth use keystone

Se definido como “true”, o Gateway de Objetos do Ceph autenticará os usuários com o Keystone. O padrão é “false”.

#### nss db path

O caminho para o banco de dados NSS.

Também é possível configurar o locatário de serviço, o usuário e a senha do Keystone (para a versão 2.0 da API do OpenStack Identity), do mesmo modo que os serviços do OpenStack costumam ser configurados. Dessa forma, você pode evitar a definição do segredo compartilhado rgw keystone admin token no arquivo de configuração, que deve ser desabilitado em ambientes de produção. As credenciais do locatário de serviço devem ter privilégios de admin. Para obter mais detalhes, consulte a [documentação oficial do OpenStack Keystone \(https://docs.openstack.org/keystone/latest/#setting-up-projects-users-and-roles\)](https://docs.openstack.org/keystone/latest/#setting-up-projects-users-and-roles)<sup>7</sup>. Veja a seguir as opções de configuração relacionadas:

#### rgw keystone admin user

Nome do usuário administrador do Keystone.

#### rgw keystone admin password

Senha do usuário administrador do Keystone.

#### rgw keystone admin tenant

Locatário do usuário administrador do Keystone versão 2.0.

Um usuário do Gateway de Objetos do Ceph é mapeado para um locatário do Keystone. Um usuário do Keystone tem funções diferentes atribuídas, possivelmente em mais do que um locatário. Quando o Gateway de Objetos do Ceph recebe o ticket, ele examina o locatário e as funções do usuário atribuídas a esse ticket e aceita ou rejeita a solicitação de acordo com a configuração da opção rgw keystone accepted roles.



## Dica: Mapeando para locatários do OpenStack

Embora os locatários do Swift sejam mapeados para o usuário do Gateway de Objetos por padrão, eles também podem ser mapeados para os locatários do OpenStack por meio da opção `rgw keystone implicit tenants`. Isso fará com que os containers usem o namespace do locatário em vez do namespace global do tipo do S3 que o Gateway de Objetos usa como padrão. É recomendável decidir sobre o método de mapeamento na fase de planejamento para evitar confusão. O motivo dessa recomendação é que alternar a opção posteriormente afeta apenas as solicitações mais recentes que são mapeadas em um locatário, enquanto os compartimentos de memória mais antigos criados antes ainda continuam em um namespace global.

Para obter a versão 3 da API do OpenStack Identity, você deve substituir a opção `rgw keystone admin tenant` por:

`rgw keystone admin domain`

Domínio do usuário administrador do Keystone.

`rgw keystone admin project`

Projeto do usuário administrador do Keystone.

## 21.12 Posicionamento do pool e classes de armazenamento

### 21.12.1 Exibindo destinos de posicionamento

Os destinos de posicionamento controlam os pools que serão associados a um determinado compartimento de memória. O destino de posicionamento de um compartimento de memória é selecionado na criação e não pode ser modificado. Você pode executar o comando a seguir para exibir a respectiva regra `placement_rule`:

```
cephuser@adm > radosgw-admin bucket stats
```

A configuração do grupo de zonas contém uma lista de destinos de posicionamento com um destino inicial chamado "default-placement". A configuração da zona mapeia cada nome de destino de posicionamento do grupo de zonas para o respectivo armazenamento local.

Essas informações de posicionamento de zona incluem o nome "index\_pool" para o índice de compartimento de memória, o nome "data\_extra\_pool" para os metadados sobre uploads de várias partes incompletos e um nome "data\_pool" para cada classe de armazenamento.

### 21.12.2 Classes de armazenamento

As classes de armazenamento ajudam a personalizar o posicionamento dos dados de objetos. As regras de Ciclo de Vida de Compartimento de Memória do S3 podem automatizar a transição dos objetos entre as classes de armazenamento.

As classes de armazenamento são definidas em termos de destinos de posicionamento. Cada destino de posicionamento do grupo de zonas lista suas classes de armazenamento disponíveis com uma classe inicial chamada "STANDARD". A configuração da zona é responsável por conceder um nome de pool "data\_pool" a cada uma das classes de armazenamento do grupo de zonas.

### 21.12.3 Configurando grupos de zonas e zonas

Use o comando **radosgw-admin** nos grupos de zonas e nas zonas para configurar o respectivo posicionamento. Você pode consultar a configuração de posicionamento do grupo de zonas usando o seguinte comando:

```
cephuser@adm > radosgw-admin zonegroup get
{
  "id": "ab01123f-e0df-4f29-9d71-b44888d67cd5",
  "name": "default",
  "api_name": "default",
  ...
  "placement_targets": [
    {
      "name": "default-placement",
      "tags": [],
      "storage_classes": [
        "STANDARD"
      ]
    }
  ],
  "default_placement": "default-placement",
  ...
}
```



Para consultar a configuração de posicionamento da zona, execute:

```
cephuser@adm > radosgw-admin zone get
{
  "id": "557cdcee-3aae-4e9e-85c7-2f86f5eddb1f",
  "name": "default",
  "domain_root": "default.rgw.meta:root",
  ...
  "placement_pools": [
    {
      "key": "default-placement",
      "val": {
        "index_pool": "default.rgw.buckets.index",
        "storage_classes": {
          "STANDARD": {
            "data_pool": "default.rgw.buckets.data"
          }
        },
        "data_extra_pool": "default.rgw.buckets.non-ec",
        "index_type": 0
      }
    }
  ],
  ...
}
```



### Nota: Sem configuração multissite anterior

Se você não fez nenhuma configuração de multissite anterior, uma zona e um grupo de zonas “padrão” são criados para você, e as mudanças feitas neles não entrarão em vigor até você reiniciar os Gateways de Objetos do Ceph. Se você criou um domínio Kerberos para multissite, as mudanças feitas na zona/grupo de zonas entrarão em vigor depois que você confirmá-las com o comando **`radosgw-admin period update --commit`**.

#### 21.12.3.1 Adicionando um destino de posicionamento

Para criar um novo destino de posicionamento chamado "temporary", comece adicionando-o ao grupo de zonas:

```
cephuser@adm > radosgw-admin zonegroup placement add \
  --rgw-zonegroup default \
  --placement-id temporary
```

Em seguida, insira as informações de posicionamento da zona para esse destino:

```
cephuser@adm > radosgw-admin zone placement add \  
  --rgw-zone default \  
  --placement-id temporary \  
  --data-pool default.rgw.temporary.data \  
  --index-pool default.rgw.temporary.index \  
  --data-extra-pool default.rgw.temporary.non-ec
```

### 21.12.3.2 Adicionando uma classe de armazenamento

Para adicionar uma nova classe de armazenamento chamada “COLD” ao destino de posicionamento padrão, comece adicionando-a ao grupo de zonas:

```
cephuser@adm > radosgw-admin zonegroup placement add \  
  --rgw-zonegroup default \  
  --placement-id default-placement \  
  --storage-class COLD
```

Em seguida, insira as informações de posicionamento da zona para essa classe de armazenamento:

```
cephuser@adm > radosgw-admin zone placement add \  
  --rgw-zone default \  
  --placement-id default-placement \  
  --storage-class COLD \  
  --data-pool default.rgw.cold.data \  
  --compression lz4
```

## 21.12.4 Personalização de posicionamento

### 21.12.4.1 Editando o posicionamento do grupo de zonas padrão

Por padrão, os novos compartimentos de memória usarão o destino default\_placement do grupo de zonas. Você pode mudar essa configuração do grupo de zonas com:

```
cephuser@adm > radosgw-admin zonegroup placement default \  
  --rgw-zonegroup default \  
  --placement-id new-placement
```

#### 21.12.4.2 Editando o posicionamento do usuário padrão

Um usuário do Gateway de Objetos do Ceph pode anular o destino de posicionamento padrão do grupo de zonas definindo um campo `default_placement` não vazio nas informações do usuário. Da mesma forma, a `default_storage_class` pode anular a classe de armazenamento `STANDARD` aplicada aos objetos por padrão.

```
cephuser@adm > radosgw-admin user info --uid testid
{
  ...
  "default_placement": "",
  "default_storage_class": "",
  "placement_tags": [],
  ...
}
```

Se o destino de posicionamento do grupo de zonas incluir tags, os usuários não poderão criar compartimentos de memória com esse destino de posicionamento, a menos que as informações de usuário deles contenham pelo menos uma tag correspondente no respectivo campo "placement\_tags". Isso pode ser útil para restringir o acesso a determinados tipos de armazenamento.

O comando `radosgw-admin` não pode modificar esses campos diretamente, portanto, você precisa editar o formato JSON manualmente:

```
cephuser@adm > radosgw-admin metadata get user:USER-ID > user.json
> vi user.json      # edit the file as required
cephuser@adm > radosgw-admin metadata put user:USER-ID < user.json
```

#### 21.12.4.3 Editando o posicionamento do compartimento de memória padrão do S3

Ao criar um compartimento de memória com o protocolo S3, é possível inserir um destino de posicionamento como parte de `LocationConstraint` para anular os destinos de posicionamento padrão do usuário e do grupo de zonas.

Normalmente, o `LocationConstraint` precisa corresponder ao `api_name` do grupo de zonas:

```
<LocationConstraint>default</LocationConstraint>
```

É possível adicionar um destino de posicionamento personalizado ao `api_name` após dois-pontos:

```
<LocationConstraint>default:new-placement</LocationConstraint>
```

#### 21.12.4.4 Editando o posicionamento do compartimento de memória do Swift

Ao criar um compartimento de memória com o protocolo Swift, você pode fornecer um destino de posicionamento em X-Storage-Policy do cabeçalho HTTP:

```
X-Storage-Policy: NEW-PLACEMENT
```

#### 21.12.5 Usando classes de armazenamento

Todos os destinos de posicionamento têm uma classe de armazenamento STANDARD, que é aplicada a novos objetos por padrão. Você pode anular esse padrão com default\_storage\_class.

Para criar um objeto em uma classe de armazenamento não padrão, insira o nome dessa classe de armazenamento em um cabeçalho HTTP com a solicitação. O protocolo S3 usa o cabeçalho X-Amz-Storage-Class, enquanto o protocolo Swift usa o cabeçalho X-Object-Storage-Class.

É possível usar o *Gerenciamento do Ciclo de Vida de Objeto do S3* para mover dados de objetos entre classes de armazenamento por meio das ações de Transição.

### 21.13 Gateways de Objetos multissite

O Ceph suporta várias opções de configuração multissite para o Gateway de Objetos do Ceph:

#### Várias zonas

Uma configuração que consiste em um grupo de zonas e várias zonas, cada uma com uma ou mais instâncias de ceph - radosgw. Cada zona é acompanhada de seu próprio Cluster de Armazenamento do Ceph. Várias zonas em um grupo de zonas fornecem recuperação de desastre para o grupo de zonas, caso uma das zonas apresente uma falha significativa. Cada zona é ativa e pode receber operações de gravação. Além da recuperação de desastre, várias zonas ativas também podem servir como base para redes de distribuição de conteúdo.

#### Vários grupos de zonas

O Gateway de Objetos do Ceph suporta vários grupos de zonas, cada um com uma ou mais zonas. Os objetos armazenados em zonas de um grupo de zonas no mesmo domínio que outro grupo de zonas compartilham um namespace de objeto global, garantindo IDs de objeto exclusivos em todos os grupos de zonas e as zonas.



## Nota

É importante observar que os grupos de zonas sincronizam *apenas* metadados entre eles. Os dados e os metadados são replicados entre as zonas do grupo de zonas. Não são compartilhados dados ou metadados em um domínio.

### Vários domínios

O Gateway de Objetos do Ceph suporta a noção de domínios: um namespace globalmente exclusivo. Vários domínios são suportados, o que pode abranger um ou diversos grupos de zonas.

Você pode configurar cada Gateway de Objetos para operar em uma configuração de zona ativa-ativa, permitindo gravações em zonas não master. A configuração multissite é armazenada em um container chamado domínio. O domínio armazena grupos de zonas, zonas e um período com várias épocas para monitorar as mudanças na configuração. Os daemons `rgw` processam a sincronização, eliminando a necessidade de um agente de sincronização separado. Essa abordagem de sincronização permite que o Gateway de Objetos do Ceph opere com uma configuração ativa-ativa, e não ativa-passiva.

### 21.13.1 Requisitos e considerações

Uma configuração multissite requer pelo menos dois clusters de armazenamento do Ceph e, no mínimo, duas instâncias do Gateway de Objetos do Ceph, uma para cada cluster de armazenamento do Ceph. A configuração a seguir considera que pelo menos dois clusters de armazenamento do Ceph estão em locais geograficamente separados. No entanto, a configuração pode funcionar no mesmo site. Por exemplo, `rgw1` e `rgw2` nomeados.

Uma configuração multissite requer um grupo de zonas master e uma zona master. Uma zona master é a fonte da verdade em relação a todas as operações de metadados em um cluster multissite. Além disso, cada grupo de zonas requer uma zona master. Os grupos de zonas podem ter uma ou mais zonas secundárias ou não master. Neste guia, o host `rgw1` atua como a zona master do grupo de zonas master, e o host `rgw2` atua como a zona secundária do grupo de zonas master.

## 21.13.2 Configurando uma zona master

Todos os gateways em uma configuração multissite recuperam sua configuração de um daemon `ceph-radosgw` em um host no grupo de zonas master e na zona master. Para configurar seus gateways em uma configuração multissite, selecione uma instância de `ceph-radosgw` para configurar o grupo de zonas master e a zona master.

### 21.13.2.1 Criando um domínio

Um domínio representa um namespace globalmente exclusivo que consiste em um ou mais grupos de zonas com uma ou mais zonas. As zonas contêm compartimentos de memória que, por sua vez, contêm objetos. Um domínio permite que o Gateway de Objetos do Ceph suporte vários namespaces e a respectiva configuração no mesmo hardware. Um domínio engloba a noção de períodos. Cada período representa o estado da configuração do grupo de zonas e da zona no tempo. Sempre que você modificar um grupo de zonas ou uma zona, atualize e confirme o período. Por padrão, o Gateway de Objetos do Ceph não cria um domínio para compatibilidade retroativa. Como melhor prática, recomendamos criar domínios para os novos clusters.

Crie um novo domínio chamado `gold` para a configuração multissite abrindo uma interface de linha de comando em um host identificado para atuar no grupo de zonas master e na zona. Em seguida, execute o seguinte:

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

Se o cluster tiver um único domínio, especifique o flag `--default`. Se `--default` for especificado, `radosgw-admin` usará esse domínio por padrão. Se `--default` não for especificado, a adição de grupos de zonas e zonas exigirá que o flag `--rgw-realm` ou `--realm-id` seja especificado para identificar o domínio ao adicionar grupos de zonas e zonas.

Após criar o domínio, `radosgw-admin` retornará a configuração do domínio:

```
{
  "id": "4a367026-bd8f-40ee-b486-8212482ddcd7",
  "name": "gold",
  "current_period": "09559832-67a4-4101-8b3f-10dfcd6b2707",
  "epoch": 1
}
```



## Nota

O Ceph gera um ID exclusivo para o domínio, o que permite renomear um domínio se houver necessidade.

### 21.13.2.2 Criando um grupo de zonas master

Um domínio deve ter pelo menos um grupo de zonas para atuar como o grupo de zonas master do domínio. Crie um novo grupo de zonas master para a configuração multissite abrindo uma interface de linha de comando em um host identificado para atuar no grupo de zonas master e na zona. Execute o seguinte comando para criar um grupo de zonas master chamado us:

```
cephuser@adm > radosgw-admin zonegroup create --rgw-zonegroup=us \
--endpoints=http://rgw1:80 --master --default
```

Se o domínio tiver apenas um grupo de zonas, especifique o flag --default. Se --default for especificado, **radosgw-admin** usará esse grupo de zonas por padrão ao adicionar novas zonas. Se --default não for especificado, a adição de zonas exigirá o flag --rgw-zonegroup ou --zonegroup-id para identificar o grupo de zonas ao adicionar ou modificar zonas.

Após criar o grupo de zonas master, **radosgw-admin** retornará a configuração do grupo de zonas. Por exemplo:

```
{
  "id": "d4018b8d-8c0d-4072-8919-608726fa369e",
  "name": "us",
  "api_name": "us",
  "is_master": "true",
  "endpoints": [
    "http://rgw1:80"
  ],
  "hostnames": [],
  "hostnames_s3website": [],
  "master_zone": "",
  "zones": [],
  "placement_targets": [],
  "default_placement": "",
  "realm_id": "4a367026-bd8f-40ee-b486-8212482ddcd7"
}
```

### 21.13.2.3 Criando uma zona master



#### Importante

As zonas precisam ser criadas em um nó do Gateway de Objetos do Ceph que estará na zona.

Crie uma nova zona master para a configuração multissite abrindo uma interface de linha de comando em um host identificado para atuar no grupo de zonas master e na zona. Execute o seguinte:

```
cephuser@adm > radosgw-admin zone create --rgw-zonegroup=us --rgw-zone=us-east-1 \
--endpoints=http://rgw1:80 --access-key=SYSTEM_ACCESS_KEY --secret=SYSTEM_SECRET_KEY
```



#### Nota

As opções `--access-key` e `--secret` não estão especificadas no exemplo acima. Essas configurações são adicionadas à zona quando o usuário é criado na próxima seção.

Após criar a zona master, `radosgw-admin` retornará a configuração da zona. Por exemplo:

```
{
  "id": "56dfabbb-2f4e-4223-925e-de3c72de3866",
  "name": "us-east-1",
  "domain_root": "us-east-1.rgw.meta:root",
  "control_pool": "us-east-1.rgw.control",
  "gc_pool": "us-east-1.rgw.log:gc",
  "lc_pool": "us-east-1.rgw.log:lc",
  "log_pool": "us-east-1.rgw.log",
  "intent_log_pool": "us-east-1.rgw.log:intent",
  "usage_log_pool": "us-east-1.rgw.log:usage",
  "reshard_pool": "us-east-1.rgw.log:reshard",
  "user_keys_pool": "us-east-1.rgw.meta:users.keys",
  "user_email_pool": "us-east-1.rgw.meta:users.email",
  "user_swift_pool": "us-east-1.rgw.meta:users.swift",
  "user_uid_pool": "us-east-1.rgw.meta:users.uid",
  "otp_pool": "us-east-1.rgw.otp",
  "system_key": {
    "access_key": "1555b35654ad1656d804",
    "secret_key": "h7GhxuBLTrlhVUyxSPUKUV8r/2EI4ngqJxD7iBdBYLhwluN30JaT3Q=="
  },
  "placement_pools": [
    {
```



```

        "key": "us-east-1-placement",
        "val": {
            "index_pool": "us-east-1.rgw.buckets.index",
            "storage_classes": {
                "STANDARD": {
                    "data_pool": "us-east-1.rgw.buckets.data"
                }
            },
            "data_extra_pool": "us-east-1.rgw.buckets.non-ec",
            "index_type": 0
        }
    },
    "metadata_heap": "",
    "realm_id": ""
}

```

#### 21.13.2.4 Apagando a zona e o grupo padrão

##### ! Importante

As etapas a seguir consideram uma configuração multissite que usa sistemas recém-instalados que ainda não estão armazenando dados. **Não apague** a zona padrão e seus pools se você já a estiver usando para armazenar dados; do contrário, os dados serão apagados de modo irreversível.

A instalação padrão do Gateway de Objetos cria o grupo de zonas padrão chamado default. Apague a zona padrão, se ela existir. Remova-a primeiro do grupo de zonas padrão.

```
cephuser@adm > radosgw-admin zonegroup delete --rgw-zonegroup=default
```

Apague os pools padrão do cluster de armazenamento do Ceph, se existirem:

##### ! Importante

A etapa a seguir considera uma configuração multissite que usa sistemas recém-instalados que não estão armazenando dados no momento. **Não apague** o grupo de zonas padrão se você já o estiver usando para armazenar dados.

```
cephuser@adm > ceph osd pool rm default.rgw.control default.rgw.control --yes-i-really-really-mean-it
```

```
cephuser@adm > ceph osd pool rm default.rgw.data.root default.rgw.data.root --yes-i-really-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.gc default.rgw.gc --yes-i-really-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.log default.rgw.log --yes-i-really-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.meta default.rgw.meta --yes-i-really-really-mean-it
```



## Atenção

Se você apagar o grupo de zonas padrão, também apagará o usuário do sistema. Se as chaves de usuário admin não forem propagadas, haverá falha na funcionalidade de gerenciamento do Gateway de Objetos do Ceph Dashboard. Avance para a próxima seção para recriar o usuário do sistema, se você prosseguir com esta etapa.

### 21.13.2.5 Criando usuários do sistema

Os daemons `ceph-radosgw` devem ser autenticados antes de extrair informações de domínio e período. Na zona master, crie um usuário do sistema para simplificar a autenticação entre daemons:

```
cephuser@adm > radosgw-admin user create --uid=zone.user \
--display-name="Zone User" --access-key=SYSTEM_ACCESS_KEY \
--secret=SYSTEM_SECRET_KEY --system
```

Anote a `access_key` e a `secret_key`, pois as zonas secundárias precisam delas para autenticação na zona master.

Adicione o usuário do sistema à zona master:

```
cephuser@adm > radosgw-admin zone modify --rgw-zone=us-east-1 \
--access-key=ACCESS-KEY --secret=SECRET
```

Atualize o período para que as mudanças entrem em vigor:

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

### 21.13.2.6 Atualize o período

Após atualizar a configuração da zona master, atualize o período:

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

Após atualizar o período, **radosgw-admin** retornará a configuração do período. Por exemplo:

```
{
  "id": "09559832-67a4-4101-8b3f-10dfcd6b2707", "epoch": 1, "predecessor_uuid": "",
  "sync_status": [], "period_map":
  {
    "id": "09559832-67a4-4101-8b3f-10dfcd6b2707", "zonegroups": [], "short_zone_ids": []
  }, "master_zonegroup": "", "master_zone": "", "period_config":
  {
    "bucket_quota": {
      "enabled": false, "max_size_kb": -1, "max_objects": -1
    }, "user_quota": {
      "enabled": false, "max_size_kb": -1, "max_objects": -1
    }
  }, "realm_id": "4a367026-bd8f-40ee-b486-8212482ddcd7", "realm_name": "gold",
  "realm_epoch": 1
}
```



#### Nota

A atualização do período muda a época e garante que as outras zonas recebam a configuração atualizada.

### 21.13.2.7 Iniciar o gateway

No host do Gateway de Objetos, inicie e habilite o serviço Gateway de Objetos do Ceph. Para identificar o FSID exclusivo do cluster, execute **ceph fsid**. Para identificar o nome do daemon do Gateway de Objetos, execute **ceph orch ps --hostname HOSTNAME**.

```
cephuser@ogw > systemctl start ceph-FSID@DAEMON_NAME
cephuser@ogw > systemctl enable ceph-FSID@DAEMON_NAME
```

### 21.13.3 Configurar zonas secundárias

As zonas dentro de um grupo de zonas replicam todos os dados para garantir que cada zona tenha os mesmos dados. Ao criar a zona secundária, execute todas as operações a seguir em um host identificado para processar a zona secundária.



#### Nota

Para adicionar uma terceira zona, siga os mesmos procedimentos da adição da zona secundária. Use um nome de zona diferente.



#### Importante

Você deve executar operações de metadados, como criação de usuário, em um host na zona master. A zona master e a zona secundária podem receber operações de compartimento de memória, mas a zona secundária redireciona essas operações para a zona master. Se a zona master estiver inativa, haverá falha nas operações de compartimento de memória.

#### 21.13.3.1 Extraindo do domínio

Usando o caminho de URL, a chave de acesso e o segredo da zona master no grupo de zonas master, extraia a configuração do domínio para o host. Para extrair de um domínio não padrão, especifique o domínio usando as opções de configuração `--rgw-realm` ou `--realm-id`.

```
cephuser@adm > radosgw-admin realm pull --url=url-to-master-zone-gateway --access-key=access-key --secret=secret
```



#### Nota

A extração do domínio também recupera a configuração do período atual remoto e também o torna o período atual neste host.

Se esse for o único domínio ou o padrão, defina o domínio como padrão.

```
cephuser@adm > radosgw-admin realm default --rgw-realm=REALM-NAME
```

### 21.13.3.2 Criando uma zona secundária

Crie uma zona secundária para a configuração multissite abrindo uma interface de linha de comando em um host identificado para atender à zona secundária. Especifique o ID do grupo de zonas, o novo nome da zona e um endpoint para a zona. *Não* use o flag `--master`. Por padrão, todas as zonas são executadas em uma configuração ativa-ativa. Se a zona secundária não aceitar operações de gravação, especifique o flag `--read-only` para criar uma configuração ativa-passiva entre a zona master e a zona secundária. Além disso, insira a `access_key` e a `secret_key` do usuário do sistema gerado armazenado na zona master do grupo de zonas master. Execute o seguinte:

```
cephuser@adm > radosgw-admin zone create --rgw-zonegroup=ZONE-GROUP-NAME\
--rgw-zone=ZONE-NAME --endpoints=URL \
--access-key=SYSTEM-KEY --secret=SECRET\
--endpoints=http://FQDN:80 \
[--read-only]
```

Por exemplo:

```
cephuser@adm > radosgw-admin zone create --rgw-zonegroup=us --endpoints=http://rgw2:80 \
--rgw-zone=us-east-2 --access-key=SYSTEM_ACCESS_KEY --secret=SYSTEM_SECRET_KEY
{
  "id": "950c1a43-6836-41a2-a161-64777e07e8b8",
  "name": "us-east-2",
  "domain_root": "us-east-2.rgw.data.root",
  "control_pool": "us-east-2.rgw.control",
  "gc_pool": "us-east-2.rgw.gc",
  "log_pool": "us-east-2.rgw.log",
  "intent_log_pool": "us-east-2.rgw.intent-log",
  "usage_log_pool": "us-east-2.rgw.usage",
  "user_keys_pool": "us-east-2.rgw.users.keys",
  "user_email_pool": "us-east-2.rgw.users.email",
  "user_swift_pool": "us-east-2.rgw.users.swift",
  "user_uid_pool": "us-east-2.rgw.users.uid",
  "system_key": {
    "access_key": "1555b35654ad1656d804",
    "secret_key": "h7GhxuBLTrlhVUyxSPUKUV8r\2EI4ngqJxD7iBdBYLhwluN30JaT3Q=="
  },
  "placement_pools": [
    {
      "key": "default-placement",
      "val": {
        "index_pool": "us-east-2.rgw.buckets.index",
        "data_pool": "us-east-2.rgw.buckets.data",
        "data_extra_pool": "us-east-2.rgw.buckets.non-ec",

```

```

        "index_type": 0
    }
}
],
"metadata_heap": "us-east-2.rgw.meta",
"realm_id": "815d74c2-80d6-4e63-8cfc-232037f7ff5c"
}

```

## Importante

As etapas a seguir consideram uma configuração multissite que usa sistemas recém-instalados que ainda não estão armazenando dados. **Não apague** a zona padrão e seus pools se você já a estiver usando para armazenar dados; do contrário, os dados serão perdidos de modo irrecuperável.

Apague a zona padrão, se necessário:

```
cephuser@adm > radosgw-admin zone delete --rgw-zone=default
```

Apague os pools padrão do cluster de armazenamento do Ceph, se necessário:

```

cephuser@adm > ceph osd pool rm default.rgw.control default.rgw.control --yes-i-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.data.root default.rgw.data.root --yes-i-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.gc default.rgw.gc --yes-i-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.log default.rgw.log --yes-i-really-mean-it
cephuser@adm > ceph osd pool rm default.rgw.users.uid default.rgw.users.uid --yes-i-really-mean-it

```

### 21.13.3.3 Atualizando o arquivo de configuração do Ceph

Atualize o arquivo de configuração do Ceph nos hosts da zona secundária adicionando a opção de configuração `rgw_zone` e o nome da zona secundária à entrada da instância.

Para isso, execute o seguinte comando:

```
cephuser@adm > ceph config set SERVICE_NAME rgw_zone us-west
```

### 21.13.3.4 Atualizando o período

Após atualizar a configuração da zona master, atualize o período:

```
cephuser@adm > radosgw-admin period update --commit
{
  "id": "b5e4d3ec-2a62-4746-b479-4b2bc14b27d1",
  "epoch": 2,
  "predecessor_uuid": "09559832-67a4-4101-8b3f-10dfcd6b2707",
  "sync_status": [ "[...]"
],
  "period_map": {
    "id": "b5e4d3ec-2a62-4746-b479-4b2bc14b27d1",
    "zonegroups": [
      {
        "id": "d4018b8d-8c0d-4072-8919-608726fa369e",
        "name": "us",
        "api_name": "us",
        "is_master": "true",
        "endpoints": [
          "http://\rgw1:80"
        ],
        "hostnames": [],
        "hostnames_s3website": [],
        "master_zone": "83859a9a-9901-4f00-aa6d-285c777e10f0",
        "zones": [
          {
            "id": "83859a9a-9901-4f00-aa6d-285c777e10f0",
            "name": "us-east-1",
            "endpoints": [
              "http://\rgw1:80"
            ],
            "log_meta": "true",
            "log_data": "false",
            "bucket_index_max_shards": 0,
            "read_only": "false"
          },
          {
            "id": "950c1a43-6836-41a2-a161-64777e07e8b8",
            "name": "us-east-2",
            "endpoints": [
              "http://\rgw2:80"
            ],
            "log_meta": "false",
            "log_data": "true",
            "bucket_index_max_shards": 0,
            "read_only": "false"
          }
        ]
      }
    ]
  }
}
```

```

    }

    ],
    "placement_targets": [
        {
            "name": "default-placement",
            "tags": []
        }
    ],
    "default_placement": "default-placement",
    "realm_id": "4a367026-bd8f-40ee-b486-8212482ddcd7"
}
],
"short_zone_ids": [
    {
        "key": "83859a9a-9901-4f00-aa6d-285c777e10f0",
        "val": 630926044
    },
    {
        "key": "950c1a43-6836-41a2-a161-64777e07e8b8",
        "val": 4276257543
    }
]
},
"master_zonegroup": "d4018b8d-8c0d-4072-8919-608726fa369e",
"master_zone": "83859a9a-9901-4f00-aa6d-285c777e10f0",
"period_config": {
    "bucket_quota": {
        "enabled": false,
        "max_size_kb": -1,
        "max_objects": -1
    },
    "user_quota": {
        "enabled": false,
        "max_size_kb": -1,
        "max_objects": -1
    }
},
"realm_id": "4a367026-bd8f-40ee-b486-8212482ddcd7",
"realm_name": "gold",
"realm_epoch": 2
}

```





## Nota

A atualização do período muda a época e garante que as outras zonas recebam a configuração atualizada.

### 21.13.3.5 Iniciando o Gateway de Objetos

No host do Gateway de Objetos, inicie e habilite o serviço Gateway de Objetos do Ceph:

```
cephuser@adm > ceph orch start rgw.us-east-2
```

### 21.13.3.6 Verificando o status da sincronização

Quando a zona secundária estiver ativa e em execução, verifique o status da sincronização. A sincronização copia os usuários e compartimentos de memória criados na zona master para a zona secundária.

```
cephuser@adm > radosgw-admin sync status
```

A saída mostra o status das operações de sincronização. Por exemplo:

```
realm f3239bc5-e1a8-4206-a81d-e1576480804d (gold)
  zonegroup c50dbb7e-d9ce-47cc-a8bb-97d9b399d388 (us)
    zone 4c453b70-4a16-4ce8-8185-1893b05d346e (us-west)
metadata sync syncing
  full sync: 0/64 shards
  metadata is caught up with master
  incremental sync: 64/64 shards
data sync source: lee9da3e-114d-4ae3-a8a4-056e8a17f532 (us-east)
  syncing
  full sync: 0/128 shards
  incremental sync: 128/128 shards
  data is caught up with source
```



## Nota

As zonas secundárias aceitam operações de compartimento de memória, mas elas redirecionam essas operações para a zona master e, em seguida, são sincronizadas com a zona master para receber o resultado das operações. Se a zona master estiver inativa, haverá falha nas operações de compartimento de memória executadas na zona secundária, mas as operações de objeto deverão ser bem-sucedidas.

### 21.13.3.7 Verificação de um objeto

Por padrão, os objetos não são verificados novamente depois que a sincronização de um objeto é bem-sucedida. Para habilitar a verificação, defina a opção `rgw_sync_obj_etag_verify` como `true`. Após a habilitação, os objetos opcionais serão sincronizados. Um checksum MD5 adicional verificará se eles foram calculados na origem e no destino. Isso é feito para garantir a integridade dos objetos buscados de um servidor remoto por HTTP, incluindo a sincronização multissite. Essa opção pode diminuir o desempenho dos RGWs, já que mais computação é necessária.

## 21.13.4 Manutenção geral do Gateway de Objetos

### 21.13.4.1 Verificando o status da sincronização

As informações sobre o status da replicação de uma zona podem ser consultadas com:

```
cephuser@adm > radosgw-admin sync status
  realm b3bc1c37-9c44-4b89-a03b-04c269bea5da (gold)
  zonegroup f54f9b22-b4b6-4a0e-9211-fa6ac1693f49 (us)
    zone adce11c9-b8ed-4a90-8bc5-3fc029ff0816 (us-west)
      metadata sync syncing
        full sync: 0/64 shards
        incremental sync: 64/64 shards
        metadata is behind on 1 shards
        oldest incremental change not applied: 2017-03-22 10:20:00.0.881361s
      data sync source: 341c2d81-4574-4d08-ab0f-5a2a7b168028 (us-east)
        syncing
          full sync: 0/128 shards
          incremental sync: 128/128 shards
          data is caught up with source
        source: 3b5d1a3f-3f27-4e4a-8f34-6072d4bb1275 (us-3)
```

```
syncing
full sync: 0/128 shards
incremental sync: 128/128 shards
data is caught up with source
```

A saída pode ser diferente dependendo do status da sincronização. Os fragmentos são descritos como dois tipos diferentes durante a sincronização:

#### Fragmentos esquecidos

Trata-se de fragmentos que precisam de uma sincronização completa de dados e de fragmentos que precisam de uma sincronização incremental de dados porque não estão atualizados.

#### Fragmentos de recuperação

Trata-se de fragmentos que encontraram um erro durante a sincronização e foram marcados para nova tentativa. Na maioria das vezes, o erro ocorre em caso de problemas secundários, como aquisição de um bloqueio em um compartimento de memória. Normalmente, esse erro se resolve sozinho.

### 21.13.4.2 Verificar os registros

Apenas para multissite, você pode verificar o registro de metadados (`mdlog`), o registro de índice do compartimento de memória (`bilog`) e o registro de dados (`datalog`). Você pode listá-los e também cortá-los. Na maioria dos casos, isso não é necessário porque a opção `rgw_sync_log_trim_interval` está definida como 20 minutos por padrão. Se isso não for definido manualmente como 0, você não precisará cortar a qualquer momento, pois poderá causar efeitos colaterais.

### 21.13.4.3 Mudando a zona master de metadados



#### Importante

Tenha cuidado ao mudar a zona que é master de metadados. Se uma zona não concluiu a sincronização de metadados da zona master atual, ela não pode processar as entradas restantes ao ser promovida a master, e essas mudanças são perdidas. Por esse motivo, recomendamos aguardar até o status da sincronização de `radosgw-admin` de uma zona concluir a sincronização de metadados antes de promovê-la a master. Da mesma forma, se as mudanças nos metadados estiverem sendo processadas pela zona master atual

enquanto outra zona for promovida a master, essas mudanças provavelmente serão perdidas. Para evitar isso, recomendamos encerrar quaisquer instâncias do Gateway de Objetos na zona master anterior. Após promover outra zona, será possível buscar seu novo período com a extração de período de `radosgw-admin` e reiniciar o(s) gateway(s).

Para promover uma zona (por exemplo, a zona `us-west` no grupo de zonas `us`) a master de metadados, execute os seguintes comandos nessa zona:

```
cephuser@ogw > radosgw-admin zone modify --rgw-zone=us-west --master
cephuser@ogw > radosgw-admin zonegroup modify --rgw-zonegroup=us --master
cephuser@ogw > radosgw-admin period update --commit
```

Isso gera um novo período, e a(s) instância(s) do Gateway de Objetos na zona `us-west` envia(m) esse período para as outras zonas.

### 21.13.5 Executando failover e recuperação de desastre

Se a zona master falhar, faça o failover para a zona secundária para recuperação de desastre.

1. Converta a zona secundária na zona master e padrão. Por exemplo:

```
cephuser@adm > radosgw-admin zone modify --rgw-zone=ZONE-NAME --master --default
```

Por padrão, o Gateway de Objetos do Ceph é executado em uma configuração ativa-ativa. Se o cluster foi configurado para ser executado em uma configuração ativa-passiva, a zona secundária é uma zona apenas leitura. Remova o status `--read-only` para permitir que a zona receba as operações de gravação. Por exemplo:

```
cephuser@adm > radosgw-admin zone modify --rgw-zone=ZONE-NAME --master --default \
--read-only=false
```

2. Atualize o período para que as mudanças entrem em vigor:

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

3. Reinicie o Gateway de Objetos do Ceph:

```
cephuser@adm > ceph orch restart rgw
```

Se a zona master anterior for recuperada, reverta a operação.

1. Da zona recuperada, extraia a configuração mais recente do domínio da zona master atual.

```
cephuser@adm > radosgw-admin realm pull --url=URL-TO-MASTER-ZONE-GATEWAY \
--access-key=ACCESS-KEY --secret=SECRET
```

2. Converta a zona recuperada na zona master e padrão:

```
cephuser@adm > radosgw-admin zone modify --rgw-zone=ZONE-NAME --master --default
```

3. Atualize o período para que as mudanças entrem em vigor:

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

4. Reinicie o Gateway de Objetos do Ceph na zona recuperada:

```
cephuser@adm > ceph orch restart rgw@rgw
```

5. Se a zona secundária precisar de uma configuração apenas leitura, atualize-a:

```
cephuser@adm > radosgw-admin zone modify --rgw-zone=ZONE-NAME --read-only
```

6. Atualize o período para que as mudanças entrem em vigor:

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

7. Reinicie o Gateway de Objetos do Ceph na zona secundária:

```
cephuser@adm > ceph orch restart@rgw
```

## 22 Ceph iSCSI Gateway

O capítulo aborda especificamente as tarefas de administração relacionadas ao iSCSI Gateway. Para ver o procedimento de implantação, consulte o *Livro “Guia de Implantação”, Capítulo 8 “Implantando os serviços principais restantes com o cephadm”, Seção 8.3.5 “Implantando Gateways iSCSI”*.

### 22.1 Destinos gerenciados pelo ceph-iscsi

Este capítulo descreve como se conectar a destinos gerenciados por `ceph-iscsi` de clientes com Linux, Microsoft Windows ou VMware.

#### 22.1.1 Conectando-se ao open-iscsi

A conexão com destinos iSCSI baseados em `ceph-iscsi` por meio do `open-iscsi` é um processo de duas etapas. Primeiramente, o iniciador deve descobrir os destinos iSCSI disponíveis no host do gateway, depois ele deve efetuar login e mapear as LUs (Logical Units – Unidades Lógicas) disponíveis.

As duas etapas exigem que o daemon `open-iscsi` esteja em execução. A maneira como você inicia o daemon `open-iscsi` depende da sua distribuição Linux:

- No SUSE Linux Enterprise Server (SLES) e nos hosts Red Hat Enterprise Linux (RHEL), execute `systemctl start iscsid` (ou `service iscsid start` se o `systemctl` não estiver disponível).
- Nos hosts do Debian e do Ubuntu, execute `systemctl start open-iscsi` (ou `service open-iscsi start`).

Se o host do seu iniciador executa o SUSE Linux Enterprise Server, consulte <https://documentation.suse.com/sles/15-SP1/single-html/SLES-storage/#sec-iscsi-initiator> para obter detalhes sobre como se conectar a um destino iSCSI.

Para qualquer outra distribuição Linux com suporte a `open-iscsi`, prossiga para a descoberta de destinos no gateway `ceph-iscsi`. Este exemplo usa `iscsi1.example.com` como endereço do portal. Para acesso de múltiplos caminhos, repita estas etapas com `iscsi2.example.com`:

```
# iscsiadm -m discovery -t sendtargets -p iscsi1.example.com
```

```
192.168.124.104:3260,1 iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol
```

Em seguida, efetue login no portal. Se o login for concluído com êxito, quaisquer unidades lógicas baseadas em RBD no portal ficarão imediatamente disponíveis no barramento SCSI do sistema:

```
# iscsiadm -m node -p iscsil.example.com --login
Logging in to [iface: default, target: iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol, portal: 192.168.124.104,3260] (multiple)
Login to [iface: default, target: iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol, portal: 192.168.124.104,3260] successful.
```

Repita esse processo para outros endereços IP ou hosts do portal.

Se o utilitário `lsscsi` estiver instalado no seu sistema, use-o para enumerar os dispositivos SCSI disponíveis no sistema:

```
lsscsi
[8:0:0:0]    disk      SUSE      RBD              4.0    /dev/sde
[9:0:0:0]    disk      SUSE      RBD              4.0    /dev/sdf
```

Em uma configuração de múltiplos caminhos (em que dois dispositivos iSCSI conectados representam a mesma LU), você também pode examinar o estado do dispositivo de múltiplos caminhos com o utilitário `multipath`:

```
# multipath -ll
360014050cf9dcfcb2603933ac3298dca dm-9 SUSE,RBD
size=49G features='0' hwhandler='0' wp=rw
|-+- policy='service-time 0' prio=1 status=active
|  `- 8:0:0:0 sde 8:64 active ready running
`-+- policy='service-time 0' prio=1 status=enabled
   `- 9:0:0:0 sdf 8:80 active ready running
```

Agora, você pode usar esse dispositivo de múltiplos caminhos como qualquer dispositivo de blocos. Por exemplo, você pode usá-lo como um Volume Físico para LVM (Logical Volume Management – Gerenciamento de Volumes Lógicos) Linux ou pode simplesmente criar um sistema de arquivos nele. O exemplo a seguir demonstra como criar um sistema de arquivos XFS no volume iSCSI de múltiplos caminhos recém-conectado:

```
# mkfs -t xfs /dev/mapper/360014050cf9dcfcb2603933ac3298dca
log stripe unit (4194304 bytes) is too large (maximum is 256KiB)
log stripe unit adjusted to 32KiB
meta-data=/dev/mapper/360014050cf9dcfcb2603933ac3298dca isize=256    agcount=17,
    agsize=799744 blks
    =                               sectsz=512   attr=2, projid32bit=1
```

	=	crc=0	finobt=0
data	=	bsize=4096	blocks=12800000, imaxpct=25
	=	sunit=1024	swidth=1024 blks
naming	=version 2	bsize=4096	ascii-ci=0 ftype=0
log	=internal log	bsize=4096	blocks=6256, version=2
	=	sectsz=512	sunit=8 blks, lazy-count=1
realtime	=none	extsz=4096	blocks=0, rtextents=0

Como o XFS é um sistema de arquivos sem cluster, você apenas pode montá-lo em um único nó do iniciador iSCSI em determinado momento.

Para descontinuar a qualquer momento o uso das LUs iSCSI associadas a determinado destino, execute o seguinte comando:

```
# iscsiadm -m node -p iscsil.example.com --logout
Logging out of session [sid: 18, iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol,
portal: 192.168.124.104,3260]
Logout of [sid: 18, target: iqn.2003-01.org.linux-iscsi.iscsi.SYSTEM-ARCH:testvol,
portal: 192.168.124.104,3260] successful.
```

Como ocorre com a descoberta e o login, você deve repetir as etapas de logout para todos os nomes de host ou endereços IP do portal.

### 22.1.1.1 Configurando múltiplos caminhos

A configuração de múltiplos caminhos é mantida nos clientes ou iniciadores e não depende de nenhuma configuração de `ceph-iscsi`. Selecione uma estratégia antes de usar o armazenamento em blocos. Após editar o `/etc/multipath.conf`, reinicie o `multipathd` com

```
# systemctl restart multipathd
```

Para uma configuração ativa-passiva com nomes amigáveis, adicione

```
defaults {
    user_friendly_names yes
}
```

ao `/etc/multipath.conf`. Após a conexão bem-sucedida com os destinos, execute

```
# multipath -ll
mpathd (36001405dbb561b2b5e439f0aed2f8e1e) dm-0 SUSE,RBD
size=2.0G features='0' hwhandler='0' wp=rw
|+- policy='service-time 0' prio=1 status=active
```



```
| ` - 2:0:0:3 sdl 8:176 active ready running
| +- policy='service-time 0' prio=1 status=enabled
| ` - 3:0:0:3 sdj 8:144 active ready running
` +- policy='service-time 0' prio=1 status=enabled
  ` - 4:0:0:3 sdk 8:160 active ready running
```

Observe o status de cada link. Para uma configuração ativa-ativa, adicione

```
defaults {
    user_friendly_names yes
}

devices {
    device {
        vendor "(LIO-ORG|SUSE)"
        product "RBD"
        path_grouping_policy "multibus"
        path_checker "tur"
        features "0"
        hardware_handler "1 alua"
        prio "alua"
        failback "immediate"
        rr_weight "uniform"
        no_path_retry 12
        rr_min_io 100
    }
}
```

ao /etc/multipath.conf. Reinicie o multipathd e execute

```
# multipath -ll
mpathd (36001405dbb561b2b5e439f0aed2f8e1e) dm-3 SUSE,RBD
size=2.0G features='1 queue_if_no_path' hwhandler='1 alua' wp=rw
` +- policy='service-time 0' prio=50 status=active
  | - 4:0:0:3 sdj 8:144 active ready running
  | - 3:0:0:3 sdk 8:160 active ready running
  ` - 2:0:0:3 sdl 8:176 active ready running
```

## 22.1.2 Conectando-se ao Microsoft Windows (iniciador iSCSI para Microsoft)

Para se conectar a um destino iSCSI do SUSE Enterprise Storage de um servidor Windows 2012, siga estas etapas:

1. Abra o Gerenciador do Servidor Windows. No Painel de Controle, selecione *Ferramentas > Iniciador iSCSI*. A caixa de diálogo *Propriedades do Iniciador iSCSI* é exibida. Selecione a guia *Descoberta*:

The screenshot shows the 'Discovery' tab of the 'iSCSI Initiator Properties' dialog box. At the top, there are tabs: 'Destinos', 'Descoberta' (selected), 'Destinos Favoritos', 'Volumes e Dispositivos', 'RADIUS', and 'Configuração'. The 'Portais de destino' section contains a text box with the instruction 'O sistema procurará os Destinos nos seguintes portais:' and an 'Atualizar' button. Below this is a table with four columns: 'Endereço', 'Porta', 'Adaptador', and 'Endereço IP'. The table is currently empty. Below the table, there is a text box with the instruction 'Para adicionar um portal de destino, clique em Descobrir Portal.' and a 'Descobrir Portal...' button. Another text box below that says 'Para remover um portal de destino, selecione o endereço acima e clique em Remover.' with a 'Remover' button. The 'Servidores iSNS' section has a text box with 'O sistema está registrado nos seguintes servidores iSNS:' and an 'Atualizar' button. Below this is a table with one column: 'Nome'. The table is empty. Below the table, there is a text box with the instruction 'Para adicionar um servidor iSNS, clique em Adicionar Servidor.' and an 'Adicionar Servidor...' button. Another text box below that says 'Para remover um servidor iSNS, selecione o servidor acima e clique em Remover.' with a 'Remover' button. At the bottom of the dialog box are three buttons: 'OK', 'Cancelar', and 'Aplicar'.

FIGURA 22.1: PROPRIEDADES DO INICIADOR iSCSI

2. Na caixa de diálogo *Descobrir Portal de Destino*, insira o nome de host ou endereço IP do destino no campo *Destino* e clique em *OK*:

Insira o endereço IP ou nome DNS e o número da porta do portal que deseja adicionar.

Para mudar as configurações padrão da descoberta do portal de destino, clique no botão Avançado.

Endereço IP ou nome DNS:	Porta: (Padrão 3260.)
<input type="text" value="192.168.124.104"/>	<input type="text" value="3260"/>

FIGURA 22.2: PORTAL DE DESCOBERTA DE DESTINO

3. Repita esse processo para todos os outros nomes de host ou endereços IP do gateway. Ao concluir, revise a lista *Portais de destino*:

Destinos | Descoberta | Destinos Favoritos | Volumes e Dispositivos | RADIUS | Configuração

**Portais de destino**

O sistema procurará os Destinos nos seguintes portais: Atualizar

Endereço	Porta	Adaptador	Endereço IP
192.168.124.104	3260	Padrão	Padrão
192.168.124.105	3260	Padrão	Padrão

Para adicionar um portal de destino, clique em Descobrir Portal. Descobrir Portal...

Para remover um portal de destino, selecione o endereço acima e clique em Remover. Remover

**Servidores iSNS**

O sistema está registrado nos seguintes servidores iSNS: Atualizar

Nome
------

Para adicionar um servidor iSNS, clique em Adicionar Servidor. Adicionar Servidor...

Para remover um servidor iSNS, selecione o servidor acima e clique em Remover. Remover

OK Cancelar Aplicar

FIGURA 22.3: PORTAIS DE DESTINO

4. Em seguida, alterne para a guia *Destinos* e revise o(s) destino(s) descoberto(s).

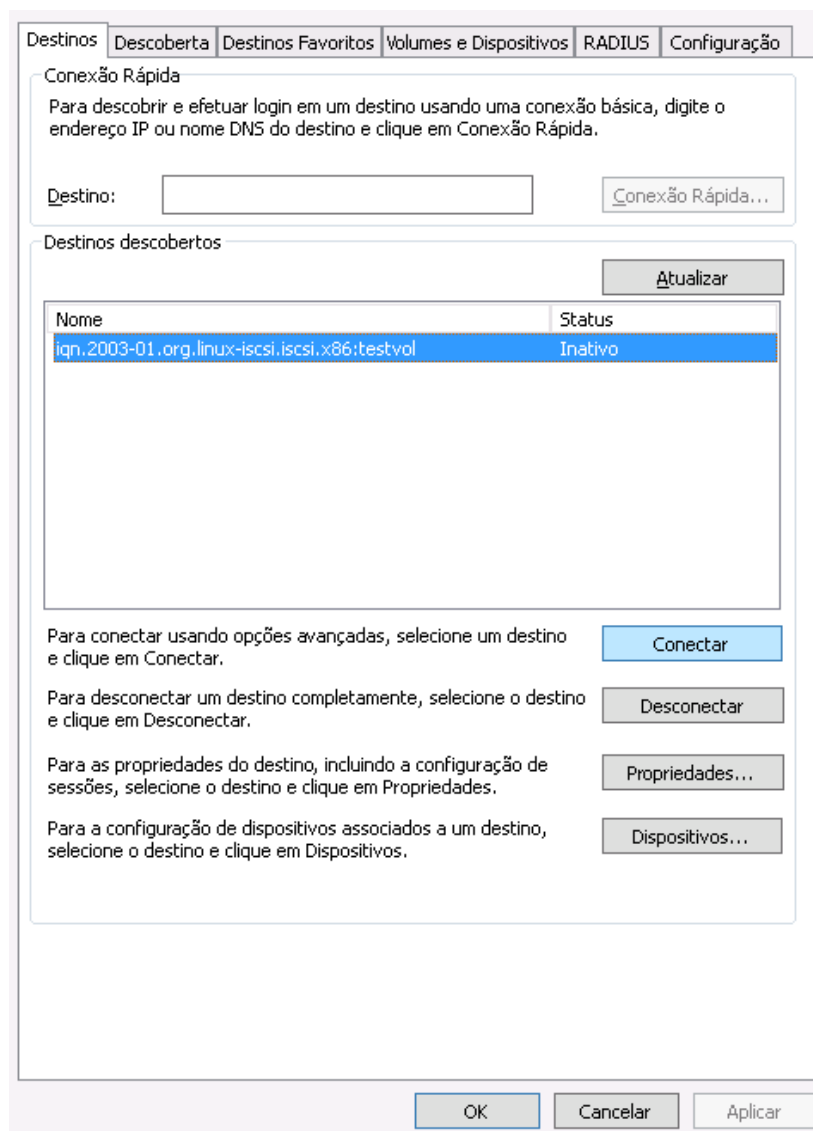


FIGURA 22.4: DESTINOS

5. Clique em *Conectar* na guia *Destinos*. A caixa de diálogo *Conectar ao Destino* é exibida. Marque a caixa de seleção *Habilitar múltiplos caminhos* para habilitar a E/S de múltiplos caminhos (MPIO) e clique em *OK*:

6. Quando a caixa de diálogo *Conectar ao Destino* for fechada, selecione *Propriedades* para revisar as propriedades do destino:

The screenshot shows the 'Conectar ao Destino' dialog box with the 'Propriedades' tab selected. The 'Sessões' tab is also visible. The 'Atualizar' button is at the top right. Below it, the 'Identificador' section contains two checked entries: 'ffffe00103669020-400001370000000f' and 'ffffe00103669020-40000137000000010'. Below the list, there are three instructions with corresponding buttons: 'Adicionar sessão', 'Desconectar', and 'Dispositivos...'. The 'Informações da Sessão' section displays session details. The 'Configurar Sessão Múltipla Conectada (MCS)' section includes an 'MCS...' button.

Informações da Sessão	
Tag do grupo de portais de destino:	1
Status:	Conectado
Total de conexões:	1
Máximo de Conexões Permitidas:	1
Autenticação:	Nada Especificado
Síntese de Cabeçalho:	Nada Especificado
Síntese de Dados:	Nada Especificado

Configurar Sessão Múltipla Conectada (MCS)  
Para adicionar outras conexões a uma sessão ou configurar a política MCS para uma sessão selecionada, clique em MCS.

MCS...

FIGURA 22.5: PROPRIEDADES DO DESTINO ISCSI

7. Selecione *Dispositivos* e clique em *MPIO* para revisar a configuração de E/S de múltiplos caminhos:

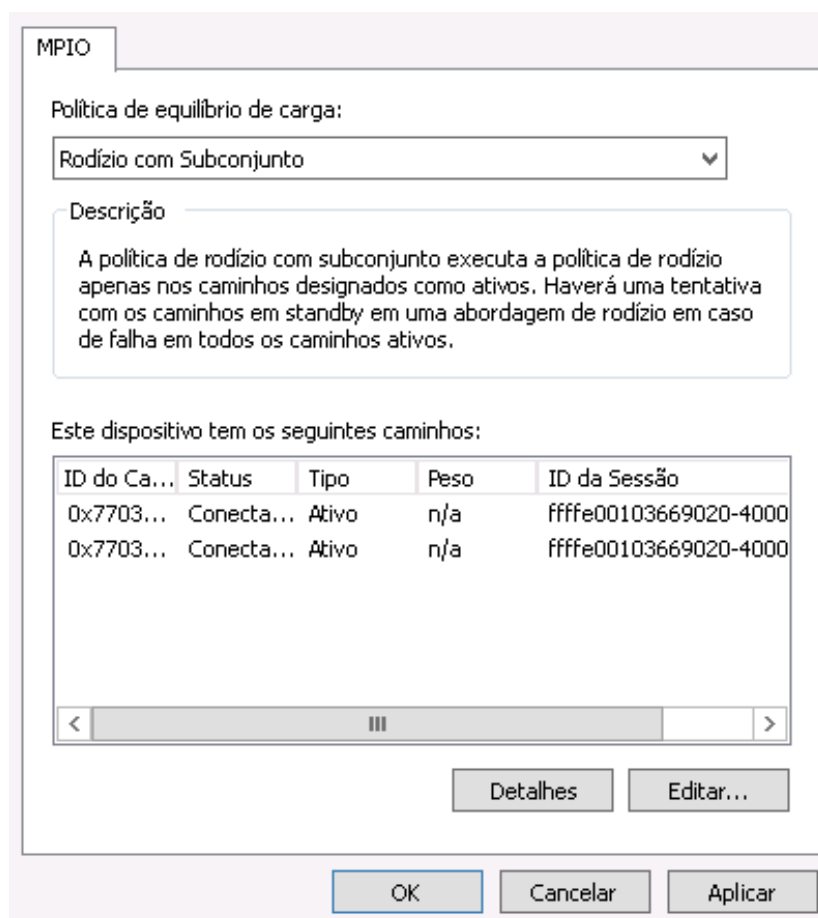


FIGURA 22.6: DETALHES DO DISPOSITIVO

A *Política de equilíbrio de carga* padrão é *Rodízio com Subconjunto*. Se você preferir uma configuração exclusivamente de failover, mude-a para *Somente Failover*.

Isso conclui a configuração do iniciador iSCSI. Agora, os volumes iSCSI estão disponíveis como qualquer outro dispositivo SCSI e podem ser inicializados para uso como volumes e unidades. Clique em *OK* para fechar a caixa de diálogo *Propriedades do Iniciador iSCSI* e prossiga com a função *Serviços de Arquivo e Armazenamento* do painel de controle do *Gerenciador do Servidor*.

Observe o volume recém-conectado. Ele é identificado como *Unidade de Múltiplos Caminhos SCSI Baseada em RBD SUSE* no barramento iSCSI e é inicialmente marcado com o status *Offline* e o tipo de tabela de partição *Desconhecido*. Se o novo volume não aparecer imediatamente, selecione *Explorar Armazenamento Novamente* na caixa suspensa *Tarefas* para explorar o barramento iSCSI novamente.

1. Clique o botão direito do mouse no volume iSCSI e selecione *Novo Volume* no menu de contexto. O *Assistente de Novo Volume* é exibido. Clique em *Próximo*, realce o volume iSCSI recém-conectado e clique em *Próximo* para começar.

#### Selecionar servidor e disco

Antes de Começar

**Servidor e Disco**

Tamanho

Letra da Unidade ou Pasta

Configurações do Sistema de Arquivos

Confirmação

Resultados

Servidor:

Provisionar para	Status	Função do Cluster	Destino
WIN-U3AILLIMUEE	Online	Sem Cluster	Local

Atualizar Explorar Novamente


Disco:

Disk	Disco Virtual	Capacidade	Espaço Livre	Subsistema
Disco 0		234 GB	233 GB	
Disco 1		234 GB	234 GB	
Disco 3		48,8 GB	48,8 GB	

< Anterior Próximo > Criar Cancelar

FIGURA 22.7: ASSISTENTE DE NOVO VOLUME

2. Inicialmente, o dispositivo está vazio e não contém uma tabela de partição. Quando solicitado, confirme a caixa de diálogo indicando que o volume será inicializado com uma tabela de partição GPT:

 O disco selecionado será colocado online e inicializado como um disco GPT. Para continuar, clique em OK, para selecionar um disco diferente ou criar um novo disco virtual, clique em Cancelar.

OK Cancelar

FIGURA 22.8: PROMPT DE DISCO OFFLINE



3. Selecione o tamanho do volume. Normalmente, você usa a capacidade total do dispositivo. Em seguida, atribua uma letra de unidade ou nome de diretório no qual o volume recém-criado estará disponível. Na sequência, selecione um sistema de arquivos para criar no novo volume e, por fim, clique em *Criar* para confirmar suas seleções e concluir a criação do volume:

#### Confirmar seleções

Antes de Começar  
Servidor e Disco  
Tamanho  
Letra da Unidade ou Pasta  
Configurações do Sistema de Arquivos  
**Confirmação**  
Resultados

Confirme se as seguintes configurações estão corretas e clique em Criar.

LOCAL DO VOLUME	
Servidor:	WIN-U3AILLIMUEE
Disco:	Disco 3
Espaço livre:	48,8 GB
PROPRIEDADES DO VOLUME	
Tamanho do volume:	48,8 GB
Letra da unidade ou pasta:	D:\
Rótulo do volume:	Novo Volume
CONFIGURAÇÕES DO SISTEMA DE ARQUIVOS	
Sistema de arquivos:	NTFS
Criação de nome de arquivo abreviado:	Desabilitado
Tamanho da unidade de alocação:	Padrão

< Anterior   Próximo >   Criar   Cancelar

FIGURA 22.9: CONFIRMAR SELEÇÕES DE VOLUME

Quando o processo for concluído, revise os resultados e clique em *Fechar* para concluir a inicialização da unidade. Quando a inicialização for concluída, o volume (e o respectivo sistema de arquivos NTFS) ficará disponível como uma unidade local recém-inicializada.

### 22.1.3 Conectando-se ao VMware

1. Para se conectar aos volumes iSCSI gerenciados por `ceph-iscsi`, você precisa de um adaptador de software iSCSI configurado. Se não houver um adaptador desse tipo disponível na configuração do vSphere, crie um selecionando *Configuração > Adaptadores de Armazenamento > Adicionar > Iniciador de Software iSCSI*.

- Quando disponível, selecione as propriedades do adaptador clicando o botão direito do mouse nele e selecionando *Propriedades* no menu de contexto:

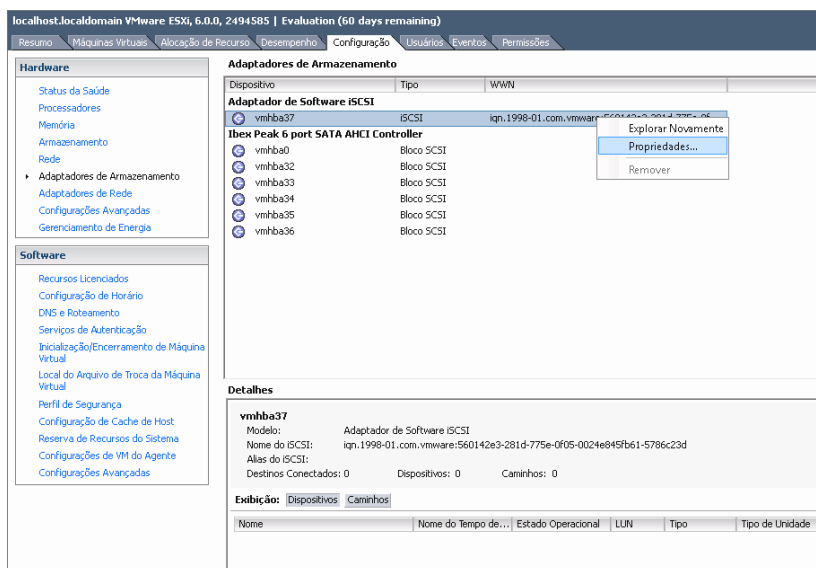


FIGURA 22.10: PROPRIEDADES DO INICIADOR ISCSI

- Na caixa de diálogo *Iniciador de Software iSCSI*, clique no botão *Configurar*. Em seguida, vá para a guia *Descoberta Dinâmica* e selecione *Adicionar*.
- Digite o endereço IP ou nome de host do seu gateway iSCSI `ceph-iscsi`. Se você executa vários gateways iSCSI em uma configuração de failover, repita essa etapa para todos os gateways que você opera.

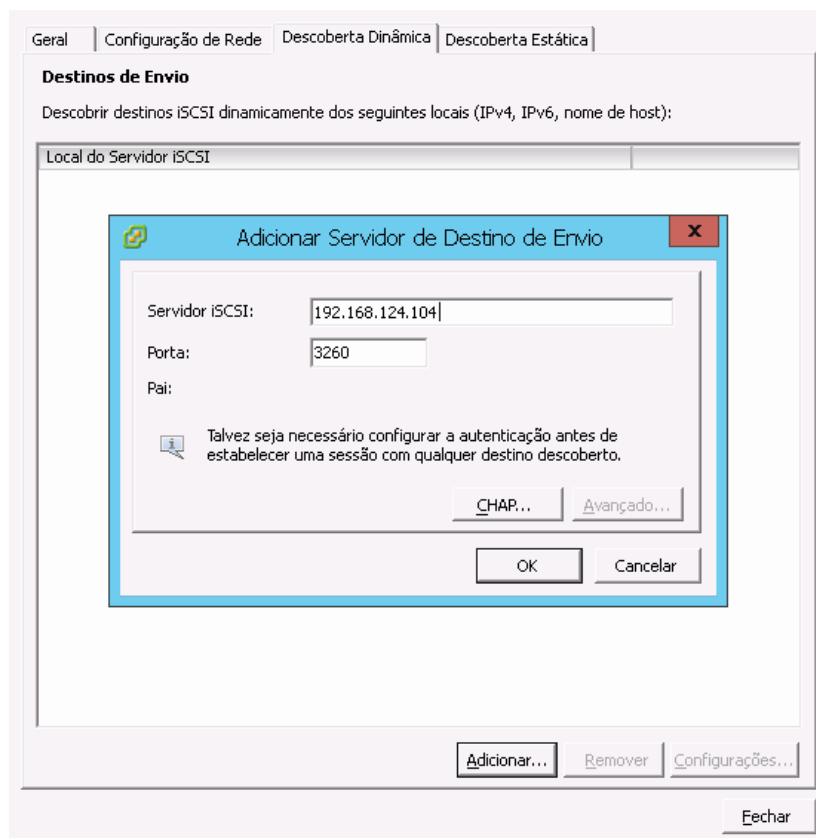


FIGURA 22.11: ADICIONAR SERVIDOR DE DESTINO

Após inserir todos os gateways iSCSI, clique em *OK* na caixa de diálogo para iniciar uma nova exploração do adaptador iSCSI.

5. Quando a nova exploração for concluída, o novo dispositivo iSCSI aparecerá abaixo da lista *Adaptadores de Armazenamento* no painel *Detalhes*. Para dispositivos de múltiplos caminhos, agora você pode clicar o botão direito do mouse no adaptador e selecionar *Gerenciar Caminhos* no menu de contexto:

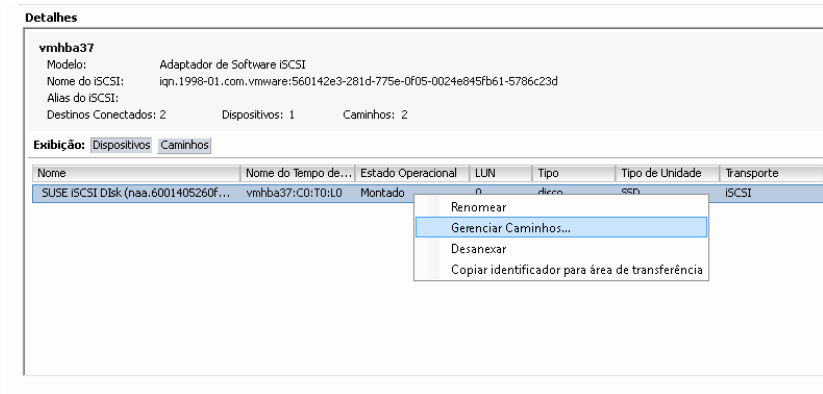


FIGURA 22.12: GERENCIAR DISPOSITIVOS DE MÚLTIPLOS CAMINHOS

Você agora deve ver todos os caminhos com um ícone verde em *Status*. Um dos seus caminhos deve estar marcado como *Ativo (E/S)*, e todos os outros apenas como *Ativo*:

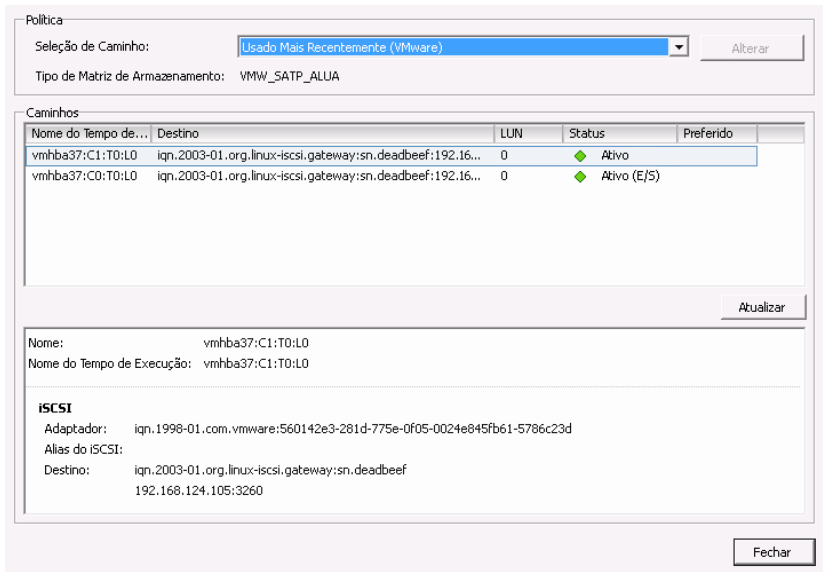


FIGURA 22.13: LISTAGEM DE CAMINHOS PARA MÚLTIPLOS CAMINHOS

6. Agora, você pode alternar de *Adaptadores de Armazenamento* para o item denominado *Armazenamento*. Selecione *Adicionar Armazenamento...* no canto superior direito do painel para exibir a caixa de diálogo *Adicionar Armazenamento*. Em seguida, selecione *Disco/LUN* e clique em *Avançar*. O dispositivo iSCSI recém-adicionado aparece na lista *Selecionar Disco/LUN*. Selecione-o e, em seguida, clique em *Avançar* para continuar:

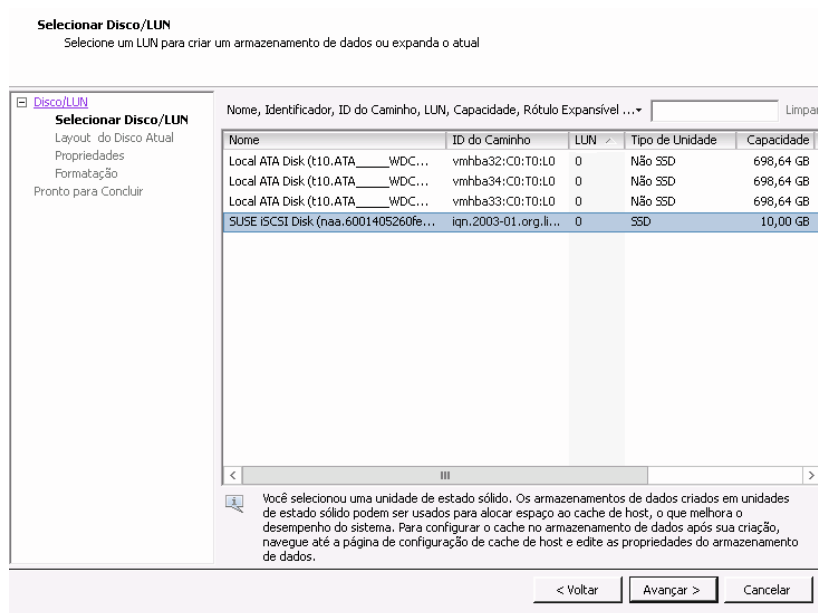


FIGURA 22.14: CAIXA DE DIÁLOGO DE ADIÇÃO DE ARMAZENAMENTO

Clique em *Avançar* para aceitar o layout de disco padrão.

7. No painel *Propriedades*, atribua um nome ao novo armazenamento de dados e clique em *Avançar*. Aceite a configuração padrão para usar todo o espaço do volume para o armazenamento de dados ou selecione *Configuração personalizada de espaço* para um armazenamento de dados menor:

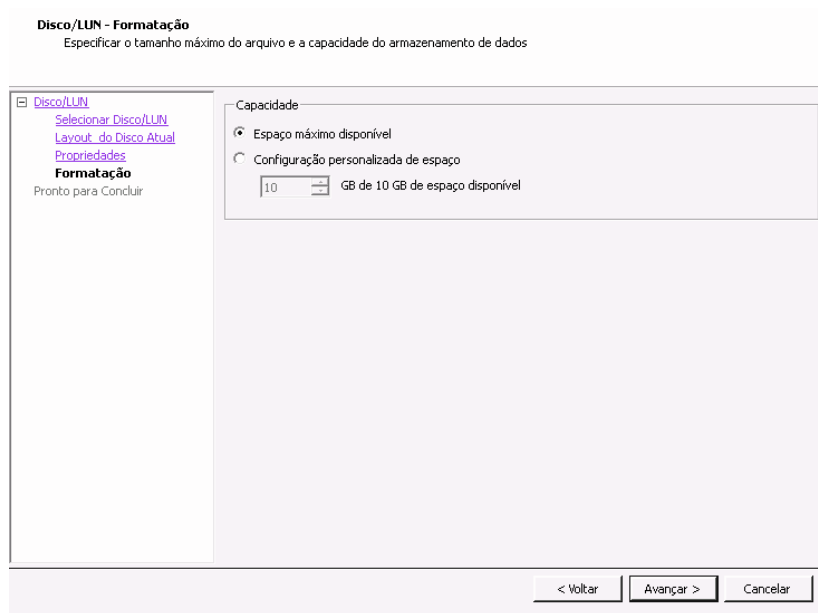


FIGURA 22.15: CONFIGURAÇÃO PERSONALIZADA DE ESPAÇO

Clique em *Concluir* para concluir a criação do armazenamento de dados.

Agora, o novo armazenamento de dados aparece na lista de armazenamentos de dados, e você pode selecioná-lo para recuperar os detalhes. Agora, você pode usar o volume iSCSI baseado em ceph-iscsi como qualquer outro armazenamento de dados vSphere.

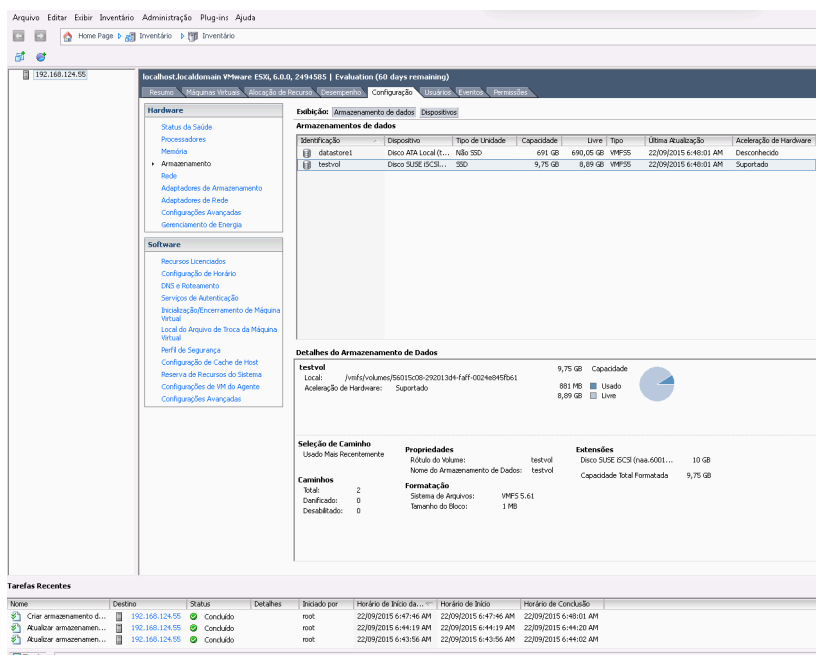


FIGURA 22.16: VISÃO GERAL DO ARMAZENAMENTO DE DADOS ISCSI

## 22.2 Conclusão

ceph-iscsi é um componente fundamental do SUSE Enterprise Storage 7.1 que concede acesso a armazenamento em blocos distribuído e altamente disponível de qualquer servidor ou cliente que reconheça o protocolo iSCSI. Ao usar o ceph-iscsi em um ou mais hosts do iSCSI Gateway, as imagens RBD do Ceph tornam-se disponíveis como Unidades Lógicas (LUs, Logical Units) associadas a destinos iSCSI, que podem ser acessados com equilíbrio de carga e alta disponibilidade.

Como a configuração de todos os ceph-iscsi é inserida no armazenamento de objetos RADOS do Ceph, os hosts do gateway ceph-iscsi são inerentemente sem estado persistente e, portanto, podem ser substituídos, aumentados ou reduzidos conforme desejado. Como resultado, o SUSE Enterprise Storage 7.1 permite que os clientes SUSE executem uma tecnologia de armazenamento empresarial verdadeiramente distribuída, altamente disponível, resiliente e autorreparável em um hardware convencional e em uma plataforma totalmente de código-fonte aberto.

## 23 Sistema de arquivos em cluster

Este capítulo descreve as tarefas de administração que normalmente são executadas depois que o cluster é configurado e o CephFS é exportado. Se você precisar de mais informações sobre como configurar o CephFS, consulte o *Livro “Guia de Implantação”, Capítulo 8 “Implantando os serviços principais restantes com o cephadm”, Seção 8.3.3 “Implantando servidores de metadados”*.

### 23.1 Montando o CephFS

Quando o sistema de arquivos é criado e o MDS está ativo, você está pronto para montar o sistema de arquivos de um host de cliente.

#### 23.1.1 Preparando o cliente

Se o host de cliente executa o SUSE Linux Enterprise 12 SP2 ou versão mais recente, o sistema está pronto para montar o CephFS “out-of-the-box”.

Se o host de cliente executa o SUSE Linux Enterprise 12 SP1, você precisa aplicar todos os patches mais recentes antes de montar o CephFS.

Em qualquer caso, tudo o que é preciso para montar o CephFS está incluído no SUSE Linux Enterprise. O produto SUSE Enterprise Storage 7.1 não é necessário.

Para suportar a sintaxe de `mount` completa, o pacote `ceph-common` (que acompanha o SUSE Linux Enterprise) deve ser instalado antes de tentar montar o CephFS.



#### Importante

Se o pacote `ceph-common` (e, portanto, sem o ajudante `mount.ceph`), os IPs dos monitores precisarão ser usados em vez dos nomes. O motivo é que o cliente do kernel não poderá executar a resolução de nome.

A sintaxe de montagem básica é:

```
# mount -t ceph MON1_IP[:PORT],MON2_IP[:PORT],...:CEPHFS_MOUNT_TARGET \
MOUNT_POINT -o name=CEPHX_USER_NAME,secret=SECRET_STRING
```



## 23.1.2 Criando um arquivo secreto

Por padrão, o cluster do Ceph é executado com a autenticação ativada. Você deve criar um arquivo que armazena sua chave secreta (não o chaveiro propriamente dito). Para obter a chave secreta para determinado usuário e, em seguida, criar o arquivo, faça o seguinte:

### PROCEDIMENTO 23.1: CRIANDO UMA CHAVE SECRETA

1. Veja a chave do usuário específico em um arquivo de chaveiro:

```
cephuser@adm > cat /etc/ceph/ceph.client.admin.keyring
```

2. Copie a chave do usuário que utilizará o sistema de arquivos Ceph FS montado. Normalmente, a chave tem a seguinte aparência:

```
AQCj2YpRiAe6CxAA7/ETt7Hcl9IxyYciVs47w==
```

3. Crie um arquivo com o nome de usuário como parte do nome de arquivo. Por exemplo, /etc/ceph/admin.secret para o usuário *admin*.
4. Cole o valor da chave no arquivo criado na etapa anterior.
5. Defina os direitos de acesso apropriados para o arquivo. O usuário deve ser a única pessoa que pode ler o arquivo. Outras pessoas não devem ter nenhum direito de acesso.

## 23.1.3 Montando o CephFS

Você pode montar o CephFS com o comando **mount**. Você precisa especificar o nome de host ou endereço IP do monitor. Como a autenticação cephx está habilitada por padrão no SUSE Enterprise Storage, você precisa especificar um nome de usuário e também o segredo relacionado:

```
# mount -t ceph ceph_mon1:6789:/ /mnt/cephfs \  
-o name=admin,secret=AQATSKdNGBnwLhAAAnNDKnH65FmVKpXZJVasUeQ==
```

Como o comando anterior permanece no histórico do shell, uma abordagem mais segura é ler o segredo de um arquivo:

```
# mount -t ceph ceph_mon1:6789:/ /mnt/cephfs \  
-o name=admin,secretfile=/etc/ceph/admin.secret
```

Observe que o arquivo de segredo deve conter apenas o segredo do chaveiro real. Em nosso exemplo, o arquivo incluirá apenas a seguinte linha:

```
AQATSKdNGBnwLhAAAnNDKnH65FmVKpXZJVasUeQ==
```



### Dica: Especificar vários monitores

Convém especificar vários monitores separados por vírgulas na linha de comando **mount** para o caso de um monitor ficar inativo no momento da montagem. Cada endereço de monitor adota o formato `host[:porta]`. Se a porta não for especificada, será usado o padrão 6789.

Crie o ponto de montagem no host local:

```
# mkdir /mnt/cephfs
```

Monte o CephFS:

```
# mount -t ceph ceph_mon1:6789:/ /mnt/cephfs \
-o name=admin,secretfile=/etc/ceph/admin.secret
```

Um subdiretório `subdir` poderá ser especificado se um subconjunto do sistema de arquivos tiver que ser montado:

```
# mount -t ceph ceph_mon1:6789:/subdir /mnt/cephfs \
-o name=admin,secretfile=/etc/ceph/admin.secret
```

Você pode especificar mais de um host de monitor no comando **mount**:

```
# mount -t ceph ceph_mon1,ceph_mon2,ceph_mon3:6789:/ /mnt/cephfs \
-o name=admin,secretfile=/etc/ceph/admin.secret
```



### Importante: Acesso de leitura ao diretório raiz

Se forem usados clientes com restrição de caminho, os recursos do MDS precisarão incluir o acesso de leitura ao diretório raiz. Por exemplo, um chaveiro pode ter a seguinte aparência:

```
client.bar
key: supersecretkey
caps: [mds] allow rw path=/barjail, allow r path=/
caps: [mon] allow r
caps: [osd] allow rwx
```

A parte `allow r path=/` indica que os clientes com restrição de caminho podem ver o volume raiz, mas não podem gravar nele. Isso pode ser um problema para casos de uso em que o isolamento completo é um requisito.

## 23.2 Desmontando o CephFS

Para desmontar o CephFS, use o comando `umount`:

```
# umount /mnt/cephfs
```

## 23.3 Montando o CephFS em `/etc/fstab`

Para montar o CephFS automaticamente na inicialização do cliente, insira a linha correspondente na respectiva tabela de sistemas de arquivos `/etc/fstab`:

```
mon1:6790,mon2:/subdir /mnt/cephfs ceph name=admin,secretfile=/etc/ceph/secret.key,noatime,_netdev 0 2
```

## 23.4 Vários daemons MDS ativos (MDS ativo-ativo)

Por padrão, o CephFS é configurado para um único daemon MDS ativo. Para aumentar o desempenho dos metadados em sistemas de grande escala, é possível habilitar vários daemons MDS ativos, que compartilharão a carga de trabalho dos metadados entre eles.

### 23.4.1 Usando o MDS ativo-ativo

Considere o uso de vários daemons MDS ativos em caso de gargalo no desempenho dos metadados no MDS único padrão.

A adição de mais daemons não aumenta o desempenho em todos os tipos de carga de trabalho. Por exemplo, um único aplicativo em execução em um só cliente não se beneficiará de um número maior de daemons MDS, a menos que o aplicativo esteja efetuando muitas operações de metadados em paralelo.

As cargas de trabalho que costumam se beneficiar de um número maior de daemons MDS ativos são aquelas com vários clientes, que podem atuar em muitos diretórios separados.

## 23.4.2 Aumentando o tamanho do cluster MDS ativo

Cada sistema de arquivos CephFS tem uma configuração `max_mds` que controla quantas classificações serão criadas. O número real de classificações no sistema de arquivos apenas será aumentado se um daemon sobressalente estiver disponível para assumir a nova classificação. Por exemplo, se houver apenas um daemon MDS em execução, e `max_mds` estiver definido como dois, não será criada uma segunda classificação.

No exemplo a seguir, definimos a opção `max_mds` como 2 para criar uma nova classificação separadamente do padrão. Para ver as mudanças, execute **ceph status** antes e depois que você definir `max_mds` e observe a linha que contém `fsmap`:

```
cephuser@adm > ceph status
[...]
```

services:

```
  [...]
  mds: cephfs-1/1/1 up {0=node2=up:active}, 1 up:standby
  [...]
```

cephuser@adm > ceph fs set cephfs max\_mds 2

cephuser@adm > ceph status

```
[...]
```

services:

```
  [...]
  mds: cephfs-2/2/2 up {0=node2=up:active,1=node1=up:active}
  [...]
```

A classificação recém-criada (1) passa pelo estado de “criação” e depois entra no estado “ativo”.



### Importante: Daemons de standby

Mesmo com vários daemons MDS ativos, um sistema altamente disponível ainda requer daemons de standby para assumir o controle em caso de falha em qualquer um dos servidores que executam um daemon ativo.

Consequentemente, o limite máximo ideal de `max_mds` para sistemas de alta disponibilidade é menor do que o número total de servidores MDS no sistema. Para se manter disponível em caso de várias falhas do servidor, aumente o número de daemons de standby no sistema para corresponder ao número de falhas do servidor que você precisa superar.

### 23.4.3 Diminuindo o número de classificações

Todas as classificações, incluindo as que devem ser removidas, devem primeiro estar ativas. Isso significa que é necessário ter pelo menos `max_mds` daemons MDS disponíveis.

Primeiramente, defina `max_mds` como um número mais baixo. Por exemplo, volte a ter um único MDS ativo:

```
cephuser@adm > ceph status
[...]
```

services:

```
  [...]
  mds: cephfs-2/2/2 up {0=node2=up:active,1=node1=up:active}
  [...]
```

cephuser@adm > ceph fs set cephfs max\_mds 1

cephuser@adm > ceph status

```
[...]
```

services:

```
  [...]
  mds: cephfs-1/1/1 up {0=node2=up:active}, 1 up:standby
  [...]
```

### 23.4.4 Fixando manualmente árvores de diretório em uma classificação

Em várias configurações de servidor de metadados ativas, um balanceador é executado, que funciona para distribuir a carga de metadados igualmente no cluster. Em geral, isso funciona bem o suficiente para a maioria dos usuários; mas, às vezes, convém anular o balanceador dinâmico com mapeamentos explícitos de metadados para classificações específicas. Isso pode permitir que o administrador ou os usuários distribuam a carga do aplicativo igualmente ou limitem o impacto das solicitações de metadados dos usuários sobre o cluster inteiro.

O mecanismo fornecido para essa finalidade é chamado “export pin”. Ele é um atributo estendido de diretórios. O nome desse atributo estendido é `ceph.dir.pin`. Os usuários podem definir esse atributo usando os comandos padrão:

```
# setfattr -n ceph.dir.pin -v 2 /path/to/dir
```

O valor (`-v`) do atributo estendido é a classificação à qual atribuir a subárvore do diretório. O valor padrão `-1` indica que o diretório não foi fixado.

O export pin de um diretório é herdado do seu pai mais próximo com um export pin definido. Portanto, a definição do export pin em um diretório afeta todos os seus filhos. No entanto, a fixação do pai pode ser anulada pela definição do export pin do diretório filho. Por exemplo:

```
# mkdir -p a/b                # "a" and "a/b" start with no export pin set.
setfattr -n ceph.dir.pin -v 1 a/ # "a" and "b" are now pinned to rank 1.
setfattr -n ceph.dir.pin -v 0 a/b # "a/b" is now pinned to rank 0
                                # and "a/" and the rest of its children
                                # are still pinned to rank 1.
```

## 23.5 Gerenciando o failover

Se um daemon MDS parar de se comunicar com o monitor, o monitor aguardará `mds_beacon_grace` segundos (o padrão é 15 segundos) antes de marcar o daemon como *lento*. Você pode configurar um ou mais daemons de “standby” para assumir o controle durante o failover do daemon MDS.

### 23.5.1 Configurando a reprodução de standby

Cada sistema de arquivos CephFS pode ser configurado para adicionar daemons de reprodução de standby. Esses daemons de standby seguem o diário de metadados do MDS ativo para reduzir o tempo de failover caso o MDS ativo se torne indisponível. Cada MDS ativo pode ter apenas um daemon de reprodução de standby o seguindo.

Configure a reprodução de standby em um sistema de arquivos com o seguinte comando:

```
cephuser@adm > ceph fs set FS-NAME allow_standby_replay BOOL
```

Quando definidos, os monitores atribuirão os daemons de standby disponíveis para seguir os MDSs ativos nesse sistema de arquivos.

Quando um MDS entrar no estado de reprodução de standby, ele apenas será usado como standby para a classificação que está seguindo. Se houver falha em outra classificação, esse daemon de reprodução de standby não será usado como substituição, mesmo que não haja outros standbys disponíveis. Por esse motivo, se a reprodução de standby for usada, será aconselhável que cada MDS ativo tenha um daemon de reprodução de standby.

## 23.6 Definindo cotas do CephFS

Você pode definir cotas em qualquer subdiretório do sistema de arquivos Ceph. A cota restringe o número de **bytes** ou de **arquivos** armazenados abaixo do ponto especificado na hierarquia de diretórios.

### 23.6.1 Limitações de cota do CephFS

O uso de cotas com o CephFS tem as seguintes limitações:

**As cotas são cooperativas e não concorrentes.**

As cotas do Ceph confiam que o cliente que está montando o sistema de arquivos pare de gravar nele quando um limite é atingido. A parte do servidor não pode evitar que um cliente mal intencionado grave a quantidade de dados que ele precisar. Não use cotas para evitar o preenchimento do sistema de arquivos em ambientes em que os clientes não são totalmente confiáveis.

**As cotas não são precisas.**

Os processos que estão gravando no sistema de arquivos serão interrompidos logo após o limite da cota ser atingido. Inevitavelmente, eles poderão gravar alguma quantidade de dados acima do limite configurado. Os gravadores do cliente serão interrompidos dentro de décimos de segundos após ultrapassar o limite configurado.

**As cotas são implementadas no cliente do kernel a partir da versão 4.17.**

As cotas são suportadas pelo cliente do espaço de usuário (libcephfs, ceph-fuse). Os clientes do kernel do Linux 4.17 e superiores suportam cotas do CephFS nos clusters do SUSE Enterprise Storage 7.1. Haverá falha nos clientes do kernel (até nas versões recentes) ao processar cotas em clusters mais antigos, mesmo que eles possam definir os atributos estendidos das cotas. Os kernels do SLE12-SP3 (e versões mais recentes) já incluem os backports necessários para administrar as cotas.

**Configure as cotas com cuidado quando forem usadas com restrições de montagem com base no caminho.**

O cliente precisa ter acesso ao inode do diretório em que as cotas são configuradas para aplicá-las. Se o cliente tiver acesso restrito a um determinado caminho (por exemplo, /home/user) com base no recurso do MDS, e se a cota for configurada em um diretório de

origem ao qual ele não tem acesso a (/home), o cliente não a aplicará. Ao usar restrições de acesso com base no caminho, certifique-se de configurar a cota no diretório que o cliente pode acessar (por exemplo, /home/user ou /home/user/quota\_dir).

## 23.6.2 Configurando cotas do CephFS

Você pode configurar cotas do CephFS usando atributos estendidos virtuais:

ceph.quota.max\_files

Configura um limite de *arquivos*.

ceph.quota.max\_bytes

Configura um limite de *bytes*.

Se os atributos aparecerem no inode de um diretório, uma cota será configurada nele. Se eles não estiverem presentes, nenhuma cota será definida nesse diretório (embora uma ainda possa ser configurada em um diretório pai).

Para definir uma cota de 100 MB, execute:

```
cephuser@mds > setfattr -n ceph.quota.max_bytes -v 100000000 /SOME/DIRECTORY
```

Para definir uma cota de 10.000 arquivos, execute:

```
cephuser@mds > setfattr -n ceph.quota.max_files -v 10000 /SOME/DIRECTORY
```

Para ver a configuração de cota, execute:

```
cephuser@mds > getfattr -n ceph.quota.max_bytes /SOME/DIRECTORY
```

```
cephuser@mds > getfattr -n ceph.quota.max_files /SOME/DIRECTORY
```



### Nota: Cota não definida

Se o valor do atributo estendido for “0”, a cota não será definida.

Para remover uma cota, execute:

```
cephuser@mds > setfattr -n ceph.quota.max_bytes -v 0 /SOME/DIRECTORY
cephuser@mds > setfattr -n ceph.quota.max_files -v 0 /SOME/DIRECTORY
```



## 23.7 Gerenciando instantâneos do CephFS

Os instantâneos do CephFS criam uma visão apenas leitura do sistema de arquivos no momento em que são capturados. Você pode criar um instantâneo em qualquer diretório. O instantâneo incluirá todos os dados no sistema de arquivos abaixo do diretório especificado. Após a criação de um instantâneo, os dados incluídos no buffer serão descarregados dos vários clientes de maneira assíncrona. Dessa forma, a criação de um instantâneo é muito rápida.



### Importante: Vários sistemas de arquivos

Se você tiver vários sistemas de arquivos CephFS compartilhando um único pool (por meio de namespaces), seus instantâneos colidirão, e a exclusão de um instantâneo resultará em dados de arquivos ausentes em outros instantâneos que compartilham o mesmo pool.

### 23.7.1 Criando instantâneos

Por padrão, o recurso de instantâneo do CephFS está habilitado nos sistemas de arquivos novos. Para habilitá-lo nos sistemas de arquivos existentes, execute:

```
cephuser@adm > ceph fs set CEPHFS_NAME allow_new_snaps true
```

Após a habilitação dos instantâneos, todos os diretórios no CephFS terão um subdiretório `.snap` especial.



### Nota

Este é um subdiretório *virtual*. Ele não aparece na listagem de diretórios do diretório pai, mas o nome `.snap` não pode ser usado como nome de arquivo ou diretório. O diretório `.snap` precisa ser acessado explicitamente, por exemplo:

```
> ls -la /CEPHFS_MOUNT/.snap/
```



## Importante: Limitação de clientes do kernel

Os clientes do kernel do CephFS têm uma limitação: eles não podem processar mais do que 400 instantâneos em um sistema de arquivos. O número de instantâneos deve ser mantido sempre abaixo desse limite, seja qual for o cliente que você usa. Se você usa clientes CephFS mais antigos, como SLE12-SP3, lembre-se de que ultrapassar 400 instantâneos é prejudicial para as operações, porque haverá falha no cliente.



## Dica: Nome personalizado do subdiretório de instantâneos

Você pode configurar um nome diferente para o subdiretório de instantâneos definindo a configuração `client snapdir`.

Para criar um instantâneo, crie um subdiretório abaixo do diretório `.snap` com um nome personalizado. Por exemplo, para criar um instantâneo do diretório `/CEPHFS_MOUNT/2/3/`, execute:

```
> mkdir /CEPHFS_MOUNT/2/3/.snap/CUSTOM_SNAPSHOT_NAME
```

## 23.7.2 Apagando instantâneos

Para apagar um instantâneo, remova seu subdiretório dentro do diretório `.snap`:

```
> rmdir /CEPHFS_MOUNT/2/3/.snap/CUSTOM_SNAPSHOT_NAME
```

## 24 Exportar dados do Ceph por meio do Samba

Este capítulo descreve como exportar os dados armazenados em um cluster do Ceph por meio de um compartilhamento do Samba/CIFS para que você possa acessá-los facilmente de máquinas cliente Windows\*. Ele também inclui informações que ajudarão você a configurar um gateway do Samba para o Ceph a fim de ingressar o Active Directory no domínio do Windows\* para autenticar e autorizar usuários.



### Nota: Desempenho do gateway do Samba

Devido ao aumento da sobrecarga do protocolo e da latência adicional causado por saltos extras de rede entre o cliente e o armazenamento, o acesso ao CephFS por meio de um Gateway do Samba pode reduzir significativamente o desempenho do aplicativo quando comparado aos clientes nativos do Ceph.

## 24.1 Exportar o CephFS por meio do compartilhamento do Samba



### Atenção: Acesso a vários protocolos

Os clientes nativos CephFS e NFS não são restritos por bloqueios de arquivos obtidos por meio do Samba, e vice-versa. Os aplicativos que dependem do bloqueio de arquivos compatível com vários protocolos poderão ter os dados corrompidos se os caminhos de compartilhamento do Samba com suporte do CephFS forem acessados por outros meios.

### 24.1.1 Configurando e exportando pacotes do Samba

Para configurar e exportar um compartilhamento do Samba, os seguintes pacotes precisam ser instalados: samba-ceph e samba-winbind. Se esses pacotes não foram instalados, instale-os:

```
cephuser@smb > zypper install samba-ceph samba-winbind
```

## 24.1.2 Exemplo de gateway único

Em preparação para exportar um compartilhamento do Samba, escolha um nó apropriado para agir como Gateway do Samba. O nó precisa ter acesso à rede de clientes do Ceph, além de CPU, memória e recursos de rede suficientes.

É possível fornecer a funcionalidade de failover com o CTDB e a SUSE Linux Enterprise High Availability Extension. Consulte a [Seção 24.1.3, “Configuring a alta disponibilidade”](#) para obter mais informações sobre configuração de Alta Disponibilidade.

1. Verifique se já existe um CephFS em execução no cluster.
2. Crie um chaveiro específico do Gateway do Samba no nó de admin do Ceph e copie-o para os dois nós do Gateway do Samba:

```
cephuser@adm > ceph auth get-or-create client.samba.gw mon 'allow r' \
  osd 'allow *' mds 'allow *' -o ceph.client.samba.gw.keyring
cephuser@adm > scp ceph.client.samba.gw.keyring SAMBA_NODE:/etc/ceph/
```

Substitua *SAMBA\_NODE* pelo nome do nó do gateway do Samba.

3. As etapas a seguir são executadas no nó do Gateway do Samba. Instale o Samba com o pacote de integração do Ceph:

```
cephuser@smb > sudo zypper in samba samba-ceph
```

4. Substitua o conteúdo padrão do arquivo `/etc/samba/smb.conf` com o seguinte:

```
[global]
  netbios name = SAMBA-GW
  clustering = no
  idmap config * : backend = tdb2
  passdb backend = tdbsam
  # disable print server
  load printers = no
  smbd: backgroundqueue = no

[SHARE_NAME]
  path = CEPHFS_MOUNT
  read only = no
  oplocks = no
  kernel share modes = no
```

O caminho `CEPHFS_MOUNT` acima deve ser montado antes de iniciar o Samba com uma configuração de compartilhamento do CephFS do kernel. Consulte a [Seção 23.3, “Montando o CephFS em /etc/fstab”](#).

A configuração de compartilhamento acima usa o cliente CephFS do kernel do Linux, que é recomendado por motivos de desempenho. Outra opção é usar o módulo `vfs_ceph` do Samba para comunicação com o cluster do Ceph. As instruções mostradas abaixo são para uso legado e não são recomendadas para novas implantações do Samba:

```
[SHARE_NAME]
path = /
vfs objects = ceph
ceph: config_file = /etc/ceph/ceph.conf
ceph: user_id = samba.gw
read only = no
oplocks = no
kernel share modes = no
```



### Dica: Oplocks e modos de compartilhamento

Os oplocks (também conhecidos como concessões do SMB2+) proporciona melhor desempenho por meio de armazenamento em cache de cliente agressivo, mas não é seguro atualmente quando o Samba é implantado com outros clientes CephFS, como `mount.ceph` do kernel, FUSE ou NFS Ganesha.

Se todo o acesso do caminho do sistema de arquivos CephFS for processado exclusivamente pelo Samba, o parâmetro oplocks poderá ser habilitado com segurança.

No momento, os modos de compartilhamento do kernel precisam ser desabilitados em um compartilhamento executado com o módulo `vfs` do CephFS para que o processamento do arquivo funcione apropriadamente.



## Importante: Permitindo acesso

O Samba mapeia usuários e grupos do SMB para contas locais. Os usuários locais podem receber uma senha para acesso ao compartilhamento do Samba por meio de:

```
# smbpasswd -a USERNAME
```

Para uma E/S bem-sucedida, a lista de controles de acesso (ACL, Access Control List) do caminho do compartilhamento precisa permitir o acesso ao usuário conectado pelo Samba. É possível modificar a ACL por uma montagem temporária por meio do cliente do kernel CephFS e usando os utilitários `chmod`, `chown` ou `setfacl` no caminho do compartilhamento. Por exemplo, para permitir o acesso de todos os usuários, execute:

```
# chmod 777 MOUNTED_SHARE_PATH
```

### 24.1.2.1 Iniciando serviços do Samba

Inicie ou reinicie serviços independentes do Samba usando os seguintes comandos:

```
# systemctl restart smb.service
# systemctl restart nmb.service
# systemctl restart winbind.service
```

Para garantir que os serviços do Samba sejam iniciados na inicialização, execute este comando para habilitá-los:

```
# systemctl enable smb.service
# systemctl enable nmb.service
# systemctl enable winbind.service
```



## Dica: Serviços nmb e winbind opcionais

Se você não precisar da pesquisa de compartilhamento de rede, não precisará habilitar e iniciar o serviço `nmb`.

O serviço `winbind` apenas é necessário quando configurado como um membro de domínio do Active Directory. Consulte a [Seção 24.2, “Ingressando no Gateway do Samba e no Active Directory”](#).

### 24.1.3 Configuring a alta disponibilidade



#### Importante: Failover transparente não suportado

Embora uma implantação de vários nós do Samba + CTDB tenha disponibilidade mais alta em comparação com o nó único (consulte o [Capítulo 24, Exportar dados do Ceph por meio do Samba](#)), o failover transparente executado no cliente não é suportado. Os aplicativos provavelmente sofrerão uma breve interrupção no momento da falha do nó do Gateway do Samba.

Esta seção apresenta um exemplo de como definir uma configuração de dois nós de alta disponibilidade dos servidores Samba. A configuração requer a SUSE Linux Enterprise High Availability Extension. Os dois nós são chamados earth (192.168.1.1) e mars (192.168.1.2). Para obter detalhes sobre a SUSE Linux Enterprise High Availability Extension, consulte <https://documentation.suse.com/sle-ha/15-SP1/>.

Além disso, dois endereços IP virtuais flutuantes permitem aos clientes se conectarem ao serviço independentemente do nó físico no qual está sendo executado. 192.168.1.10 é usado para administração do cluster com Hawk2, e 192.168.2.1 é usado exclusivamente para exportações CIFS. Isso facilita aplicar as restrições de segurança mais tarde.

O procedimento a seguir descreve a instalação de exemplo. Mais detalhes podem ser encontrados em <https://documentation.suse.com/sle-ha/15-SP1/single-html/SLE-HA-install-quick/>.

1. Crie um chaveiro específico do Gateway do Samba no Nó de Admin e copie-o para os dois nós:

```
cephuser@adm > ceph auth get-or-create client.samba.gw mon 'allow r' \
    osd 'allow *' mds 'allow *' -o ceph.client.samba.gw.keyring
cephuser@adm > scp ceph.client.samba.gw.keyring earth:/etc/ceph/
cephuser@adm > scp ceph.client.samba.gw.keyring mars:/etc/ceph/
```

2. A configuração de SLE-HA requer um dispositivo de fencing para evitar uma situação de *split brain* quando os nós do cluster ativos perdem a sincronização. Para essa finalidade, você pode usar uma imagem RBD do Ceph com o Dispositivo de Blocos Stonith (SBD, Stonith Block Device). Consulte <https://documentation.suse.com/sle-ha/15-SP1/single-html/SLE-HA-guide/#sec-ha-storage-protect-fencing-setup> para obter mais detalhes.

Se ele ainda não existir, crie um pool RBD chamado `rbd` (consulte a [Seção 18.1, “Criando um pool”](#)) e associe-o ao `rbd` (consulte a [Seção 18.5.1, “Associando pools a um aplicativo”](#)). Em seguida, crie uma imagem RBD relacionada chamada `sbd01`:

```
cephuser@adm > ceph osd pool create rbd
cephuser@adm > ceph osd pool application enable rbd rbd
cephuser@adm > rbd -p rbd create sbd01 --size 64M --image-shared
```

3. Prepare o `earth` e o `mars` para hospedar o serviço do Samba:

- a. Verifique se os seguintes pacotes estão instalados antes de continuar: `ctdb`, `tdb-tools` e `samba`.

```
# zypper in ctdb tdb-tools samba samba-ceph
```

- b. Verifique se os serviços Samba e CTDB estão parados e desabilitados:

```
# systemctl disable ctdb
# systemctl disable smb
# systemctl disable nmb
# systemctl disable winbind
# systemctl stop ctdb
# systemctl stop smb
# systemctl stop nmb
# systemctl stop winbind
```

- c. Abra a porta `4379` do seu firewall em todos os nós. Isso é necessário para o CTDB se comunicar com outros nós do cluster.

4. No `earth`, crie os arquivos de configuração para o Samba. Mais tarde, eles serão sincronizados automaticamente com o `mars`.

- a. Insira uma lista de endereços IP particulares dos nós do Gateway do Samba no arquivo `/etc/ctdb/nodes`. Encontre mais detalhes na página de manual do `ctdb` (**man 7 ctdb**).

```
192.168.1.1
192.168.1.2
```



- b. Configure o Samba. Adicione as seguintes linhas à seção `[global]` do `/etc/samba/smb.conf`. Use o nome de host de sua escolha em vez `CTDB-SERVER` (todos os nós no cluster aparecerão como um nó grande com esse nome). Adicione também uma definição do compartilhamento. Considere `SHARE_NAME` como um exemplo:

```
[global]
    netbios name = SAMBA-HA-GW
    clustering = yes
    idmap config * : backend = tdb2
    passdb backend = tdbsam
    ctdbd socket = /var/lib/ctdb/ctdb.socket
    # disable print server
    load printers = no
    smbd: backgroundqueue = no

[SHARE_NAME]
    path = /
    vfs objects = ceph
    ceph: config_file = /etc/ceph/ceph.conf
    ceph: user_id = samba.gw
    read only = no
    oplocks = no
    kernel share modes = no
```

Observe que os arquivos `/etc/ctdb/nodes` e `/etc/samba/smb.conf` precisam ser iguais em todos os nós do Gateway do Samba.

## 5. Instale e inicialize o cluster da SUSE Linux Enterprise High Availability.

- a. Registre a SUSE Linux Enterprise High Availability Extension em `earth` e `mars`:

```
root@earth # SUSEConnect -r ACTIVATION_CODE -e E_MAIL
```

```
root@mars # SUSEConnect -r ACTIVATION_CODE -e E_MAIL
```

- b. Instale o `ha-cluster-bootstrap` nos dois nós:

```
root@earth # zypper in ha-cluster-bootstrap
```

```
root@mars # zypper in ha-cluster-bootstrap
```

- c. Mapeie a imagem RBD `sbd01` nos dois Gateways do Samba usando `rbdmap.service`.

Edite `/etc/ceph/rbdmap` e adicione uma entrada para a imagem SBD:

```
rbd/sbd01 id=samba.gw,keyring=/etc/ceph/ceph.client.samba.gw.keyring
```

Habilite e inicie o `rbdmap.service`:

```
root@earth # systemctl enable rbdmap.service && systemctl start rbdmap.service
root@mars # systemctl enable rbdmap.service && systemctl start rbdmap.service
```

O dispositivo `/dev/rbd/rbd/sbd01` deve estar disponível em ambos os Gateways do Samba.

d. Inicialize o cluster no `earth` e permita que `mars` ingresse nele.

```
root@earth # ha-cluster-init
```

```
root@mars # ha-cluster-join -c earth
```

## Importante

Durante o processo de inicialização e ingresso no cluster, será questionado se você deseja usar o SBD. Confirme com `y` e depois especifique `/dev/rbd/rbd/sbd01` como o caminho para o dispositivo de armazenamento.

6. Verifique o status do cluster. Você deve ver dois nós adicionados ao cluster:

```
root@earth # crm status
2 nodes configured
1 resource configured

Online: [ earth mars ]

Full list of resources:

admin-ip          (ocf::heartbeat:IPaddr2):      Started earth
```

7. Execute os seguintes comandos no `earth` para configurar o recurso CTDB:

```
root@earth # crm configure
crm(live)configure# primitive ctdb ocf:heartbeat:CTDB params \
    ctdb_manages_winbind="false" \
    ctdb_manages_samba="false" \
    ctdb_recovery_lock="!/usr/lib64/ctdb/ctdb_mutex_ceph_rados_helper
    ceph client.samba.gw cephfs_metadata ctdb-mutex"
```

```

ctdb_socket="/var/lib/ctdb/ctdb.socket" \
  op monitor interval="10" timeout="20" \
  op start interval="0" timeout="200" \
  op stop interval="0" timeout="100"
crm(live)configure# primitive smb systemd:smb \
  op start timeout="100" interval="0" \
  op stop timeout="100" interval="0" \
  op monitor interval="60" timeout="100"
crm(live)configure# primitive nmb systemd:nmb \
  op start timeout="100" interval="0" \
  op stop timeout="100" interval="0" \
  op monitor interval="60" timeout="100"
crm(live)configure# primitive winbind systemd:winbind \
  op start timeout="100" interval="0" \
  op stop timeout="100" interval="0" \
  op monitor interval="60" timeout="100"
crm(live)configure# group g-ctdb ctdb winbind nmb smb
crm(live)configure# clone cl-ctdb g-ctdb meta interleave="true"
crm(live)configure# commit

```



### Dica: Primitivos nmb e winbind opcionais

Se você não precisar da pesquisa de compartilhamento de rede, não precisará adicionar o primitivo nmb.

O primitivo winbind apenas é necessário quando configurado como membro do domínio do Active Directory. Consulte a [Seção 24.2, "Ingressando no Gateway do Samba e no Active Directory"](#).

O binário /usr/lib64/ctdb/ctdb\_mutex\_ceph\_rados\_helper na opção de configuração ctdb\_recovery\_lock tem os parâmetros CLUSTER\_NAME, CEPHX\_USER, RADOS\_POOL e RADOS\_OBJECT, nessa ordem.

Um parâmetro de tempo de espera de bloqueio extra pode ser anexado para anular o valor padrão usado (10 segundos). Um valor mais alto aumentará o tempo de failover do master de recuperação CTDB, enquanto um valor menor pode resultar na detecção incorreta do master de recuperação como inativo, provocando oscilação de failovers.

#### 8. Adicione um endereço IP em cluster:

```

crm(live)configure# primitive ip ocf:heartbeat:IPaddr2
  params ip=192.168.2.1 \
  unique_clone_address="true" \
  op monitor interval="60" \

```

```

meta resource-stickiness="0"
crm(live)configure# clone cl-ip ip \
    meta interleave="true" clone-node-max="2" globally-unique="true"
crm(live)configure# colocation col-with-ctdb 0: cl-ip cl-ctdb
crm(live)configure# order o-with-ctdb 0: cl-ip cl-ctdb
crm(live)configure# commit

```

Se `unique_clone_address` for definido como `true`, o agente de recurso IPAddr2 adicionará um ID de clone ao endereço especificado, resultando em três endereços IP diferentes. Geralmente, eles não são necessários, mas ajudam no equilíbrio de carga. Para obter mais informações sobre este tópico, consulte <https://documentation.suse.com/sle-ha/15-SP1/single-html/SLE-HA-guide/#cha-ha-lb>.

#### 9. Verifique o resultado:

```

root@earth # crm status
Clone Set: base-clone [dlm]
    Started: [ factory-1 ]
    Stopped: [ factory-0 ]
Clone Set: cl-ctdb [g-ctdb]
    Started: [ factory-1 ]
    Started: [ factory-0 ]
Clone Set: cl-ip [ip] (unique)
    ip:0      (ocf:heartbeat:IPAddr2):      Started factory-0
    ip:1      (ocf:heartbeat:IPAddr2):      Started factory-1

```

10. Faça o teste de uma máquina cliente. Em um cliente Linux, execute o seguinte comando para ver se você pode copiar arquivos do sistema e para o sistema:

```
# smbclient //192.168.2.1/myshare
```

#### 24.1.3.1 Reiniciando os recursos de HA do Samba

Após qualquer mudança de configuração do Samba ou CTDB, os recursos de HA talvez tenham de ser reiniciados para que as mudanças entrem em vigor. Isso pode ser feito pelo comando:

```
# crm resource restart cl-ctdb
```

## 24.2 Ingressando no Gateway do Samba e no Active Directory

Você pode configurar o gateway do Samba para Ceph para se tornar um membro do domínio do Samba com suporte ao Active Directory (AD). Como membro do domínio do Samba, você pode utilizar usuários e grupos de domínio em listas de acesso (ACLs) locais em arquivos e diretórios do CephFS exportado.

### 24.2.1 Preparando a instalação do Samba

Esta seção apresenta as etapas preparatórias que você precisa executar antes de configurar o próprio Samba. Começar com um ambiente limpo ajuda você a evitar confusão e verificar se nenhum arquivo da instalação do Samba anterior está misturado com a nova instalação do membro do domínio.



#### Dica: Sincronizando relógios

Todos os relógios dos nós do Gateway do Samba precisam ser sincronizados com o controlador de Domínio do Active Directory. Uma divergência no relógio pode resultar em falhas na autenticação.

Verifique se não há processos de cache de nome ou do Samba em execução:

```
cephuser@smb > ps ax | egrep "samba|smbd|nmbd|winbindd|nscd"
```

Se a saída listar quaisquer processos `samba`, `smbd`, `nmbd`, `winbindd` ou `nscd`, pare-os.

Se você já executou uma instalação do Samba neste host, remova o arquivo `/etc/samba/smb.conf`. Remova também todos os arquivos de banco de dados Samba, como `*.tdb` e `*.ldb`. Para listar os diretórios que contêm os bancos de dados Samba, execute:

```
cephuser@smb > smbd -b | egrep "LOCKDIR|STATEDIR|CACHEDIR|PRIVATE_DIR"
```

### 24.2.2 Verificando o DNS

O Active Directory (AD) usa o DNS para localizar outros controladores de domínio (DCs) e serviços, como o Kerberos. Portanto, os membros do domínio e os servidores do AD precisam ser capazes de resolver as zonas DNS do AD.

Verifique se o DNS está configurado corretamente e se ambas as pesquisas avançada e reversa são resolvidas corretamente, por exemplo:

```
cephuser@adm > nslookup DC1.domain.example.com
Server:          10.99.0.1
Address:         10.99.0.1#53

Name:   DC1.domain.example.com
Address: 10.99.0.1
```

```
cephuser@adm > 10.99.0.1
Server:          10.99.0.1
Address:         10.99.0.1#53

1.0.99.10.in-addr.arpa name = DC1.domain.example.com.
```

### 24.2.3 Resolvendo registros SRV

O AD usa os registros SRV para localizar serviços, como Kerberos e LDAP. Para verificar se os registros SRV foram resolvidos corretamente, use o shell interativo **nslookup**. Por exemplo:

```
cephuser@adm > nslookup
Default Server:  10.99.0.1
Address:         10.99.0.1

> set type=SRV
> _ldap._tcp.domain.example.com.
Server:  UnKnown
Address: 10.99.0.1

_ldap._tcp.domain.example.com  SRV service location:
        priority      = 0
        weight        = 100
        port          = 389
        svr hostname   = dc1.domain.example.com
domain.example.com    nameserver = dc1.domain.example.com
dc1.domain.example.com internet address = 10.99.0.1
```

## 24.2.4 Configurando o kerberos

O Samba suporta os back ends Heimdal e MIT do Kerberos. Para configurar o Kerberos no membro do domínio, defina o seguinte no arquivo `/etc/krb5.conf`:

```
[libdefaults]
default_realm = DOMAIN.EXAMPLE.COM
dns_lookup_realm = false
dns_lookup_kdc = true
```

O exemplo anterior configura o Kerberos para o domínio Kerberos DOMAIN.EXAMPLE.COM. Não recomendamos definir nenhum outro parâmetro no arquivo `/etc/krb5.conf`. Se o `/etc/krb5.conf` contém uma linha `include`, ele não funcionará. Você **deve** remover essa linha.

## 24.2.5 Resolvendo o nome de host local

Quando você ingressa um host no domínio, o Samba tenta registrar o nome de host na zona DNS do AD. Para isso, o utilitário **net** precisa ser capaz de resolver o nome de host usando DNS ou uma entrada correta no arquivo `/etc/hosts`.

Para verificar se o nome de host foi resolvido corretamente, use o comando **getent hosts**:

```
cephuser@adm > getent hosts example-host
10.99.0.5      example-host.domain.example.com    example-host
```

O nome de host e o FQDN não devem ser resolvidos para o endereço IP 127.0.0.1 nem para qualquer endereço IP diferente do que foi usado na interface LAN do membro do domínio. Se nenhuma saída for exibida ou se o host for resolvido para o endereço IP incorreto, e você não estiver usando DHCP, defina a entrada correta no arquivo `/etc/hosts`:

```
127.0.0.1      localhost
10.99.0.5      example-host.samdom.example.com    example-host
```



### Dica: DHCP e `/etc/hosts`

Se você estiver usando DHCP, confira se `/etc/hosts` contém apenas a linha “127.0.0.1”. Se os problemas persistirem, contate o administrador do seu servidor DHCP.

Se você precisa adicionar aliases ao nome de host da máquina, adicione-os ao fim da linha que começa com o endereço IP da máquina, e não à linha “127.0.0.1”.

## 24.2.6 Configurando o Samba

Esta seção apresenta informações sobre opções de configuração específicas que você precisa incluir na configuração do Samba.

A participação no domínio do Active Directory é configurada principalmente definindo `security = ADS` junto com os parâmetros apropriados de mapeamento de domínio e ID do Kerberos na seção `[global]` do `/etc/samba/smb.conf`.

```
[global]
security = ADS
workgroup = DOMAIN
realm = DOMAIN.EXAMPLE.COM
...
```

### 24.2.6.1 Escolhendo o back end para mapeamento de ID no winbindd

Se você precisa que seus usuários tenham shells de login e/ou caminhos de diretório pessoal do Unix diferentes, ou se você deseja que eles tenham o mesmo ID em todos os lugares, será necessário usar o back end “ad” do winbind e adicionar os atributos RFC2307 ao AD.



#### Importante: Atributos RFC2307 e Números de ID

Os atributos RFC2307 não são adicionados automaticamente quando os usuários ou grupos são criados.

Os números de ID encontrados em um DC (números na faixa de 3000000) *não* são atributos RFC2307 e não serão usados nos Membros do Domínio Unix. Se você precisa ter os mesmos números de ID em todos os lugares, adicione os atributos `uidNumber` e `gidNumber` ao AD e use o back end “ad” do winbind nos Membros do Domínio Unix. Se você adicionar os atributos `uidNumber` e `gidNumber` ao AD, não use números na faixa de 3000000.

Se os seus usuários utilizarão apenas o DC do AD para Samba para autenticação e não armazenarão dados nem efetuarão login nele, você poderá usar o back end “rid” do winbind. Isso calcula os IDs do usuário e do grupo do RID do Windows\*. Se você usar a mesma seção `[global]` do `smb.conf` em cada membro do domínio Unix, obterá os mesmos IDs. Se você usar o back end “rid”, não precisará adicionar nada ao AD, e os atributos RFC2307 serão ignorados. Ao usar o back end “rid”, defina os parâmetros `template shell` e `template homedir` no `smb.conf`.



Essas configurações são globais, e todos recebem o mesmo shell de login e caminho de diretório pessoal do Unix (ao contrário dos atributos RFC2307, em que você pode definir shells e caminhos de diretório pessoal do Unix individuais).

Há outra maneira de configurar o Samba: quando você precisa que seus usuários e grupos tenham o mesmo ID em todos os lugares, mas precisa apenas que seus usuários tenham o mesmo shell de login e usem o mesmo caminho de diretório pessoal do Unix. Para fazer isso, use o back end “ad” do winbind e as linhas de gabarito no `smb.conf`. Dessa forma, você precisa apenas adicionar os atributos `uidNumber` e `gidNumber` ao AD.



### Dica: Mais Informações sobre Back Ends para Mapeamento de ID

Encontre informações mais detalhadas sobre os back ends de mapeamento de ID disponíveis nas páginas de manual relacionadas: [`man 8 idmap\_ad`](#), [`man 8 idmap\_rid`](#) e [`man 8 idmap\_autorid`](#).

#### 24.2.6.2 Definindo faixas de IDs de usuário e grupo

Após decidir qual back end do winbind será usado, você precisará especificar as faixas a serem usadas com a opção `idmap config` em `smb.conf`. Por padrão, há vários blocos de IDs de usuário e grupo reservados em um membro do domínio Unix:

TABELA 24.1: BLOCOS DE IDS PADRÃO DE USUÁRIOS E GRUPOS

IDs	Faixa
0-999	Usuários e grupos do sistema local.
A partir de 1000	Usuários e grupos do Unix local.
A partir de 10000	Usuários e grupos do DOMÍNIO.

Como é possível ver nas faixas acima, você não deve definir as faixas “\*” ou “DOMÍNIO” para começar em 999 ou menos, pois elas interferem nos usuários e grupos do sistema local. Você também deve deixar um espaço para quaisquer usuários e grupos do Unix local. Dessa forma, começar as faixas `idmap config` em 3000 parece ser um bom compromisso.

Você precisa decidir o quanto seu "DOMÍNIO" poderá crescer e se você pretende ter quaisquer domínios confiáveis. Em seguida, você poderá definir as faixas `idmap config` da seguinte maneira:

TABELA 24.2: FAIXAS DE ID

Domínio	Faixa
*	3000-7999
DOMÍNIO	10000-999999
CONFIÁVEL	1000000-9999999

### 24.2.6.3 Mapeando a conta de administrador de domínio para o usuário `root` local

O Samba permite mapear contas de domínio para uma conta local. Use esse recurso para executar operações de arquivo no sistema de arquivos do membro do domínio como um usuário diferente daquele da conta que solicitou a operação no cliente.



#### Dica: Mapeando o administrador de domínio (opcional)

O mapeamento do administrador de domínio para a conta `root` local é opcional. Configure o mapeamento apenas se o administrador de domínio precisa ter a capacidade de executar operações de arquivo no membro de domínio usando as permissões de `root`. Esteja ciente de que o mapeamento do Administrador para a conta `root` não permite que você efetue login nos membros do domínio Unix como “Administrador”.

Para mapear o administrador de domínio para a conta `root` local, siga estas etapas:

1. Adicione o seguinte parâmetro à seção `[global]` do arquivo `smb.conf`:

```
username map = /etc/samba/user.map
```

2. Crie o arquivo `/etc/samba/user.map` com o seguinte conteúdo:

```
!root = DOMAIN\Administrator
```

## ! Importante

Ao usar o back end de mapeamento de ID “ad”, não defina o atributo `uidNumber` para a conta de administrador de domínio. Se a conta tiver o atributo definido, o valor anulará o UID “0” local do usuário `root` e, portanto, haverá falha no mapeamento.

Para obter mais detalhes, consulte o parâmetro `username map` na página de manual `smb.conf` ([man 5 smb.conf](#)).

## 24.2.7 Ingressando no domínio do Active Directory

Para que o host ingresse em um Active Directory, execute:

```
cephuser@smb > net ads join -U administrator
Enter administrator's password: PASSWORD
Using short domain name -- DOMAIN
Joined EXAMPLE-HOST to dns domain 'DOMAIN.example.com'
```

## 24.2.8 Configurando o Name Service Switch

Para disponibilizar usuários e grupos de domínio ao sistema local, você precisa habilitar a biblioteca NSS (Name Service Switch). Anexe a entrada `winbind` aos seguintes bancos de dados no arquivo `/etc/nsswitch.conf`:

```
passwd: files winbind
group:  files winbind
```

## ! Importante: Pontos a Serem Considerados

- Mantenha a entrada `files` como a primeira origem para os dois bancos de dados. Desse modo, o NSS pode procurar os usuários e grupos de domínio dos arquivos `/etc/passwd` e `/etc/group` antes de consultar o serviço `winbind`.
- Não adicione a entrada `winbind` ao banco de dados de `sombra` do NSS. Isso pode causar uma falha no utilitário `wbinfo`.
- Não use no domínio os mesmos nomes de usuário do arquivo local `/etc/passwd`.

## 24.2.9 Iniciando os serviços

Após as mudanças de configuração, reinicie os serviços do Samba de acordo com a [Seção 24.1.2.1, “Iniciando serviços do Samba”](#) ou a [Seção 24.1.3.1, “Reiniciando os recursos de HA do Samba”](#).

## 24.2.10 Testar a conectividade de winbindd

### 24.2.10.1 Enviando um ping de winbindd

Para verificar se o serviço `winbindd` pode se conectar aos Controladores de Domínio (DC, Domain Controllers) do AD ou a um PDC (Primary Domain Controller), digite:

```
cephuser@smb > wbinfo --ping-dc
checking the NETLOGON for domain[DOMAIN] dc connection to "DC.DOMAIN.EXAMPLE.COM"
succeeded
```

Se houver falha no comando anterior, verifique se o serviço `winbindd` está em execução e se o arquivo `smb.conf` está configurado corretamente.

### 24.2.10.2 Pesquisando usuários e grupos de domínio

A biblioteca `libnss_winbind` permite pesquisar usuários e grupos de domínio. Por exemplo, para pesquisar usuário de domínio “DOMAIN\demo01”:

```
cephuser@smb > getent passwd DOMAIN\\demo01
DOMAIN\demo01:*:10000:10000:demo01:/home/demo01:/bin/bash
```

Para pesquisar o grupo de domínio “Domain Users”:

```
cephuser@smb > getent group "DOMAIN\\Domain Users"
DOMAIN\domain users:x:10000:
```

### 24.2.10.3 Atribuindo permissões de arquivo a usuários e grupos de domínio

A biblioteca NSS (Name Service Switch) permite usar contas de usuário e grupos de domínio em comandos. Por exemplo, para definir o proprietário de um arquivo como o usuário de domínio “demo01”, e o grupo como o grupo de domínio “Domain Users”, digite:

```
cephuser@smb > chown "DOMAIN\\demo01:DOMAIN\\domain users" file.txt
```

## 25 NFS Ganesha

NFS Ganesha é um servidor NFS executado em um espaço de endereço do usuário, e não como parte do kernel do sistema operacional. Com o NFS Ganesha, você pode conectar seu próprio mecanismo de armazenamento, como Ceph, e acessá-lo de qualquer cliente NFS. Para obter instruções, consulte o Livro *“Guia de Implantação”, Capítulo 8 “Implantando os serviços principais restantes com o cephadm”, Seção 8.3.6 “Implantando o NFS Ganesha”*.



### Nota: Desempenho do NFS Ganesha

Devido ao aumento da sobrecarga do protocolo e da latência adicional causado por saltos extras de rede entre o cliente e o armazenamento, o acesso ao Ceph por meio de um NFS Gateway pode reduzir significativamente o desempenho do aplicativo quando comparado ao CephFS nativo.

Cada serviço do NFS Ganesha consiste em uma hierarquia de configuração que contém:

- Um `ganesha.conf` de boot
- Um objeto de configuração comum RADOS por serviço
- Um objeto de configuração RADOS por exportação

A configuração de boot é a configuração mínima para iniciar o daemon `nfs-ganesha` em um container. Cada configuração de boot incluirá uma diretiva `%url` com qualquer configuração adicional do objeto de configuração comum RADOS. O objeto de configuração comum pode incluir diretivas `%url` adicionais para cada uma das exportações NFS definidas nos objetos de configuração RADOS de exportação.

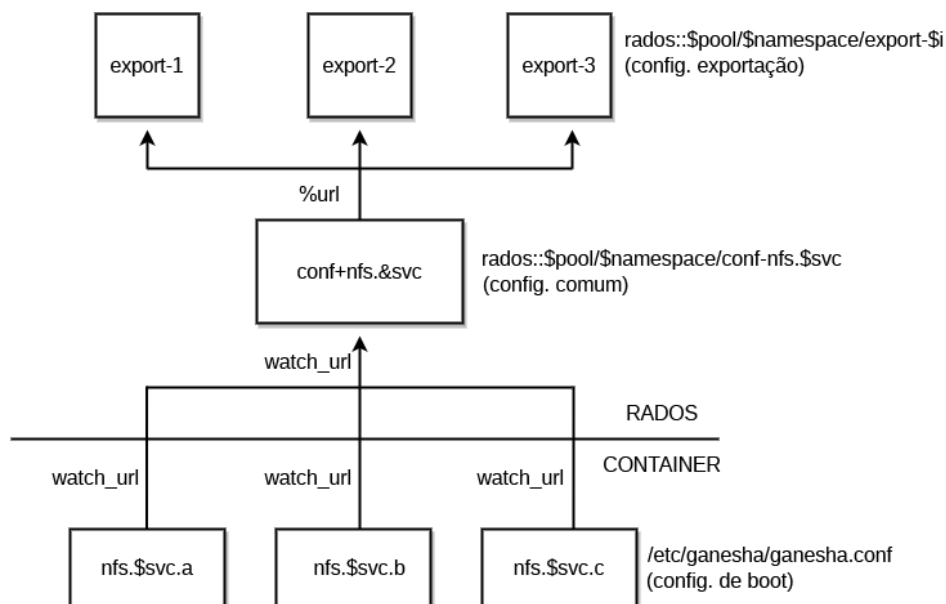


FIGURA 25.1: ESTRUTURA DO NFS GANESHA

## 25.1 Criando um serviço do NFS

A maneira recomendada de especificar a implantação dos serviços do Ceph é criar um arquivo no formato YAML com a especificação dos serviços que você pretende implantar. Você pode criar um arquivo de especificação separado para cada tipo de serviço ou especificar vários (ou todos) tipos de serviços em um arquivo.

Dependendo do que você escolheu fazer, será necessário atualizar ou criar um arquivo relevante no formato YAML para criar um serviço do NFS Ganesha. Para obter mais informações sobre como criar o arquivo, consulte o *Livro "Guia de Implantação", Capítulo 8 "Implantando os serviços principais restantes com o cephadm", Seção 8.2 "Especificação de serviço e posicionamento"*.

Após atualizar ou criar o arquivo, execute o seguinte comando para criar um serviço `nfs-ganesha`:

```
cephuser@adm > ceph orch apply -i FILE_NAME
```

## 25.2 Iniciando ou reiniciando o NFS Ganesha

### Importante

A inicialização do serviço NFS Ganesha não exporta automaticamente um sistema de arquivos CephFS. Para exportar um sistema de arquivos CephFS, crie um arquivo de configuração de exportação. Consulte a [Seção 25.4, “Criando uma exportação do NFS”](#) para obter mais detalhes.

Para iniciar o serviço do NFS Ganesha, execute:

```
cephuser@adm > ceph orch start nfs.SERVICE_ID
```

Para reiniciar o serviço do NFS Ganesha, execute:

```
cephuser@adm > ceph orch restart nfs.SERVICE_ID
```

Para reiniciar apenas um daemon do NFS Ganesha, execute:

```
cephuser@adm > ceph orch daemon restart nfs.SERVICE_ID
```

Quando o NFS Ganesha é iniciado ou reiniciado, ele tem um tempo de espera extra de 90 segundos para o NFS v4. Durante o período extra, as novas solicitações dos clientes são ativamente rejeitadas. Portanto, os clientes podem enfrentar lentidão nas solicitações durante o período extra do NFS.

## 25.3 Listando objetos no pool de recuperação do NFS

Execute o seguinte comando para listar os objetos no pool de recuperação do NFS:

```
cephuser@adm > rados --pool POOL_NAME --namespace NAMESPACE_NAME ls
```

## 25.4 Criando uma exportação do NFS

Você pode criar uma exportação do NFS no Ceph Dashboard ou manualmente por linha de comando. Para criar a exportação usando o Ceph Dashboard, consulte a [Capítulo 7, Gerenciar o NFS Ganesha](#), mais especificamente a [Seção 7.1, “Criando exportações do NFS”](#).

Para criar uma exportação do NFS manualmente, crie um arquivo de configuração para a exportação. Por exemplo, um arquivo `/tmp/export-1` com o seguinte conteúdo:

```
EXPORT {
    export_id = 1;
    path = "/";
    pseudo = "/";
    access_type = "RW";
    squash = "no_root_squash";
    protocols = 3, 4;
    transports = "TCP", "UDP";
    FSAL {
        name = "CEPH";
        user_id = "admin";
        filesystem = "a";
        secret_access_key = "SECRET_ACCESS_KEY";
    }
}
```

Depois de criar e gravar o arquivo de configuração para a nova exportação, execute o seguinte comando para criar a exportação:

```
rados --pool POOL_NAME --namespace NAMESPACE_NAME put EXPORT_NAME EXPORT_CONFIG_FILE
```

Por exemplo:

```
cephuser@adm > rados --pool example_pool --namespace example_namespace put export-1 /tmp/
export-1
```



### Nota

O bloco FSAL deve ser modificado para incluir o ID de usuário e a chave de acesso secreta do `cephx` desejado.

## 25.5 Verificando a exportação do NFS

O NFS v4 criará uma lista de exportações na raiz de um pseudo sistema de arquivos. Você pode verificar se os compartilhamentos NFS foram exportados montando `/` do nó do servidor NFS Ganesha:

```
# mount -t nfs nfs_ganesha_server_hostname:/ /path/to/local/mountpoint
```



```
# ls /path/to/local/mountpoint cephfs
```



### Nota: O NFS Ganesha é apenas v4

Por padrão, o `cephadm` configurará um servidor NFS v4. O NFS v4 não interage com os daemons `rpcbind` ou `mountd`. As ferramentas do cliente NFS, como `showmount`, não mostrarão nenhuma exportação configurada.

## 25.6 Montando a exportação do NFS

Para montar o compartilhamento NFS exportado em um host de cliente, execute:

```
# mount -t nfs nfs_ganesha_server_hostname:/ /path/to/local/mountpoint
```

## 25.7 Vários clusters do NFS Ganesha

É possível definir vários clusters do NFS Ganesha. Esse procedimento permite:

- Clusters separados do NFS Ganesha para acessar o CephFS.

## V Integração com ferramentas de virtualização

26    libvirt e Ceph **369**

27    Ceph como back end para instância de KVM QEMU **375**

## 26 libvirt e Ceph

A biblioteca `libvirt` cria uma camada de abstração de máquina virtual entre as interfaces de hipervisor e os aplicativos de software que as utilizam. Com a `libvirt`, os desenvolvedores e administradores de sistema podem se dedicar a uma estrutura comum de gerenciamento, API comum e interface de shell comum (`virsh`) com vários hipervisores diferentes, incluindo QEMU/KVM, Xen, LXC ou VirtualBox.

Os dispositivos de blocos do Ceph suportam o QEMU/KVM. Você pode usá-los com um software que estabeleça interface com a `libvirt`. A solução de nuvem usa a `libvirt` para interagir com o QEMU/KVM, e o QEMU/KVM interage com os dispositivos de blocos do Ceph pela `librbd`.

Para criar VMs que usam os dispositivos de blocos do Ceph, siga os procedimentos nas seções abaixo. Nos exemplos, usamos `libvirt-pool` para o nome do pool, `client.libvirt` para o nome de usuário e `new-libvirt-image` para o nome da imagem. Você pode usar qualquer valor desejado, mas deve substituí-los durante a execução dos comandos nos procedimentos seguintes.

### 26.1 Configurando o Ceph com libvirt

Para configurar o Ceph para uso com a `libvirt`, execute as seguintes etapas:

1. Crie um pool. O exemplo a seguir usa o nome do pool `libvirt-pool` com 128 grupos de posicionamento.

```
cephuser@adm > ceph osd pool create libvirt-pool 128 128
```

Verifique se o pool existe.

```
cephuser@adm > ceph osd lspools
```

2. Crie um Usuário do Ceph. O exemplo a seguir utiliza o nome de usuário do Ceph `client.libvirt` e faz referência ao `libvirt-pool`.

```
cephuser@adm > ceph auth get-or-create client.libvirt mon 'profile rbd' osd \
'profile rbd pool=libvirt-pool'
```

Verifique se o nome existe.

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



### Nota: Nome de usuário ou ID

A `libvirt` acessará o Ceph usando o ID `libvirt`, não o nome do Ceph `client.libvirt`. Consulte a [Seção 30.2.1.1, “Usuário”](#) para ver uma explicação detalhada da diferença entre ID e nome.

3. Use o QEMU para criar uma imagem em seu pool RBD. O exemplo a seguir usa o nome da imagem `new-libvirt-image` e faz referência ao `libvirt-pool`.



### Dica: Local do arquivo de chaveiro

A chave de usuário `libvirt` é armazenada em um arquivo de chaveiro salvo no diretório `/etc/ceph`. O arquivo de chaveiro precisa ter um nome apropriado que inclua o nome do cluster do Ceph ao qual ele pertence. Para o nome do cluster padrão “ceph”, o nome do arquivo de chaveiro é `/etc/ceph/ceph.client.libvirt.keyring`.

Se o chaveiro não existir, crie-o com:

```
cephuser@adm > ceph auth get client.libvirt > /etc/ceph/  
ceph.client.libvirt.keyring
```

```
# qemu-img create -f raw rbd:libvirt-pool/new-libvirt-image:id=libvirt 2G
```

Verifique se a imagem existe.

```
cephuser@adm > rbd -p libvirt-pool ls
```

## 26.2 Preparando o gerenciador de VM

É possível usar a `libvirt` sem um gerenciador de VM, mas talvez você ache mais simples criar primeiro o domínio com **`virt-manager`**.

1. Instale um gerenciador de máquina virtual.

```
# zypper in virt-manager
```

2. Prepare/faça download de uma imagem de OS do sistema que deseja virtualizar.

3. Inicie o gerenciador de máquina virtual.

```
virt-manager
```

## 26.3 Criando uma VM

Para criar uma VM com **virt-manager**, execute as seguintes etapas:

1. Escolha a conexão na lista, clique o botão direito do mouse nela e selecione *New* (Novo).
2. *Importe a imagem de disco existente* informando o caminho para o armazenamento existente. Especifique o tipo de OS, as configurações de memória e *nomeie* a máquina virtual. Por exemplo, `libvirt-virtual-machine`.
3. Conclua a configuração e inicie a VM.
4. Verifique se o domínio recém-criado existe com `sudo virsh list`. Se necessário, especifique a string de conexão, como

```
virsh -c qemu+ssh://root@vm_host_hostname/system list
Id      Name                                     State
-----
[...]
```

9	libvirt-virtual-machine	running
---	-------------------------	---------

5. Efetue login na VM e pare-a antes de configurá-la para uso com o Ceph.

## 26.4 Configurando a VM

Neste capítulo, vamos nos concentrar na configuração de VMs para integração com o Ceph usando **virsh**. Geralmente, os comandos **virsh** exigem privilégios de root (**sudo**) e não retornarão os resultados apropriados nem o notificarão sobre a necessidade dos privilégios de root. Para uma referência dos comandos **virsh**, consulte `man 1 virsh` (requer a instalação do pacote `libvirt-client`).

1. Abra o arquivo de configuração com `virsh edit vm-domain-name`.

```
# virsh edit libvirt-virtual-machine
```

2. Em `<devices>`, deve haver uma entrada `<disk>`.

```
<devices>
  <emulator>/usr/bin/qemu-system-SYSTEM-ARCH</emulator>
  <disk type='file' device='disk'>
    <driver name='qemu' type='raw' />
    <source file='/path/to/image/recent-linux.img' />
    <target dev='vda' bus='virtio' />
    <address type='drive' controller='0' bus='0' unit='0' />
  </disk>
```

Substitua `/path/to/image/recent-linux.img` pelo caminho para a imagem do OS.

## ! Importante

No lugar de um editor de texto, use **`sudo virsh edit`**. Se você editar o arquivo de configuração em `/etc/qemu` com um editor de texto, a `libvirt` poderá não reconhecer a mudança. Em caso de qualquer diferença entre o conteúdo do arquivo XML em `/etc/libvirt/qemu` e o resultado de **`sudo virsh dumpxml vm-domain-name`**, a VM pode não funcionar apropriadamente.

3. Adicione a imagem RBD do Ceph que você já criou como uma entrada `<disk>`.

```
<disk type='network' device='disk'>
  <source protocol='rbd' name='libvirt-pool/new-libvirt-image'>
    <host name='monitor-host' port='6789' />
  </source>
  <target dev='vda' bus='virtio' />
</disk>
```

Substitua `monitor-host` pelo nome do seu host e substitua o nome do pool e/ou da imagem, conforme necessário. Você pode adicionar várias entradas `<host>` aos Ceph Monitors. O atributo `dev` é o nome do dispositivo lógico que aparecerá no diretório `/dev` da VM. O atributo de barramento opcional indica o tipo de dispositivo de disco a ser emulado. As configurações válidas são específicas do driver (por exemplo, `ide`, `scsi`, `virtio`, `xen`, `usb` ou `sata`).

4. Grave o arquivo.
5. Se a autenticação estiver habilitada no cluster do Ceph (que é o padrão), você deverá gerar um segredo. Abra o editor de sua preferência e crie um arquivo chamado `secret.xml` com o seguinte conteúdo:

```
<secret ephemeral='no' private='no'>
```

```
<usage type='ceph'>
    <name>client.libvirt secret</name>
</usage>
</secret>
```

6. Defina o segredo.

```
# virsh secret-define --file secret.xml
<uuid of secret is output here>
```

7. Obtenha a chave `client.libvirt` e grave a string de chave em um arquivo.

```
cephuser@adm > ceph auth get-key client.libvirt | sudo tee client.libvirt.key
```

8. Defina o UUID do segredo.

```
# virsh secret-set-value --secret uuid of secret \
--base64 $(cat client.libvirt.key) && rm client.libvirt.key secret.xml
```

Você também deve definir o segredo manualmente adicionando a seguinte entrada `<auth>` ao elemento `<disk>` que você inseriu antes (substituindo o valor do uuid pelo resultado do exemplo de linha de comando acima).

```
# virsh edit libvirt-virtual-machine
```

Em seguida, adicione o elemento `<auth></auth>` ao arquivo de configuração do domínio:

```
...
</source>
<auth username='libvirt'>
    <secret type='ceph' uuid='9ec59067-fdbc-a6c0-03ff-df165c0587b8' />
</auth>
<target ...
```



## Nota

O ID do exemplo é `libvirt`, e não o nome do Ceph `client.libvirt`, conforme gerado na etapa 2 da [Seção 26.1, “Configurando o Ceph com libvirt”](#). Use o componente de ID do nome do Ceph que você gerou. Se, por algum motivo, você precisar gerar novamente o segredo, terá que executar **`sudo virsh secret-undefine uuid`** antes de executar **`sudo virsh secret-set-value`** outra vez.

## 26.5 Resumo

Após configurar a VM para uso com o Ceph, você poderá iniciá-la. Para verificar se a VM e o Ceph estão se comunicando, você pode executar os procedimentos a seguir.

1. Verifique se o Ceph está em execução:

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

2. Verifique se a VM está em execução:

```
# virsh list
```

3. Verifique se a VM está se comunicando com o Ceph. Substitua vm-domain-name pelo nome do domínio da sua VM:

```
# virsh qemu-monitor-command --hmp vm-domain-name 'info block'
```

4. Verifique se o dispositivo de &target dev='hdb' bus='ide' /> aparece em /dev ou em /proc/partitions:

```
> ls /dev  
> cat /proc/partitions
```



## 27 Ceph como back end para instância de KVM QEMU

O caso de uso mais frequente do Ceph envolve fornecer imagens de dispositivos de blocos para máquinas virtuais. Por exemplo, um usuário pode criar uma imagem “perfeita” com um OS e qualquer software relevante em uma configuração ideal. Em seguida, ele captura um instantâneo da imagem. Por fim, ele clona o instantâneo (normalmente, várias vezes. Consulte a [Seção 20.3, “Instantâneos”](#) para obter detalhes). A capacidade de criar clones de cópia em gravação de um instantâneo significa que o Ceph pode provisionar imagens de dispositivos de blocos para máquinas virtuais rapidamente, pois o cliente não precisa fazer download de uma imagem inteira cada vez que capturar uma nova máquina virtual.

Os dispositivos de blocos do Ceph podem ser integrados às máquinas virtuais QEMU. Para obter mais informações sobre a KVM QEMU, consulte <https://documentation.suse.com/sles/15-SP1/single-html/SLES-virtualization/#part-virt-qemu>.

### 27.1 Instalando qemu-block-rbd

Para usar os dispositivos de blocos do Ceph, o QEMU precisa ter o driver apropriado instalado. Verifique se o pacote `qemu-block-rbd` está instalado, e instale-o se necessário:

```
# zypper install qemu-block-rbd
```

### 27.2 Usando o QEMU

A linha de comando do QEMU espera você especificar o nome do pool e da imagem. Você também pode especificar o nome de um instantâneo.

```
qemu-img command options \  
rbd:pool-name/image-name@snapshot-name:option1=value1:option2=value2...
```

Por exemplo, se você especificar as opções `id` e `conf`, a aparência deverá ser a seguinte:

```
qemu-img command options \  
rbd:pool_name/image_name:id=glance:conf=/etc/ceph/ceph.conf
```

## 27.3 Criando imagens com o QEMU

Você pode criar uma imagem de dispositivo de blocos no QEMU. Você deve especificar `rbd`, o nome do pool e o nome da imagem que deseja criar. Você também deve especificar o tamanho da imagem.

```
qemu-img create -f raw rbd:pool-name/image-name size
```

Por exemplo:

```
qemu-img create -f raw rbd:pool1/image1 10G
Formatting 'rbd:pool1/image1', fmt=raw size=10737418240 nocow=off cluster_size=0
```



### Importante

O formato de dados brutos é realmente a única opção de formato sensível para usar com o RBD. Tecnicamente, você pode usar outros formatos compatíveis com o QEMU, como `qcow2`, mas isso aumentará o overhead e também tornará o volume não seguro para migração dinâmica da máquina virtual quando o cache estiver habilitado.

## 27.4 Redimensionando imagens com o QEMU

É possível redimensionar uma imagem de dispositivo de blocos usando o QEMU. Você deve especificar `rbd`, o nome do pool e o nome da imagem que deseja redimensionar. Você também deve especificar o tamanho da imagem.

```
qemu-img resize rbd:pool-name/image-name size
```

Por exemplo:

```
qemu-img resize rbd:pool1/image1 9G
Image resized.
```

## 27.5 Recuperando informações da imagem com o QEMU

Você pode recuperar informações da imagem do dispositivo de blocos usando o QEMU. Você deve especificar `rbd`, o nome do pool e o nome da imagem.

```
qemu-img info rbd:pool-name/image-name
```

Por exemplo:

```
qemu-img info rbd:pool1/image1
image: rbd:pool1/image1
file format: raw
virtual size: 9.0G (9663676416 bytes)
disk size: unavailable
cluster_size: 4194304
```

## 27.6 Executando o QEMU com o RBD

O QEMU pode acessar uma imagem como dispositivo de blocos virtual diretamente pelo `librbd`. Isso evita um switch de contexto adicional e pode ser melhor do que o cache do RBD.

Você pode usar `qemu-img` para converter as imagens existentes de máquinas virtuais em imagens de dispositivos de blocos do Ceph. Por exemplo, se você tem uma imagem `qcow2`, pode executar:

```
qemu-img convert -f qcow2 -O raw sles12.qcow2 rbd:pool1/sles12
```

Para executar uma inicialização de máquina virtual dessa imagem, você pode executar:

```
# qemu -m 1024 -drive format=raw,file=rbd:pool1/sles12
```

O cache do RBD pode melhorar o desempenho significativamente. As opções de cache do QEMU controlam o cache do `librbd`:

```
# qemu -m 1024 -drive format=rbd,file=rbd:pool1/sles12,cache=writeback
```

Para obter mais informações sobre cache do RBD, consulte a [Seção 20.5, “Configurações de cache”](#).

## 27.7 Habilitando descarte e TRIM

Os dispositivos de blocos do Ceph suportam a operação de descarte. Isso significa que um convidado pode enviar solicitações TRIM para permitir que um dispositivo de blocos do Ceph reaproveite o espaço não usado. Para habilitar esse recurso, monte o `XFS` com a opção de descarte.

Para que ele fique disponível ao convidado, é necessário habilitá-lo explicitamente no dispositivo de blocos. Para fazer isso, você deve especificar `discard_granularity` associado à unidade:

```
# qemu -m 1024 -drive format=raw,file=rbd:pool1/sles12,id=drive1,if=none \
```

```
-device driver=ide-hd,drive=drive1,discard_granularity=512
```



## Nota

O exemplo acima usa o driver IDE. O driver virtio não suporta descarte.

Se for usar o `libvirt`, edite o arquivo de configuração do domínio `libvirt` usando `virsh edit` para incluir o valor `xmlns:qemu`. Em seguida, adicione `qemu:commandline` block como filho desse domínio. O exemplo a seguir mostra como definir dois dispositivos com `qemu id=` com valores `discard_granularity` diferentes.

```
<domain type='kvm' xmlns:qemu='http://libvirt.org/schemas/domain/qemu/1.0'>
  <qemu:commandline>
    <qemu:arg value='-set' />
    <qemu:arg value='block.scsi0-0-0.discard_granularity=4096' />
    <qemu:arg value='-set' />
    <qemu:arg value='block.scsi0-0-1.discard_granularity=65536' />
  </qemu:commandline>
</domain>
```

## 27.8 Definindo as opções de cache do QEMU

As opções de cache do QEMU correspondem às seguintes configurações de Cache do Ceph RBD. Writeback:

```
rbd_cache = true
```

Writethrough:

```
rbd_cache = true
rbd_cache_max_dirty = 0
```

Nenhuma:

```
rbd_cache = false
```

As configurações de cache do QEMU anulam as configurações padrão do Ceph (configurações que não são explicitamente definidas no arquivo de configuração do Ceph). Se você definir explicitamente as configurações de Cache do RBD no arquivo de configuração do Ceph (consulte

a [Seção 20.5, “Configurações de cache”](#)), as configurações do Ceph anularão as configurações de cache do QEMU. Se você definir as configurações de cache na linha de comando do QEMU, elas anularão as configurações do arquivo de configuração do Ceph.

## VI Configurando um cluster

- 28 Configuração do cluster do Ceph **381**
- 29 Módulos do Ceph Manager **402**
- 30 Autenticação com cephx **407**

## 28 Configuração do cluster do Ceph

Este capítulo descreve como configurar o cluster do Ceph por meio das opções de configuração.

### 28.1 Configurar o arquivo `ceph.conf`

O `cephadm` usa um arquivo `ceph.conf` básico que contém apenas um conjunto mínimo de opções para conexão com MONs, autenticação e busca de informações de configuração. Na maioria dos casos, ele está limitado à opção `mon_host` (embora isso possa ser evitado com o uso de registros DNS SRV).



#### Importante

O `ceph.conf` não serve mais como um local central para armazenar a configuração do cluster, sendo preferível o banco de dados de configuração (consulte o [Seção 28.2, “Banco de dados de configuração”](#)).

Se você ainda precisa mudar a configuração do cluster por meio do arquivo `ceph.conf`, por exemplo, porque usa um cliente que não suporta opções de leitura do banco de dados de configuração, você precisa executar o comando a seguir e cuidar da manutenção e da distribuição do arquivo `ceph.conf` em todo o cluster:

```
cephuser@adm > ceph config set mgr mgr/cephadm/manage_etc_ceph_ceph_conf false
```

#### 28.1.1 Acessando o `ceph.conf` dentro das imagens de container

Embora os daemons do Ceph sejam executados dentro de containers, você ainda pode acessar o arquivo de configuração `ceph.conf` deles. Ele é *montado por vínculo* como o seguinte arquivo no sistema host:

```
/var/lib/ceph/CLUSTER_FSID/DAEMON_NAME/config
```

Substitua `CLUSTER_FSID` pelo FSID exclusivo do cluster em execução, conforme retornado pelo comando `ceph fsid`, e `DAEMON_NAME` pelo nome do daemon específico, conforme listado pelo comando `ceph orch ps`. Por exemplo:

```
/var/lib/ceph/b4b30c6e-9681-11ea-ac39-525400d7702d/osd.2/config
```

Para modificar a configuração de um daemon, edite seu arquivo `config` e reinicie-o:

```
# systemctl restart ceph-CLUSTER_FSID-DAEMON_NAME
```

Por exemplo:

```
# systemctl restart ceph-b4b30c6e-9681-11ea-ac39-525400d7702d-osd.2
```



## Importante

Todas as configurações personalizadas serão perdidas depois que o cephadm implantar o daemon novamente.

## 28.2 Banco de dados de configuração

Os Ceph Monitors gerenciam um banco de dados central de opções de configuração que afetam o comportamento de todo o cluster.

### 28.2.1 Configurando seções e máscaras

As opções de configuração armazenadas pelo MON podem residir em uma seção *global*, de *tipo de daemon* ou de *daemon específico*. Além disso, as opções também podem ter uma *máscara* associada a elas para restringir ainda mais os daemons ou clientes aos quais a opção é aplicada. As máscaras têm dois formatos:

- `TYPE:LOCATION`, em que `TYPE` é uma propriedade CRUSH, como `rack` ou `host`, e `LOCATION` é o valor dessa propriedade.  
Por exemplo, `host:example_host` limitará a opção apenas aos daemons ou clientes executados em um host específico.
- `CLASS:DEVICE_CLASS`, em que `DEVICE_CLASS` é o nome de uma classe de dispositivo CRUSH, como `hdd` ou `ssd`. Por exemplo, `class:ssd` limitará a opção apenas aos OSDs suportados por SSDs. Essa máscara não tem efeito para daemons ou clientes não OSD.



## 28.2.2 Definindo e lendo as opções de configuração

Use os comandos a seguir para definir ou ler as opções de configuração do cluster. O parâmetro *WHO* pode ser um nome de seção, uma máscara ou uma combinação dos dois, separados por uma barra (/). Por exemplo, *osd/rack:foo* representa todos os daemons OSD no rack chamado *foo*.

### **`ceph config dump`**

Despeja todo o banco de dados de configuração de um cluster inteiro.

### **`ceph config get WHO`**

Despeja a configuração de um daemon ou cliente específico (por exemplo, *mds.a*), conforme armazenado no banco de dados de configuração.

### **`ceph config set WHO OPTION VALUE`**

Define a opção de configuração como o valor especificado no banco de dados de configuração.

### **`ceph config show WHO`**

Mostra a configuração de execução relatada para um daemon em execução. Essas configurações poderão ser diferentes das armazenadas pelos monitores se também houver arquivos de configuração locais em uso ou se as opções tiverem sido anuladas na linha de comando ou em tempo de execução. A origem dos valores de opção é relatada como parte da saída.

### **`ceph config assimilate-conf -i INPUT_FILE -o OUTPUT_FILE`**

Importa um arquivo de configuração especificado como *INPUT\_FILE* e armazena todas as opções válidas no banco de dados de configuração. Quaisquer configurações não reconhecidas, inválidas ou que não possam ser controladas pelo monitor serão retornadas em um arquivo abreviado armazenado como *OUTPUT\_FILE*. Esse comando é útil para a transição de arquivos de configuração legados para a configuração centralizada baseada no monitor.

## 28.2.3 Configurando daemons em tempo de execução

Na maioria dos casos, o Ceph permite que você faça mudanças na configuração de um daemon em tempo de execução. Isso é útil, por exemplo, quando você precisa aumentar ou diminuir a quantidade de saída de registro ou para executar a otimização do cluster em tempo de execução.

Você pode atualizar os valores das opções de configuração com o seguinte comando:

```
cephuser@adm > ceph config set DAEMON OPTION VALUE
```

Por exemplo, para ajustar o nível de registro de depuração em um OSD específico, execute:

```
cephuser@adm > ceph config set osd.123 debug_ms 20
```



## Nota

Se a mesma opção também for personalizada em um arquivo de configuração local, a configuração do monitor será ignorada porque tem prioridade mais baixa do que o arquivo de configuração.

### 28.2.3.1 Anulando valores

Você pode modificar temporariamente o valor de uma opção usando os subcomandos **tell** ou **daemon**. Essa modificação afeta apenas o processo em execução e é descartada após a reinicialização do daemon ou do processo.

Há duas maneiras de anular valores:

- Usar o subcomando **tell** para enviar uma mensagem a um daemon específico de qualquer nó do cluster:

```
cephuser@adm > ceph tell DAEMON config set OPTION VALUE
```

Por exemplo:

```
cephuser@adm > ceph tell osd.123 config set debug_osd 20
```



## Dica

O subcomando **tell** aceita curingas como identificadores de daemons. Por exemplo, para ajustar o nível de depuração em todos os daemons OSD, execute:

```
cephuser@adm > ceph tell osd.* config set debug_osd 20
```

- Usar o subcomando **daemon** para se conectar a um processo de daemon específico por meio de um soquete em `/var/run/ceph` do nó onde o processo está sendo executado:

```
cephuser@adm > cephadm enter --name osd.ID -- ceph daemon DAEMON config  
set OPTION VALUE
```

Por exemplo:

```
cephuser@adm > cephadm enter --name osd.4 -- ceph daemon osd.4 config set debug_osd  
20
```



## Dica

Ao ver as configurações de tempo de execução com o comando **ceph config show** (consulte o [Seção 28.2.3.2, “Visualizando as configurações de tempo de execução”](#)), os valores temporariamente anulados serão mostrados com a origem override.

### 28.2.3.2 Visualizando as configurações de tempo de execução

Para ver todas as opções definidas para um daemon:

```
cephuser@adm > ceph config show-with-defaults osd.0
```

Para ver todas as opções não padrão definidas para um daemon:

```
cephuser@adm > ceph config show osd.0
```

Para inspecionar uma opção específica:

```
cephuser@adm > ceph config show osd.0 debug_osd
```

Você também pode se conectar a um daemon em execução do nó onde seu processo está sendo executado e observar sua configuração:

```
cephuser@adm > cephadm enter --name osd.0 -- ceph daemon osd.0 config show
```

Para ver apenas as configurações não padrão:

```
cephuser@adm > cephadm enter --name osd.0 -- ceph daemon osd.0 config diff
```

Para inspecionar uma opção específica:

```
cephuser@adm > cephadm enter --name osd.0 -- ceph daemon osd.0 config get debug_osd
```

## 28.3 config-key armazenar

config-key é um serviço de finalidade geral oferecido pelos Ceph Monitors. Ele simplifica o gerenciamento de chaves de configuração armazenando pares de chave-valor permanentemente. O serviço config-key é usado principalmente pelas ferramentas e daemons do Ceph.



### Dica

Após adicionar uma nova chave ou modificar uma existente, reinicie o serviço afetado para que as mudanças entrem em vigor. Encontre mais detalhes sobre a operação dos serviços do Ceph no *Capítulo 14, Operação de serviços do Ceph*.

Use o comando `ceph config-key` para operar o armazenamento config-key. O comando **config-key** usa os seguintes subcomandos:

**ceph config-key rm KEY**

Apaga a chave especificada.

**ceph config-key exists KEY**

Verifica a existência da chave especificada.

**ceph config-key get KEY**

Recupera o valor da chave especificada.

**ceph config-key ls**

Lista todas as chaves.

### **`ceph config-key dump`**

Despeja todas as chaves e seus valores.

### **`ceph config-key set KEY VALUE`**

Armazena a chave especificada com o valor fornecido.

## 28.3.1 iSCSI Gateway

O Gateway iSCSI usa o armazenamento `config-key` para gravar ou ler suas opções de configuração. Todas as chaves relacionadas ao Gateway iSCSI recebem a string `iscsi` como prefixo, por exemplo:

```
iscsi/trusted_ip_list
iscsi/api_port
iscsi/api_user
iscsi/api_password
iscsi/api_secure
```

Por exemplo, se você precisar de dois conjuntos de opções de configuração, estenda o prefixo com outra palavra-chave descritiva, como `datacenterA` e `datacenterB`:

```
iscsi/datacenterA/trusted_ip_list
iscsi/datacenterA/api_port
[...]
iscsi/datacenterB/trusted_ip_list
iscsi/datacenterB/api_port
[...]
```

## 28.4 Ceph OSD e BlueStore

### 28.4.1 Configurando o dimensionamento automático do cache

É possível configurar o BlueStore para redimensionar automaticamente os caches quando `tc_malloc` está configurado como o alocador de memória e a configuração `bluestore_cache_autotune` está habilitada. Por padrão, essa opção está habilitada atualmente. O BlueStore tentará manter o uso de memória heap OSD abaixo de um tamanho de destino designado por meio da opção de configuração `osd_memory_target`. Esse é um algoritmo de

melhor esforço, e os caches não terão um tamanho menor do que o valor especificado por `osd_memory_cache_min`. Os índices do cache serão escolhidos com base em uma hierarquia de prioridades. Se as informações de prioridade não estiverem disponíveis, as opções `bluestore_cache_meta_ratio` e `bluestore_cache_kv_ratio` serão usadas como fallbacks.

#### `bluestore_cache_autotune`

Ajusta automaticamente os índices atribuídos aos diferentes caches do BlueStore respeitando os valores mínimos. O padrão é `True` (Verdadeiro).

#### `osd_memory_target`

Quando `tc_malloc` e `bluestore_cache_autotune` estão habilitados, tenta manter essa quantidade de bytes mapeada na memória.



#### Nota

Talvez esse valor não corresponda exatamente ao uso de memória RSS do processo. Embora a quantidade total de memória heap mapeada pelo processo deva geralmente ficar próxima a esse destino, não há garantia de que o kernel recuperará de fato a memória que não foi mapeada.

#### `osd_memory_cache_min`

Quando `tc_malloc` e `bluestore_cache_autotune` estão habilitados, define a quantidade mínima de memória usada para os caches.



#### Nota

Se for definido um valor muito baixo, o resultado poderá ser uma ultrapaginação do cache.

## 28.5 Gateway de Objetos do Ceph

Você pode influenciar o comportamento do Gateway de Objetos por meio de várias opções. Se uma opção não for especificada, seu valor padrão será usado. Veja a seguir uma lista completa das opções do Gateway de Objetos:

### 28.5.1 Configurações gerais

#### `rgw_frontends`

Configura o(s) front end(s) HTTP. Especifique vários front ends em uma lista delimitada por vírgula. Cada configuração de front end pode incluir uma lista de opções separadas por espaços, com cada opção no formato “chave = valor” ou “chave”. O padrão é `beast port=7480`.

#### `rgw_data`

Define o local dos arquivos de dados para o Gateway de Objetos. O padrão é `/var/lib/ceph/radosgw/CLUSTER_ID`.

#### `rgw_enable_apis`

Habilita as APIs especificadas. O padrão é “s3, swift, swift\_auth, admin All APIs”.

#### `rgw_cache_enabled`

Habilita ou desabilita o cache do Gateway de Objetos. O padrão é `true`.

#### `rgw_cache_lru_size`

O número de entradas no cache do Gateway de Objetos. O padrão é 10000.

#### `rgw_socket_path`

O caminho do soquete de domínio. `FastCgiExternalServer` usa esse soquete. Se você não especificar um caminho de soquete, o Gateway de Objetos não será executado como um servidor externo. O caminho que você especifica aqui precisa ser o mesmo especificado no arquivo `rgw.conf`.

#### `rgw_fcgi_socket_backlog`

A lista de pendências do soquete para o fcgi. O padrão é 1024.

#### `rgw_host`

O host para a instância do Gateway de Objetos. Pode ser um endereço IP ou um nome de host. O padrão é `0.0.0.0`.

**rgw\_port**

O número da porta que a instância usa para escutar as solicitações. Se não for especificado, o Gateway de Objetos executará o FastCGI externo.

**rgw\_dns\_name**

O nome DNS do domínio atendido.

**rgw\_script\_uri**

O valor alternativo para o SCRIPT\_URI, se não for definido na solicitação.

**rgw\_request\_uri**

O valor alternativo para o REQUEST\_URI, se não for definido na solicitação.

**rgw\_print\_continue**

Habilite 100-continue se for operacional. O padrão é true.

**rgw\_remote\_addr\_param**

O parâmetro de endereço remoto. Por exemplo, o campo HTTP com o endereço remoto ou o endereço X-Forwarded-For, se um proxy reverso for operacional. O padrão é REMOTE\_ADDR.

**rgw\_op\_thread\_timeout**

O tempo de espera em segundos para threads abertos. O padrão é 600.

**rgw\_op\_thread\_suicide\_timeout**

O tempo de espera em segundos para o processo do Gateway de Objetos ser encerrado. Desabilitado se definido como 0 (padrão).

**rgw\_thread\_pool\_size**

Número de threads para o servidor Beast. Aumente para um valor mais alto se você precisa atender a mais solicitações. O padrão é 100 threads.

**rgw\_num\_rados\_handles**

O número de manipuladores do cluster RADOS para o Gateway de Objetos. Agora, cada thread do worker do Gateway de Objetos precisa selecionar um manipulador do RADOS para sua vida útil. Essa opção pode ser descontinuada e removida em versões futuras. O padrão é 1.

**rgw\_num\_control\_oids**

O número de objetos de notificação usados para sincronização de cache entre instâncias diferentes do Gateway de Objetos. O padrão é 8.

**rgw\_init\_timeout**

O número de segundos até o Gateway de Objetos desistir da inicialização. O padrão é 30.



#### **rgw\_mime\_types\_file**

O caminho e o local dos tipos MIME. Usado para detecção automática de tipos de objeto do Swift. O padrão é /etc/mime.types.

#### **rgw\_gc\_max\_objs**

O número máximo de objetos que podem ser processados pela coleta de lixo em um ciclo de processamento de coleta de lixo. O padrão é 32.

#### **rgw\_gc\_obj\_min\_wait**

O tempo mínimo de espera antes que o objeto possa ser removido e processado pela coleta de lixo. O padrão é 2\*3600.

#### **rgw\_gc\_processor\_max\_time**

O tempo máximo entre o início dos dois ciclos consecutivos de processamento da coleta de lixo. O padrão é 3600.

#### **rgw\_gc\_processor\_period**

O tempo do ciclo para o processamento da coleta de lixo. O padrão é 3600.

#### **rgw\_s3\_success\_create\_obj\_status**

A resposta alternativa de status de êxito para create-obj. O padrão é 0.

#### **rgw\_resolve\_cname**

Se o Gateway de Objetos deve usar o registro DNS CNAME do campo de nome de host da solicitação (se o nome de host não for igual ao nome name DNS do Gateway de Objetos). O padrão é false.

#### **rgw\_obj\_stripe\_size**

O tamanho de uma faixa de objetos do Gateway de Objetos. O padrão é 4 << 20.

#### **rgw\_extended\_http\_attrs**

Adicione um novo conjunto de atributos que podem ser definidos em uma entidade (por exemplo, um usuário, um compartimento de memória ou um objeto). Esses atributos extras podem ser definidos por meio de campos de cabeçalho HTTP ao especificar a entidade ou modificá-la usando o método POST. Se definidos, esses atributos serão retornados como campos HTTP ao solicitar GET/HEAD na entidade. O padrão é content\_foo, content\_bar, x-foo-bar.

#### **rgw\_exit\_timeout\_secs**

Por quantos segundos esperar por um processo antes de sair incondicionalmente. O padrão é 120.

#### `rgw_get_obj_window_size`

O tamanho da janela em bytes para uma solicitação única de objeto. O padrão é 16 << 20.

#### `rgw_get_obj_max_req_size`

O tamanho máximo da solicitação de uma única operação GET enviada para o Cluster de Armazenamento do Ceph. O padrão é 4 << 20.

#### `rgw_relaxed_s3_bucket_names`

Habilita regras de nome de compartimento de memória S3 flexíveis para compartimentos de memória na região EUA. O padrão é false.

#### `rgw_list_buckets_max_chunk`

O número máximo de compartimentos de memória a serem recuperados em uma única operação ao listar compartimentos de memória de usuário. O padrão é 1000.

#### `rgw_override_bucket_index_max_shards`

Representa o número de fragmentos para o objeto Índice do compartimento de memória. A configuração 0 (padrão) indica que não há fragmentação. Não é recomendado definir um valor muito grande (por exemplo, 1000), pois isso aumenta o custo para listagem de compartimentos de memória. Essa variável deve ser definida no cliente ou nas seções globais para ser automaticamente aplicada aos comandos `radosgw-admin`.

#### `rgw_curl_wait_timeout_ms`

O tempo de espera em milissegundos para determinadas chamadas `curl`. O padrão é 1000.

#### `rgw_copy_obj_progress`

Habilita a saída do progresso do objeto durante operações longas de cópia. O padrão é true.

#### `rgw_copy_obj_progress_every_bytes`

O mínimo de bytes entre a saída do progresso da cópia. O padrão é 1024\*1024.

#### `rgw_admin_entry`

O ponto de entrada para um URL de solicitação de admin. O padrão é admin.

#### `rgw_content_length_compat`

Habilita o processamento de compatibilidade de solicitações FCGI com os dois comandos `CONTENT_LENGTH` e `HTTP_CONTENT_LENGTH` definidos. O padrão é false.

#### `rgw_bucket_quota_ttl`

Por quanto tempo em segundos as informações de cota armazenadas em cache são confiáveis. Após esse tempo, as informações de cota serão buscadas novamente do cluster. O padrão é 600.

#### `rgw_user_quota_bucket_sync_interval`

Por quanto tempo em segundos as informações de cota de compartimento de memória são acumuladas antes da sincronização com o cluster. Durante esse tempo, outras instâncias do Gateway de Objetos não verão as mudanças nas estatísticas de cota de compartimento de memória relacionadas às operações nesta instância. O padrão é 180.

#### `rgw_user_quota_sync_interval`

Por quanto tempo em segundos as informações de cota de usuário são acumuladas antes da sincronização com o cluster. Durante esse tempo, outras instâncias do Gateway de Objetos não verão as mudanças nas estatísticas de cota de usuário relacionadas às operações nesta instância. O padrão é 180.

#### `rgw_bucket_default_quota_max_objects`

Número máximo padrão de objetos por compartimento de memória. Ele será definido com base nos novos usuários, se nenhuma outra cota for especificada, e não terá efeito sobre os usuários existentes. Essa variável deve ser definida no cliente ou nas seções globais para ser automaticamente aplicada aos comandos `radosgw-admin`. O padrão é -1.

#### `rgw_bucket_default_quota_max_size`

Capacidade máxima padrão por compartimento de memória em bytes. Ela será definida com base nos novos usuários, se nenhuma outra cota for especificada, e não terá efeito sobre os usuários existentes. O padrão é -1.

#### `rgw_user_default_quota_max_objects`

Número máximo padrão de objetos para um usuário. Isso inclui todos os objetos em todos os compartimentos de memória de propriedade do usuário. Ele será definido com base nos novos usuários, se nenhuma outra cota for especificada, e não terá efeito sobre os usuários existentes. O padrão é -1.

#### `rgw_user_default_quota_max_size`

O valor da cota de tamanho máximo do usuário em bytes definido com base nos novos usuários, se nenhuma outra cota for especificada. Ele não tem efeito sobre os usuários existentes. O padrão é -1.

#### rgw\_verify\_ssl

Verificar os certificados SSL ao fazer as solicitações. O padrão é true.

#### rgw\_max\_chunk\_size

Tamanho máximo de um pacote de dados que será lido em uma única operação. Aumentar o valor para 4 MB (4194304) proporcionará um melhor desempenho ao processar objetos grandes. O padrão é 128 KB (131072).

### CONFIGURAÇÕES MULTISITE

#### rgw\_zone

O nome da zona para a instância do gateway. Se nenhuma zona for definida, um padrão do tamanho do cluster poderá ser configurado com o comando **radosgw-admin zone default**.

#### rgw\_zonegroup

O nome do grupo de zonas para a instância do gateway. Se nenhum grupo de zonas for definido, um padrão do tamanho do cluster poderá ser configurado com o comando **radosgw-admin zonegroup default**.

#### rgw\_realm

O nome do domínio Kerberos para a instância do gateway. Se nenhum domínio Kerberos for definido, um padrão do tamanho do cluster poderá ser configurado com o comando **radosgw-admin realm default**.

#### rgw\_run\_sync\_thread

Se houver outras zonas no domínio Kerberos das quais sincronizar, gere threads para processar a sincronização dos dados e metadados. O padrão é true.

#### rgw\_data\_log\_window

As janelas de entradas de registro de dados em segundos. O padrão é 30.

#### rgw\_data\_log\_changes\_size

O número de entradas na memória a serem armazenadas para o registro de mudanças de dados. O padrão é 1000.

#### rgw\_data\_log\_obj\_prefix

O prefixo do nome do objeto para o registro de dados. O padrão é "data\_log".

#### rgw\_data\_log\_num\_shards

O número de fragmentos (objetos) nos quais manter o registro de mudanças de dados. O padrão é 128.

#### `rgw_md_log_max_shards`

O número máximo de fragmentos para o registro de metadados. O padrão é 64.

### CONFIGURAÇÕES DO SWIFT

#### `rgw_enforce_swift_acls`

Impõe as configurações da Lista de Controle de Acesso (ACL, Access Control List) do Swift. O padrão é `true`.

#### `rgw_swift_token_expiration`

O tempo em segundos para expirar um token Swift. O padrão é 24\*3600.

#### `rgw_swift_url`

O URL para a API Swift do Gateway de Objetos do Ceph.

#### `rgw_swift_url_prefix`

O prefixo de URL para o StorageURL do Swift que fica na frente da parte `/v1`. Isso permite executar várias instâncias do Gateway no mesmo host. Para compatibilidade, se essa variável de configuração for definida como vazia, o `/swift` padrão será usado. Use o prefixo `/` explícito para iniciar o StorageURL na raiz.



### Atenção

Se essa opção for definida como `/`, ela não funcionará se a API do S3 estiver habilitada. Saiba que a desabilitação do S3 impossibilita a implantação do Gateway de Objetos na configuração multissite.

#### `rgw_swift_auth_url`

URL padrão para verificar os tokens de autenticação v1 quando a autenticação interna do Swift não é usada.

#### `rgw_swift_auth_entry`

O ponto de entrada para um URL de autenticação do Swift. O padrão é `auth`.

#### `rgw_swift_versioning_enabled`

Habilita o Controle de Versão de Objeto da API de Armazenamento de Objetos do OpenStack. Isso permite que os clientes insiram o atributo `X-Versions-Location` nos containers que devem ter o controle de versão. O atributo especifica o nome do container que armazena as versões arquivadas. Ele deve ser de propriedade do mesmo usuário que

o container com controle de versão, por motivos de verificação de controle de acesso. As ACLs *não* são levadas em consideração. Não é possível controlar a versão desses containers usando o mecanismo de controle de versão de objetos do S3. O padrão é false.

## CONFIGURAÇÕES DE REGISTRO

### rgw\_log\_nonexistent\_bucket

Permite que o Gateway de Objetos registre uma solicitação para um compartimento de memória não existente. O padrão é false.

### rgw\_log\_object\_name

O formato de registro para um nome de objeto. Consulte a página de manual [man 1 date](#) para obter detalhes sobre especificadores de formato. O padrão é %Y-%m-%d-%H-%i-%n.

### rgw\_log\_object\_name\_utc

Se um nome de objeto registrado inclui horário UTC. Se definido como false (padrão), o horário local será usado.

### rgw\_usage\_max\_shards

O número máximo de fragmentos para registro de uso. O padrão é 32.

### rgw\_usage\_max\_user\_shards

O número máximo de fragmentos usados para registro de uso de um único usuário. O padrão é 1.

### rgw\_enable\_ops\_log

Habilitar o registro para cada operação bem-sucedida do Gateway de Objetos. O padrão é false.

### rgw\_enable\_usage\_log

Habilitar o registro de uso. O padrão é false.

### rgw\_ops\_log\_rados

Se o registro de operações deve ser gravado no back end do Cluster de Armazenamento do Ceph. O padrão é true.

### rgw\_ops\_log\_socket\_path

O soquete de domínio do Unix para gravação de registros de operações.

### rgw\_ops\_log\_data\_backlog

O tamanho máximo dos dados da lista de pendências para registros de operações gravados em um soquete de domínio do Unix. O padrão é 5 < < 20.

#### `rgw_usage_log_flush_threshold`

O número de entradas fundidas modificadas no registro de uso antes do descarregamento sincronizado. O padrão é 1024.

#### `rgw_usage_log_tick_interval`

Descarregar os dados de registro de uso pendentes a cada “n” segundos. O padrão é 30.

#### `rgw_log_http_headers`

Lista delimitada por vírgula de cabeçalhos HTTP para incluir nas entradas de registro. Os nomes de cabeçalho não diferenciam maiúsculas de minúsculas e usam o nome completo do cabeçalho com palavras separadas por sublinhados. Por exemplo, “http\_x\_forwarded\_for”, “http\_x\_special\_k”.

#### `rgw_intent_log_object_name`

O formato de registro para o nome do objeto de registro de intenções. Consulte a página de manual **man 1 date** para obter detalhes sobre especificadores de formato. O padrão é “%Y-%m-%d-%i-%n”.

#### `rgw_intent_log_object_name_utc`

Se o nome do objeto de registro de intenções inclui horário UTC. Se definido como false (padrão), o horário local será usado.

### CONFIGURAÇÕES DO KEYSTONE

#### `rgw_keystone_url`

O URL para o servidor Keystone.

#### `rgw_keystone_api_version`

A versão (2 ou 3) da API OpenStack Identity que deve ser usada para comunicação com o servidor Keystone. O padrão é 2.

#### `rgw_keystone_admin_domain`

O nome do domínio do OpenStack com o privilégio de administrador ao usar a API OpenStack Identity v3.

#### `rgw_keystone_admin_project`

O nome do projeto do OpenStack com o privilégio de administrador ao usar a API OpenStack Identity v3. Se não for definido, o valor de **rgw keystone admin tenant** será usado.

#### `rgw_keystone_admin_token`

O token de administrador do Keystone (segredo compartilhado). No Gateway de Objetos, a autenticação com o token de administrador tem prioridade em relação à autenticação com as credenciais de administrador (opções `rgw keystone admin user`, `rgw keystone admin password`, `rgw keystone admin tenant`, `rgw keystone admin project` e `rgw keystone admin domain`). O recurso de token de administrador é considerado descontinuado.

#### `rgw_keystone_admin_tenant`

O nome do locatário do OpenStack com o privilégio de administrador (Locatário de Serviço) ao usar a API OpenStack Identity v2.

#### `rgw_keystone_admin_user`

O nome do usuário do OpenStack com o privilégio de administrador para autenticação do Keystone (Usuário de Serviço) ao usar a API OpenStack Identity v2.

#### `rgw_keystone_admin_password`

A senha para o usuário administrador do OpenStack ao usar a API OpenStack Identity v2.

#### `rgw_keystone_accepted_roles`

As funções necessárias para atender às solicitações. O padrão é “Member, admin”.

#### `rgw_keystone_token_cache_size`

O número máximo de entradas em cada cache de token do Keystone. O padrão é 10000.

#### `rgw_keystone_revocation_interval`

O número de segundos entre as verificações de revogação de token. O padrão é 15\*60.

#### `rgw_keystone_verify_ssl`

Verificar os certificados SSL ao fazer as solicitações de token para o Keystone. O padrão é `true`.

### 28.5.1.1 Notas adicionais

#### `rgw_dns_name`

Permite que os clientes usem compartimentos de memória no estilo `vhost`.

O acesso no estilo `vhost` indica o uso de `bucketname.s3-endpoint/object-path`. Esta é uma comparação com o acesso no estilo `path`: `s3-endpoint/bucket/object`

Se `rgw dns name` for definido, verifique se o cliente S3 está configurado para direcionar as solicitações ao endpoint especificado por `rgw dns name`.



## 28.5.2 Configurando front ends HTTP

### 28.5.2.1 Beast

#### porta, ssl\_port

Números de porta de escuta IPv4 e IPv6. Você pode especificar vários números de porta:

```
port=80 port=8000 ssl_port=8080
```

O padrão é 80.

#### endpoint, ssl\_endpoint

Os endereços de escuta no formato “endereço[:porta]”, em que o endereço é uma string de endereço IPv4 no formato decimal com pontos ou um endereço IPv6 na notação hexadecimal entre colchetes. Se for especificado um endpoint IPv6, a escuta será apenas no IPv6. O número da porta opcional considera o padrão de 80 para endpoint e 443 para ssl\_endpoint. Você pode especificar vários endereços:

```
endpoint=[::1] endpoint=192.168.0.100:8000 ssl_endpoint=192.168.0.100:8080
```

#### ssl\_private\_key

Caminho opcional para o arquivo de chave privada usado para endpoints habilitados para SSL. Se não for especificado, o arquivo ssl\_certificate será usado como uma chave privada.

#### tcp\_nodelay

Se especificado, a opção de soquete desabilitará o algoritmo Nagle na conexão. Isso significa que os pacotes serão enviados o mais rápido possível, em vez de esperar por um buffer completo ou o tempo de espera se esgotar.

“1” desabilita o algoritmo Nagle para todos os soquetes.

“0” mantém o algoritmo Nagle habilitado (padrão).

#### EXEMPLO 28.1: EXEMPLO DE CONFIGURAÇÃO DO BEAST

```
cephuser@adm > ceph config set rgw.myrealm.myzone.ses-min1.kwwazo \  
rgw_frontends beast port=8000 ssl_port=443 \  
ssl_certificate=/etc/ssl/ssl.crt \  
error_log_file=/var/log/radosgw/beast.error.log
```

### 28.5.2.2 CivetWeb

#### port

O número da porta de escuta. Para as portas habilitadas para SSL, adicione um sufixo "s" (por exemplo, "443s"). Para vincular um endereço IPv4 ou IPv6 específico, use o formato "endereço:porta". Você pode especificar vários endpoints adicionando "+" ou inserindo várias opções:

```
port=127.0.0.1:8000+443s
port=8000 port=443s
```

O padrão é 7480.

#### num\_threads

O número de threads gerados pelo Civetweb para processar as conexões HTTP recebidas. Efetivamente, isso limita o número de conexões simultâneas que o front end pode atender. O padrão é o valor especificado pela opção `rgw_thread_pool_size`.

#### request\_timeout\_ms

Por quanto tempo em milissegundos o Civetweb aguardará por mais dados recebidos antes de desistir.

O padrão é de 30.000 milissegundos.

#### access\_log\_file

Caminho para o arquivo de registro de acessos. Você pode especificar um caminho completo ou um caminho relativo ao diretório de trabalho atual. Se não for especificado (padrão), os acessos não serão registrados.

#### error\_log\_file

Caminho para o arquivo de registro de erros. Você pode especificar um caminho completo ou um caminho relativo ao diretório de trabalho atual. Se não for especificado (padrão), os erros não serão registrados.

#### EXEMPLO 28.2: EXEMPLO DE CONFIGURAÇÃO DO CIVETWEB EM /etc/ceph/ceph.conf

```
cephuser@adm > ceph config set rgw.myrealm.myzone.ses-min2.ingabw \
  rgw_frontends civetweb port=8000+443s request_timeout_ms=30000 \
  error_log_file=/var/log/radosgw/civetweb.error.log
```

### 28.5.2.3 Opções comuns

#### **ssl\_certificate**

Caminho para o arquivo de certificado SSL usado para endpoints habilitados para SSL.

#### **prefix**

Uma string de prefixo que é inserida no URI de todas as solicitações. Por exemplo, um front end apenas Swift pode inserir um prefixo de URI /swift.

## 29 Módulos do Ceph Manager

A arquitetura do Ceph Manager (consulte o *Livro “Guia de Implantação”, Capítulo 1 “SES e Ceph”, Seção 1.2.3 “Nós e daemons do Ceph”* para ver uma breve introdução) permite estender a funcionalidade dele por meio de *módulos*, como “dashboard” (consulte o *Parte I, “Ceph Dashboard”*), “prometheus” (consulte o *Capítulo 16, Monitoramento e alerta*) ou “balancer”.

Para listar todos os módulos disponíveis, execute:

```
cephuser@adm > ceph mgr module ls
{
  "enabled_modules": [
    "restful",
    "status"
  ],
  "disabled_modules": [
    "dashboard"
  ]
}
```

Para habilitar ou desabilitar um módulo específico, execute:

```
cephuser@adm > ceph mgr module enable MODULE-NAME
```

Por exemplo:

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

Para listar os serviços que os módulos habilitados oferecem, execute:

```
cephuser@adm > ceph mgr services
{
  "dashboard": "http://myserver.com:7789/",
  "restful": "https://myserver.com:8789/"
}
```

### 29.1 Balanceador

O módulo balanceador otimiza a distribuição do grupo de posicionamento (PG, placement group) entre os OSDs para uma implantação mais equilibrada. Embora o módulo esteja ativado por padrão, ele está inativo. Ele suporta estes dois modos: crush-compatible e upmap.



## Dica: Status e configuração atuais do balanceador

Para ver o status e as informações de configuração atuais do balanceador, execute:

```
cephuser@adm > ceph balancer status
```

### 29.1.1 O modo “crush-compatible”

No modo “crush-compatible”, o balanceador ajusta os conjuntos de reweight dos OSDs para obter uma melhor distribuição dos dados. Ele move os PGs entre os OSDs, causando temporariamente um estado de cluster HEALTH\_WARN resultante dos PGs deslocados.



## Dica: Ativação do Modo

Embora o “crush-compatible” seja o modo padrão, recomendamos ativá-lo explicitamente:

```
cephuser@adm > ceph balancer mode crush-compatible
```

### 29.1.2 Planejando e executando o balanceamento de dados

Usando o módulo balanceador, você pode criar um plano para balanceamento de dados. Em seguida, você pode executar o plano manualmente ou permitir que o balanceador equilibre os PGs continuamente.

A decisão de executar o balanceador no modo manual ou automático depende de vários fatores, como desequilíbrio dos dados atuais, tamanho do cluster, contagem de PGs ou atividade de E/S. Recomendamos criar um plano inicial e executá-lo durante um momento de carga baixa de E/S no cluster. A razão para isso é que o desequilíbrio inicial provavelmente será considerável, e é uma boa prática manter um baixo impacto nos clientes. Após uma execução manual inicial, considere ativar o modo automático e monitorar o tráfego de redistribuição sob carga normal de E/S. As melhorias na distribuição de PGs precisam ser ponderadas em relação ao tráfego de redistribuição causado pelo balanceador.



## Dica: Fração Móvel de Grupos de Posicionamento (PGs)

Durante o processo de balanceamento, o módulo balanceador limita os movimentos de PG para que apenas uma fração configurável de PGs seja movida. O padrão é 5%, e você pode ajustar a fração para 9%, por exemplo, executando o seguinte comando:

```
cephuser@adm > ceph config set mgr target_max_misplaced_ratio .09
```

Para criar e executar um plano de balanceamento, siga estas etapas:

1. Confira a pontuação atual do cluster:

```
cephuser@adm > ceph balancer eval
```

2. Crie um plano. Por exemplo, "great\_plan":

```
cephuser@adm > ceph balancer optimize great_plan
```

3. Veja que mudanças o “great\_plan” fará:

```
cephuser@adm > ceph balancer show great_plan
```

4. Confira a possível pontuação do cluster se você decidir aplicar o “great\_plan”:

```
cephuser@adm > ceph balancer eval great_plan
```

5. Execute o “great\_plan” apenas uma vez:

```
cephuser@adm > ceph balancer execute great_plan
```

6. Observe o balanceamento do cluster com o comando **ceph -s**. Se você estiver satisfeito com o resultado, ative o balanceamento automático:

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

Posteriormente, se você decidir desativar o balanceamento automático, execute:

```
cephuser@adm > ceph balancer off
```



## Dica: Balanceamento Automático sem Plano Inicial

Você pode ativar o balanceamento automático sem executar um plano inicial. Neste caso, espere uma redistribuição possivelmente longa dos grupos de posicionamento.

## 29.2 Habilitando o módulo de telemetria

O plug-in de telemetria envia os dados anônimos do projeto do Ceph sobre o cluster no qual o plug-in está sendo executado.

Esse componente (aceitação) contém contadores e estatísticas sobre como o cluster foi implantado, a versão do Ceph, a distribuição dos hosts e outros parâmetros que ajudam o projeto a obter uma melhor compreensão da forma como o Ceph é usado. Ele não contém dados confidenciais, como nomes de pool, nomes de objeto, conteúdo de objeto ou nomes de host.

O objetivo do módulo de telemetria é fornecer um loop de feedback automatizado para os desenvolvedores para ajudar a quantificar as taxas de adoção, o monitoramento ou apontar elementos que precisam ser mais bem explicados ou validados durante a configuração para evitar resultados indesejáveis.



### Nota

O módulo de telemetria exige que os nós do Ceph Manager consigam transmitir dados por HTTPS para os servidores de upstream. Verifique se os firewalls corporativos permitem essa ação.

1. Para habilitar o módulo de telemetria:

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



### Nota

Esse comando apenas permite que você veja seus dados localmente. Esse comando não compartilha seus dados com a comunidade do Ceph.

2. Para permitir que o módulo de telemetria comece a compartilhar dados:

```
cephuser@adm > ceph telemetry on
```

3. Para desabilitar o compartilhamento de dados de telemetria:

```
cephuser@adm > ceph telemetry off
```

4. Para gerar um relatório JSON que pode ser impresso:

```
cephuser@adm > ceph telemetry show
```

5. Para adicionar um contato e uma descrição ao relatório:

```
cephuser@adm > ceph config set mgr mgr/telemetry/contact John Doe  
john.doe@example.com  
cephuser@adm > ceph config set mgr mgr/telemetry/description 'My first Ceph cluster'
```

6. Por padrão, o módulo compila e envia um novo relatório a cada 24 horas. Para ajustar esse intervalo:

```
cephuser@adm > ceph config set mgr mgr/telemetry/interval HOURS
```



## 30 Autenticação com cephx

Para identificar clientes e proteger-se contra ataques man-in-the-middle, o Ceph oferece o sistema de autenticação cephx. Neste contexto, os *clientes* são pessoas, como o usuário admin, ou serviços/daemons relacionados ao Ceph, por exemplo, OSDs, monitores ou Gateways de Objetos.



### Nota

O protocolo cephx não atende à criptografia de dados em transporte, como TLS/SSL.

### 30.1 Arquitetura de autenticação

O cephx usa chaves secretas compartilhadas para autenticação, o que significa que tanto o cliente quanto os Ceph Monitors têm uma cópia da chave secreta do cliente. O protocolo de autenticação permite que ambas as partes comprovem uma para a outra que têm uma cópia da chave sem precisar revelá-la. Isso permite uma autenticação mútua: o cluster tem certeza de que o usuário possui a chave secreta, e o usuário também tem certeza de que o cluster tem uma cópia da chave secreta.

Um recurso de escalabilidade importante do Ceph é para evitar uma interface centralizada com o armazenamento de objetos do Ceph. Isso significa que os clientes do Ceph podem interagir diretamente com os OSDs. Para proteger os dados, o Ceph oferece o sistema de autenticação cephx, que autentica clientes do Ceph.

Cada monitor pode autenticar clientes e distribuir chaves, portanto, não há nenhum ponto único de falha ou gargalo ao usar o cephx. O monitor retorna uma estrutura de dados de autenticação que contém uma chave de sessão para uso na obtenção dos serviços do Ceph. Essa chave de sessão é autocriptografada com a chave secreta permanente do cliente para que apenas o cliente possa solicitar serviços dos Ceph Monitors. Em seguida, o cliente usa a chave de sessão para solicitar os serviços desejados do monitor, e o monitor emite um ticket para o cliente que lhe autenticará nos OSDs que realmente processam os dados. Os Ceph Monitors e OSDs compartilham um segredo, portanto, o cliente pode usar o ticket emitido pelo monitor com qualquer OSD ou servidor de metadados no cluster. Os tickets do cephx expiram para que um invasor não consiga usar um ticket expirado ou uma chave de sessão obtida indevidamente.

Para usar o cephx, um administrador deve primeiro configurar clientes/usuários. No diagrama a seguir, o usuário `client.admin` invoca **`ceph auth get-or-create-key`** da linha de comando para gerar um nome de usuário e a chave secreta. O subsistema auth do Ceph gera o nome de

usuário e a chave, armazena uma cópia com o(s) monitor(es) e transmite o segredo do usuário de volta ao usuário `client.admin`. Isso significa que o cliente e o monitor compartilham uma chave secreta.

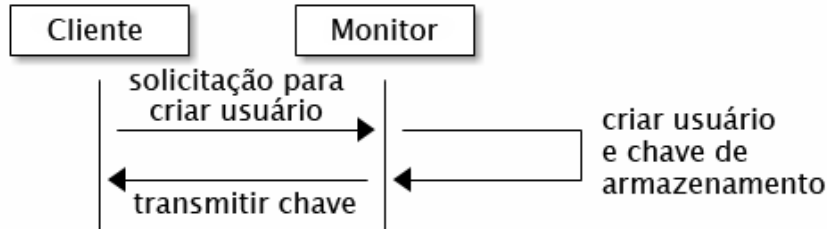


FIGURA 30.1: AUTENTICAÇÃO BÁSICA DO cephx

Para autenticar-se no monitor, o cliente envia o nome de usuário ao monitor. O monitor gera uma chave de sessão e a criptografa com a chave secreta associada ao nome de usuário e transmite o ticket criptografado de volta para o cliente. Em seguida, o cliente decodifica os dados com a chave secreta compartilhada para recuperar a chave de sessão. A chave de sessão identifica o usuário da sessão atual. Em seguida, o cliente solicita um ticket relacionado ao usuário, que é assinado pela chave de sessão. O monitor gera um ticket, criptografa-o com a chave secreta do usuário e o transmite de volta para o cliente. O cliente decodifica o ticket e o utiliza para assinar solicitações para OSDs e servidores de metadados em todo o cluster.

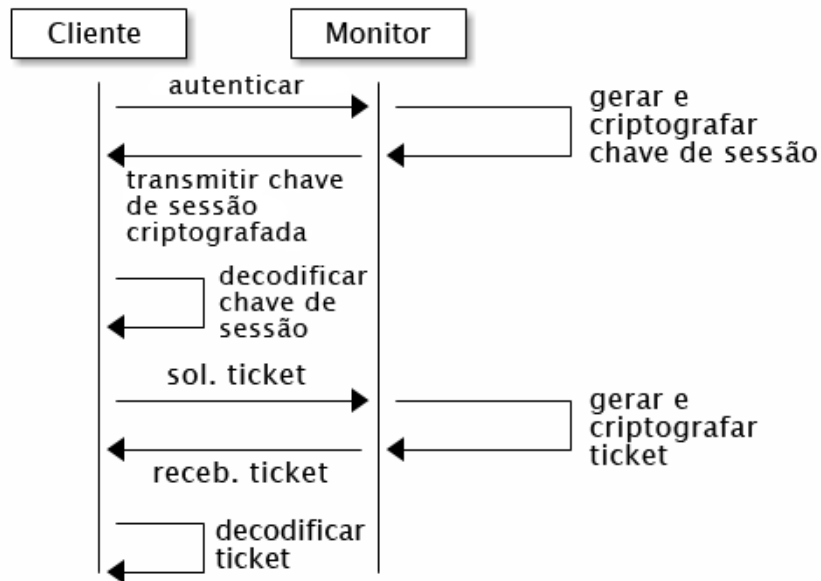


FIGURA 30.2: cephx AUTENTICAÇÃO

O protocolo cephx autentica as constantes comunicações entre a máquina cliente e os servidores Ceph. Cada mensagem enviada entre um cliente e um servidor após a autenticação inicial é assinada usando um ticket que os monitores, OSDs e servidores de metadados podem verificar com o segredo compartilhado.

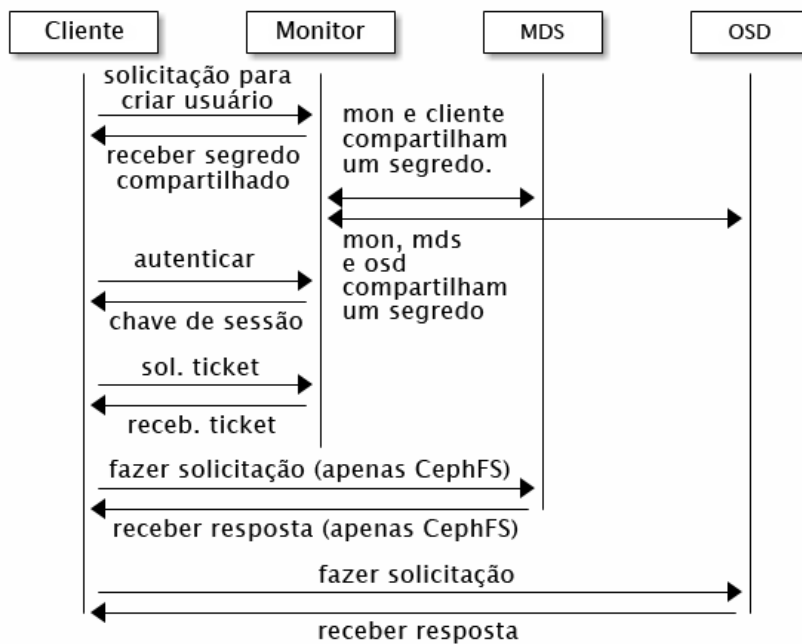


FIGURA 30.3: AUTENTICAÇÃO DO cephx: MDS E OSD

## ! Importante

A proteção oferecida por essa autenticação ocorre entre o cliente do Ceph e os hosts de cluster do Ceph. A autenticação não ultrapassa o cliente do Ceph. Se um usuário acessar o cliente do Ceph de um host remoto, a autenticação do Ceph não será aplicada à conexão entre o host do usuário e do cliente.

## 30.2 As principais áreas de

Esta seção descreve os usuários de cliente do Ceph e a autenticação e autorização no cluster de armazenamento do Ceph. *Usuários* são pessoas ou mecanismos de sistema, como aplicativos, que usam os clientes do Ceph para interagir com os daemons do cluster de armazenamento do Ceph. Quando o Ceph é executado com a autenticação e a autorização habilitadas (padrão), você deve especificar um nome de usuário e um chaveiro que contém a chave secreta do usuário especificado (geralmente por meio da linha de comando). Se você não especificar um nome de usuário, o Ceph usará o `client.admin` como padrão. Se você não especificar um chaveiro,

o Ceph procurará um na configuração de chaveiros no arquivo de configuração do Ceph. Por exemplo, se você executar o comando **ceph health** sem especificar um nome de usuário ou chaveiro, o Ceph interpretará o comando da seguinte forma:

```
cephuser@adm > ceph -n client.admin --keyring=/etc/ceph/ceph.client.admin.keyring health
```

Se preferir, você poderá usar a variável de ambiente `CEPH_ARGS` para não ter que redigitar o nome de usuário e o segredo.

## 30.2.1 Informações de referência

Seja qual for o tipo de cliente do Ceph (por exemplo, dispositivo de blocos, armazenamento de objetos, sistema de arquivos ou API nativa), o Ceph armazena todos os dados como objetos em *pools*. Os usuários do Ceph precisam ter acesso aos pools para ler e gravar dados. Os usuários do Ceph também devem ter permissões de execução para utilizar os comandos administrativos do Ceph. Os conceitos a seguir ajudarão você a entender o gerenciamento de usuários do Ceph.

### 30.2.1.1 Usuário

Um usuário é uma pessoa ou um mecanismo de sistema, como um aplicativo. A criação de usuários permite controlar quem (ou o quê) pode acessar o cluster de armazenamento do Ceph, os pools e os dados dos pools.

O Ceph usa *tipos* de usuários. Para fins de gerenciamento de usuários, o tipo sempre será `client`. O Ceph identifica os usuários no formato delimitado por ponto (.), que consiste no tipo e ID de usuário. Por exemplo, `TYPE.ID`, `client.admin` ou `client.user1`. O motivo da definição de tipo do usuário é que os Ceph Monitors, OSDs e servidores de metadados também usam o protocolo cephx, mas eles não são clientes. A distinção do tipo de usuário ajuda a diferenciar os usuários que são clientes dos demais, otimizando o controle de acesso, o monitoramento de usuários e o rastreamento.

Às vezes, o tipo de usuário do Ceph pode parecer confuso, porque a linha de comando do Ceph permite especificar um usuário com ou sem o tipo, dependendo do seu uso da linha de comando. Se você especificar `--user` ou `--id`, poderá omitir o tipo. Portanto, é possível inserir `client.user1` simplesmente como `user1`. Se você especificar `--name` ou `-n`, deverá especificar o tipo e o nome, como `client.user1`. Recomendamos o uso do tipo e do nome como uma melhor prática, sempre que possível.



## Nota

Um usuário de cluster de armazenamento do Ceph não é o mesmo que um usuário de armazenamento de objetos ou de sistema de arquivos do Ceph. O Gateway de Objetos do Ceph utiliza um usuário de cluster de armazenamento do Ceph para comunicação entre o daemon do gateway e o cluster de armazenamento, mas o gateway tem sua própria funcionalidade de gerenciamento para usuários finais. O sistema de arquivos do Ceph usa semânticas do POSIX. O espaço do usuário associado a ele não é o mesmo de um usuário de cluster de armazenamento do Ceph.

### 30.2.1.2 Autorização e recursos

O Ceph usa o termo "recursos" (caps) para descrever a autorização de um usuário autenticado para executar as funcionalidades dos monitores, OSDs e servidores de metadados. Os recursos também podem restringir o acesso aos dados em um pool ou namespace do pool. Um usuário administrador do Ceph define os recursos do usuário ao criá-lo ou atualizá-lo.

A sintaxe de recurso segue o formato:

```
daemon-type 'allow capability' [...]
```

Veja a seguir uma lista de recursos para cada tipo de serviço:

#### Recursos do monitor

incluem r, w, x e allow profile cap.

```
mon 'allow rwx'
mon 'allow profile osd'
```

#### Recursos do OSD

incluem r, w, x, class-read, class-write e profile osd. Os recursos do OSD também permitem configurações de pool e namespace.

```
osd 'allow capability' [pool=poolname] [namespace=namespace-name]
```

#### Recurso do MDS

requer apenas allow ou fica em branco.

```
mds 'allow'
```

As entradas a seguir descrevem cada recurso:

#### **allow**

Antecede as configurações de acesso para um daemon. Implica apenas no rw para MDS.

#### **r**

Concede o acesso de leitura ao usuário. Necessário com monitores para recuperar o mapa CRUSH.

#### **w**

Concede ao usuário acesso de gravação em objetos.

#### **x**

Permite que o usuário chame métodos de classe (tanto de leitura quanto de gravação) e execute operações do auth em monitores.

#### **class-read**

Permite que o usuário chame métodos de leitura de classe. Subconjunto do x.

#### **class-write**

Permite que o usuário chame métodos de gravação de classe. Subconjunto do x.

#### **\***

Concede ao usuário permissões de leitura, gravação e execução para determinado daemon/pool e permite executar comandos de admin.

#### **profile osd**

Concede a um usuário permissões para conectar-se como OSD a outros OSDs ou monitores. Atribuído aos OSDs para permitir que eles processem o tráfego de heartbeat de replicação e o relatório de status.

#### **profile mds**

Concede a um usuário permissões para conectar-se como MDS a outros MDSs ou monitores.

#### **profile bootstrap-osd**

Concede a um usuário permissões para inicializar um OSD. Delegado a ferramentas de implantação para que elas tenham permissões para adicionar chaves ao inicializar um OSD.

#### **profile bootstrap-mds**

Concede a um usuário permissões para inicializar um servidor de metadados. Delegado a ferramentas de implantação para que elas tenham permissões para adicionar chaves ao inicializar um servidor de metadados.

### 30.2.1.3 Pools

Um pool é uma partição lógica em que os usuários armazenam dados. No caso das implantações do Ceph, é comum criar um pool como partição lógica para tipos de dados semelhantes. Por exemplo, ao implantar o Ceph como back end para o OpenStack, uma implantação típica tem pools para volumes, imagens, backups, máquinas virtuais e usuários como `client.glance` ou `client.cinder`.

## 30.2.2 Gerenciando usuários

A funcionalidade de gerenciamento de usuários permite aos administradores de cluster do Ceph criar, atualizar e apagar usuários diretamente do cluster do Ceph.

Ao criar ou apagar usuários do cluster do Ceph, talvez você tenha que distribuir chaves aos clientes para que elas possam ser adicionadas aos chaveiros. Consulte a [Seção 30.2.3, “Gerenciando chaveiros”](#) para obter os detalhes.

### 30.2.2.1 Listando usuários

Para listar os usuários em seu cluster, execute o seguinte:

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

O Ceph listará todos os usuários em seu cluster. Por exemplo, em um cluster com dois nós, a saída de `ceph auth list` tem esta aparência:

```
installed auth entries:

osd.0
    key: AQCvCbtToC6MDhAATtuT70Sl+DymPCfDSsyV4w==
    caps: [mon] allow profile osd
    caps: [osd] allow *
osd.1
    key: AQC4CbtTCFJBChAAVq5spj0ff4eHZICxIOVZeA==
    caps: [mon] allow profile osd
    caps: [osd] allow *
client.admin
    key: AQBHCbtT6APDHhAA5W00cBchwKQjh3dkKsyPjw==
    caps: [mds] allow
    caps: [mon] allow *
    caps: [osd] allow *
client.bootstrap-mds
```



```
key: AQBICbtT0K9uGBAAdbE5zcIGHZL3T/u2g6EBww==
caps: [mon] allow profile bootstrap-mds
client.bootstrap-osd
key: AQBHCbtT4Gxq0RAADE5u7RkpCN/oo4e5W0uBtw==
caps: [mon] allow profile bootstrap-osd
```



### Nota: Notação TYPE.ID

Observe que a notificação `TYPE.ID` para usuários é aplicada de modo que `osd.0` especifique um usuário do tipo `osd` e o ID seja `0`. `client.admin` é um usuário do tipo `client` e o ID é `admin`. Observe também que cada entrada tem uma entrada `key: value`, e uma ou mais entradas `caps:`.

Você pode usar a opção `-o nome_de_arquivo` com `ceph auth list` para gravar a saída em um arquivo.

#### 30.2.2.2 Obtendo informações sobre usuários

Para recuperar um usuário, chave e recursos específicos, execute o seguinte:

```
cephuser@adm > ceph auth get TYPE.ID
```

Por exemplo:

```
cephuser@adm > ceph auth get client.admin
exported keyring for client.admin
[client.admin]
key = AQA19uZUqIwkHxAAFuUwvq0eJD4S173oFRxe0g==
caps mds = "allow"
caps mon = "allow *"
caps osd = "allow *"
```

Os desenvolvedores também podem executar o seguinte:

```
cephuser@adm > ceph auth export TYPE.ID
```

O comando `auth export` é idêntico a `auth get`, mas também imprime o ID de autenticação interno.

#### 30.2.2.3 Adicionando usuários

A adição de um usuário cria um nome de usuário (`TYPE.ID`), uma chave secreta e quaisquer recursos incluídos no comando que você usa para criar o usuário.

A chave do usuário permite que ele se autentique no cluster de armazenamento do Ceph. Os recursos do usuário lhe autorizam a ler, gravar ou executar Ceph Monitors (mon), Ceph OSDs (osd) ou servidores de metadados do Ceph (mds).

Há alguns comandos disponíveis para adicionar um usuário:

#### **ceph auth add**

Esse comando é a forma canônica de adicionar um usuário. Ele criará o usuário, gerará uma chave e adicionará quaisquer recursos especificados.

#### **ceph auth get-or-create**

Geralmente, esse comando é o método mais prático de criar um usuário, pois ele retorna um formato de arquivo de chaves com o nome de usuário (entre parênteses) e a chave. Se o usuário já existir, esse comando simplesmente retornará o nome de usuário e a chave no formato de arquivo de chaves. Você pode usar a opção -o *nomedearquivo* para gravar a saída em um arquivo.

#### **ceph auth get-or-create-key**

Esse comando é um método prático de criar um usuário e retornar a chave dele (apenas). Ele é útil para clientes que precisam apenas da chave (por exemplo, libvirt). Se o usuário já existir, esse comando retornará apenas a chave. Você pode usar a opção -o *nomedearquivo* para gravar a saída em um arquivo.

Ao criar usuários de cliente, você pode criá-los sem recursos. Um usuário sem recursos pode apenas se autenticar, nada mais. Esse tipo de cliente não pode recuperar o mapa de cluster do monitor. No entanto, você pode criar um usuário sem recursos para adiar a adição de recursos usando o comando **ceph auth caps**.

Um usuário comum tem pelo menos recursos de leitura no Ceph Monitor e recursos de leitura e gravação nos Ceph OSDs. Além disso, as permissões de OSD do usuário costumam limitar-se ao acesso a determinado pool.

```
cephuser@adm > ceph auth add client.john mon 'allow r' osd \
'allow rw pool=liverpool'
cephuser@adm > ceph auth get-or-create client.paul mon 'allow r' osd \
'allow rw pool=liverpool'
cephuser@adm > ceph auth get-or-create client.george mon 'allow r' osd \
'allow rw pool=liverpool' -o george.keyring
cephuser@adm > ceph auth get-or-create-key client.ringo mon 'allow r' osd \
'allow rw pool=liverpool' -o ringo.key
```

## ! Importante

Se você conceder a um usuário recursos para OSDs, mas *não* restringir o acesso a determinados pools, o usuário terá acesso a *todos* os pools no cluster.

### 30.2.2.4 Modificando recursos do usuário

O comando **ceph auth caps** permite especificar um usuário e mudar os recursos dele. A definição de novos recursos sobregravará os atuais. Para ver os recursos atuais, execute **ceph auth get *USERTYPE.USERID***. Para adicionar recursos, você também precisa especificar os recursos existentes quando usar o formato a seguir:

```
cephuser@adm > ceph auth caps USERTYPE.USERID daemon 'allow [r|w|x|*|...] \
    [pool=pool-name] [namespace=namespace-name]' [daemon 'allow [r|w|x|*|...] \
    [pool=pool-name] [namespace=namespace-name']
```

Por exemplo:

```
cephuser@adm > ceph auth get client.john
cephuser@adm > ceph auth caps client.john mon 'allow r' osd 'allow rw pool=prague'
cephuser@adm > ceph auth caps client.paul mon 'allow rw' osd 'allow r pool=prague'
cephuser@adm > ceph auth caps client.brian-manager mon 'allow *' osd 'allow *'
```

Para remover um recurso, você pode redefini-lo. Para que o usuário não tenha acesso a determinado daemon já definido, especifique uma string vazia:

```
cephuser@adm > ceph auth caps client.ringo mon ' ' osd ' '
```

### 30.2.2.5 Apagando usuários

Para apagar um usuário, execute **ceph auth del**:

```
cephuser@adm > ceph auth del TYPE.ID
```

em que *TYPE* é *client*, *osd*, *mon* ou *mds*, e *ID* é o nome de usuário ou o ID do daemon.

Se você criou usuários com permissões estritamente para um pool que não existe mais, convém apagá-los também.

### 30.2.2.6 Imprimindo uma chave do usuário

Para imprimir a chave de autenticação do usuário em uma saída padrão, execute o seguinte:

```
cephuser@adm > ceph auth print-key TYPE.ID
```

em que *TYPE* é *client*, *osd*, *mon* ou *mds*, e *ID* é o nome de usuário ou o ID do daemon.

A impressão da chave do usuário é útil quando você precisa preencher o software cliente com a chave do usuário (como *libvirt*), conforme mostrado neste exemplo:

```
# mount -t ceph host:/ mount_point \  
-o name=client.user,secret=`ceph auth print-key client.user`
```

### 30.2.2.7 Importando usuários

Para importar um ou mais usuários, execute **ceph auth import** e especifique um chaveiro:

```
cephuser@adm > ceph auth import -i /etc/ceph/ceph.keyring
```



#### Nota

O cluster de armazenamento do Ceph adicionará novos usuários, as chaves e os recursos deles e atualizará os usuários existentes, as chaves e os recursos deles.

## 30.2.3 Gerenciando chaveiros

Quando você acessa o Ceph por um cliente, esse cliente procura um chaveiro local. Por padrão, o Ceph predefine a configuração de chaveiro com os quatro nomes de chaveiro a seguir, portanto, você não precisa defini-la em seu arquivo de configuração do Ceph, a menos que queira anular os padrões:

```
/etc/ceph/cluster.name.keyring  
/etc/ceph/cluster.keyring  
/etc/ceph/keyring  
/etc/ceph/keyring.bin
```

A metavariável *cluster* é o nome do cluster do Ceph conforme definido pelo nome do arquivo de configuração do Ceph. *ceph.conf* significa que o nome do cluster é *ceph*, portanto, *ceph.keyring*. A metavariável *name* é o tipo e o ID de usuário. Por exemplo, *client.admin*, portanto, *ceph.client.admin.keyring*.

Após criar um usuário (por exemplo, `client.ringo`), você deverá obter a chave e adicioná-la a um chaveiro no cliente do Ceph para que o usuário possa acessar o cluster de armazenamento do Ceph.

A [Seção 30.2, “As principais áreas de”](#) apresenta os detalhes de como listar, obter, adicionar, modificar e apagar usuários diretamente do cluster de armazenamento do Ceph. No entanto, o Ceph também oferece o utilitário `ceph-authtool` para que você possa gerenciar chaveiros de um cliente do Ceph.

### 30.2.3.1 Criando um chaveiro

Ao usar os procedimentos na [Seção 30.2, “As principais áreas de”](#) para criar usuários, você precisa fornecer as chaves de usuário ao(s) cliente(s) do Ceph para permitir a recuperação da chave do usuário especificado e a autenticação no cluster de armazenamento do Ceph. Os clientes do Ceph acessam os chaveiros para pesquisar um nome de usuário e recuperar a chave do usuário:

```
cephuser@adm > ceph-authtool --create-keyring /path/to/keyring
```

Durante a criação de um chaveiro com vários usuários, é recomendável usar o nome do cluster (por exemplo, `cluster.keyring`) para o nome de arquivo do chaveiro e gravá-lo no diretório `/etc/ceph` para que a configuração padrão do chaveiro obtenha o nome do arquivo sem que você tenha que especificá-lo na cópia local do seu arquivo de configuração do Ceph. Por exemplo, crie `ceph.keyring` executando o seguinte:

```
cephuser@adm > ceph-authtool -C /etc/ceph/ceph.keyring
```

Durante a criação de um chaveiro com um único usuário, é recomendável usar o nome do cluster, o tipo de usuário e o nome de usuário e gravá-lo no diretório `/etc/ceph`. Por exemplo, `ceph.client.admin.keyring` para o usuário `client.admin`.

### 30.2.3.2 Adicionando um usuário a um chaveiro

Ao adicionar um usuário ao cluster de armazenamento do Ceph (consulte a [Seção 30.2.2.3, “Adicionando usuários”](#)), você pode recuperar o usuário, a chave e os recursos e gravá-lo em um chaveiro.

Para usar apenas um usuário por chaveiro, o comando **ceph auth get** com a opção **-o** gravará a saída no formato de arquivo do chaveiro. Por exemplo, para criar um chaveiro para o usuário `client.admin`, execute o seguinte:

```
cephuser@adm > ceph auth get client.admin -o /etc/ceph/ceph.client.admin.keyring
```

Para importar usuários para um chaveiro, você pode usar **ceph-authtool** para especificar o chaveiro de destino e de origem:

```
cephuser@adm > ceph-authtool /etc/ceph/ceph.keyring \  
--import-keyring /etc/ceph/ceph.client.admin.keyring
```



## Importante

Se o chaveiro estiver comprometido, apague sua chave do diretório `/etc/ceph` e recrie uma chave seguindo as mesmas instruções da [Seção 30.2.3.1, “Criando um chaveiro”](#).

### 30.2.3.3 Criando um usuário

O Ceph inclui o comando **ceph auth add** para criar um usuário diretamente no cluster de armazenamento do Ceph. No entanto, você também pode criar um usuário, as chaves e os recursos diretamente em um chaveiro de cliente do Ceph. Em seguida, você pode importar o usuário para o cluster de armazenamento do Ceph:

```
cephuser@adm > ceph-authtool -n client.ringo --cap osd 'allow rwx' \  
--cap mon 'allow rwx' /etc/ceph/ceph.keyring
```

Você também pode criar um chaveiro e adicionar um novo usuário a ele simultaneamente:

```
cephuser@adm > ceph-authtool -C /etc/ceph/ceph.keyring -n client.ringo \  
--cap osd 'allow rwx' --cap mon 'allow rwx' --gen-key
```

Nos cenários anteriores, o novo usuário `client.ringo` está apenas no chaveiro. Para adicionar o novo usuário ao cluster de armazenamento do Ceph, você ainda deve adicioná-lo ao cluster:

```
cephuser@adm > ceph auth add client.ringo -i /etc/ceph/ceph.keyring
```

### 30.2.3.4 Modificando usuários

Para modificar os recursos do registro de um usuário em um chaveiro, especifique o chaveiro e o usuário seguidos dos recursos:

```
cephuser@adm > ceph-authtool /etc/ceph/ceph.keyring -n client.ringo \
--cap osd 'allow rwx' --cap mon 'allow rwx'
```

Para atualizar o usuário modificado no ambiente de cluster do Ceph, você deve importar as mudanças do chaveiro para a entrada do usuário no cluster do Ceph:

```
cephuser@adm > ceph auth import -i /etc/ceph/ceph.keyring
```

Consulte a [Seção 30.2.2.7, “Importando usuários”](#) para obter detalhes sobre como atualizar um usuário do cluster de armazenamento do Ceph de um chaveiro.

## 30.2.4 Uso da linha de comando

O comando **ceph** suporta as seguintes opções relacionadas à manipulação de nome de usuário e segredo:

### --id ou --user

O Ceph identifica os usuários com um tipo e um ID (*TYPE.ID*, como client.admin ou client.user1). As opções id, name e -n permitem especificar a parte do ID do nome de usuário (por exemplo, admin ou user1). Você pode especificar o usuário com --id e omitir o tipo. Por exemplo, para especificar o usuário client.foo, digite o seguinte:

```
cephuser@adm > ceph --id foo --keyring /path/to/keyring health
cephuser@adm > ceph --user foo --keyring /path/to/keyring health
```

### --name ou -n

O Ceph identifica os usuários com um tipo e um ID (*TYPE.ID*, como client.admin ou client.user1). As opções --name e -n permitem especificar o nome completo do usuário. Você deve especificar o tipo de usuário (normalmente client) com o ID de usuário:

```
cephuser@adm > ceph --name client.foo --keyring /path/to/keyring health
cephuser@adm > ceph -n client.foo --keyring /path/to/keyring health
```

## --keyring

O caminho para o chaveiro que contém um ou mais nomes de usuário e segredos. A opção --secret tem a mesma funcionalidade, mas não funciona com o Gateway de Objetos, que usa --secret para outra finalidade. Você pode recuperar um chaveiro com ceph auth get-or-create e armazená-lo localmente. Essa é a abordagem preferencial, pois você pode alternar nomes de usuário sem mudar o caminho do chaveiro:

```
cephuser@adm > rbd map --id foo --keyring /path/to/keyring mypool/myimage
```



## A Atualizações de manutenção do Ceph baseadas nos point releases de upstream do “Pacific”

Vários pacotes importantes no SUSE Enterprise Storage 7.1 são baseados na série de lançamentos Pacific do Ceph. Quando o projeto do Ceph (<https://github.com/ceph/ceph>) publica novos point releases na série Pacific, o SUSE Enterprise Storage 7.1 é atualizado para garantir que o produto aproveite as correções de bug de upstream e os backports de recursos mais recentes.

Este capítulo contém resumos das mudanças importantes contidas em cada point release de upstream que foi, ou está planejado para ser incluído no produto.

# Glossário

## Geral

### **Alertmanager**

Um binário único que processa os alertas enviados pelo servidor Prometheus e notifica o usuário final.

### **Armazenamento de Objetos do Ceph**

O “produto”, o serviço ou os recursos de armazenamento de objetos, que consistem em um Cluster de Armazenamento do Ceph e um Gateway de Objetos do Ceph.

### **Árvore de roteamento**

Um termo que representa qualquer diagrama que mostra as várias rotas que um receptor pode executar.

### **Ceph Dashboard**

Um aplicativo incorporado de monitoramento e gerenciamento do Ceph baseado na Web que administra vários aspectos e objetos do cluster. O painel de controle é implementado como um módulo do Ceph Manager.

### **Ceph Manager**

O Ceph Manager ou MGR é o software de gerenciador do Ceph, que coleta o estado completo de todo o cluster em um único local.

### **Ceph Monitor**

O Ceph Monitor ou MON é o software de monitoração do Ceph.

### **ceph-salt**

Inclui ferramentas para implantação de clusters do Ceph gerenciados pelo cephadm por meio do Salt.

### **cephadm**

O cephadm implanta e gerencia um cluster do Ceph conectando-se aos hosts do daemon do gerenciador por SSH para adicionar, remover ou atualizar os containers de daemons do Ceph.

## CephFS

O sistema de arquivos do Ceph.

## CephX

O protocolo de autenticação do Ceph. O CephX opera como o Kerberos, mas não tem um ponto único de falha.

## Cliente do Ceph

A coleção de componentes do Ceph que podem acessar um Cluster de Armazenamento do Ceph. Eles incluem o Gateway de Objetos, o Dispositivo de Blocos do Ceph, o CephFS e as bibliotecas, os módulos de kernel e os clientes FUSE correspondentes.

## Cluster de Armazenamento do Ceph

O conjunto principal do software de armazenamento que armazena os dados do usuário. Esse conjunto consiste em Ceph Monitors e OSDs.

## Compartimento de memória

Um ponto que agrega outros nós em uma hierarquia de locais físicos.

## Conjunto de Regras

Regras para determinar o posicionamento de dados em um pool.

## CRUSH, Mapa CRUSH

*Controlled Replication Under Scalable Hashing*: Um algoritmo que determina como armazenar e recuperar dados calculando os locais de armazenamento de dados. O CRUSH requer um mapa de cluster para armazenar e recuperar dados de forma pseudo-aleatória nos OSDs com uma distribuição uniforme dos dados pelo cluster.

## Daemon Ceph OSD

O daemon **ceph-osd** é o componente do Ceph responsável por armazenar objetos em um sistema de arquivos local e por conceder acesso a eles pela rede.

## Dispositivo de Blocos RADOS (RBD)

O componente de armazenamento de blocos do Ceph. Também conhecido como dispositivo de blocos do Ceph.

## DriveGroups

DriveGroups são uma declaração de um ou mais layouts de OSD que podem ser mapeados para unidades físicas. Um layout de OSD define como o Ceph aloca fisicamente o armazenamento OSD na mídia correspondente aos critérios especificados.

## **Gateway de Objetos**

O componente de gateway do S3/Swift do Armazenamento de Objetos do Ceph. Também conhecido como RADOS Gateway (RGW).

## **Gateway do Samba**

O Gateway do Samba ingressa no Active Directory no domínio do Windows para autenticar e autorizar usuários.

## **Grafana**

Análise de banco de dados e solução de monitoramento.

## **Módulo de sincronização de arquivos**

Módulo que permite a criação de uma zona do Gateway de Objetos para manter o histórico das versões de objeto do S3.

## **Nó**

Qualquer máquina ou servidor único em um cluster do Ceph.

## **Nó de admin**

O host do qual você executa os comandos relacionados ao Ceph para administrar os hosts do cluster.

## **Nó OSD**

Um nó do cluster que armazena dados, processa a replicação, a recuperação, o preenchimento e a redistribuição de dados e fornece algumas informações de monitoramento aos Ceph Monitors examinando outros daemons Ceph OSD.

## **OSD**

*Dispositivo de Armazenamento de Objetos:* Uma unidade de armazenamento física ou lógica.

## **PG**

Grupo de Posicionamento: uma subdivisão de um *pool*, usado para ajuste de desempenho.

## **Point Release**

Qualquer versão Ad Hoc que inclua apenas correções de bug ou segurança.

## **Pool**

Partições lógicas para armazenamento de objetos, como imagens de disco.

## **Prometheus**

Kit de ferramentas de monitoramento e alertas de sistemas.

## **Regra CRUSH**

A regra de posicionamento de dados CRUSH que se aplica a um ou mais pools específicos.

## **Reliable Autonomic Distributed Object Store (RADOS)**

O conjunto principal do software de armazenamento que armazena os dados do usuário (MON + OSD).

## **Samba**

Software de integração do Windows.

## **Servidor de Metadados**

O Servidor de Metadados ou MDS é o software de metadados do Ceph.

## **Várias zonas**

## **zonegroup**